Proceedings of the
5th International Workshop on
EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA

ES³LOD 2014

Workshop Programme

26 May 2014

14:00 – 15:00 **Keynote I** (Chair: Björn Schuller)

Walter Daelemans (University of Antwerp, The Netherlands), *Profiling and sentiment mining for detecting threatening situations in social networks: the AMiCA project*

15:00 – 16:00 **Session 1: Markup and Linked Data** (Chair: Walter Daelemans)

Felix Burkhardt, Christian Becker-Asano, Edmon Begoli, Roddy Cowie, Gerhard Fobe, Patrick Gebhard, Abe Kazemzadeh, Ingmar Steiner and Tim Llewellyn, *Application of EmotionML*

Gabriela Vulcu, Paul Buitelaar, Sapna Negi, Bianca Pereira, Mihael Arcan, Barry Coughlan, Fernando J. Sanchez and Carlos A. Iglesias, *Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources*

16:00 – 16:30 Coffee break

16:30 – 18:00 **Session 2: Spoken Language** (Chair: Laurence Devillers)

Anna Prokofieva and Julia Hirschberg, *Hedging and Speaker Commitment*

Björn Schuller, Yue Zhang, Florian Eyben and Felix Weninger, *Intelligent Systems' Holistic Evolving Analysis of Real-life Universal Speaker Characteristics*

Zixing Zhang, Florian Eyben, Jun Deng and Björn Schuller, *An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena*

27 May 2014

09:00 – 10:30 **Keynote II and Plenary Discussion** (Chair: Paul Buitelaar)

Carlos Iglesias (Universidad Politécnica de Madrid, Spain), *A linked data approach for describing sentiments and emotions*

Plenary Discussion: *W3C Community Group on Linked Data Models for Emotion and Sentiment Analysis*

10:30 – 11:00 Coffee break

11:00 – 13:00 **Session 3: Corpora and Data Collection** (Chair: Thierry Declerck)

Véronique Aubergé, Yuko Sasa, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Antunes, Gilles De Biasi, Sybille Caffiau and Florian Nebout, *The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform*

Mohamed A. Sehili, Fan Yang, Violaine Leynaert and Laurence Devillers, *A corpus of social interaction between NAO and elderly people*

Kateřina Veselovská, *Fear and Trembling: Annotating Emotions in Czech Holocaust Testimonies*

Heather Pon-Barry, *Using Ambiguous Handwritten Digits to Induce Uncertainty*

13:00 – 14:00 Lunch break

14:00 – 16:00 **Session 4: Social Networks** (Chair: Carlos Iglesias)

Eshrag Refaee and Verena Rieser, *Can We Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds*

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti and Emilio Sulis, *Detecting Happiness in Italian Tweets: Towards an Evaluation Dataset for Sentiment Analysis in Felicittà*

Erik Tjong Kim Sang, *Using Tweets for Assigning Sentiments to Regions*

Francisco Rangel, Irazú Hernández, Paolo Rosso and Antonio Reyes, *Emotions and Irony per Gender in Facebook*

16:00 – 16:30 Coffee break

16:30 – 18:00 **Session 5: Written Language** (Chair: Chloé Clavel)

Ekaterina Volkova and Betty J. Mohler, *On-line Annotation System and New Corpora for Fine-Grained Sentiment Analysis of Text*

Elizabeth Baran, *Correlating Document Sentiment Scores with Web-Sourced Emotional Response Polls for a More Realistic Measure of Sentiment Performance*

Caroline Langlet and Chloé Clavel, *Modelling user's attitudinal reactions to the agent utterances: focus on the verbal content*

Editors / Workshop Organising Committee

Björn Schuller	Imperial College London, UK
Paul Buitelaar	NUI Galway, Ireland
Laurence Devillers	U. Sorbonne / CNRS-LIMSI, France
Catherine Pelachaud	CNRS-LTCI, France
Thierry Declerck	DFKI, Germany
Anton Batliner	FAU/TUM, Germany
Paolo Rosso	PRHLT, U. Politèc. Valencia, Spain
Seán Gaines	Vicomtech-IK4, Spain

Workshop Programme Committee

Rodrigo Agerri	EHU, Spain
Noam Amir	Tel-Aviv U., Isreal
Elisabeth André	U. Augsburg, Germany
Alexandra Balahur-Dobrescu	ISPRA, Italy
Cristina Bosco	U. Torino, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Carlos Busso	UT Dallas, USA
Rafael Calvo	U. Sydney, Australia
Erik Cambria	NUS, Singapore
Antonio Camurri	U. Genova, Italy
Mohamed Chetouani	UPMC, France
Montse Cuadros	VicomTech, Spain
Francesco Danza	Expert System, Italy
Thierry Dutoit	U. Mons, Belgium
Julien Epps	NICTA, Australia
Anna Esposito	IIASS, Italy
Francesca Frontini	CNR, Italy
Hatice Gunes	Queen Mary U., UK
Hayley Hung	TU Delft, the Netherlands
Carlos Iglesias	UPM, Spain
Isa Maks	VU, the Netherlands
Daniel Molina	Paradigma Tecnologico, Spain
Monica Monachini	CNR, Italy
Shrikanth Narayanan	USC, USA
Viviana Patti	U. Torino, Italy
German Rigau	EHU, Spain
Fabien Ringeval	U. Fribourg, Switzerland
Massimo Romanelli	Attensity EUROPE, Germany
Albert Ali Salah	Boğaziçi University, Turkey
Metin Sezgin	Koc U., Turkey
Carlo Strapparava	FBK, Italy
Jianhua Tao	CAS, P.R. China
Tony Veale	UCD, Ireland
Michel Valstar	U. Nottingham, UK
Alessandro Vinciarelli	U. Glasgow, UK
Piek Vossen	VU, the Netherlands

Table of contents

SOCIAL NETWORKS

WRITTEN LANGUAGE

Author Index

Emotion, Social Signals, Sentiment & Linked Open Data: A Short Introduction

The fifth instalment of the highly successful series of Corpora for Research on Emotion held at the last LRECs (2006, 2008, 2010, 2012) aims to help further bridging the gap between research on human emotion, social signals and sentiment from speech, text, and further modalities, and low availability of language and multimodal resources and labelled data for learning and testing.

As usually rather labels than the actual data are sparse, this year emphasis was put also on efficient community-shared and computer-supported labelling approaches and on cross-corpora experiments. Following LREC 2014's hot topics of Big Data and Linked Open Data in particular also methods for semi-automated and collaborative labelling of large data archives such as by efficient combinations of active learning and crowd sourcing are featured in this edition – in particular also for combined annotations of emotion, social signals, and sentiment. Multi- and cross-corpus studies (transfer learning, standardisation, corpus quality assessment, etc.) were further considered as highly relevant, given their importance in order to test the generalisation power of models.

A further main motivation for this year's workshop was to survey and promote the uptake of Linked Data in emotion, sentiment & social signal analysis research and applications. Linked Open Data is an increasingly wide-spread methodology for the publishing, sharing and interlinking of data sets. In the context of this workshop we were also interested in reports on and experiences with the use of Linked Open Data in the context of emotion, social signals, and sentiment in analysis projects and applications.

As before, also the multimodal community was invited and encouraged to contribute new corpora, perspectives and findings – emotion, sentiment, and social behaviour are multimodal and complex and there is still an urgent need for sufficient naturalistic uni- and multimodal data in different languages and from different cultures.

From the papers received, 16 were selected for the final programme (rejecting six) by the 36 members of the technical programme committee and the eight organisers. The accepted contributions were all selected as oral presentation and come from a total of 65 authors. They were grouped into the five groups *markup (languages) and linked data* (two papers), *spoken language* (three papers), *corpora and data collection* (four papers), *social networks* (four papers), and *written language* (three papers). Obviously, several of the papers fall under multiple of these headings and other groupings could have been thought off.

From the 16 accepted contributions one was selected as best paper by the technical program committee and organisers based on the review results and a rigorous second screening – contributions including members of the organising committee were not eligible for fairness reasons. This best paper award was given to Véronique Aubergé, Yuko Sasa, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Antunes, Gilles De Biasi, Sybille Caffiau and Florian Nebout for their outstanding and inspiring introduction and efforts of and around *The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform*.

Two keynote speeches by distinguished researchers crossing the communities further focused on the above named topics of particular interest: Walter Daelemans's (University of Antwerp, The Netherlands) talk *Profiling and sentiment mining for detecting threatening situations in social*

networks: the AMiCA project introduced findings from a larger project. The second speech given by Carlos Iglesias (Universidad Politécnica de Madrid, Spain) was entitled *A linked data approach for describing sentiments and emotions*, and followed by a plenary discussion around the *W3C Community Group on Linked Data Models for Emotion and Sentiment Analysis*.

The organisers are further grateful for the sponsorship of the Association for the Advancement of Affective Computing (AAAC, former HUMAINE Association) and the SSPNet. The workshop was further partially organised in the context of and received funding from the following European projects: ASC-Inclusion (http://www.asc-inclusion.eu), EuroSentiment (http://eurosentiment.eu), iHEARu (http://www.ihearu.eu), ilhaire (http://www.ilhaire.eu/), LIDER (http://lider-project.eu/), OpeNER (http://www.opener-project.org), TARDIS (http://www.tardis-project.eu), TrendMiner (http://www.trendminer-project.eu), and WiQ-Ei. The responsibility lies with the organisers and authors.

To conclude, we would like to thank all the dedicated members of the technical program committee, the sponsors, ELRA, and of course all authors for an inspiring and exciting workshop and proceedings.

Björn Schuller, Paul Buitelaar, Laurence Devillers, Catherine Pelachaud,
Thierry Declerck, Anton Batliner, Paolo Rosso, Seán Gaines

Organisers of ES³LOD 2014

Application of EmotionML

Felix Burkhardt[1], Christian Becker-Asano[2], Edmon Begoli[3], Roddy Cowie[4], Gerhard Fobe[5], Patrick Gebhard[6], Abe Kazemzadeh[7], Ingmar Steiner[6,8], Tim Llewellyn[9]

[1]Deutsche Telekom Laboratories, Berlin, Germany, [2]Albert-Ludwigs-Universität, Freiburg, Germany, [3]University of Tennessee, Knoxville, USA, [4]Queen's University Belfast, UK, [5]Technische Universität Chemnitz, Germany, [6]DFKI, Saarbrücken, Germany, [7]University of Southern California, USA, [8]Saarland University, Saarbrücken, Germany, [9]nViso, Lausanne, Switzerland

Felix.Burkhardt@telekom.de

Abstract

We present EmotionML, a new W3C recommendation to represent emotion related states in data processing systems, by first introducing the language and then discussing a series of concrete implementations that utilize EmotionML.

Keywords: emotionml, applications, sentiment

1. Introduction

We present EmotionML[1], a new W3C recommendation to represent emotion related states in data processing systems as well as a series of concrete implementations that utilize EmotionML.

EmotionML was developed by a subgroup of the W3C MMI (Multimodal Interaction) Working Group chaired by Deborah Dahl in a first version from approximately 2005 until 2013, most of this time the development was lead by Marc Schröder.

In the scientific literature on emotion research, there is no single agreed description of emotions, not even a clear consensus on the use of terms like affect, emotion or other related phenomena. For a markup language representing emotional phenomena it therefore appears important to allow the representation of their most relevant aspects in the wider sense. Given the lack of agreement in the literature on the most relevant aspects of emotion, it is inevitable to provide a relatively rich set of descriptive mechanisms.

The working group iteratively extracted requirements on the markup language from a number of 39 collected use cases[2]. Based on these requirements, a syntax for EmotionML has been produced.

It is possible to use EmotionML both as a standalone markup and as a plug-in annotation in different contexts. Emotions can be represented in terms of four types of descriptions taken from the scientific literature: categories, dimensions, appraisals, and action tendencies, with a single <emotion> element containing one or more of such descriptors.

The first part of the paper deals with a short summary of EmotionML by describing selected aspects and the procedure and thinking behind its development. The second half introduces a number of applications that integrated EmotionML and were submitted as implementation reports during the W3C recommendation track process.

2. Overview of EmotionML

Based on the requirements, a syntax for EmotionML (Schröder et al., 2012) has been produced in a sequence of steps.

The following snippet exemplifies the principles of the EmotionML syntax (Burkhardt et al., 2013).

```
<sentence id="sent1">
  Do I have to go to the dentist?
</sentence>
<emotion xmlns="http://www.w3.org/
  2009/10/emotionml" category-set=
  "http://.../xml#everyday-categories">
  <category name="afraid" value="0.4"/>
  <reference role="expressedBy"
     uri="#sent1"/>
</emotion>
```

The following properties can be observed.

- The emotion annotation is self-contained within an <emotion> element;
- all emotion elements belong to a specific namespace;
- it is explicit in the example that emotion is represented in terms of categories;
- it is explicit from which category set the category label is chosen;
- the link to the annotated material is realized via a reference using a URI, and the reference has an explicit role.

2.1. Design principles: self-contained emotion annotation

EmotionML is conceived as a plug-in language, with the aim to be usable in many different contexts. Therefore, proper encapsulation is essential. All information concerning an individual emotion annotation is contained within a single <emotion> element. All emotion markup belongs to a unique XML namespace. EmotionML differs from many other markup languages in the sense that it does not *enclose* the annotated material. In order to link the emotion markup with the annotated material, either the reference mechanism in EmotionML or another mechanism external to EmotionML can be used.

A top-level element <emotionml> enables the creation of stand-alone EmotionML documents, essentially grouping a number of emotion annotations together, but also providing document-level mechanisms for annotating global

[1]http://www.w3.org/TR/emotionml/

[2]http://www.w3.org/2005/Incubator/emotion/XGR-emotion/#AppendixUseCases

meta data and for defining emotion vocabularies (see below). It is thus possible to use EmotionML both as a standalone markup and as a plug-in annotation in different contexts.

2.2. Representations of emotion

Emotions can be represented in terms of four types of descriptions taken from the scientific literature (Schröder et al., 2011): `<category>`, `<dimension>`, `<appraisal>`, and `<action-tendency>`. An `<emotion>` element can contain one or more of these descriptors; each descriptor must have a `name` attribute and can have a `value` attribute indicating the intensity of the respective descriptor. For `<dimension>`, the `value` attribute is mandatory, since a dimensional emotion description is always a position on one or more scales; for the other descriptions, it is possible to omit the `value` to only make a binary statement about the presence of a given category, appraisal or action tendency.
The following example illustrates a number of possible uses of the core emotion representations.

```
<category name="affectionate"/>
<dimension name="valence" value="0.9"/>
<appraisal name="agent-self"/>
<action-tendency name="approach"/>
```

2.3. Mechanism for referring to an emotion vocabulary

Since there is no single agreed-upon vocabulary for each of the four types of emotion descriptions, EmotionML provides a mandatory mechanism for identifying the vocabulary used in a given `<emotion>`. The mechanism consists in attributes of `<emotion>` named `category-set`, `dimension-set`, etc., indicating which vocabulary of descriptors for annotating categories, dimensions, appraisals and action tendencies are used in that emotion annotation. These attributes contain a URI pointing to an XML representation of a vocabulary definition. In order to verify that an emotion annotation is valid, an EmotionML processor must retrieve the vocabulary definition and check that every `name` of a corresponding descriptor is part of that vocabulary.
Some vocabularies are suggested by the W3C (Schröder et al., 2012) and to make EmotionML documents interoperable users are encouraged to use them.

2.4. Meta-information

Several types of meta-information can be represented in EmotionML.
First, each emotion descriptor (such as `<category>`) can have a `confidence` attribute to indicate the expected reliability of this piece of the annotation. This can reflect the confidence of a human annotator or the probability computed by a machine classifier. If several descriptors are used jointly within an `<emotion>`, each descriptor has its own `confidence` attribute. For example, it is possible to have high confidence in, say, the arousal dimension but be uncertain about the pleasure dimension:

```
<emotion dimension-set="http://www.w3.
org/TR/emotion-voc/xml#pad-dimensions">
    <dimension name="arousal"
       value="0.7" confidence="0.9"/>
    <dimension name="pleasure"
       value="0.6" confidence="0.3"/>
</emotion>
```

Each `<emotion>` can have an `expressed-through` attribute providing a list of modalities through which the emotion is expressed. Given the open-ended application domains for EmotionML, it is naturally difficult to provide a complete list of relevant modalities. The solution provided in EmotionML is to propose a list of human-centric modalities, such as `gaze`, `face`, `voice`, etc., and to allow arbitrary additional values. The following example represents a case where an emotion is recognized from, or to be generated in, face and voice:

```
<emotion category-set="http://.../xml
  #everyday-categories"
     expressed-through="face voice">
    <category name="satisfaction"/>
</emotion>
```

For arbitrary additional meta data, EmotionML provides an `<info>` element which can contain arbitrary XML structures. The `<info>` element can occur as a child of `<emotion>` to provide local meta data, i.e. additional information about the specific emotion annotation; it can also occur in standalone EmotionML documents as a child of the root node `<emotionml>` to provide global meta data, i.e. information that is constant for all emotion annotations in the document. This can include information about sensor settings, annotator identities, situational context, etc.

2.5. References to the "rest of the world"

Emotion annotation is always *about* something. There is a subject "experiencing" (or simulating) the emotion. This can be a human, a virtual agent, a robot, etc. There is observable behavior expressing the emotion, such as facial expressions, gestures, or vocal effects. With suitable measurement tools, this can also include physiological changes such as sweating or a change in heart rate or blood pressure. Emotions are often caused or triggered by an identifiable entity, such as a person, an object, an event, etc. More precisely, the appraisals leading to the emotion are triggered by that entity. And finally, emotions, or more precisely the emotion-related action tendencies, may be directed towards an entity, such as a person or an object.
EmotionML considers all of these external entities to be out of scope of the language itself; however, it provides a generic mechanism for referring to such entities. Each `<emotion>` can use one or more `<reference>` elements to point to arbitrary URIs. A `<reference>` has a `role` attribute, which can have one of the following four values: `expressedBy` (default), `experiencedBy`, `triggeredBy`, and `targetedAt`. Using this mechanism, it is possible to point to arbitrary entities filling the above-mentioned four roles; all that is required is that these entities be identified by a URI.

2.6. Time

Time is relevant to EmotionML in the sense that it is necessary to represent the time during which an emotion annotation is applicable. In this sense, temporal specification complements the above-mentioned reference mechanism.
Representing time is an astonishingly complex issue. A number of different mechanisms are required to cover the range of possible use cases. First, it may be necessary to link to a time span in media, such as video or audio recordings. For this purpose, the `<reference role="expressedBy">` mechanism can use a so-called Media Fragment URI to point to a time span within the media. Second, time may be represented on an absolute or relative scale. Absolute time is represented in milliseconds since 1 January 1970, using the attributes `start`, `end` and `duration`. Absolute times are useful for applications such as affective diaries, which record emotions throughout the day, and whose purpose it is to link back emotions to the situations in which they were encountered. Other applications require relative time, for example time since the start of a session. Here, the mechanism borrowed from EMMA is the combination of `time-ref-uri` and `offset-to-start`. The former provides a reference to the entity defining the meaning of time 0; the latter is time, in milliseconds, since that moment.

2.7. Representing continuous values and dynamic changes

A mentioned above, the emotion descriptors `<category>`, `<dimension>`, etc. can have a `value` attribute to indicate the position on a scale corresponding to the respective descriptor. In the case of a dimension, the value indicates the position on that dimension, which is mandatory information for dimensions; in the case of categories, appraisals and action tendencies, the value can be optionally used to indicate the extent to which the respective item is present.
In all cases, the `value` attribute contains a floating-point number between 0 and 1. The two end points of that scale represent the most extreme possible values, for example the lowest and highest possible positions on a dimension, or the complete absence of an emotion category vs. the most intense possible state of that category.
The `value` attribute thus provides a fine-grained control of the position on a scale, which is constant throughout the temporal scope of the individual `<emotion>` annotation. It is also possible to represent changes over time of these scale values, using the `<trace>` element which can be a child of any `<category>`, `<dimension>`, `<appraisal>`, or `<action-tendency>` element. This makes it possible to encode trace-type annotations of emotions as produced.

3. Selected Applications

This section discusses several implementations that integrated EmotionML. Common to them is that they were submitted as an implementation report to the W3C during the recommendation track process[3]. The implementations concern very different aspects of emotion related machine processing, which reflects the diversity of the field. We categorize them in four areas: research related, core libraries, frameworks, and commercial applications.

3.1. Research related

These applications deal primarily with research questions on the nature of emotion related states.

3.1.1. EMO20Q

Emotion twenty questions (EMO20Q) is an experimental dialog game that is used to study how people describe emotions with language. By gamifying the question-asking discourse and collecting large amounts of data, EMO20Q aims to to define emotion words through crowd-sourcing (Kazemzadeh et al., 2011). Storing the belief state in EmotionML makes it possible to persist the agent's belief state in cases where the dialog is implemented in a transactional setting, such as HTTP where the agent's context must be reloaded for each request.

3.1.2. Gtrace

Gtrace (General Trace program) by the Queen's University Belfast is the successor to FEELtrace and the tools used to label the HUMAINE database (Cowie and Douglas-Cowie, 2012). It allows users to play a video of a person and create "traces" which show how the person's emotions appear to be changing over time. It includes over 50 scales, and also allows users to create their own. Alternative ways of using the scales are provided. It runs on current versions of Windows. A manual provides broad background as well as instructions for use. The system currently implements EmotionML by tracing for category and dimensional descriptors.

3.2. Libraries

Some libraries for different programming languages have already been developed by the community. In addition, there is also one for Java from Alexandre Denis at LORIA (Nancy, France)[4] and a library to check on the validity of EmotionML documents by Marc Schröder[5].

3.2.1. C# library

The EmotionML C# library[6] was developed at the University of Chemnitz as part of a project dealing with emoticons like smileys or emojis and the issues of this kind of emotion representation during the interaction in an intercultural text based chat (Fobe, 2012).
With the help of the integrated EmotionML-parser it is possible to create related object instances automatically. Furthermore object instances can be converted to EmotionML as well (in DOM and XML mode). Beside a standalone EmotionML document the plug-in version for the inclusion of emotions in other languages is supported.

[3] http://www.w3.org/2002/mmi/2013/emotionml-ir/

[4] http://code.google.com/p/loria-synalp-emotionml/

[5] https://github.com/marc1s/emotionml-checker-java

[6] https://github.com/gfobe/EmotionML-Lib-CSharp

3.2.2. EMLPy

EMLPy is a generator library for EmotionML documents[7]. It is a Python based library intended as a utility to be invoked from other EmotionML programs. EMLPy generates EmotionML documents by transforming the user specified and populated Python object tree into a XML representation. EMLPy performs EmotionML checks covered in assertions while executing this object to XML transformation. From an API perspective, the user interacts with an object tree hierarchy that maps directly to an EmotionML hierarchy of elements and attributes. EMLPy validates the object tree and its properties against the EmotionML schema and specification rules.

3.3. Frameworks

The following examples illustrate the use as part of a larger framework used in different contexts.

3.3.1. ALMA

ALMA EmotionML is an extension extension to the ALMA computational model of affect. ALMA allows the the real-time simulation of three basic types of affective features that humans can experience: (1) emotions reflect discrete short-term affect that decays after a short period of time; (2) moods reflect continuous medium-term affect, which is generally not related to a concrete event, action, or object; and (3) personality reflects discrete individual differences in mental characteristics and affective dispositions. The simulation is based on situational appraisal of the current situation according to the cognitive model of emotions created by Ortony, Clore, and Collins (OCC) (Ortony et al., 1988). ALMA combines this with the Big Five model of personality (McCrae and John, 1992) and a simulation of mood based on the Pleasure, Arousal, and Dominance (PAD) model (Mehrabian, 1996).

The ALMA EmotionML[8] implementation supports most of the EmotionML standard: (1) appraisal representation, (2) discrete and continuous emotion and mood representation, and (3) PAD and OCC emotional vocabularies. All computational output, e.g. intensities of current active emotions, or the current mood are described in an EmotionML representation. The EmotionML extension allows a fine-grained control of affect related body behavior of virtual characters, like emotional facial expressions or mood related posture control.

3.3.2. WASABI

WASABI[9] is an architecture for affect simulation for believable interactivity (Becker-Asano., 2008). It was initially developed to enhance the believability of the virtual human MAX at University of Bielefeld. Since then, it was integrated into several virtual and robotic agent systems (Becker-Asano, 2014). It realizes the concurrent simulation of emotion dynamics based on the interaction between emotion and mood and it utilizes the PAD emotional vocabulary. Its specification uses EmotionML extended by several `<info>` elements to define WASABI-specific parameters. Its UDP-based network output can be configured to represent its internal dynamics in terms of `<dimension>` elements in combination with the `<trace>` element. Thereby, it has proven easy to adjust WASABI's configuration to the project's needs and to interface it with other soft- and hardware modules, such as MARY TTS.

3.3.3. MARY TTS

MARY TTS[10] is an open-source, multilingual text-to-speech synthesis platform that includes modules for expressive speech synthesis (Charfuelan and Steiner, 2013). Particularly the support for both categorical and dimensional representations of emotions by EmotionML is important to MARY's expressive speech synthesis. These categories and dimensions are implemented by modifying the predicted pitch contours, pitch level, and speaking rate.

Using this approach, expressive synthesis is most effective when using HMM-based voices, since the statistical parametric synthesis framework allows appropriate prosody to be realized with consistent quality. Expressive unit-selection voices support EmotionML best if they are built from multiple-style speech databases (Steiner et al., 2013), which preserves intonation and voice quality better than when applying signal manipulation to conventional unit-selection output.

3.4. Applications

Lastly, the following lists commercial applications that utilize EmotionML to represent emotion related models.

3.4.1. NViso

NViso uses emotion detection to analyze customer reaction on brands and (web) interfaces (nViso, 2011). It provides a cloud service to measure instantaneous emotional reactions of consumers in online environments and thus provides real-time information for Market Research, Brands, Creative Agencies and R&D Product Development.

The NViso 3D Facial Imaging API is an online service for the recognition of emotions depicted through facial expressions in still images and videos. The focus of the integration of EmotionML into the tool is on using the media type and URI time for video.

3.4.2. Speechalyzer

The Speechalyzer by Deutsche Telekom Laboratories is an open source project[11] for analysis, annotation and transcription of speech files (Burkhardt, 2011). It can be used to rapidly judge large numbers of audio files emotionally, an automatic classification is integrated. The Speechalyzer was part of a project to identify disgruntled customers in an automated voice service portal (Burkhardt et al., 2009) with two use cases in mode: a) transfer angry users to a trained human agent, and b) gain some statistic insight on the number of angry customers at the end of each day. It utilizes EmotionML as an exchange format to import and export emotionally annotated speech data.

[7] https://github.com/ebegoli/EMLPy

[8] ALMA is freely available for download: http://www.dfki.de/~gebhard/alma

[9] https://github.com/CBA2011

[10] http://mary.dfki.de/ and https://github.com/marytts

[11] https://github.com/dtag-dbu/speechalyzer

4. Conclusions

We presented EmotionML, a new W3C recommendation to represent emotion related states. The first part of the paper deals with a short summary of EmotionML and the second half introduces a number of applications that integrated EmotionML and were submitted as implementation reports during the W3C recommendation track process. We hope this article encourages the reader to use EmotionML in own projects and give feedback to the W3C to pave the way towards EmotionML version 2.0.

5. References

Becker-Asano., C. (2008). *WASABI: Affect Simulation for Agents with Believable Interactivity*. Ph.D. thesis, University of Bielefeld.

Becker-Asano, C. (2014). WASABI for affect simulation in human-computer interaction. In *Proc. International Workshop on Emotion Representations and Modelling for HCI Systems*.

Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., and Huber, R. (2009). Detecting real life anger. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

Burkhardt, F., Schröder, M., Baggia, P., Pelachaud, C., Peter, C., and Zovato, E. (2013). W3C Emotion Markup Language (EmotionML) 1.0 proposed recommendation. http://www.w3.org/TR/emotionml/.

Burkhardt, F. (2011). Speechalyzer: a software tool to process speech data. In *Proc. Elektronische Sprachsignalverarbeitung*.

Charfuelan, M. and Steiner, I. (2013). Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In *Proc. Interspeech*.

Cowie, Roddy, G. M. and Douglas-Cowie, E. (2012). Tracing emotion: An overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1):1–17.

Fobe, G. (2012). Serialisierung von Emotionen in der textuellen Kommunikation. Master's thesis, Technical University of Chemnitz.

Kazemzadeh, A., Lee, S., Georgiou, P. G., and Narayanan, S. S. (2011). Emotion twenty questions: Toward a crowd-sourced theory of emotions. In *Proc. Affective Computing and Intelligent Interaction (ACII)*.

McCrae, R. and John, O. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

nViso. (2011). http://nviso.ch.

Ortony, A., Clore, G. L., and Collins., A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Schröder, M., Pirker, H., Lamolle, M., Burkhardt, F., Peter, C., and Zovato, E. (2011). Representing emotions and related states in technological systems. In Petta, P., Cowie, R., and Pelachaud, C., editors, *Emotion-Oriented Systems – The Humaine Handbook*, pages 367–386. Springer.

Schröder, M., Pelachaud, C., Ashimura, K., Baggia, P., Burkhardt, F., Oltramari, A., Peter, C., and Zovato, E. (2012). Vocabularies for EmotionML. http://www.w3.org/TR/emotion-voc/.

Steiner, I., Schröder, M., and Klepp, A. (2013). The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech. In *Proc. Phonetik & Phonologie 9*.

Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources

Gabriela Vulcu, Paul Buitelaar, Sapna Negi, Bianca Pereira, Mihael Arcan, Barry Coughlan, J. Fernando Sanchez, Carlos A. Iglesias

Insight, Centre for Data Analytics, Galway, Ireland

gabriela.vulcu@insight-center.org, paul.buitelaar@insight-center.org, sapna.negi@insight-center.org, bianca.pereira@insight-center.org, mihael.arcan@insight-center.org, b.coughlan2@gmail.com,

Universidad Politecnica de Madrid, Spain

jfernando@gsi.dit.upm.es, cif@dit.upm.es

Abstract

We present a methodology for legacy language resource adaptation that generates domain-specific sentiment lexicons organized around domain entities described with lexical information and sentiment words described in the context of these entities. We explain the steps of the methodology and we give a working example of our initial results. The resulting lexicons are modelled as Linked Data resources by use of established formats for Linguistic Linked Data (lemon, NIF) and for linked sentiment expressions (Marl), thereby contributing and linking to existing Language Resources in the Linguistic Linked Open Data cloud.

Keywords: domain specific lexicon, entity extraction and linking, sentiment analysis

1. Introduction

In recent years, there has been a high increase in the use of commercial websites, social networks and blogs which permitted users to create a lot of content that can be reused for the sentiment analysis task. However the development of systems for sentiment analysis which exploit these valuable resources is hampered by difficulties to access the necessary language resources for several reasons: (i) language resource owners fear for losing competitiveness; (ii) lack of agreed language resource schemas for sentiment analysis and not normalised magnitudes for measuring sentiment strength; (iii) high costs for adapting existing language resources for sentiment analysis; (iv) reduced visibility, accessibility and interoperability of the language resources with other language or semantic resources like the Linguistic Linked Open Data cloud (i.e. LLOD). In this paper we are focusing on the second and the forth challenges by describing a methodology for the conversion, enhancement and integration of a wide range of legacy language and semantic resources into a common format based on the lemon[1](McCrae et al., 2012) and Marl [2] (Westerski et al., 2011) Linked Data formats.

1.1. Legacy Language Resources

We identified several categories of legacy language resources with respect to our methodology: domain-specific English review corpora, non-English review corpora, sentiment annotated dictionaries and Wordnets. The existing legacy language resources (gathered in the EUROSENTIMENT project [3]) are available in many formats and they contain several types of annotations that are relevant for the sentiment analysis task. The language resources formats range from plain text with or without custom made annotations, HTML, XML, EXCEL, TSV, CSV to RDF/XML. The language resources annotations are all or a subset of: *domain* - the broad context of the review corpus (i.e. 'hotel' is the domain for the TripAdvisor corpus); *language* - the language of the language resource; *context entities* - relevant entities in the corpus; *lemma* - lemma annotations of the relevant entities; *POS* - part-of-speach annotations of the relevant entities; *WordNet synset* - annotations with existing synsets from Wordnet of the relevant entities; *sentiment* - positive or negative sentiment annotation both at sentence level and or at entity level; *emotion* - more fine grained polarity values both expressed as numbers or as concepts from well defined ontologies; *inflections* - morphosyntactic annotations of the relevant entities.

1.2. Methodology for LR Adaptation and Sentiment Lexicon Generation

Our method generates domain-specific sentiment lexicons from legacy language resources and enriching them with semantics and additional linguistic information from resources like DBpedia and BabelNet. The language resources adaptation pipeline consists of four main steps highlighted by dashed rectangles in Figure 1: (i) the Corpus Conversion step normalizes the different language resources to a common schema based on Marl and NIF[4]; (ii) the Semantic Analysis step extracts the domain-specific entity classes and named entities and identifies links between these entities and concepts from the LLOD Cloud; (iii) the Sentiment Analysis step extracts contextual sentiments and identifies SentiWordNet synsets corresponding to these contextual sentiment words; (iv) the Lexicon Generator step uses the results of the previous steps, enhances them with multilingual and morphosyntactic information and converts the results into a lexicon based on the lemon and Marl formats. Different language resources are processed with variations of the given adaptation pipeline. For example the domain-specific English review corpora are

[1] http://lemon-model.net/lexica/pwn/

[2] http://www.gi2mo.org/marl/0.1/ns.html

[3] http://eurosentiment.eu/

[4] http://persistence.uni-leipzig.org/nlp2rdf/

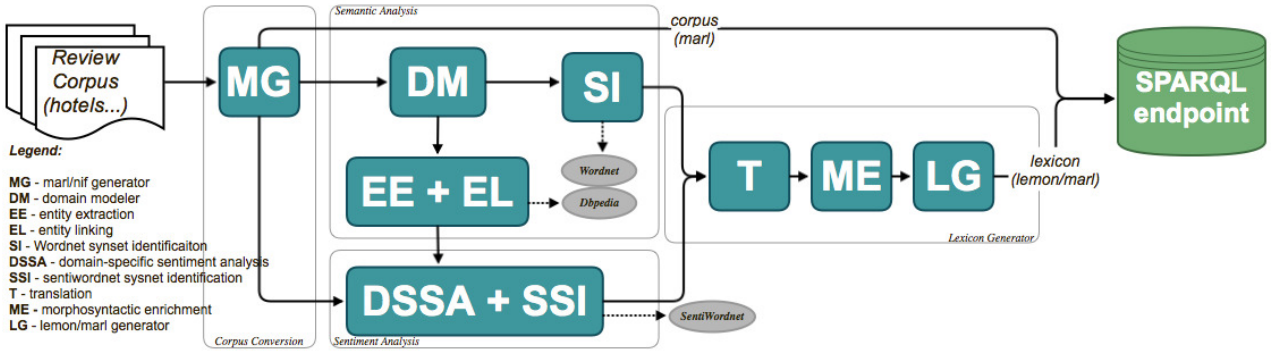


Figure 1: Methodology for Legacy Language Resources Adaptation for Sentiment Analysis.

processed using the pipeline described in Figure 1 while the sentiment annotated dictionaries are converted to the lemon/Marl format using the Lexicon Generator step. We detail these steps in the subsequent sections.

2. Corpus conversion

Due to the formats heterogeneity of the legacy language resources we need a common model that captures all the existing annotations in a structural way. The Corpus Conversion step adapts corpus resources to a common schema. We defined a schema based on the NIF and Marl formats that structures the annotations from the corpora reviews. For example each review in the corpus is an entry that can have overall sentiment annotations or annotations at the substring level. The Corpus Generator has been designed to be extensible and to separate the technical aspects from the content and formats being translated.

3. Semantic analysis

The Semantic Analysis step consists of: Domain Modeller (DM), Entity Extraction (EE), Entity Linking (EL) and Synset Identification (SI) components. The DM extracts a set of entity class using a pattern-based term extraction algorithm with a generic domain model (Bordea, 2013) on each document, aggregates the lemmatized terms and computes their ranking in the corpus(Bordea et al., 2013). The EE and EL components are based on AELA (Pereira et al., 2013) framework for Entity Linking that uses a Linked Data dataset as reference for entity mentioning identification, extraction and disambiguation. By default, DBPedia and DBPedia Lexicalization (Mendes et al., 2011) are used as reference sources but domain-specific datasets could be used as well. The SI identifies and disambiguates WordNet synsets that match with the extracted entity classes. It extends each candidate synset with their direct hyponym and hypernym synsets. Then we compute the occurrence of a given entity class in each of these bag of words. We choose the synset with the highest occurrence score for an entity class.

4. Sentiment analysis

The Sentiment Analysis step consists of: Domain-Specific Sentiment Polarity Analysis (DSSA) and Sentiment Synset Identification (SSI) components. The DSSA component identifies a set of sentiment words and their polarities in the context of the entities identified in the Semantic Analysis step. The clause in which a entity mention occurs is considered the span for a sentiment word/phrase in the context of that entity. The DSSA is based on earlier research on sentiment analysis for identifying adjectives or adjective phrases (Hu and Liu, 2004), adverbs (Benamara et al., 2007), two-word phrases (Turney and Littman, 2005) and verbs (Subrahmanian and Reforgiato, 2008). Particular attention is given to the sentiment phrases which can represent an opposite sentiment than what they represent if separated into individual words. For example, 'ridiculous bargain' represents a positive sentiment while 'ridiculous' could represent a negative sentiment. Sentiment words/phrases in individual reviews are assigned polarity scores based on the available user ratings. In case of language resources with no ratings we use a bootstrapping process based on Sentiwordnet that will rate the domain aspects in the review. We select the most frequent scores as the final sentiment score for a sentiment word/phrase candidate based on its occurrences in all the reviews. The SSI component identifies SentiWordNet synsets for the extracted contextual sentiment words. The sentiment phrases however, are not assigned any synset. Linking the sentiment words with those of SentiWordNet further enhances their semantic information. We identify the nearest SentiWordNet sense for a sentiment candidate using Concept-Based Disambiguation (Raviv and Markovitch, 2012) which utilizes the semantic similarity measure 'Explicit Semantic Analysis' (Gabrilovich and Markovitch, 2006) to represent senses in a high-dimensional space of natural concepts. Concepts are obtained from large knowledge resources such as Wikipedia, which also covers domain specific knowledge. We compare the semantic similarity scores obtained by computing semantic similarity of a bag of words containing domain name, entity and sentiment word with bags of words which contain members of the synset and the gloss for each synset of that SentiWordNet entry. The synset with the highest similarity score above a threshold is considered.

5. Lexicon generator

The Lexicon Generator step consists of: MorphoSyntactic Enrichment (ME), Machine Translation(T) and lemon/Marl Generator(LG) components. As WordNet does not provide

Sentiment	PolarityValue	Context
”good”@en	”1.0”	”alarm”@en
”damaged”@en	”-2.0”	”apple”@en
”amazed”@en	”2.0”	”flash”@en
”expensive”@en	”-1.0”	”flash”@en
”annoying”@en	”-1.5”	”player”@en

Table 1: Sentiment words the 'electronics' domain.

Sentiment	PolarityValue	Context
”warm”@en	”2.0”	”pastries”@en
”warm”@en	”2.0”	”comfort”@en
”warm”@en	”1.80”	”restaurant”@en
”warm”@en	”1.73”	”service”@en
”warm”@en	”0.98”	”hotel”@en

Table 2: Sentiment word 'warm' in the 'hotel' domain.

any morphosyntactic information (besides part of speech), such as inflection and morphological or syntactic decomposition, the ME provides a further process for the conversion and integration of lexical information for selected synsets from other legacy language resources like CELEX [5]. Next, the T component translates extracted entity classes and sentiment words in other languages using a domain-adaptive machine translation approach (Arcan et al., 2013). This way we can build sentiment lexicons in other languages. It uses the SMT toolkit Moses (Koehn et al., 2007). Word alignments are built with the GIZA++ toolkit (Och and Ney, 2003), where a 5-gram language model was built by SRILM with Kneser-Ney smoothing (Stolcke, 2002). We use two different parallel resources: the JRC-Acquis (Steinberger et al., 2006) available in almost every EU official language (except Irish) and the OpenSubtitles2013 (Tiedemann, 2012) which contains fan-subtitled text for the most popular language pairs. The LG component converts the results of the previous components (named entities and entity classes linked to LOD and sentiment words with polarity values) to a domain-specific sentiment lexicon represented as RDF in the lemon/Marl format. The lemon model was developed in the Monnet project to be a standard for sharing lexical information on the semantic web. The model draws heavily from earlier work, in particular from LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006). The Marl model is a standardised data schema designed to annotate and describe subjective opinions.

6. Working Example

Figure 2 shows an example of a generated lexicon for the domain 'hotel' in English. It shows 3 *lemon:LexicalEntries*: 'room ' (entity class), 'Paris' (named entity) and 'small' (sentiment word) which in the context of the lexical entry 'room' has negative polarity. Each of them consists of senses, which are linked to DBpedia and/or Wordnet concepts.

We applied our methodology on an annotated corpus of 10.000 reviews for the hotel domain and an annotated corpus of 600 reviews for the electronics domain. Table 1 shows an example of sentiment words from the 'electronics' domain, while Table 2 shows an example of different contexts of the sentiment word 'warm' with their corresponding polarities in the 'hotel' domain.

7. Future Work

We are currently working on evaluating the Semantic Analysis and Sentiment Analysis components by participating in the SemEval challenge [6] on aspect-based sentiment analysis. We also plan to investigate ways of linking the extracted named entities with other Linked Data datasets like Yago or Freebase. A next step for the use of our results is to aggregate sentiment lexicons obtained from Language Resources on the same domain.

8. Conclusions

In this paper we presented a methodology for creating domain-specific sentiment lexicons from legacy Language Resources, described the components of our methodology and provided example results.

9. Acknowledgements

This work has been funded by the European project EUROSENTIMENT under grant no. 296277.

10. References

Arcan, M., Thomas, S. M., Brandt, D. D., and Buitelaar, P. (2013). Translating the FINREP taxonomy using a domain-specific corpus. Poster presented at the Machine Translation Summit XIV, Nice, France.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM'07.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, TIA'13, Paris, France.

Bordea, G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. Ph.D. thesis, National University of Ireland, Galway.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006). Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia. ACL.

Gabrilovich, E. and Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, AAAI'06. AAAI Press.

[5] http://celex.mpi.nl/

[6] http://alt.qcri.org/semeval2014/

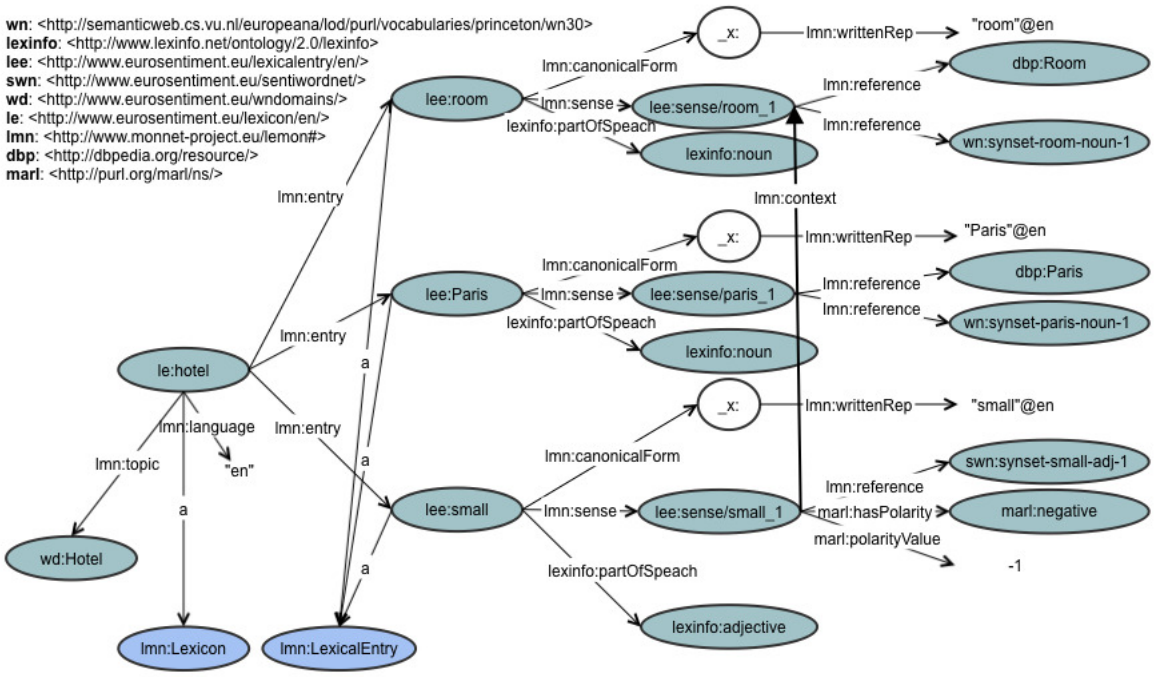


Figure 2: Example lexicon for the domain 'hotel' in English.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, New York, NY, USA. ACM.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Stroudsburg, PA, USA. ACL.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, New York, NY, USA. ACM.

Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., and Peters, W. (2008). Modelling multilinguality in ontologies. In *Poster at COLING'10*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, March.

Pereira, B., Aggarwal, N., and Buitelaar, P. (2013). Aela: An adaptive entity linking approach. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW'13, Republic and Canton of Geneva, Switzerland.

Raviv, A. and Markovitch, S. (2012). Concept-based approach to word-sense disambiguation. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufis, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'06.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing (ICSLP 2002*.

Subrahmanian, V. and Reforgiato, D. (2008). Ava: Adjective-verb-adverb combinations for sentiment analysis. *Intelligent Systems*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12, Istanbul, Turkey. ELRA.

Turney, P. D. and Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*.

Westerski, A., Iglesias, C. A., and Tapia, F. (2011). Linked Opinions: Describing Sentiments on the Structured Web of Data. In *Proceedings of the 4th International Workshop Social Data on the Web*.

Hedging and Speaker Commitment

Anna Prokofieva, Julia Hirschberg
Columbia University
prokofieva@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Hedging is a behavior wherein speakers or writers attempt to distance themselves from the proposition they are communicating. Hedge terms include items such as "***I think*** *X*" or "*It's* ***sort of*** *Y*". Identifying such behaviors is important for extracting meaning from speech and text, and can also reveal information about the social and power relations between the conversants. Yet little research has been done on the automatic identification of hedges since the CONLL 2010 Shared Task. In this paper, we present our newly expanded and generalized guidelines for the annotation of hedge expressions in text and speech. We describe annotation and automatic extraction experiments using these guidelines and describe future work on the automatic identification of hedges.

Keywords: hedging, annotation guidelines, crowd-sourced annotation

1. Introduction

Hedging is a phenomenon in which a speaker communicates a lack of commitment to what they are saying. For example:

(1) "I ***think*** *it's a* ***little*** *odd."*

This phrase contains two hedges,"think" and "a little"; one indicating the speaker's lack of commitment to the proposition "it's a little odd" and the other indicating lack of commitment to the quality of oddness.

Hedges occur quite commonly in text and speech: Prince et al. (1982) noted that hedges occurred about every 15 seconds in their 12-hour medical corpus. Since people may hedge for many reasons - for example, to save face (Prince et al., 1982), to show politeness (Ardissono et al., 1999), or to appear more cooperative (Vasilieva, 2004) - the study of hedging behaviors can give us important insight into conversational dynamics. They are also thought to correlate with power relations between conversational participants in domains such as the medical hierarchy. Our goal is to develop procedures for automatically classifying hedges in text and speech corpora so that we can better define speaker commitments and relationships. To this end we have developed hedging annotation guidelines expanding upon previous work, which we are using for semi-automated corpus annotation.

2. Previous work

Lakoff (1975) originally defined hedges as words "whose job it is to make things fuzzier". Prince et al. (1982) noted that this 'fuzziness' could be manifested in two ways: as fuzziness within the propositional content, or as fuzziness in the relationship between the propositional content and the speaker. These two types of hedges are thus termed *propositional* and *relational*.

Others have expanded this notion of 'fuzziness' to encompass words that signal uncertainty, a lack of precision or non-specificity, or an attempt to downplay speakers' commitment to elements in an utterance. Previous studies of hedging have found that the phenomenon is correlated with many discourse functions, such as attempting to evade questions and avoid criticism (Crystal, 1988). de Figueiredo-Silva (2001) proposed viewing hedging as a manifestation of the speaker's attitude towards a claim and towards their audience. As such, hedging can be viewed as an expression of the speaker's inner state.

On the other hand, we can also look at hedging from the listeners' perspective, since the use of hedge words (or the lack thereof) can shape the listeners' opinion of the speaker and of their argument (Blankenship and Holtgraves, 2005; Hosman and Siltanen, 2006; Erickson et al., 1978). In this way, hedges are part of a feedback loop in conversational dynamics.

To date, most of the exploration of hedging in text has been focused on the domain of academic writing (Meyer, 1997; Hyland, 1998; Varttala, 1999). The organizers of the CONLL 2010 Shared Task investigated hedging in the BioScope corpus, which contains abstracts and articles in the biomedical field. This corpus, along with a Wikipedia corpus annotated for "weasel words" (words that equivocate without communicating a precise claim), were used in the Shared Task to investigate techniques for the automatic detection of hedges (Farkas et al., 2010). This Shared Task produced the first set of detailed guidelines on hedge annotation. However, these guidelines are somewhat domain and genre-dependent.

There has also been some investigation of hedging in other corpora, although to date no additional hedge annotations have been made public (Aijmer, 1986; Poos and Simpson, 2002). There has been little work on hedging in speech, beyond Prince et al. (1982)'s study of conversations between medical personnel and patients; even in that study, the audio data was not made available to the researchers so no specific analysis of the speech itself was possible.

3. Defining Hedges

Given the prevalence and importance of hedging behavior to the interpretation of speaker commitment and other social aspects of dialogue, we have begun a study of hedging behavior with the goal of creating a more general tool for identifying hedges in text and speech. Ultimately, we want to create a corpus annotated for hedging. To this end, we have created a new set of Hedging Annotation Guidelines which are more comprehensive than the CONLL 2010

Guidelines and are applicable to both text and speech from various domains and of various levels of formality.

3.1. Domain and Genre Specificity

These guidelines have been developed and refined using several diverse corpora: the CONLL BioScope Corpus (Vincze et al., 2008), the SCOTUS Supreme Court Corpus, and the NIST Meeting Corpus (Garofolo et al., 2004). In the process, we have explored a number of challenges faced in identifying and annotating the phenomenon.

Our investigations of hedging in multiple domains and genres have shown that many terms clearly used as hedges in other corpora were not included in the CONLL guidelines. Some of the hedge terms we discovered appear to be specific to the domains our corpora represent and the linguistic conventions in those domains. In our new Guidelines, we have thus considerably expanded the set of potential hedge terms based on the hedging behaviors we have observed in these different corpora. For example, "in my opinion" is not mentioned in the CONLL guidelines as a hedge, probably because it did not appear in the corpus, but appears quite frequently in the SCOTUS Corpus as a hedge. This is due to the fact that the CONLL guidelines were meant for annotation on academic text, where expressing a personal opinion is often discouraged, whereas in the Supreme Court arguments of the SCOTUS corpus, the lawyers often hedged their views by stating something as opinion rather than fact in order to avoid criticism from the judges.

Additionally, it became clear that other hedge terms found in our corpora were specific to spoken conversation. We thus added our own observations from the SCOTUS Corpus together with those observed in other speech-focused studies to the guidelines (Prince et al., 1982). A pilot annotation on the more informal NIST meeting corpus (Garofolo et al., 2004) led us to further broaden the guidelines to include hedging instances from other selections of conversational speech. In particular, we were able to add many new multi-word hedge constructions, such as "and all that" and "something or other" to our list of hedges; these were not present in the more formal SCOTUS or BioScope corpora. This illustrates our finding that hedging is quite domain-specific and depends on the level of formality, as well as any established conventions of the domain.

3.2. Hedging and Disfluency

The CONLL Guidelines, developed for text annotation, did not include mechanisms for dealing with speech phenomena such as hesitations, self-repairs, and other disfluencies.

> *(3) "**I think** it's – I **think** it's an extremist group that's trying to make us move faster."*

In (3), there is a repetition of the hedge word; to be consistent with the standard for disfluency annotation, both instances would be marked as hedges. Our pilot annotations of the Supreme Court Corpus showed that these conversational phenomena and others, including interruptions, ungrammatical phrases and incomplete utterances, all require special handling in the annotation guidelines.

Specifically, we annotate the hedge word wherever it is at least partially formed, based on the speaker's intention as far as we can determine such from the context. It is the hope that broadening the scope of our annotation in such a way will allow a more in-depth investigation into the relationship between disfluency and hedging.

3.3. Relational vs Propositional Hedges

Based on Prince et al. (1982), we have expanded and clarified distinctions between relational and propositional hedges. Using Prince et al. (1982)'s definitions, we identify relational hedges as those that have to do with the speaker's relation to the propositional content, and propositional hedges as those that introduce uncertainty into the propositional content itself. Since these distinctions themselves can sometimes be confusing, we have provided additional questions annotators may ask themselves to make such a determination. In particular, the annotator can try to preface a potentially hedged sentence with "I'm certain" to see whether the hedge contained therein is relational or propositional.

> (4) "I'm certain that ... *his feet are **sort of** blue." (propositional hedge)*

> (5) # "I'm certain that ... *I **guess** John is right." (relational hedge)*

In (4), inserting "I'm certain" does not change the meaning of the sentence; however, in (5), such an insertion is infelicitous.

However, there is one type of relational hedge for which this test fails: this is the *attributive* hedge. In attributive hedges, a speaker attributes information to some other source in order to downplay its force (as in (6)) or to garner authoritative power for their statement (as in (7)).

> *(6) "**People I've talked to say** "Lincoln" was okay."*

> *(7) "Well, **the Encyclopedia Britannica says** that, so it must be true."*

We mark these as relational hedges, since in either case such attribution indicates a lack of commitment on the part of the speaker with respect to an entire proposition. These sorts of hedges are difficult to annotate automatically, but are nonetheless important for showing a lack of the speaker's personal investment in what they are saying.

3.4. Multi-word Hedges

Hedges can be single cue words or combinations of words. In some cases words which would not normally function as hedges do so in combination with other words. For example, the phrase "in my understanding" can serve as a hedge even though each individual word, when placed in a different context, would not. "In my mind", "my thinking is" and "if I'm understanding you correctly" are other examples of multi-word relational hedges. Multi-word propositional hedges include "and so forth" and "or something like that". Attributive hedges are most often multi-word hedges as well, since both the source to which the information is being attributed, along with the accompanying verb, are included in the hedge.

3.5. Ambiguity

One of the major difficulties in detecting hedges is that potential hedge words are inherently ambiguous. For example:

*(1) "I **think** it's a **little** odd."*

(2) "I think about you all the time."

In (1),"think" is a hedge, but not so in (2). This is true for most hedge verbs and distinguishing whether the verb is being used in a hedging context is a difficult task even for trained annotators. Moving forward, we plan to address these issues using word sense disambiguation techniques. Yarowsky (2000) successfully utilized hierarchical decision lists for a word sense disambiguation task and achieved a precision of 78.9%; we believe that such an approach, which would use lexical and syntactic features to distinguish hedge senses from non-hedge senses, would be adequate to resolve this issue.

3.6. Hedges in Questions

Due to the inherent uncertainty that questions themselves convey, the CONLL 2010 guidelines did not mark hedges in questions. However, we have found that it is in fact possible to find hedges that are independent of the overall uncertainty conveyed by the question. For example:

*(6) "What about the argument that the plaintiff **may not** have been harmed by the disclosure?"*

*(7) "Is this the type of statute that depends **largely** on private enforcement to implement it?"*

We find hedges in both wh- and yes-no questions. In (6), the speaker is questioning the validity of "the argument", but the argument itself contains a hedge ("may") that is independent of the overall uncertainty inherent in the question. In (7), the question itself expresses the speaker's uncertainty about the type of the statute, but the presence of the hedge "largely" is independent of that uncertainty.

In general, hedges should be identified in questions when the hedge words themselves do not identify the statement as a question. For example, auxiliaries that might serve as hedges in statements are not marked in questions, because their use in questions is dictated by rules of grammar rather than a desire to hedge. For example, in: *"Could you clarify this for me?"*, "could" is not marked as a hedge.

In the specific case of statements followed by tag questions, such as: *"It **might** rain, might it not?"*, "might" would be marked as a hedge in the first part of the statement (which can stand as a statement by itself), but not in the tag.

4. Data

Major revisions were necessary to make the guidelines appropriate for annotating text as well as speech, which suggests that hedging may be domain specific. To that end, we wanted to compare whether hedging was more or less prevalent in formal speech as compared to informal speech. We obtained gold standard annotations as per our latest iteration of the annotation guidelines on the Supreme Court Corpus (an instance of less conversational, more formal speech) to compare the presence of hedging therein to the hedging found in the NIST Meeting Corpus (arguably a much more informal, conversational setting).

	SCOTUS	NIST
% Turns with Hedges	38.5%	23.5%
% Sentences with Hedges	23.0%	16.9%
% hRel	71.4%	53.4%
% hProp	28.6%	46.6%

Table 1: Presence of hedges in the SCOTUS and NIST Meeting corpora.

These results were surprising given that we expected more hedging in informal speech. However, the high percentage of relational hedges in the SCOTUS corpus can be explained by the fact that lawyers frequently used "I think" when responding to the judges' queries; this can also account for the higher percentage of hedging in general in that corpus.

5. Automatic Hedge Detection

While our guidelines focus on the lexical items which **may** serve as hedges, they rely upon human interpretation of the context in which potential hedge terms occur in order to determine whether an item is being used as a hedge or not. To understand the importance of this disambiguation process to the identification of hedges, we performed a small experiment in automatic hedge detection.

Our pilot annotation of meetings from the NIST Meeting Corpus has given us a small seed of gold standard data. To motivate the necessity of creating a smart algorithm for the automatic detection of hedges, as opposed to a keyword-search approach, we ran a simple lexical-based search for potential hedges on those meetings. The keywords used were hedges mentioned in the CONLL 2010 Guidelines and those found in a previous annotation exercise we had done on the Supreme Court Corpus.

	NIST Corpus
Precision	0.45
Recall	0.66
F-score	0.53

Table 2: Keyword search approach to hedge detection.

These results provide some evidence that hedge detection requires more than simple key-word search. In the majority of cases, words that are identified by the lexical search as hedges are actually not hedges in that particular context. Moreover, only two-thirds of the hedge terms identified by our labelers in the NIST Meeting Corpus had been previously seen in other corpora. Thus, successful hedge detection will need to involve not only disambiguation of potential hedge terms but also methods to identify new ways of expressing this phenomenon.

Given that annotating hedging can be complicated and time-consuming, we are exploring the potential for crowd-

sourcing hedge annotation, using Amazon Mechanical Turk (AMT). However, as with any complex task, this will require careful planning in order to obtain reliable annotations from untrained labelers. Currently we are developing a multi-stage strategy to incorporate crowd-sourcing into the process of creating a large corpus annotated for hedging. We are building a rule-based algorithm from our guidelines to identify potential hedges syntactically, using terms identified by simple keyword search. These can then be checked by AMT labelers to distinguish hedge uses from non-hedge uses in a series of simple word sense disambiguation tasks. Specifically, annotators would be presented with a sentence containing a potential hedge and asked whether that word could be replaced by a synonym representing one of its potential senses.

*(1) "It's **sort of** diagonal here."*

*Does **sort of** in this sentence mean 'type of'?*

In this case, the correct answer would be 'no' and that would inform us that "sort of" was being used in a hedging sense in this sentence.

Snow et al. (2008) conducted a similar word sense disambiguation task on Amazon Mechanical Turk and were able to obtain 100% accuracy using majority voting based on 10 annotations of each word. Those sentences that are verified by multiple labelers as containing hedges in this first stage will then be passed along to the second stage of annotation. In this stage, annotators will be asked to identify the type of hedge, relational or propositional, by answering questions about the role of the hedge in the matrix sentence. We also hope to reduce the amount of annotation necessary in the first verification stage by using an active learning algorithm trained on a small seed set of gold standard annotated data in order to select the most ambiguous and difficult cases for annotation. We plan to use this additional annotated data to train a statistical classifier to disambiguate hedge uses automatically.

6. Conclusion

In this paper, we have described newly expanded and generalized guidelines for the annotation of hedge expressions in text and speech. We present a more detailed description of this phenomenon, some preliminary experimental results on annotation and automatic detection of hedges, and discuss future plans for disambiguating potential hedge terms using crowd-sourcing and, eventually, automatic machine learning methods.

7. References

Aijmer, K. (1986). Discourse variation and hedging. *Corpus Linguistics II. New studies in the analysis and exploitation of computer corpora*, pages 1–18.

Ardissono, L., Boella, G., and Lesmo, L. (1999). Politeness and speech acts. *Proc. Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, pages 41–55.

Blankenship, K. and Holtgraves, T. (2005). The role of different markers of linguistic powerlessness in persuasion. *Journal of Language and Social Psychology*, 24(1):3–24.

Crystal, D. (1988). On keeping one's hedges in order. *English Today*, 15:46–47.

de Figueiredo-Silva, M. I. R. (2001). Teaching academic reading: Some initial findings from a session on hedging. *Postgraduate Conference of the University of Edinburgh*.

Erickson, B., Lind, E., Johnson, B., and O'Barr, W. (1978). Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of Experimental Social Psychology*, 14(3):266–279.

Farkas, R., Vincze, V., Mora, G., Csirik, J., and Szarvas, G. (2010). The conll-2010 shared task: learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 1–12.

Garofolo, J. S., Laprun, C., Michel, M., Stanford, V., and Tabassi, E. (2004). The nist meeting room pilot corpus. *LREC*.

Hosman, L. A. and Siltanen, S. (2006). Powerful and powerless language forms their consequences for impression formation, attributions of control of self and control of others, cognitive responses, and message memory. *Journal of Language and Social Psychology*, 25(1):33–46.

Hyland, K. (1998). *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Lakoff, G. (1975). *Hedges: A study in meaning criteria and the logic of fuzzy concepts*. Springer, Netherlands.

Meyer, P. G. (1997). Hedging strategies in written discourse: Strengthening the argument by weakening the claim. In *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*, Berlin. Walter de Gruyter.

Poos, D. and Simpson, R. (2002). Cross-disciplinary comparisons of hedging. *Using corpora to explore linguistic variation*, 9(1).

Prince, E. F., Frader, J., and Bosk, C. (1982). On hedging in physician-physician discourse. *Linguistics and the Professions*, pages 83–97.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Varttala, T. (1999). Remarks on the communicative functions of hedging in popular scientific and specialist research articles on medicine. *English for Specific Purposes*, 18(2):177–200.

Vasilieva, I. (2004). Gender-specific use of boosting and hedging adverbs in english computer-related texts–a corpus-based study. *International Conference on Language, Politeness and Gender*, pages 2–5.

Vincze, V., Szarvas, G., Farkas, R., Mora, G., and Csirik, J. (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities*, 34(1-2):179–186.

Intelligent Systems' Holistic Evolving Analysis of Real-Life Universal Speaker Characteristics

Björn Schuller[1,2], Yue Zhang[2], Florian Eyben[2], Felix Weninger[2]

[1]Department of Computing, Imperial College London, London, U. K.
[2]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Munich, Germany
{schuller,y.zhang,eyben,weninger}@tum.de

Abstract

In this position paper we present the FP7 ERC starting grant project iHEARu (Intelligent systems' Holistic Evolving Analysis of Real-life Universal speaker characteristics). This project addresses several fundamental shortcomings in state of the art methods for computational paralinguistics, by introducing holistic analysis, evolving learning of features and models, and collection of real-life, large-scale data annotated in multiple dimensions ('universally'). We discuss the first aspect of the project, holistic analysis, in more detail, and give benchmark results using state of the art multi-target learning methods on the INTERSPEECH 2012 Speaker Trait Challenge dataset (Likability Sub-Challenge). The results clearly indicate the need for improved machine learning methods and data collection to learn holistic speaker classification.

Keywords: Computational Paralinguistics, Holistic Analysis, Multi-target Learning

1. Introduction

With recent technology advances, automatic speech recognition and synthesis have matured to the degree that they are used on a daily basis by millions of people, e.g., on their smart phones or in call services. During the next years, it is expected that speech processing technology will move to a new level of social awareness to make interaction more intuitive, speech retrieval more efficient, and lend additional competence to computer-mediated communication and speech analysis services in the commerce, health, security, and further sectors. To reach this goal, rich speaker traits and states such as age, height, personality and physical and mental states as carried by the tone of the voice and the spoken words must be reliably identified by machines.

The **iHEARu** project aims to push the limits of **i**ntelligent systems for computational paralinguistics by considering **H**olistic analysis of multiple speaker attributes at once, **E**volving and self-learning, deeper **A**nalysis of acoustic parameters - all on **R**ealistic data on a large scale, ultimately progressing from individual analysis tasks towards **u**niversal speaker characteristics analysis, which can be easily learnt about and can be adapted to new, previously unexplored characteristics.

In this paper, the state of the art in the field is described in Section 2. Next, we will introduce our long-term goals and describe the methodologies of the iHEARu project in Section 3. An in-depth discussion of holistic analysis of multiple speaker attributes is given in Section 4. Further, a first attempt on multi-target classification to improve on three paralinguistics tasks by jointly learning age, gender, and subjective likability of the voice, is presented and evaluated in Sections 5 and 6. We conclude with a summary and outlook on future research topics in Section 7.

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 289021 (STREP ASC-Inclusion).

2. State of the Art

Analysing 'the voice behind the words' has been an active topic in many fields of research for more than two decades now (Wu and Childers, 1991; Cowie et al., 2001; Schuller and Batliner, 2014). Early studies have emerged from research in phonetics and automatic speech recognition (ASR), and have focussed on simple characteristics such as gender (Wu and Childers, 1991). Research on recognizing human emotion from speech started at the beginning of this century (Cowie et al., 2001). As a matter of fact, the related paradigm of 'affective computing', that focusses on emotional aspects of natural human-machine interaction, has driven speech technology research throughout the last decade. In the recent years, a new major field of speech recognition research investigating the speaker characteristics beyond affective states is evolving: 'computational paralinguistics' (Schuller and Batliner, 2014). Research in this field has delivered highly promising results and tools for the community including the first widely used open-source affect analysis toolkit openEAR (Eyben et al., 2009) and its large-scale acoustic feature extractor openSMILE (Eyben et al., 2013) which both have become standard tools and references in the field. Furthermore, researchers from all over the world have reviewed their speech analysis systems in the light of the INTERSPEECH Challenges that have targeted a multitude of tasks such as emotion (2009), interest, age and gender (2010), sleepiness and alcohol intoxication (2011), as well as the OCEAN five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism), voice pathology and likability (2012), emotion, autism, and social signals (2013), and cognitive and physical load (2014). An overview of the evaluation campaigns up to 2012 is given in (Schuller, 2012).

From a methodological point of view, today's speaker characteristics recognition mostly relies on standard machine learning techniques that have been proven successful for various audio recognition tasks including speech and

speaker recognition. Most established techniques are static modelling with Support Vector Machines (SVMs) and dynamic modelling with Hidden Markov Models (HMMs). Generally, one starts with standard low-level descriptors (LLDs) such as (Mel-frequency) spectrum, Cepstrum, pitch, or voicing probability, extracted from short overlapping frames of fixed length. Static modelling is then performed by computing statistics of the LLD contours. Combining static modelling of utterances with context knowledge, Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) have successfully been introduced for affect recognition (Eyben et al., 2010). As a more recent approach to machine learning from unsupervisedly generated features, Deep Belief Networks (DBNs) have been applied to affect and likability recognition (Stuhlsatz et al., 2011; Brueckner and Schuller, 2012; Le and Mower, 2013). Despite the manifold work done for a plethora of speaker characteristics, the methodology has converged to a degree of standardisation, and major breakthroughs have been lacking in the past years. For many studies, it remains largely unclear to what extent their findings can be transferred to actual systems 'in the wild', for reasons outlined below.

Most importantly, today's studies consider speaker characteristics in isolation, i. e., single or only few speaker characteristics are considered at once (cf. Figure 1). There is very little exploitation of the interplay and synergies between different characteristics, yet in reality, strong interdependencies between bits of paralinguistic information exist. For example, it is intuitively clear that acoustic models for gender classification (male vs. female) should be different by age, since arguably the most important feature, pitch, is also influenced by age. Still, before interdependencies can be exploited in a more generic fashion, i.e., be learnt from data, richly annotated data sets will have to be created: at present, databases provide labels for only one or a few speaker characteristics at the same time. Another significant limitation of today's systems can be seen in their usage of acoustic features. These are mostly chosen ad-hoc because 'they seem to work well', and are often simply borrowed from neighbouring disciplines in audio processing such as ASR, instead of being tailored to the modelling of speaker characteristics. Apart from features, the limited transferability of most of today's studies to real-life applications is a more generic issue. First of all, this is because they are mostly carried out on hand-segmented, often manually transcribed utterances recorded from noise-free channels or in the presence of artificial noise and reverberation, and often prompted speech. To cope with real-life conditions in retrieval applications, however, robust single-channel automatic speech detection, segmentation and enhancement of spontaneous utterances in real acoustic environments, transmitted over arbitrary channels, must be addressed. Furthermore, all but a very few studies overlook the issue of potential malicious system use, such as faking of age, alcohol intoxication, or affective states; in fact, this phenomenon has only lately received some attention in speaker verification (Alegre et al., 2013). Finally, meaningful confidence measures (i. e., beyond simple posterior probabilities or distances in the feature space) have only been attempted recently (Deng and Schuller, 2012) despite them being crucial for real-life applications such as retrieval, dialogue systems and computer-mediated human-to-human conversation.

All these shortcomings are the starting point for the research envisioned in the iHEARu project.

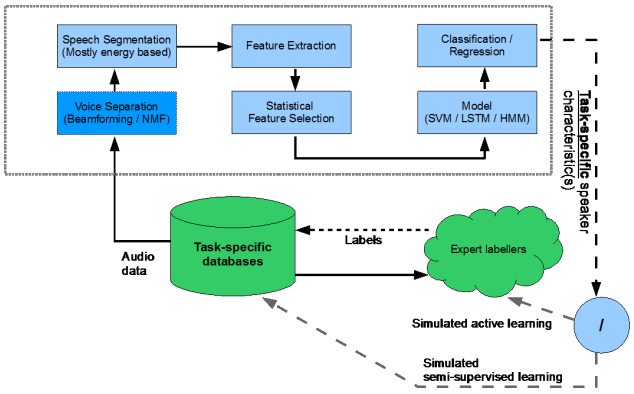


Figure 1: State of the art method for recognition of individual speaker characteristics. A standard machine learning pipeline is applied, consisting of pre-processing (voice separation and segmentation), feature extraction and selection, and classification/regression. Labels for (rather) small task specific databases are supplied by expert labellers. Simulated active and simulated semi-supervised learning are only considered by omitting labels from those expert labelled databases.

3. Methodology

To realise its ambitious goals, the iHEARu project aims to leverage novel techniques for multi-target (multi-task), semi-supervised and unsupervised learning. It is envisioned to overcome today's sparseness of annotated realistic speech data by large-scale speech and meta-data mining from public sources such as social media, crowdsourcing for labelling and quality control, and shared semi-automatic annotation. Furthermore, by utilising feedback, deep, and evolutionary learning methods, all stages from pre-processing and feature extraction to the statistical modelling can be subject to 'life-long learning' according to new data. Finally, human-in-the-loop system validation and novel perception studies are expected to help understanding both of system behaviour and human interpretation in a large variety of speaker classification tasks.

3.1. Holistic Processing

The iHEARu project intends to advance the state of the art by investigating novel methodology for holistic analysis of established speaker attributes, such as age and gender, in conjunction with currently under-researched characteristics, such as speech in different physiological and mental states. Large-scale speech and meta data mining from public sources (e.g., social media), combined with semi-automatic annotation methods (e.g., active learning) will be an essential means for building large, realistic, richly annotated and transcribed data sets.

3.2. Evolving "Life-Long" Learning for Self-Improvement

Self-learning and self-improvement in the iHEARu project will not be limited to iterative data collection. Rather, iHEARu will consider self-optimising feature extraction and self-organising classifiers: The whole process of speaker characteristics learning and analysis shall be self-optimising, as depicted in the flow chart above. For realising these ambitious goals, deep learning (Hinton et al., 2012) combined with neuroevolutionary methods and non-parametric Bayesian learning will play an essential role. This provides promising means for creating self-optimising statistical models and hierarchical input representations with very little amount of supervision.

3.3. Analysis with Deeper Understanding and Context-Dependent Speech Features

The iHEARu project approaches the acoustic feature generation and selection issue by trying to understand human reasoning in challenging conditions, from very low SNR, application of voice conversion algorithms, and speech compression, all the way to deliberate faking of voice or speaker states by the subjects. As a consequence, the iHEARu project will not only address environmental (technical) robustness, but more importantly also robustness against fraud.

3.4. Real-Life

To automatically obtain robust speech detection and segmentation into meaningful units, the iHEARu project aims to improve all of the pre-processing algorithms including speech separation, noise reduction, voice activity detection, and segmentation in a loop with the subsequent analysis algorithms and the confidence scores given by these (cf. Fig. 2). Further, dealing with real-life data also means coping with various transmission channels.

3.5. Universal Analysis

The iHEARu project addresses the automatic recognition of speaker attributes and speaking styles that can be clearly identified by humans. However, the iHEARu approach to universal analysis is not to simply define more and more new recognition tasks that are chosen 'ad hoc'; conversely, it is aimed at developing data-driven methods for a framework which is able to automatically identify characteristics of interest by looking at crowd-sourced resources, such as tag collections, opinions in textual comments, or explicitly collected annotations from paid click-workers.

4. Holistic Speaker Analysis with Multi-Task Learning

Integrating the concept of holistic analysis into automatic systems demands enhanced machine learning methods for context-aware learning. The first step toward a holistic analysis of speaker attributes is to consider multiple speaker attributes simultaneously and jointly in existing learning methods. One encounters many terms and buzz-words in this respect in the literature, which all refer to different concepts: multi-class, multi-label, multi-target, multi-task, multi-instance, and others. Therefore, it is important to

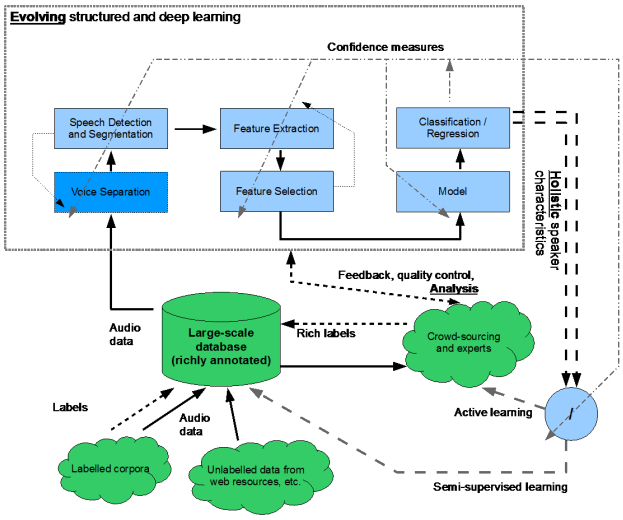


Figure 2: Flowchart of the proposed concept for holistic evolving analysis of realistic universal speaker characteristics. A large-scale collection of richly annotated data is created and extended by semi-supervised and active learning. Confidence measures of system components as well as humans in the loop are used to give feedback to components in the processing chain in order to implement evolving holistic learning.

first clarify the definitions of these terms at this point. Traditional *single-label* or *single-target* learning is concerned with learning from examples, where each example is associated with a single label l from a set of disjoint labels $L, |L| > 1$. For $|L| = 2$, the learning problem is called a binary classification problem or filtering in the case of textual and web data, while for $|L| > 2$, it is referred to as a multi-class problem (Tsoumakas and Katakis, 2007; Madjarov et al., 2012). In contrast, *multi-label learning* is concerned with learning from examples, where each training example is associated with zero, one, or more labels taken form a finite set of labels $Y \subseteq L$ (Zhang and Zhou, 2013).

During the past decade, the multi-label problem has received significant attention due to its wide variety of applications including text categorization, automatic annotation for multimedia contents (e.g., images, music, video), bioinformatics, web and rule mining, information retrieval, tag recommendation, etc. (Zhang and Zhou, 2013). Tsoumakas and Katakis (Tsoumakas and Katakis, 2007) were the first to group the multi-label learning approaches into two main categories: a) problem transformation methods, and b) algorithm adaption methods. The problem transformation methods refer to methods which transform the multi-label classification problem into either one or more single-label classification problems, for which there exits a plethora of machine learning algorithms. The algorithm adaption methods refer to multi-label methods where an existing machine learning algorithm is adapted, extended and customised in order to handle multi-label data directly. Furthermore, besides these two categories of methods for multi-label learning, Madjarov et al. (Madjarov et al., 2012) have introduced a third category: en-

semble methods. The most well known problem transformation ensemble methods are the RAKEL system by Tsoumakas et al. (Tsoumakas and Vlahavas, 2007), ensembles of pruned sets (EPS) (Read et al., 2008) and ensembles of classifier chains (Read et al., 2011) (ECC). The ECC method iteratively trains a multi-target classifier (or regressor) $(y_1, \ldots, y_{|L|}) = h(\mathbf{x})$, where $\mathbf{x}$ is a feature vector. For $l = 1, \ldots, |L|$, a single-target base classifier $y_l = h_l(\mathbf{x}, y_1, \ldots, y_{l-1})$ is trained, i.e., the estimates of the other targets are included as features. Since the order of labels clearly affects the results, bagging is performed to create an ensemble of classifiers using different label orders (and instance weights). An advantage of ECC over multi-task methods based on regularization (Evgeniou and Pontil, 2004), which presumes task similarity, is that not only correlations among labels but also correlations of labels with label-feature combinations can be effectively exploited, and that the method does not saturate with large amounts of training data (Read, 2010).

In a broad sense, multi-label learning can be regarded as a special case of *multi-target* learning, i. e., multi-dimensional learning. In multi-target learning, an example (a data instance) is associated with more than one target variable (as opposed to single-target learning, where only one target value is associated). Each target variable can take multiple numeric (regression) or nominal values (discrete classes). The multi-label case can now be seen as a special case of multi-target learning, where all target variables are binary and each target variable corresponds to a label being present or not.

Multi-target learning is often also referred to as multi-task learning. Besides learning multiple tasks/targets in parallel, information of related tasks is used as an inductive bias to improve the generalization performance of other tasks (Caruana, 1997).

Going back to multi-label learning, the differences between multi-label and multi-task learning are not conceptually based, but given by the different nature of the problems and use-cases addressed. Thus, in multi-label learning often a large space of labels is handled while in standard multi-task or multi-target learning a small set of labels is handled. For the holistic analysis in the iHEARu project both methods will be considered and investigated. Given the fact that they are closely related might result in novel, beneficial combinations of algorithms from both areas (Mencía, 2010).

Completely different from the problems of multi-label and multi-task learning, is *multi-instance learning*, where label sparseness is the core issue: for a bag of multiple instances, only one label exists for the whole bag and information on labels for the individual instances is lacking (Maron and Lozano-Pérez, 1998). In the most primitive case the label is only a binary label (positive and negative instances) and positively labelled bags have to contain at least one instance with a positive label, and negatively labelled bags contain only instances with negative labels (Maron and Lozano-Pérez, 1998). In the context of computational paralinguistics, potential applications of multi-instance learning can be found, e.g., in emotion detection: For example, if a speaker displays negative emotion, this usually affects a few short-time observations, while the remaining observations are similar to a 'neutral' state; in turn, manual annotation of each short-time observation is too cumbersome to perform on a large scale, in contrast to labelling whole utterances.

Table 1: *Partitioning of Speaker Likability Database (L: likable / NL: non-likable); Age (Y: young / A: adult / O: old); Gender (M: male / F: female)*

Task	SLD #	Train	Devel	Test	Σ
Likability	L	189	94	117	400
	NL	205	84	111	400
Age	Y	116	47	70	233
	A	131	58	76	265
	O	147	73	82	302
Gender	M	195	89	113	397
	F	199	89	115	403

5. Experimental Setup

5.1. Selected Database

This section introduces multi-task learning experiments for the joint classification of speaker age, gender, and the average subjective likability of the speaker's voice by others. For that purpose, we use the database of the *Likability Sub-Challenge* of the INTERSPEECH 2012 Speaker Trait Challenge and perform multi-task learning with the MEKA toolkit, which is an extension to the WEKA machine learning framework by adding support for multi-label and multi-target classification (Hall et al., 2009).

In the *Likability Sub-Challenge*, the "Speaker Likability Database" (SLD) was used (Burkhardt et al., 2011). The SLD is a subset of the German Agender database (Burkhardt et al., 2010), which was originally recorded to study automatic age and gender recognition from telephone speech. The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. The database contains 18 utterance types taken from a set listed in detail in (Burkhardt et al., 2010). An age and gender balanced set of 800 speakers is selected. While the annotation provides likability in multiple levels, the classification task is binarised into 'likable' (L) and 'non-likable' (NL). The data are partitioned into a training, development, and test exactly as in the INTERSPEECH 2012 Speaker Trait Challenge (cf. Table 1).

5.2. Feature Extraction

The acoustic feature set used in this experiment corresponds to the baseline feature set of the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012). The open-source openSMILE feature extractor is used (Eyben et al., 2013) to 'brute-force' a high-dimensional feature set by applying statistical functionals to frame-wise LLDs, which comprise energy, spectral and voicing related low-level descriptors (LLDs). The chosen set of LLDs is shown in Table 2. Regarding functionals, we aim at a compromise between a broad variety of functionals, and careful selection so as not to include meaningless features, such as the arithmetic mean of delta coefficients, which is expected to be zero. The set of applied functionals is given in detail

Table 2: *64 provided low-level descriptors (LLD).*

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
54 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing
logarithmic HNR, Jitter (local, delta), Shimmer (local)

Table 3: *Applied functionals.* [1]*: arithmetic mean of LLD / positive* Δ *LLD.* [2]*: only applied to voice related LLD.* [3]*: not applied to voice related LLD except F0.* [4]*: only applied to F0.*

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile ($\approx$ min), 99 % percentile ($\approx$ max)
position of min / max
percentile range 1 %–99 %
arithmetic mean[1], root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90% range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature[2]
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length[3]
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames[4]

in Table 3. Altogether, the 2012 Speaker Trait Challenge feature set contains 6 125 features, which is roughly a 40 % increase over previous year's feature set.

5.3. Single- and Multi-Target Learning

To assess the potential of multi-target learning, we compare the following learning schemes, all of which can be found in MEKA.

- Single-target learning (ST), i.e., independent training of single-target classifiers – linear support vector machines (SVMs) trained by sequential minimal optimization (SMO) are chosen;
- Multi-target learning by the ECC method, using SMO-trained SVMs as the base classifier;
- Multi-target learning by the Ensembles of Class Relevance (ECR) method, using SMO-trained SVMs as the base classifier – this corresponds to bagging of single-target SVM classifiers;
- 'Oracle' multi-target learning with SMO-trained SVMs (OMT), where each single-task classifier uses the correct labels for the other tasks as features.

In contrast to ECC, ECR breaks down the multi-target learning problem by considering each l independently, i.e., $y_l = h_l(\mathbf{x})$. However, in contrast to ST, an ensemble of classifiers is trained with different instance weights (bagging). Finally, the OMT method can be written as $y_l = h_l(\mathbf{x}, \hat{y}_1, \ldots, \hat{y}_{l-1}, \hat{y}_{l+1}, \ldots, \hat{y}_{|L|})$, where $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_{|L|})$ is a vector of ground truth labels.
For the parameter instantiation, we choose the complexity parameter $C \in \{10^{-4}, 10^{-3}, \ldots, 1\}$ for the SMO algorithm that achieves best UA recall on the development set, while the rest of the parameters are set as default values recommended by MEKA.
As evaluation measure, we use unweighted average (UA) recall (UAR) as used in the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012).

Table 4: Classification results for likability, age, and gender targets for single target classification (ST), multi-target classification by Ensembles of Classifier Chains (ECC) or Class Relevance (ECR), and "oracle" single target classification with the other two labels included as features (OMT). SVM with SMO training, complexity C optimised on the development set between 0.0001 and 1.0.

UAR [%]	**ST**	**ECC**	**ECR**	**OMT**
Development set				
Likability	58.9	55.4	54.9	**60.0**
Age	49.7	**51.9**	**51.9**	50.2
Gender	94.4	94.9	94.9	**95.5**
Test set				
Likability	**58.1**	52.8	57.5	57.3
Age	**46.9**	46.0	45.3	**46.9**
Gender	**96.9**	**96.9**	96.0	96.9

6. Results and Discussion

Table 4 shows the results obtained for single and multi-task classification, as well as for the oracle single-task experiment where the ground truths of the other labels are included as features in the training and development/test sets. Let us first look at the results of the oracle experiment, which hint at the performance attainable by the ECC approach, which is based on iterative classification using *estimated* class labels for the other tasks. It can be seen that only a few slight (statistically insignificant, according to a

z-test) performance improvements on the development set are obtained when including the ground truth labels for the other two tasks (OMT). Unsurprisingly, this greatly limits the performance of the ECC multi-task learning approach. Comparing with the ECR results, the slight performance improvement observed in age classification by the ECC approach might as well be attributed to bagging, not multi-target learning as such. On the test set, none of the multi-target methods can improve over the single-target baseline (ST).

Overall, but particularly for the likability task, we found that performance heavily depended on the complexity parameter, and parameter selection on the development set did not generalise to the test set. As the complexity parameter controls the feature weights in the SVM, this indicates that the features deemed most important on the development set do not model well the test set. For instance, if we tuned the complexity for the likability task on the test set, we could attain 61.4 % UAR with ECC and 61.0 % with ECR, instead of 52.8 / 57.5 %.

7. Conclusions

In this paper, we introduced the iHEARu project, which addresses some of the shortcomings of current research in computational paralinguistics, one of them being looking at speaker attributes in isolation. A few initial experiments with state of the art multi-target learning methods could not demonstrate improvements over conventional methods. As there are clear signs of overfitting, poor performance can also be attributed to very limited amounts of training data, and failure to extract features that generalise across different speakers. Furthermore, since even the inclusion of ground truth labels from other tasks could not improve performance, it is obvious that there is still large room for improvement in existing machine learning methods for multi-target learning, as foreseen in the iHEARu project. For example, the combination of large-scale, continuous valued feature sets with small-scale, discrete valued label sets in a linear or kernel feature space is arguably sub-optimal; a more suited alternative could lie in novel architectures of Bayesian networks or decision forests. Besides, it seems that multi-target learning can only be successful if considerable progress is also made in the other research challenges addressed by the iHEARu project: large-scale data collection with truly multi-dimensional ('universal') labels, but also unsupervised and semi-supervised feature learning, as well as features inspired by human perception, which are expected to lead to better generalisation. For example, to address the scarcity of multi-target databases (where all instances are labelled in multiple dimensions), and alleviate overfitting, we can investigate large-scale unsupervised feature learning followed by discriminative fine-tuning, using semi-supervised learning to determine missing labels.

8. References

F. Alegre, A. Amehraye, and N. Evans. 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proc. of ICASSP*, pages 3068–3072, Vancouver, Canada. IEEE.

R. Brueckner and B. Schuller. 2012. Likability Classification-A not so Deep Neural Network Approach. In *Proc. of INTERSPEECH*, Portland, OR, USA.

F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. 2010. A database of age and gender annotated telephone speech. In *LREC*.

F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger. 2011. 'Would You Buy A Car From Me?'—On the Likability of Telephone Voice. In *Proc. Interspeech*.

R. Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S.K.W. Fellenz, and J. Taylor. 2001. Emotion Recognition in Human-computer Interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80.

J. Deng and B. Schuller. 2012. Confidence Measures in Speech Emotion Recognition Based on Semi-supervised Learning. In *Proc. INTERSPEECH*.

T. Evgeniou and M. Pontil. 2004. Regularized multi-task learning. In *Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*. HUMAINE Association, IEEE, September.

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. 2010. On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues. *Journal on Multimodal User Interfaces, Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots*, 3(1–2):7–12, March.

F. Eyben, F. Weninger, F. Groß, and B. Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. ACM Multimedia*, Barcelona, Spain. ACM.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. 2009. The WEKA Data Mining Software: an Update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

D. Le and E. Mower. 2013. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic.

G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski. 2012. An Extensive Experimental Comparison of Meth-

ods for Multi-label Learning. *Pattern Recognition*, 45(9):3084–3104.

O. Maron and T. Lozano-Pérez. 1998. A Framework for Multiple-instance Learning. *Advances in neural information processing systems*, pages 570–576.

E.L. Mencía. 2010. Multilabel Classification in Parallel Tasks. *Working Notes*, page 29.

J. Read, B. Pfahringer, and G. Holmes. 2008. Multi-label Classification Using Ensembles of Pruned Sets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 995–1000. IEEE.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. 2011. Classifier Chains for Multi-label Classification. *Machine learning*, 85(3):333–359.

J. Read. 2010. *Scalable Multi-label Classification*. Ph.D. thesis, The University of Waikato, New Zealand.

B. Schuller and A. Batliner. 2014. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, November. to appear.

B. Schuller, S. Steidl, A. Batliner, E. Nth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012. The INTERSPEECH 2012 speaker trait challenge. In *Proc. of INTERSPEECH*, Portland, OR, USA. ISCA.

B. Schuller. 2012. The Computational Paralinguistics Challenge. *IEEE Signal Processing Magazine*, 29(4):97–101, July.

A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller. 2011. Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In *Proc. ICASSP*. IEEE.

G. Tsoumakas and I. Katakis. 2007. Multi-label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

G. Tsoumakas and I. Vlahavas. 2007. Random k-labelsets: An Ensemble Method for Multilabel Classification. In *Machine Learning: ECML 2007*, pages 406–417. Springer.

K. Wu and D. Childers. 1991. Gender Recognition from Speech. Part I: Coarse Analysis. *The Journal of the Acoustical society of America*.

M.L. Zhang and Z.H. Zhou. 2013. A Review On Multi-Label Learning Algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, Preprints(99):1–43.

An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena

Zixing Zhang[1], **Florian Eyben**[1], **Jun Deng**[1], **and Björn Schuller**[2,1]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Department of Computing, Imperial College London, United Kingdom
zixing.zhang@tum.de

Abstract

Redundant instances in subjective speech phenomena may cause increased training time and performance degradation of a classifier like in other pattern recognition tasks. Instance selection, aiming at discarding some 'troublesome' instances and choosing the most informative ones, is a way to solve this issue. We thus propose a tailored algorithm based on human *A*greement levels of labelling and class *S*parseness for learning *I*nstance *S*election – ASIS for short. Extensive experiments on a standard speech emotion recognition task show the effectiveness of ASIS, indicating that by selecting only 30 % of the training set, the system performance significantly outperforms training on the whole training set without instance balancing. In terms of performance it remains comparable to the classifier trained with instance balancing, but at a fraction of the training material.

Keywords: Instance Selection, Subjective Speech Phenomena, Human Agreement Level, Sparse Instance Tracking

1. Introduction

Instance selection is important in many pattern recognition tasks. This includes in particular also the field of automatic recognition of (often highly) *subjective* paralinguistic speech phenomena, such as speakers' emotion, interest, sleepiness, intoxication, or voice likability. There are three main reasons why instance selection is worth considering: Following the idea of "there is no data like more data", many efforts have recently been undertaken to collect and/or create large amounts of data with the aim to improve recognition performance: more manual annotations, aggregation of multiple corpora (Schuller et al., 2011), and semi-supervised learning or co-training (Zhang et al., 2011; Zhang et al., 2013). However, as the size of the data set which is used to train a classifier increases, the complexity of the models and the training time increases (Schuller et al., 2012b). Even though, for most commercial applications classifiers training is done once and is not a time critical operation, faster training times gives companies a competitive advantage. Researchers, on the other hand, will train many models when optimising parameters and testing new methods. Thus, they largely benefit from reduced training times.

Another main reason is the subjectivity of the paralinguistic phenomena. Unlike traditional pattern recognition tasks where a true 'ground truth' is available, those tasks only have 'gold standard' labels, which are often assigned by (sometimes weighted) majority voting over multiple human ratings. In fact, instance labelling for such tasks highly depends on the labellers' personal judgement. For music mood, for example, some would consider a musical piece more sad or happy than others or even have opposing views due to personal associations with a song. The same holds for speaker emotion or likability recognition (cf. (Sneddon et al., 2012; Schuller, 2013)). Instances with high labeller uncertainty could potentially cause the model to over-fit these 'noisy' instances resulting in increased complexity (Angelova, 2004). This thus would deteriorate the generalization performance.

The last reason relates to the imbalance of the number of instances among classes, which is most pronounced in databases with natural and spontaneous speech, where 'neutral' speech is much more frequent than clear cut cases of emotional or other target speech. This leads to the fact, that some models tend to favour the majority classes and thus show a bad performance on the sparse (minority) classes. However, these sparse classes are usually of most interest in practical applications.

Therefore, a reduction of the amount of training instances is beneficial if the following two criteria are met: 1) Equal or improved performance. That is, the model trained on a subset should perform equally to or better than the model trained on all instances; and 2) Reduction of training time. To this end, we propose an instance selection method in this paper which is based on human labeller agreement level and class sparseness. The two main contributions of this paper are: 1) We investigate whether pruning of the instances with the lowest labeller agreement improves performance; and 2) After pruning we select an equal amount of instances from each class in order to produce a set with a balanced number of instances of each class.

1.1. Related Work

In pattern recognition, numerous methods have been proposed and investigated in the literature for solving the data selection problem. Most of them can be assigned to one of the following two groups from a technical point of view (Liu and Motoda, 2002; Olvera-López et al., 2010):

The first group is *wrapper-based selection*, where the selection criterion is based on the accuracy obtained by a classifier (Olvera-López et al., 2010). Those instances that do not improve predictive performance of classification will be discarded from the training set. Most of the wrapper-based selection methods are related to the k-nearest neighbour classifier (Cover and Hart, 1967) like the Condensed

Nearest Neighbour (CNN) (Hart, 1968), Selective Nearest Neighbour rule (SNN) (Ritter et al., 1975), or Decremental Reduction Optimisation Procedure (DROP) (Wilson and Martinez, 2000). With CNN, for example, the instances misclassified by the classifier will be selected and added into the initial training set.

Unlike wrapper-based selection methods, *filter-based selection* methods in the second group attempt to select the instances by means of sampling or clustering, without depending on a prediction of a classifier (Olvera-López et al., 2010). Among them, a prominent algorithm is RANdom SAmple Consensus (RANSAC) proposed by Fishler and Bolles (Fischler and Bolles, 1981). It uses a data set as small as possible to determine model parameters – mostly used for estimating homography transformation matrices in computer vision. Then, other data are tested against the estimated model and those data which fit the model within a predefined tolerance ϵ will be considered part of a consensus set. Whenever the ratio of the number of consensus data to the total number data in the set exceeds a predefined threshold, the model parameters are re-estimated using all consensus and all initial data. This procedure is repeated a fixed number of times. Another example is the Pattern by Ordered Projections (POP) (Riquelme et al., 2003) which discards interior instances and selects some border instances, where a border instance is defined by its nearest neighbour belonging to other classes, and an interior instance is defined by its nearest neighbour belonging to the same class. In addition, to address the issue of class imbalance, e. g., Garcia et al. proposed a scalable instance selection method in (García-Pedrajas et al., 2013).

However, most of these methods are developed for objective pattern recognition tasks with a definite ground truth, such as face recognition (Angelova et al., 2005), textual news classification into groups (Fragoudis et al., 2002), speech recognition, or language translation (Wu et al., 2007; Lu et al., 2012). Even though there is some work dealing with subjective pattern recognition tasks (e. g., (Erdem et al., 2010) which selects a training subset by RANSAC for emotion recognition), the influence of labelling uncertainty on recognition performance has not been considered directly nor has the class imbalance problem been addressed. These two issues are the focus of the work presented in this paper.

In the following, we introduce the details of our proposed instance selection algorithm in Section 2.. Then, we describe the databases used for the experiments and discuss the results of the proposed instance selection algorithm in Section 3.. Finally, we draw the conclusions in Section 4..

2. Methodology

The main idea of our algorithm is to discard the instances with low labelling agreement and afterwards sub-sample the data set by selecting an equal amount of instances for each class from the remaining instances.

2.1. Human Agreement Levels

To measure human inter-rater agreement levels, we employ Fleiss' frequently used Kappa coefficient, which is expressed as:

$$\kappa := \frac{p_0 - p_c}{1 - p_c}, \tag{1}$$

where p_0 is the observed agreement of labellers, and p_c is chance-level agreement. In the case of a single instance, the probability of p_0 can be simplified by estimating the proportion of cases in which labellers agree on a common category:

$$p_0 = \sum_{m=1}^{M} \frac{\eta_m}{M}, \tag{2}$$

where $\eta_m \in (0, 1)$ stands for a binary annotation of a specific category, and M is the number of labellers. Thus, the difference $p_0 - p_c$ indicates the proportion of cases where 'beyond-chance agreement' occurs. It is normalized by the probability of disagreement $1 - p_c$ which is expected by chance.

2.2. ASIS: Agreement and Sparseness-based Instance Selection

The details of the proposed algorithm are presented in Algorithm 1. It includes two steps: Agreement-based Instance Selection (AIS) and Sparseness-based Instance Selection (SIS).

The AIS step aims at discarding the most noisy instances mainly caused by high disagreement of human labelling. In this process, we prefer to *proportionally* discard instances across classes. On the one hand, it prevents the case of potential maldistribution of instances which might result in discarding such instances mainly belonging to certain classes, especially the sparse ones. On the other hand, it probably improves the separability of classes by potentially removing instances close to the class boundary in the feature space. Therefore, the larger we choose the discarded subset ($P_D[\%]$) to be, the fewer instances – relatively seen – might be located near the class boundaries, and the less complex the model becomes.

The SIS step randomly selects an equal number of instances from each class set with the aim of coping with the class sparseness problem. Note, that in the case of a class balanced task the size of the selected subset ($P_S[\%]$) will satisfy $P_D + P_S \leq 1$, while in the case of a class imbalanced task, P_S is limited by the instance count of the most sparse class. This problem could be eased to some extent by loosening the constraint of ending up with a balanced distribution of instances after sub-sampling, i. e., the missing amount of instances of the sparse classes can be filled in with instances from the abundant classes. In this paper, however, we adhere to the strict rule of balanced selection and evaluate binary tasks only as straightforward examples.

3. Experiments

To evaluate the effectiveness of our algorithm, we selected two well-standardised machine learning tasks and according data from the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) and the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012a). Both are of highly subjective nature and together they cover the spectrum from short-term (emotion) to long-term (likability) speaker traits. In the following, the two according

Algorithm 1: ASIS: The proposed agreement and sparseness-based instance selection algorithm.

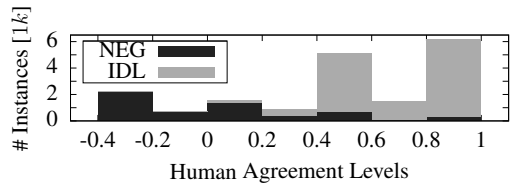
Input:
$\mathcal{D}$: Database of N instances annotated in classes C_i $(i = 1, \cdots, k)$ and corresponding human agreement levels l;
P_D: Size of discarded subset with low human agreement levels (percentage of the full training set);
P_S: Size of selected subset (percentage of the full training set);
k: number of classes;
Output:
$\mathcal{S}$: Subset of database $\mathcal{D}$;

Process
Obtain the proportional distribution of each class R_i $(i = 1, \cdots, k)$ in the training set of $\mathcal{D}$;
(Step: Agreement-based Instance Selection (AIS))
for $i = 1, \cdots, k$ **do**
    Sort the instances that are annotated as class C_i by human agreement levels l from low to high, producing queue Q_i;
    Delete $n_{Di} = N \times P_D \times R_i$ instances which are at the beginning of Q_i;
end
(Step: Sparsness-based Instance Selection (SIS))
for $i = 1, \cdots, k$ **do**
    Randomly select $n_{Si} = N \times P_S / k$ instances belonging to class C_i;
end
Fuse n_{Si} $(i = 1, \cdots, k)$ into one output subset $\mathcal{S}$.

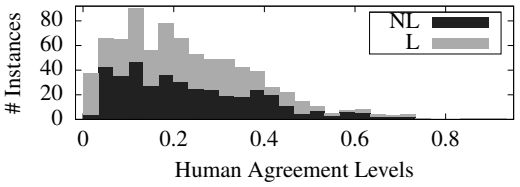
databases are introduced and then the results obtained on these sets are described.

3.1. Emotion: FAU Aibo Emotion Corpus

The FAU Aibo Emotion Corpus (Steidl, 2009) is the official corpus of the INTERSPEECH 2009 Emotion Challenge (EC) (Schuller et al., 2009). It contains recordings of children interacting with Sony's pet robot Aibo. The language is German. The Wizard-of-Oz controlled Aibo robot sometimes disobeyed children's commands, thereby provoking various emotional reactions. The recording was done at two different schools – MONT and OHM –, and features 51 children with 21 boys and 30 girls at ages ranging from 10 to 13 years. Five labellers listened to the turns in sequential order and labelled each word independently from each other as neutral or as belonging to one of ten other emotion classes. In the challenge, the final labelling and human agreement levels for chunks are determined by majority voting on labels of the five labellers on the word level onto one label for the whole chunk. Then, chunks were grouped into the 2-class labelling: **NEG**ative (subsuming $angry$, $touchy$, $reprimanding$, and $emphatic$) and **IDL**e (consisting of all other states). Fig. 1 (a) shows the instance distribution of the training set with human agreement levels. For our experiments, we use the whole corpus consisting of 18 216 chunks, where the training set includes 3 358 'NEG' and 6 601 'IDL' instances, and the test set consists of 2 465 'NEG' and 5 792 'IDL' instances. Note that, for the sake of balancing categories, some instances with negative human agreement level also belong to the class 'NEG'.



(a) Emotion: AEC



(b) Likability: SLD

Figure 1: Number of instances with human agreement levels in the AEC (a), and SLD (b).

3.2. Likability: Speaker Likability Database

The Speaker Likability Database (SLD) (Burkhardt et al., 2010) was the official corpus of the Likability Sub-Challenge in the INTERSPEECH 2012 Speaker Trait Challenge (STC) (Schuller et al., 2012a). The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. An age and gender balanced set of 800 speakers is selected. For each speaker, the longest sentence (consisting of a command embedded in a free sentence) was selected. Likability rating was executed by 32 labellers according to how well they personally liked the voices. They were asked not to take into account the linguistic content or the transmission quality. The rating was done on a seven point Likert scale ([-3, -2, -1, 0, 1, 2, 3]). To establish a coherent consensus from the highly individual likability ratings, the evaluator weighted estimator (EWE) (Grimm and Kroschel, 2005) was used in the challenge. It uses higher weights for more agreeable labellers. Based on the median EWE rating of all stimuli in the SLD, the data was discretised at the threshold of 0.108 into the classes—'likable' (**L**, $EWE > 0.108$) and 'non-likable' (**NL**, $EWE < 0.108$). The final challenge set of 800 instances is partitioned as follows: training set (L, 189; NL, 205), development set (L, 92; NL, 86), and test set (L, 119; NL, 109). Fig. 1 (b) shows the instance distribution along with human agreement levels. Here, we slightly modify Kappa as follows:

$$Adapted \quad Fleiss' \quad \kappa := \left| \frac{p_{ewe} - p_t}{p_{ewe_{max}} - p_t} \right|, \qquad (3)$$

by replacing the p_c with the threshold of 'L' and 'NL' p_t at 0.108, p_o with the EWE values p_{ewe}, and 1 with the maximum EWE value of p_{ewe}. Therefore, the instances with an EWE value near 0.108 are considered as low agreement and vice versa.

3.3. Protocol and Results

As in the challenge tasks, we evaluate performance in terms of unweighted average recall (UAR). In addition, we use the original challenge feature sets for the tasks of emotion and likability recognition in our experiments. Thus, for emotion recognition, we use 384 features resulting from a systematic combination of 16 low-level-descriptors (LLDs) and corresponding first order delta coefficients with 12 functionals (Schuller et al., 2009); for likability recognition, we utilize 6 125 features by brute-forcing based on 64 LLDs and 61 functionals (Schuller et al., 2012a) – all features are extracted with the open-source toolkit openSMILE (Eyben et al., 2010). In the same vein, we keep the classifiers, their implementations, and parameters as in Challenges: for emotion recognition, Support Vector Machines (SVMs) trained by Sequential Minimal Optimization (SMO) with polynomial kernel (degree 1) and a complexity constant of 0.05; for likability recognition, Random Forests (RF) with a number of trees $N = 1\,000$ and a feature subspace size of $P = .02$. The Weka toolkit (Hall et al., 2009) is used in both cases. Note, that the instance selection algorithm is only applied on the training set. The test set is not modified and kept the same as in the original Challenge setup in order to allow for a direct comparison.

3.4. Emotion

The following experiments were executed for emotion recognition with different variations of the instance selection algorithm: 1) only agreement-based instance selection ('AIS') based on discarding low-agreement instances (cf. Step 'AIS' in Algorithm 1); 2) only sparseness-based instance selection ('SIS') by selecting sparse instances (cf. Step 'SIS' in Algorithm 1); 3) both steps (ASIS) at the same time (random selection with balancing of instances across classes).

For comparison, we denote the control methods of Random Instance Selection (RIS) as randomly selecting a predefined number of instances from the whole set without other constraints.

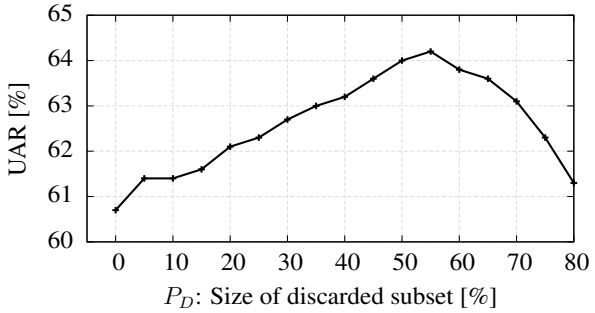


Figure 2: Agreement-based Instance Selection (AIS): UAR on the AEC test set after discarding low agreement training instances (no balancing).

Fig. 2 gives an overview on performances after discarding a certain ratio of instances with low human agreement (AIS). Note, that the human agreement levels by discarding 5, 10, 20, 30, 40, 50, 60 % of the instances for the class IDL are

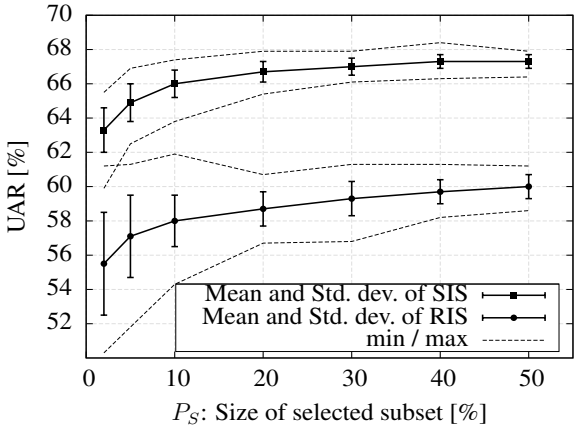


Figure 3: Sparseness-based Instance Selection (SIS): UAR mean, standard deviation (std. dev.), minimum (min), and maximum (max) on the AEC test set over 40 independent runs. Comparison of balanced SIS and random instance selection (RIS) from the training set. No discarding of instances with low agreement ($P_D = 0$).

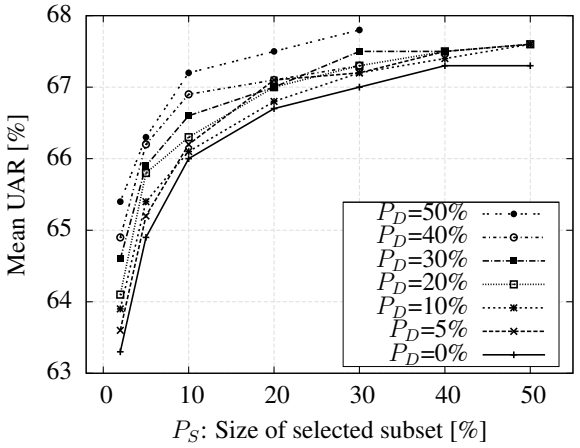


Figure 4: Agreement and Sparseness-based Instance Selection (ASIS): Mean UAR on the AEC test set in 40 independent runs of *balanced sub-sampling* after discarding P_D instances with lowest labeller agreement from the training set of the AEC.

0.4, 0.52, 0.60, 0.60, 0.60, 0.72, 0.90, and for the class NEG are -0.28, -0.2, -0.2, -0.2, -0.08, 0.06, 0.2, respectively. No instance balancing is performed here. The performance of the classifier improves continuously and significantly (one-sided z-test) until 55 % of the training set instances with human agreement are discarded (from 60.7 % to 64.2 % UAR). Fig. 3 compares the performance of two instance sub-sampling strategies (with (SIS) and without balancing), both without any prior discarding of low agreement instances. As expected, UAR is increased by about 8 % absolute when balancing is performed, showing the importance of a balanced distribution for SVM (and further) classifiers. Fig. 4 shows results obtained when randomly sub-sampling the training set and balancing after discard-

ing low agreement instances (ASIS). At a certain ratio of discarded instances, increasing the number of selected instances enhances the system robustness. As more instances are added, however, the increase of UAR converges. At a certain amount of sub-sampling, discarding up to 50 % of low agreement instances improves UAR. Note that this improvement is more obvious for a small subset size, as in this case the disturbing influence of the low agreement instances has a larger relative impact on the model. The best result of 67.8 % of UAR is achieved by discarding 50 % of lowest agreement instances and selecting only 30 % of instances (relative to the whole set) for model building. This is equivalent to the baseline (67.7 % of UAR) in (Schuller et al., 2009) where the whole training set with Synthetic Minority Oversampling TEchnique (SMOTE) is considered (for balancing). Note that, for this experiment the amount of sub-sampling is limited by the size of the minority class 'NEG'.

3.5. Likability

We further evaluate the potential of our algorithm for a secondary task: automatic speaker's voice likability recognition. Fig. 5 visualises the performance after discarding instances with low agreement levels (AIS). Table 1 shows the relationship between the percentage of discarded instances and the human agreement levels of the classes 'L' and 'NL'. Due to the way the classes 'L' and 'NL' have been defined (by median), the instances are already balanced among the two classes. Thus, no balancing is therefore necessary (i. e., no SIS). By discarding the lowest 10 % agreement levels, the UAR is raised from 59.0 % to 62.0 %. One notices that discarding more instances does not bring additional improvement. This might be due to the small size of the dataset with only 600 instances in the training set.

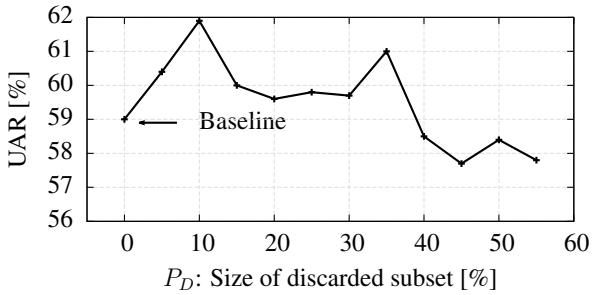


Figure 5: Agreement-based Instance Selection (AIS): UAR on the SLD (likability) test set after discarding low agreement training instances from the training set.

Table 1: Relationship percentage of discarded instances and agreement levels

	Percentage discarded				
Levels	10 %	20 %	30 %	40 %	10 %
L	0.01	0.05	0.09	0.15	0.18
NL	0.07	0.10	0.13	0.17	0.20

4. Conclusions

We proposed ASIS – agreement and sparseness-based instance selection which exploits labeller agreement levels and the concept of sparse class learning by random subsampling of the training space. We demonstrated the potential of this algorithm for two standard machine learning challenge tasks for speech emotion and voice likability recognition. For the emotion recognition experiments on the FAU AEC set, we observe obvious improvement of performance by balancing the instance distribution among both classes through random sub-sampling (SIS). Yet, discarding the instances with low agreement levels (AIS) brings a further improvement. A performance comparable with the baseline of the INTERSPEECH 2009 Emotion Challenge is achieved when only 30 % of the whole training set – selected by the proposed method – are used for training. The experiments on the Speaker Likability Database further prove the effectiveness of AIS in the case of discarding training instances.

In future work, the discarding instance number needs to be more discussed when in the blind of test set and type of tasks which are with different distribution of labelling agreement.

5. References

A. Angelova, Y. Abu-Mostafam, and P. Perona. 2005. Pruning training sets for learning of object categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 494–501, San Diego, CA.

A. Angelova. 2004. Data Pruning. M. sci. thesis, California Institute of Tchnology.

F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. 2010. A Database of Age and Gender Annotated Telephone Speech. In *Proc. of LREC*, pages 1562–1565, Valletta, Malta.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

C. E. Erdem, E. Bozkurt, E. Erzin, and A T. Erdem. 2010. Ransac-based training data selection for emotion recognition from spontaneous speech. In *the 3rd international workshop on Affective interaction in natural environments*, pages 9–14, New York, NY.

F. Eyben, M. Wöllmer, and B. Schuller. 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, Florence, Italy.

M. A Fischler and R. C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

D. Fragoudis, D. Meretakis, and S. Likothanassis. 2002. Integrating feature and instance selection for text classification. In *Proc. the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 501–506, Edmonton, Canada.

N. García-Pedrajas, J. Pérez-Rodríguez, and A. de Haro-García. 2013. OligoIS: Scalable Instance Selection for

Class-Imbalanced Data Sets. *IEEE Transactions on Cybernetics*, 43(1):332–346.

M. Grimm and K. Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 381–385, Cancun, Mexico.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

P. Hart. 1968. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516.

H. Liu and H. Motoda. 2002. On issues of instance selection. *Data Mining and Knowledge Discovery*, 6(2):115–130.

S. Lu, W. Wei, X. Fu, L. Fan, and B. Xu. 2012. Phrase-based data selection for language model adaptation in spoken language translation. In *2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 193–196, Hong Kong, China.

J A. Olvera-López, J A. Carrasco-Ochoa, J F. Martínez-Trinidad, and J. Kittler. 2010. A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143.

J. C Riquelme, J. S Aguilar-Ruiz, and M. Toro. 2003. Finding representative patterns with ordered projections. *Pattern Recognition*, 36(4):1009–1018.

G. Ritter, H. Woodruff, S. Lowry, and T. Isenhour. 1975. An algorithm for a selective nearest neighbor decision rule. *IEEE Transactions on Information Theory*, 21(6):665–669.

B. Schuller, S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. of INTERSPEECH*, pages 312–315, Brighton, UK.

B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll. 2011. Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? In *Proc. INTERSPEECH 2011*, pages 1553–1556, Florence, Italy.

B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. V. Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. 2012a. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. of INTERSPEECH*, Portland, OR. 4 pages.

B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. 2012b. AVEC 2012 – The Continuous Audio/Visual Emotion Challenge. In *Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012*, pages 449–456, Santa Monica, CA.

B. Schuller. 2013. Multimodal Affect Databases - Collection, Challenges & Chances. In Rafael A. Calvo, Sidney DMello, Jonathan Gratch, and Arvid Kappas, editors, *Handbook of Affective Computing*. Oxford University Press.

I. Sneddon, Ma. McRorie, G. McKeown, and J. Hanratty. 2012. The belfast induced natural emotion database. *IEEE Transaction on Affective Computing*, 3(1):32–41.

S. Steidl. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin.

D R. Wilson and T. R Martinez. 2000. Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286.

Y. Wu, R. Zhang, and A. Rudnicky. 2007. Data selection for speech recognition. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 562–565, Kyoto, Japan.

Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller. 2011. Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition. In *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 523–528, Big Island, HY.

Z. Zhang, J. Deng, and B. Schuller. 2013. Co-Training Succeeds in Computational Paralinguistics. In *Proc. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8505–8509, Vancouver, Canada.

The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform

Véronique Aubergé[1], Yuko Sasa [1], Nicolas Bonnefond [1], Brigitte Meillon [1], Tim Robert [1], Jonathan Rey-Gorrez[1], Adrien Schwartz[1,2], Leandra Antunes[1,3], Gilles De Biasi [1], Sybille Caffiau [1], Florian Nebout[4]

1 LIG-lab CNRS, Grenoble University, France
2 Floralis Company, Gières, France
3 University of Ouro Preto, Brazil
4 Awabot Company, Lyon, France
E-mail: Veronique.Auberge@imag.fr

Abstract

The aim of this preliminary study of feasibility is to give a glance at interactions in a Smart Home prototype between the elderly and a companion robot that is having some socio-affective language primitives as the only vector of communication. The paper particularly focuses on the methodology and the scenario made to collect a spontaneous corpus of human-robot interactions. Through a Wizard of Oz platform (EmOz), which was specifically developed for this issue, a robot is introduced as an intermediary between the technological environment and some elderly who have to give vocal commands to the robot to control the Smart Home. The robot vocal productions increases progressively by adding prosodic levels: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. The elderly subjects' speech behaviours confirm the hypothesis that the socio-affective "glue" effect increase towards the prosodic levels, especially for socio-isolated people. The actual corpus is still on recording process and is motivated to collect data from socio-isolated elderly in real need.

Keywords: socio-affective "glue", human-robot interaction, socio-affective prosody, elderly, Smart Home, spontaneous corpus

1. Introduction

It is supposed here that whatever the social role created or borrowed for a robot entering the social sphere of the human, its role competences can be integrated in the human social space only if the relational link allows the dialog architecture by building the relevant "socio-affective glue" in a co-construction processing. The hypothesis underlying this work is that the material of the "socio-affective glue" is sufficiently non-lexical sounds and mimicry with "glue" prosody (Aubergé, 2012). Non lexical sounds - non phonological but prosodically relevant items - produced during or outside the talk turn, like onomatopoeias, interjections, fillers, grunts, bursts have been studied for their emotional functions [affects bursts (Scherer, 1994; Schröder 2003)], or for their pragmatic functions in dialog (Fonagy & Target, 1997). Non-lexical sounds have been observed both in listener feedbacks in backchannel (see Humaine D6d works) and in the feedbacks of the speaker, implied in a human/human or human/machine interaction (Morlec, 2001; Mairesse & al., 2007; Morency, 2010). They can express emotions, intentions, attitudes and cognitive/mental states and processing (like concentration, hesitation about an answer, etc.) that we name *Feeling of Thinking – FoT* (Aubergé, 2012). From a large spontaneous corpus (Aubergé, Rilliard, Audibert, 2005) some such functional lexical words have been selected (Vanpé & Aubergé, 2011) and perceively measured (De Biasi, 2012; Sasa & al. 2013). These non lexical sounds have been perceively classified in increasing "glue" competence: in order to long term develop an application of a "socio-affective glue trainer" for a robot with relationally isolated human, an experiment is presented is that shows that these selected sounds and mimicry, given to a butler robot in smart home, build a strong socio-affective glue with some isolated elderly.

2. Elderly situation

Gerontechnology emerged from a society challenge due to the demographic evolution of elderly (Bouma & al., 2007). The number of aged people living at home becomes higher every year in Western countries: over 80 years, 6/10 people live at home (4/10 in nursing homes), 25% of them with low dependency while only 2,5% are strongly dependent (Harrington & Harrington, 2000). The Smart Homes are often presented as convenient (and economical) issues to help elderly to stay longer at home. In a such socioeconomic situation, one main vector of elderly frailty is now the socio-affective isolation: it was observed in many studies [see the last ISG http://www.gerontechnology.info conferences] that the affective and organizational dimensions of isolation have direct and very strong consequences on the physical and mental health [6,30], which allows to keep elderly living at home. The main cue pointed by all those studies [see IAAG http://www.iagg.info/index.php and IUGMS http://www.eugms2014.org Congresses] is the isolated ones' socio-affective interactional competences degradation. It means that socio-affective interactional "coaching" would be a main issue, that starts to be taken into account by some professionals of elderly caregiving [www.bienalamaison.com]. The socio-affective interactional degradation occurring for elderly can appear in other societal areas, like the hikikomori syndrome described in Japan for young people (Furlong, 2008). Of course, it becomes a central issue for the pathologies including communication diseases, like Alzheimer or autistic syndrome.

That is why the present study prior goal is to collect a large spontaneous corpus of ecological situations

implying elderly and a companion robot in order to further design technologies of human-robot interactions specifically devoted to elderly living in a Smart Home. This socio-interactional robotic technology (Interabot Project[1]) will be built to train the elderly to communicate (socio-affective prosthesis) with a robot while this tool is presented as the Smart Home's butler.

Some theoretical objectives motivate this study too: using a robot is here a method to evaluate some hypotheses on the interactions primitives that build what we call the socio-affective "glue". Prosody carries emotional, socio-affective and interactional information where each language has its own values (Decety, 2007). This communicative information appears in different prosodic levels as in non-lexical sounds. Those can be non-phonetic sounds like grunts, affect bursts or mouth noises (Schröder, 2006, Poggi, 2008), phonological as fillers, mind markers or interjections (Amecka, 1992), or onomatopoeia, widely studied in Japanese (Shibatani, 1990). These sounds that we can consider as pure prosodic tools, were studied for a specific and supposed emotional (Aubergé, 2012) and pragmatic (Fonagy & Target, 1997) functions, as well as moods, emotions, intentions, attitudes, cognitive processes and mental states also known as "Feeling of Thinking" (Aubergé, 2012). Moreover lexicons, sentences and paraphrases prosodic form also support various socio-affective values (Decety, 2007). These cues can be extended from simple sounds to sentences produced in a same context, which have been tested in synthesis (Morlec, 2001; Mairesse & al., 2007; Morency, 2010). Lately, the prosody carrying this communicative information was introduced as a way to develop "socio-affective glue" (Aubergé & al., 2013) that allows interlocutors to build dynamically the communicative channel depending on the interaction context. Furthermore, imitation has been studied as a basic process to create the same kind of "glue" in children language acquisition (Tomasello & al., 2005) or as a primitive of robots learning (Schaal, 1999).

By the way, since the 90's, Affecting Computing and multidisciplinary communities have been focusing their work on the face-to-face interactions, especially on facial, gestural and vocal expressions using virtual agents and robots as in various studies in social computing (Schaal, 1999; Breazeal & al., 2002). It is interesting to see that when a robot is not explicitly humanoid, human creates by himself a socio-affective relationship with this device toward its « pet » stance (Sasa & al. 2012). Because all these different prosodic levels have not been studied together particularly to see their functions in the "socio-affective glue" building, our work will test them progressively thanks to a robot interacting with elderly towards gradual vocal productions: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. This will be tested with the EmOz wizard of Oz platform developed for this project (Aubergé & al. 2013) in the experimental Living Lab Domus of the LIG lab. Domus is a completely authentic Smart Home with hidden equipment and control room (Niitamo & al. 2006). A complex script is held to collect comparable data for many senior subjects (more than 40 are under recording), in the increasing levels of "glue", for the EMOX robot (developed by Awabot www.awabot.com/) playing the role of Domus' butler. The resulting corpus is EEE (Elderly EmOz Expressions).

[1] The Interabot project is financed by the French Industry ministry (Investissements d'Avenir) and held together by some industrial companies and academic partners,

3. The EEE script with EmOz

3.1 Experimental tools: EmOz – a Wizard of Oz

3.1.1 Domus: a Smart Home prototype

The LIG developed a living-lab into the Multicom Platform where we can record high quality sounds in a specific room (*see A on figure 1*), film a recording set which looks like a meeting room but can also be arranged to look like every other kind of environments for experiments (*see B on figure 1*), and which all the devices can be controlled from a control room (*see C on figure 1*).

Figure 1: Multicom Platform of the LIG-lab.

In this platform, Domus (*see D on figure 1*), where our study takes place, is designed like a 40m^2 flat with a kitchen, a bedroom and a living room equipped with two cameras and two microphones in each room, and a shower room with a microphone (*see figure 2*). It has sensors and actuators conforming to the automation standards KNX (Konnex) that group a heterogeneous set of protocols exchanging information outside the building through an OSGI gateway.

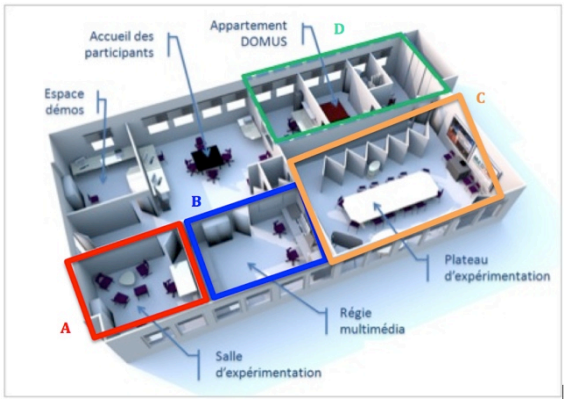


Figure 2: Illustrations of Domus Smart Home.

In our work, we selected few possible actions to execute into DOMUS and we proposed 30 vocal commands (*see Table 1*) that we can simulate from the platform control room.

Kitchen	Monter/descendre/arrêter les stores *To up/down/stop blinds* Allumer/éteindre la lumière *To turn on/off the light* Mettre/éteindre la bouilloire *To turn on/off the kettle*
Bedroom	Monter/descendre/arrêter les stores *To up/down/stop blinds* Ouvrir/fermer les rideaux *To open/close curtains* Allumer/éteindre la lumière *To turn on/off the light* Allumer/éteindre les lampes *To turn on/off lamps* Mettre la lumière verte/bleue /jaune *To turn on the green/blue/yellow light* Allumer/éteindre cette lumière *To turn on/off this light* Allumer/éteindre la télé *To turn on/off* Moins/plus fort la télé *Lower/louder TV*
Living Room	Monter/descendre/arrêter les stores *To up/down/stop blinds* Moins/plus fort la radio To lower/louder the radio

Table 1: The vocal commands available in each room.

3.1.2 Emox: a non-anthropomorphic robot

For this study we chose a non-anthropomorphic robot, Emox (*see Figure 3*), develop by Awabot Company. It used an Urbi system. Ethically, the fact that this robot neither look like a human nor an animal avoids the induction of the way people picture the device and would create artefacts that cannot be controlled or be misinterpreted. Its voice is also non-human, a choice that was motivated from a previous study (Sasa, Aubergé, Franck, Guillaume, Moujtahid, 2012), which tested different types of aesthetics for the robot voice by only changing the Fundamental Frenquency (F0). At the same time, we asked people which voice they prefer and checked if the information carried by some "mouth noises" (non phonetic nor phonologic sounds; e.g. laughs and various vocalizations or breaths) were recognized. Finally, the robot has a voice pitch increased by 1.52 from the original female speaker's F0, using Voxal software[2] for voice conversion. That gives robot a "cartoon-like voice", reducing the anthropomorphism to the minimal information carried by the speech.

Figure 3: The Emox robot – Awabot company.

3.1.3 EmOz : an interface for non-programmer to control Emox and Domus

In this study, we created a Wizard of Oz interface to control both Domus and Emox, using java-programming language. In order to facilitate the use by non-programmer researchers, this interface generates buttons based on excel files in which you fill in simple parameters as sounds file name, basic moves characteristics and Domus automation actions (*see figure 4 for instance*). One excel file corresponds to one button on the interface. Each time you create a button on the interface, it is possible to drag and drop it wherever you want, and the last positions of all the buttons are saved which allows displaying previous versions of the interface.

Auteur				
Auteur du script	Yuko			
Description				
Description du script				
Paramètre	**Valeur**			
AFFICHER	1			
ZONE	2			
	P3.4 MonterStores			
Groupe	**Action**	**P1**	**P2**	**P3**
1	moveFront	25	15	
2	domus	cuisine	store	up
2	speaker	2-ok2		
2	moveBack	25	15	

Figure 4: Example of an excel script to create a button named "P2.4 MonterStores" to move Emox forward and backward, up the blinds in the kitchen while he is playing the "2-ok2" sound.

That is how it is possible to create buttons in A and D zone of Figure 5 which shows the final aspect of the interface. In A, we placed our Emox stimuli in a specific order to graduate different levels of prosody while we follow with accuracy the script that carries our hypothesis. In D, there are some complex stimuli associated to some moves or moves and sounds. The B zone generates automatically all the audio stimuli that we use while we have to do some improvisation, depending on the subjects reactions. The tool in C allows us to record our voice in live, increased the F0 to have the same voice aesthetics as the other sounds and play it on Emox if needed, because some reactions are unpredictable during the experiment. Finally, we can control all Domus automation in the E zone.

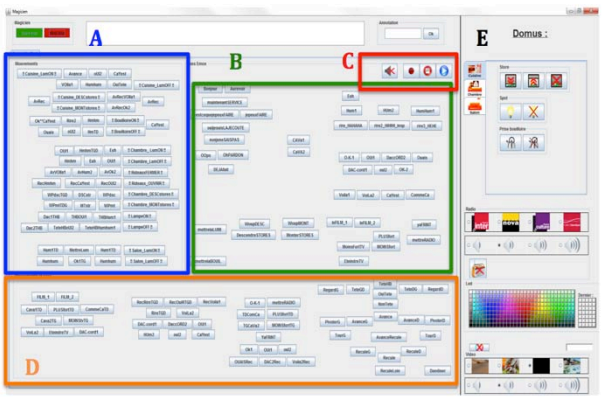


Figure 5: EmOz interface illustration.

In the left and up corner of the interface, we have got start and stop buttons generating (1) a form where you can fill in the subject's characteristics to anonymize the data; and (2) create a .csv format file with a timestamp, saving all the tracks of actions you did on the interface during the experiment.

[2] www.nchsoftware.com/voicechanger

3.2 The Elderly EmOz Expressions script

3.2.1 Communication lack appearing with aging

Aging process depends on a physical, neuropsychological, social and environmental factors (Markle-Reid & Browne, 2003) and differs among individuals. In fact, some studies showed that our age is far different from biological and cognitive age (Anstey & Smith, 1999). Generally we talk about « elderly » over 75 years old, but their frailty or non-frailty is not equivalent and not related to their age, which is difficult to focus on the subjects who are interesting to observe. This kind of persons who likely become socio-isolated by loosing progressively their role in society, communicate less frequently to finally find more and more difficulties to engage in efficient interactions with their kinfolks or other persons they are in contact with. Each times their interactions fail, the elderly loose confidence which strengthening their lack of communication abilities (Segrin, 1994), while diseases and physical problems, directly related to communication failures appears. Thereby this loneliness and the loss of social relationships are strongly related to mortality (Holt-Lunstad, Smith & Layton, 2010; Luo, Hawkley, Waite & Cacioppo, 2012). To find really needed elderly, we asked a partnership with a national home caregivers company, Bien à la Maison, to help us find people who are still living in their own but start being frail and isolated. This choice is based on the opinion of the caregivers (mostly women) who visit regularly the elderly and who accepted to be the experimenters' accomplices in our study. The company measures the frailty with their own tools and that allowed us to base on person who are scaled GIR 5 or 6, a French standard to illustrate elderly frailty and dependency (Coutton, 2001). Once the caregiver or the organism manager find a subject corresponding to our criteria, an experimenter visits the elderly for a first interview.

3.2.2 Pretext task to bring the subjects into Domus

During the first interview, the goal is to know better the subjects and to motivate them to come in Domus, our living-lab. The experimenter who is doing the recruitment introduce himself as a gerontechnology student who wants to know how people over 75 years old live and what kind of opinion they have got on technologies over a questionnaire. This gives an overall knowledge on the subjects' profile. As transition on technologies, the student says that some works have been done on a Smart Home prototype to study how we can allow seniors (who start having some physical but not too serious problems), to live as long as possible in their own house. He continues telling that in order to ease elderly's life, some researcher created technologies associated to the Smart Home but which cannot be tooled up yet at their home. So we expect the elderly to test these technologies in our flat prototype. However, in "previous studies" we observed some difficulties: when elderly change their living environment (e.g. move into a retirement/nursing home or a hospital) they mostly have trouble to accustom to this new place. Moreover, when there are technologies in this environment, people get completely confused. One of our "so-called hypotheses" to avoid this phenomena consist to ask people to bring some personal items (e.g. books, trinkets, decorative objects…etc.) and to arrange the new environment they have to handle with these items, so they can get used to the place more easily. This justification follows the idea of transitional objects sometimes used to help Alzheimer patients to be less lost (Habernas & Paha, 2002; Loboprabhu, Molinari, & Lomax, 2007). Finally, if the elderly do agree to come to the Smart Home, the student asks them to bring around ten items they care about and place their objects in Domus while evaluating it and its technologies. To help them choosing their objects, the experimenter gives a sheet where the elderly have to fill in the items they want to bring. As subjects, they will spend about two or three hours in the Smart Home. If they accept to be accompanied, we also ask them to come with their caregiver, which can ensure security. At last, the day they come to the living-lab, the student proposes to give a lift to the elderly and their caregiver, creating a situation that will facilitate the experimental scenario. He also tells that he has not visited yet the Smart Home as it was his adviser who took care of reserving Domus, so he will discover it at the same time as the subject.

3.2.3 Scenario to introduce the Emox robot and Elderly

On the experiment day, the Smart Home engineer welcomes the student, the elderly and the caregiver in a reception room. They spend some time discussing about the study context to let the elderly calm down and feel comfortable. As the engineer pretends not knowing the student and the real purpose of his work, the student explains his "hypotheses" based on the personal items that allow elderly getting used to unknown and technologized environment more easily. Once the subject is ready and he is convinced of the pretext task, the engineer introduces the Smart Home and its different rooms. Very quickly, once everyone is in Domus, a third experimenter, waiting in the control room, calls the elderly's caregiver on her mobile phone, pretending he is the home help services company manager giving a mission to his employee that cannot be refused, as it is an emergency. At this time, the caregiver is aware of every details of the experiment because she passed a private interview with the experimenters before the experiment day in which she was told how to react precisely in each step of the experiment as accomplice. So she pretends having got an urgent mission very near the Smart Home that takes less than an hour and that she has to leave a moment. As she came by the student car, she asks for a ride because her mission is very important. The student understanding the emergency proposes to accompany her. He then asks to the engineer if it is possible for him to take care of the elderly subject for a while. The engineer says that he cannot stay all the time because he has got other work to do but that he will stay as long as

necessary to explain how to use the Smart Home because its features are special. The student then demands the subject to start placing the personal items wherever the elderly wants to in the Smart Home, especially if both him and the caregiver have not returned yet after the engineer presentation. In addition, neither the student nor the caregiver know how the Smart Home works, so they ask to listen carefully to the engineer's explanation so the subject can describe all the operation while they will be back. Then, both student and caregiver leave Domus to go in the control room. Once they are gone, the engineer tells the subject that the Smart Home has not got any switches but it can be handle by vocal commands. At this moment he calls for "Emox", a robot he introduces as the butler of Domus and which will listen then execute the vocal commands. However, the engineer explains that at first, the robot has to learn the elderly voice for the effectiveness of the system (which is not true because both robot and Smart Home are controlled with a Wizard of Oz). The engineer then proposes a list of 30 possible vocal commands to the subject so he trains the robot to recognize his voice. The subject is asked to test at least once all the commands. When the elderly understood and starts giving the first commands, the engineer says that he has to go and he leaves the elderly to get into the control room, saying he will come back later to see if everything is fine.

3.2.4 Scenario for Emox and Elderly interactions

In the control room, there are two or three Wizard of Oz experimenters who: (1) drive Emox with a joystick to follow the elderly while he is moving around Domus, (2) activate Domus automation while the subject is giving a vocal command, (3) play the vocal stimulus on Emox that carries our hypotheses on the "socio-affective glue". As the subject starts giving the first vocal commands, the Wizards are just executing Domus automation without speech from the robot. Then after three of four commands, we play some "mouth noises" that illustrate pure prosody, without any lexical information (Scherer, 1994; Campbell, 2004; Schröder & al., 2006) that we supposed to be the tools and selected from a database of noise collected (Aubergé, Rilliard, Audibert, 2005), described (Aubergé, Loyau 2006; Vanpé, Aubergé, 2011) and measured (Signorello, Aubergé, Vanpé, Granjon, Audibert, 2010; De Biasi, Aubergé, Granjon, Vanpé 2012; Sasa, Aubergé, Rilliard, 2013) from previous studies. We think these noises able to engage people in the glue process to converge with Emox. Then after some of these noises, we let Emox interact with lexicons as interjections (Ameka, 1992; Poggi, 2008), carrying also glue prosody. Finally, we introduce commands imitations, always with supposed glue prosody, to reinforce the eventual established relationship as described in the chameleon effect (Schaal, 1999; Decety, 2007). The Table 2 shows the 30 stimuli used and supposed to create and reinforce the "socio-affective glue" between the elderly and the robot.

These stimuli follow an accurate order in response to each Domus command, described in a script. Nonetheless, we sometime skip some sounds for more graduated form, whether because the elderly do not follow exactly the order of the commands list, or because very naively, as human, the wizards tend to react differently when they see some specific reactions from the subjects. If these modifications are objectively observed while analyzing the corpus, this could be a model that the robot has to follow and fit to make the "glue" with elderly as human do.

Mouth noises	3 types of laughs
	euh - hum1 - hum2 - humhum
	2 onomatopoeia « woup » (associated to the blinds up/down movements)
Interjections	Sounds to play before Domus actions - prosody1: d'accord1 - ok1 - oui1 / *all right1, okay1, yes1*
	Sounds to play before Domus actions - prosody2: d'accord2 - ok2 - oui2 / *all right2, okay2, yes2*
	Sounds to play after Domus actions : ça va1 - ça va2 - comme ça - ça y est / *fine1-fine2-like this-that it*
	If Emox is mistaken : oups / *oops*
Commands Imitations	Infinitive form of the vocal commands + interjections used before
Other sentences	Some predictable minimal dialogue sentences used only if needed. E.g. : je peux faire quelque chose / *can I do something*, mais c'est déjà fait / *but it is already done*, oh pardon / *oh sorry*

Table 2: Emox robot audio stimuli.

4. The EEE Corpus

4.1 Overall description of the corpus

The subjects are, for now, from 68 to 92 years old (*see Figure 6*). It is still on recording processing, as we will try to collect around 40 subjects interactions. This corpus is composed by ten experiments lasting from an hour and a half to two hours each. For each subject, we have six videos (two per rooms) and an audio file collected by the subjects' lapel microphone. We have nearly 456 interactions between EMOX and the elderly (from 43 to 52 per subject), throughout the full experiment. Each interaction lasts about 10 to 50 seconds, showing a sequence of exchanges around one voice command. For the analysis we divided the results in three steps while the subjects: (1) are learning the commands with the engineer, (2) are alone with Emox, (3) are explaining how DOMUS and Emox work to the helper and then to the recruiter. We quote the commands forms used by the subjects, count and store them in the chronological order of appearance. Those commands are associated with punctual or gradual reactions of both the robot and the subjects, which illustrate the "socio-affective glue" degree between the robot and the elderly.

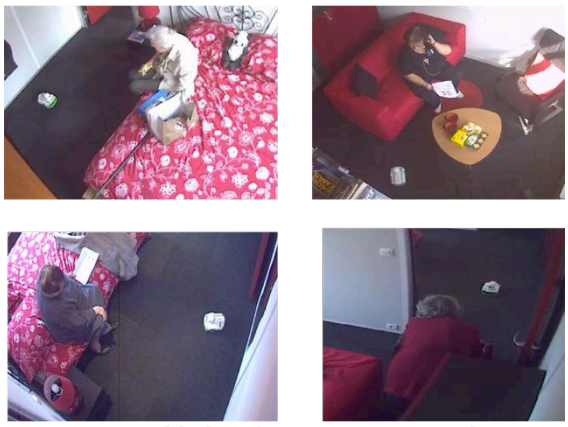


Figure 6: Some elderly subject interacting with Emox.

4.2 First analyses of EEE

There are of course some variations concerning the subject's behaviors during the experiment first steps, but some common main characteristics emerged as features of the "glue" building increasing steps: (1) declarative commands without paraphrasing; (2) the same original form commands but with a positive attitude prosody (in particular fundamental frequency arise which systematically appear at the end of the sentences, with a breathy voice); (3) commands paraphrased variations (used in synergy with a "we") with a globally high fundamental frequency and a great arise at the end of the sentences; and finally (4) multiple prosodic focuses of support terms with a higher fundamental frequency. These phenomena are observed as well as a voice quality becoming more and more breathy. This elderly's voice quality breathiness seems to vary particularly while the robot produced a feedback based on pure prosodic vocal micro-sounds.

The elderly's speech behaviors confirm that the effect of the socio-affective "glue" increases towards the prosodic levels, especially for socio-isolated people. Moreover, to allow a precise control of the robot reactions timing and order, we need an efficient interface so the cognitive effort of the Wizard of Oz experimenter is the same as the effort the robot "seems to produce" to execute the commands. Consequently an HRI technology will be specifically developed for the EEE situation, thinking of useful features add to the Smart Home. These technologies will then need important ethical considerations, leading to a functional system with a theoretical background focused on the practice of socio-affective interactions competences for socio-isolated people.

5. Conclusion

This work had both a theoretical and a technological goal: (1) to show that a strong "socio-affective glue" is built by carefully selected non lexical sounds and selected prosody on mimicry and that this glue is the base of any relation and ensures the relevance and the acceptability of the social role (here to control the smart home) (2) at least for isolated person, like elderly, whatever the role allowed to the robot, the really crucial expected role is to build a glue: the robot can train the human to relational performances and consequently help the isolated person and more efficient in the human-human communication. This has been validated both by the collected subjects expressions, the subjects' request and the professional of elderly car who assisted this experiment. The EEE corpus will be completed to a large panel of subjects, in order to build by machine learning, within hybrid system (rules on non lexical sounds hypotheses enriched and adapted by stochastic data learning). This will carry on a minimal dialog system for elderly in smart home that will be completed and augmented in active learning by telecare.by professional of care. It must be noted that the choice of a non-humanoid and non-animal like robot (to avoid the uncanny valley effect) is largely validated both by elderly and professional of care.

6. Acknowledgements

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) granted to Yuko Sasa and the Interabot project (Investissements d'Avenir, DGCIS) collaboration with Awabot (robotics) company. We thank Bien à la Maison company (elderly personal services) for their active participation to collect the corpus. Thanks also to CAPES Foundation, Brazil, for the postdoctoral scholarship granted to Leandra Antunes – BEX 18020-12-7. This work has been partly supported by the Major Program for the National Social Science Found of China (13&ZD189).

7. References

Ameka, F. (1992). "Interjections: The universal yet neglected part of speech", Journal of Pragmatics, 18, 101-118, 1992.

Anstey, K. J., & Smith, G. A. (1999). Interrelationships among biological markers of aging, health, activity, acculturation, and cognitive performance in late adulthood. Psychology and aging, 14(4), 605.

Aubergé, V., Rilliard, A., & Audibert, N. (2005). De E-Wiz à E-Clone: méthodologie expérimentale pour la modélisation des émotions et affects authentiques. Proceedings of WACA Grenoble France.

Aubergé, V., Loyau, F. (2006) Expressions d'un agent humain entre ses tours de parole.

Aubergé V., Sasa Y., Robert T., Bonnefond N., Meillon B. (2013) "Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot". In proceedings of WASSS 2013, Grenoble, France.

Aubergé, V. (2012) "Attitude vs. Emotion: A Question of Voluntary vs. Involuntary Control." In GSCP. Belo Horizonte, Brazil

Bayles K. A. and Kaszniak A. (1987) "Communication and Cognition: Normal Aging and Dementia", Little, Brown, Boston.

Bouma, H., Fozard, J.L., Bouwhuis, D.G., Taipale V.T. (2007) Gerontechnology in perspective. In Gerontechnology, Vol. 16, No 4, pp.190-216

Breazeal, C. and Aryananda, L. (2002) "Recognition of affective communicative intent in Robot-Directed speech", Autonomous Robots 12, pp 83-104.

Cacioppo, J. T., & Patrick, B. (2008). Loneliness: human nature and the need for social connection. New York: W. W. Norton & Company.

Campbell, N (2004). "Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language", Languages Resources and Evaluation, 39, 109-118.

Chaby, L.; Chetouani, M.; Plaza, M.; Cohen, D. (2012) "Exploring multimodal social-emotional behaviors in autism spectrum disorders". In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust; IEEE Computer Society: Washington, DC, USA, 2012, pp. 950–954.

Coutton, V. (2001). Évaluer la dépendance à l'aide de groupes iso-ressources (GIR): une tentative en France avec la grille aggir. Gérontologie et société, (4), 111-129.

Darling K. (2012). “Extending Legal Rights to Social Robots”. We Robot Conference, University of Miami, April 2012.

De De Biasi, G, Aubergé V, and Granjon L (2012). “Perception of Social Affects from Non Lexical Sounds.” In GSCP. Belo Horizonte, Brazil.

Decety, J. (2007). “A social cognitive neuroscience model of human empathy”. In E. Harmon-Jones & P. Winkielman (Eds.), Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior (pp. 246-270). New York: Guilford Publications.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). “Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines”. IEEE Transactions on Affective Computing 3(3), pp. 349-365.

Fonagy, P., & Target, M. (1997). “Attachment and reflective function: Their role in self-organization”. Development and Psychopathology, 9, 679-700.

Furlong, A. (2008). The Japanese hikikomori phenomenon: acute social withdrawal among young people. The Sociological Review, 56(2), pp. 309-325.

Greenberg Y., Tsuzaki M., Kato H. and Sagisaka Y. (2006) “A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech”. In Proceedings of Speech Prosody 2006, pp. 37-40.

Habermas, T., & Paha, C. (2002). Souvenirs and Other Personal Objects: Reminding of Past Events and Significant Others in. Critical advances in reminiscence work: From theory to application, 123.

Harrington, T.L., Harrington, M.K. (2000) Gerontechnology : Why and How ? Eindoven : Shaker Publishing.

Ladd D.R., & Cuttler A. (1983). Models and measurements in the study of prosody. In A. Cutler & D. R. Ladd (Eds.), Prosody: Models and measurements (pp. 1-10). Heidelberg: Springer-Verlag.

Loboprabhu, S., Molinari, V., & Lomax, J. (2007). The transitional object in dementia: Clinical implications. International Journal of Applied Psychoanalytic Studies, 4(2), 144-169.

Loyau, F., Aubergé, V. (2006). “Expressions outside the talk turn: ethograms of the Feeling of Thinking”, 5th LREC, pp.47-50.

Mac D, Castelli E, Aubergé V. (2012). “Modeling the prosody of Vietnamese attitudes for expressive speech synthesis”. Workshop of Spoken Languages Technologies for Under-resourced Languages (SLTU 2012), Cape Town, South Africa.

Mairesse, F., A. Walker, M., R. Mehl Matthias and K. Moore, R. (2007) “Using linguistic cues for the automatic recognition of personality in conversation and text”. In Journal of Artificial Intelligence Research 30, pp. 457-500.

Markle-Reid, M., Browne, G. (2003) Conceptualizations of frailty in relation to older adults. In Journal of Advanced Nursing, Vol. 44, 58-68.

Morency, L.P (2010) “Modeling human communication dynamics”, Social Sciences. Signal Processing Magazine, IEEE, 27(5), pp. 112 -116.

Morlec Y., Bailly G. and Aubergé V. (2001). “Generating prosodic attitudes in French: data, model and evaluation”. Speech Communication: 357-371.

Niitamo, V.-P.; Kulkki, S.; Eriksson, M.; Hribernik, K. A. (2006) “State-of-the-art and good practice in the field of living labs”. In Proceedings of the 12th International Conference on Concurrent Enterprising: Innovative Products and Services through Collaborative Networks, Milan, Italy, pp.349-357.

Poggi I (2008). “The language of interjections”, In COST 2012 School: 170–186.

Renault S. (2004). “Du concept de fragilité et de l'efficacité de la grille AGGIR”, in Gérontologie et société. 2004, n° 109, pp.83-107.

Sasa Y., Aubergé V., Franck P., Guillaume L., Moujtahid S. (2012). “Des micro-expressions au service de la macro-communication pour le robot compagnon EMOX”, Actes du WACAI 2012, Grenoble, pp.54-59.

Sasa, Y., Aubergé V, and Rilliard A. (2013) “Audio-Visual Micro-Expressions within Japanese-French Contrast.” In WASSS 2013. Grenoble, France

Schaal, S. (1999) “Is imitation learning the route to humanoid robots?” Trends Cognit. Sci.3, 233-242.

Scherer, K.R., “Affect bursts” (1994). In S.H.M. van Goozen, N. E. van de Poll & J.A. Sergeant (Eds.), Emotions, Hillsdale (NJ, USA), Lawrence Erlbaum, 161-193

Schröder M., Heylen D., Poggi I. (2006) “Perception of non-verbal emotional listener feedback”. In Proceedings of Speech Prosody 2006.

Schröder, M., “Experimental study of affect bursts”, Speech Communication, 40(1-2), 99-116, 2003

Segrin, C. (1994). Social skills and psychosocial problems among the elderly. Research on Aging, 16(3), 301-321.

Shibatani. M. (1990). “The languages of Japan”.

Signorello, R., Aubergé, V., Vanpé, A., Granjon, L., & Audibert, N. (2011). Indices de langue et de culture dans les micro-événements audibles et visibles de l'interaction face à face. In 9ème Rencontres des Jeunes Chercheurs en Parole 2011 (RJCP).

Sun X., K. Truong, A. Nijholt, and M. Pantic, (2012) "Automatic Visual Mimicry Expression Analysis in Interpersonal Interaction," in Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition (CVPR-W’11), Workshop on CVPR for Human Behaviour Analysis, pp. 40-46.

Tomaka, J., Thompson, S., & Palacios, R. (2006). The relation of social isolation, loneliness, and social support to disease outcomes among the elderly. Journal of Aging and Health, 18(3), 359-384.

Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005) “Understanding and sharing intentions: The origins of cultural cognition”. Behavioral and Brain Sciences 28(5), pp. 675–91.

Vacher M., A. Fleury, F. Portet, J.-F. Serignat, N. Noury (2010) Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, New Developments in Biomedical Engineering, Intech Book, pp. 645-673.

Vanpé, A., & Aubergé, V. (2010). Prosodie expressive audio-visuelle de l'interaction personne-machine. Techniques et Sciences Informatiquess, 29(spécial Agents Conversationels Animés), 880-832.

Ward, N. “Non-lexical conversational sounds in American English”, Pragmatics & Cognition, 14(1), 129-182 (54), 2006.

Wichmann A. (2000). “The attitudinal effects of prosody, and how they relate to emotion”. In proceedings of ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK.

A corpus of social interaction between Nao and elderly people

Mohamed A. Sehili[1], Fan Yang[1,2], Violaine Leynaert[4], Laurence Devillers [1,3]

[1] Department of Human Communication, LIMSI-CNRS, Orsay, France

[2] Department of Computer Sciences, University Paris 11, Orsay, France

[3] University Paris-Sorbonne, Paris, France

[4] Approche, Propara, Montpellier, France

E-mail: {sehili, fan.yang, devil}@limsi.fr

Abstract

This paper presents a corpus featuring social interaction between elderly people in a retirement home and the humanoid robot Nao. This data collection is part of the French project ROMEO2 that follows the ROMEO project. The goal of the project is to develop a humanoid robot that can act as a comprehensive assistant for persons suffering from loss of autonomy. In this perspective, the robot is able to assist a person in their daily tasks when they are alone. The aim of this study is to design an affective interactive system driven by interactional, emotional and personality markers. In this paper we present the data collection protocol and the interaction scenarios designed for this purpose. We will then describe the collected corpus (27 subjects, average age: 85) and discuss the results obtained from the analysis of two questionnaires (a satisfaction questionnaire, and the Big-five questionnaire).

Keywords: Human-Robot interaction, emotions recognition, elderly people interaction corpus

1. Introduction

To effectively understand and model the behavioral patterns of very old people in presence of a robot, relevant data is needed. This kind of data is however not easy to obtain and to share due to many factors. Databases recorded with young people might be easier to create but they do not meet the requirements of studies like the one presented in this paper.

This study takes place within the ROMEO2 research project (http://projetromeo.com/) which follows its precursor ROMEO [Delaborde & Devillers, 2010, Buendia & Devillers, 2013] and whose goals are to design a humanoid robot that can be used as an assistant for persons with loss of autonomy. The targeted population are elderly people living alone. In this work, we present the first steps toward the design of an interaction-driven system. We will present the data collection protocol used to record conversations between Nao and 27 subjects, the dialogue strategy and an analysis of two questionnaires used with each subject after each interaction .

Since ELIZA's success [Weizenbaum, 1966] most chatterbots have been emulating the same principles to overcome the Turing test. The Turing test is used as criterion of intelligence of a computer program and it assesses the ability of the program to embody a human agent in a real-time conversation with a real human and mislead them so that they cannot realize that they have actually been talking to a machine. The basic idea of ELIZA is the recognition of key words or phrases in the input of the human subject and the effective use of these words (or phrases) within preprepared or predefined replies in order to push the conversation forward in a way that seems meaningful to the human. When an input contains the words “mother” or “son”, for example, the program's respond is typically “Tell me more about your family” [Weizenbaum, 1966].

Thus, our human – robot dialog has been designed in the same spirit of ELIZA. The main challenges are to give the conversation a fairly good level of meaningfulness and make the elderly subject stick to the dialog as long as possible. However, unlike ELIZA and all chatterbots in general, as the robot is in the same room as the person, and hence visible to them, we have focused on the fact

that it should be viewed as an intelligent machine, not a human.

The goals of this work are:

- Get a first feedback of elderly people
- Validate and enhance the envisaged scenarios
- Getting a corpus for future research in interaction

This paper will present the data collection protocol and used scenarios in Section 2, the collected data (27 people, average age: 85) in Section 3, the results based on the analysis of the two questionnaires (Satisfaction, Big-five) in Section 4 and our first annotations of the commitment level of the subjects (laughs, smiles, gazes, etc.) (Section 5). Conclusion and perspectives will be reported in Section 6.

2. Data collection protocol

2.1 Targeted type of data

To effectively carry out a study on elderly people – robot social interaction, some relevant data is needed. In fact, this kind of data is rather rare and can neither be collected in a laboratory nor from TV shows or phone conversations [Castelano et al., 2010][Devillers, Schuller et al., 2012]. Furthermore, due to ethical and linguistic issues, this type of content cannot be easily shared with other researchers. Relevant corpora must, from our perspective, depict elderly subjects having spontaneous conversations with the robot.

To fulfill these requirements, our strategy was to seek the desired population in an old people's retirement home, to design a few interesting conversational scenarios that would encourage people to cooperate with the robot and to use a Wizard of Oz (WoZ) scheme to control the robot so that its behavior adapts seamlessly and quickly to most situations. The retirement home is a French EHPAD (a public accommodation for non-autonomous old people) in Montpellier. We also have focused on the fact that the robot should be viewed as an intelligent machine, not a human. Thus, many sentences are deliberately used by Nao to emphasize this such as “I have just come out of my box”, “I have just left the factory”, “I have many robot-friends” or “I need to charge my batteries”.

2.2 Conversational scenarios

The conversation is split up into many independent scenarios that must be run in a specific order. Figure 1 depicts an example of a social interaction between an elderly person and the robot Nao.

The scenarios of social interactions were:

- Greetings
- Reminder events: take medicine
- Social interaction: call a relative
- Cognitive stimulation: song recognition game

Figure 1: Example of a social interaction between an elderly people and the robot Nao

In the first scenario (Introduction - greetings) Nao introduces itself and announces its capacities (that it can speak, sing and move) to spark the person's curiosity and make them want to talk. It then asks the subject a few personal questions that include name, age, how long they have been in this accommodation, if the person has a family and so on. In the second scenario, Nao tries to draw the person into more common social conversation subjects such as today's weather and their favorite games. In this scenario it asks them about the last meal they have had and which medication they should take. In the third scenario, Nao tries to talk about family and children for encouraging the person to call parents. In the last scenario, whose main goal is to cognitively stimulate the person, Nao tries to identify the subjects that might be interesting for people such as movies, cooking, and TV programs. It then plays about thirty seconds of a few famous old French songs and asks them if they have

recognized the song's title or the performer's name.

Using a Wizard of Oz, Nao is obviously remotely controlled by a human who observes the course of the conversions and reacts accordingly. The content of each scenario is predefined and Nao (that is, the human wizard) follow a conversation tree to perform the next action (uttering a text, playing a song or performing a gesture). Furthermore, thanks to Nao's text-to-speech facility, the wizard can dynamically type and send new text such as the person's name during the conversation.

In case of tricky situations when the person insists or not following the conversation tree, the wizard uses generic sentences (46 generic sentences such as “it is true”, “yes”). The average number of phrases per session was 82 sentences. The different number of sentences of the WOZ is 265 (including the generic sentences) with lot of empathic sentences such as “I like you name”.

2.3 Wizard of OZ

The main goal of the WoZ is to take advantage of Nao's communication abilities and to build a social interaction between the robot and elderly people. Therefore, the tool we used consists in a software with a GUI and is globally designed to send the text utterances to Nao, perform gestures and play sounds (e.g. old songs). For the sake of spontaneity and quickness in Nao's reactions, almost all speech utterances are encoded beforehand. Moreover, the human wizard can dynamically update a few snippets of text (e.g. the name of the person) or add a free text to keep an appropriately high level of conversation and match the subject's current theme if they do not stick to the scenario. To make the use of free text as low as possible, many generic utterances (e.g. “Yes”, “No”, “I see”, “Can you hear me”, “I'm sorry”, etc.) were made available for the wizard. Each scene in a scenario is built as dialogue tree. At each node the wizard has, according to the subject's reaction, to choose the next node of dialogue to visit.

3. Corpus description

For this data collection we have mainly been focusing on two modalities: audio and video. A log file is also available for each conversation. It contains all Nao's timestamped actions and can be used to rebuild the dialog tree. Furthermore, it can be used to extract some useful information such repeated utterances and the time spent by the person to react to an action of Nao etc.

Beside Nao's video camera and 4 microphones, we have used an HD webcam to capture facial expression (a white screen was set up behind the person), a standard HD camera to record the whole interaction from a profile perspective and a lavaliere microphone to get an isolated high quality audio track.

The number of subjects is 27 (3 men and 24 women), recorded over two sessions (14 subjects in November 2013 and 13 in January 2014 respectively) making up around 9 hours of signal. The same hardware has been used for both sessions though each session has taken place in a different room. We also used two questionnaires for each subject.

This study has been conducted with people who are not under tutorship. They all agreed to participate to the study and signed an authorization to use and reproduce the collected images and voice. To meet the researchers, each person was individually hosted in a room within the retirement home and was made aware of the ability to stop the experiment at any time.

4. Questionnaires

After each interaction, two questionnaires have also been used: a first satisfaction questionnaire meant to evaluate the quality of the interaction with the robot and then a short version of the well known Big-five questionnaire.

4.1 Personality questionnaire

A very brief measure of the Big-Five personality domains based upon the Ten-Item Personality Inventory (TIPI) [Gosling et al., 2003] has been used. The questions relied on the own perception of oneself in a variety of situations. The subject is given a set of statements and replies by indicating the strength of his agreement with each statement on a scale from 1 to 7 (1 denotes a strong disagreement, 7 denotes a strong agreement, and the other values represent intermediate judgments). For each subject, we computed a value for each of the five

dimensions which are Emotional Stability, Extroversion, Openness, Conscientiousness and Agreeableness

4.2 Satisfaction questionnaire

As for the satisfaction questionnaire, closed-ended questions have been used. The subjects were also asked to supply answers using a 7-scale evaluation scheme. In the following we report the satisfaction questions and the respective average scores for the 27 persons between parentheses:

- (Q1) Did Nao understand you well? (5.2)
- (Q2) Did it show any empathy? (6.3)
- (Q3) Was it nice to you? (6.2)
- (Q4) Was it polite? (6.4)

For the open questions, we give a list of example answers below. For convenience, the answers have then been encoded into numerical values using different strategies. For example, for Q6 we use 1 for human names and 0 for other names. Numerical values are used to calculate correlations between the satisfaction answers and personality traits:

- (Q5) What would be the best adjective to describe the robot? (right, comic, nice, very nice, surprising, friendly, funny, sweet, pleasant)
- (Q6) Which name would you give to the robot? Some of the proposed names (only 4 persons were not able to give a name): Pierre, Michel, Alfred, rigolo (comic), Zizou, Toto, Nano, Nicolin, Jo, gentil (nice), patachou, a name of an extraterrestrial, Mikey.
- (Q7) Would you like it to address you as “tu” (using the familiar form) or as “vous” (using the formal form)? 55% of the subjects prefer the familiar form and 45% say that they have no preference. None prefers the familiar form.
- (Q8) Would you agree to redo the test with the robot? 81.5 % of the subjetcs agree.
- (Q9) Would you like to own a robot? Only 26 % of the subjects agree.
- (Q10) Would you prefer a robot that looks like a robot or a human? 55% of the subjects prefer a human-like robot.
- (Q11) Doyou consider the robot as a machine or as a friend or a companion (human)? The answer was 52% for a machine and 48 % for a friend and/or companion.

5. Analysis of the questionnaires

For a better understanding and interpretation of the collected answers, a score of correlation is calculated between the personality traits and a few of the satisfaction questions. Correlations are calculated between the satisfaction questions as well.

We used the Pearson product-moment correlation coefficient with a permutation test (using the R language). The most interesting correlation was with Emotional Stability (see Table 1).

Table1 shows the correlation between the emotional stability of a subject and a number of questions that reveal how the subject perceives the robot.

Question	Corr.	P-value
Human/Non-human name (Q6)	-0.31	0.1
Would you like to own a robot? (Q9)	-0.63	0.0003
Do you consider the robot as a machine or as a friend or a companion (human)? (Q11)	0.43	0.02

Table 1: Correlation between the “Emotional Stability” personality trait and a few of the satisfaction questions.

It should be noted that a p-value under 0.05 indicates a high significance level of the reliability of the correlation between the two variables. The negative correlation between the Q6 and the emotional stability suggests that subjects with a high emotional stability tend to give a non-human name to the robot. Although no useful conclusion can be learned from this correlation, due to a high p-value, we can interestingly observe that subjects with a high emotional stability view the robot as machine, not as a human (Q11, third row of the table).

There is also a strong correlation between the emotional stability and the fact that the subject does not want to own a robot (row two).

Among the correlations found between the satisfaction questions, we mention three pairs of questions (Table 2).

From the first row in Table 2 one can find and obvious link between the understanding level of the robot and a tendency to accept to redo the test. From the second row, we can conclude that the more the robot is viewed as a machine, the less people want to have one. As for the third row, we can learn that if people give a human name to the robot, they tend to agree when it comes to owning a robot.

Pair of questions	Corr.	P-value
Did Nao understand you well? (Q1) – Would you agree to redo the test (Q8)	0.38	0.04
Would you prefer a robot that looks like a robot or a human? (Q10) – Would you like to own a robot? (Q9)	-0.62	0.0004
Human/Non-human name (Q6) – Would like to have own a robot? (Q9)	0.41	0.03

Table 2: Correlation between a couple of satisfaction questions.

6. Annotations

Given the content of the corpus, there are many strategies to annotate its content. Each strategy may apply to different levels of information. Annotations can apply to both audio and video streams. For audio streams for instance, we can focus on non-verbal cues such as laughter, or cues that suggest that the person does not understand what the robot is saying. For video streams, attention can be given to visual cues on the face. Such detailed annotations require both time and human effort.

In this work we have followed a behavioral annotation scheme. Thought has been given to the commitment level of the subject during conversation. In this regard, we have been interested on how much a subject looked at the robot, how well did they understand it and how much did they imitate it. Furthermore, many non-verbal cues such laughter, smile, surprise have been annotated. Annotations have been carried out by two experts on the data collected in the first session (14 persons). As a first result, we report the presence of laughs and smiles from all subjects through the conversation. A complete annotation of the whole corpus is being carried out. A more detailed annotation should also be done afterwards. It will include a more comprehensive annotation of audio and image cues.

7. Conclusion and future work

In this paper we present a data corpus of a social interaction between a humanoid robot and elderly people. This work is part of the Romeo 2 project. The corpus contains 27 conversations with an average duration of 20 minutes.

To the best of our knowledge This kind of data is relatively rare as it is a very challenging task to record people over 80 years old. Furthermore, it could be very difficult to share this kind of data.

The recorded subjects were furthermore asked to answer a set of questions from two different questionnaires. They collected answers are used in this paper as first evaluation data.

As a first result, we can conclude from the answers that the majority of the subjects shown an interest in the experiment. This could be backed by the presence of laughs and smiles as observed in the first annotations.

Moreover, many interesting correlation could be shown, be that between the elements belonging to the same questionnaire or to two different questionnaires. One of the major conclusions is that the way an elderly person interacts with the robot depends on their personality. Therefore we intend in future work to dynamically build a user profile and adapt the robot's behaviour accordingly.

The analysis of social interaction between elderly people and a robot allowed us to get a first feedback of the concerned people and to validate and enhance our interaction scenarios. This corpus will be used as an evaluation data for further experiments.

8. Thanks

Thanks are due to all the participants to these experiments (especially to the EHPAD of Montpellier) and to the partners of the ROMEO2 project for the interest they expressed for this data collection.

9. References

A.Buendia, L. Devillers (2013). From informative cooperative dialogues to long-term social relation with a robot. Towards a Natural Interaction with *Robots*, Knowbots and Smartphones, IWSDS 2013, Springer.

G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P.W. McOwan (2010). Affect recognition for interactive companions: challenges and design in real world scenarios. Multimodal User Interfaces, 2010.

A.. Delaborde, L. Devillers. (2010) Use of Nonverbal Speech Cues in Social Interaction between Human and Robot: Emotional and Interactional Markers, in AFFINE '10: Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, ACM, October 2010.

L. Devillers, B., Schuller, A., Batliner, P., Rosso, E., Douglas-Cowie, R., Cowie, and C. Pelachaud, editors (2012). Proceedings of the 4th International Workshop on EMOTION SENTIMENT & SOCIAL SIGNALS 2012 (ES 2012) Corpora for Research on Emotion, Sentiment & Social Signals, Istanbul, Turkey. ELRA, ELRA. held in conjunction with LREC 2012.

S. D. Gosling, S. D., P. J. Rentfrow, P. J., W. B Swann, W. B., Jr. (2003). Ten-Item Personality Inventory (TIPI). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528

J. Weizenbaum, (1966), "ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine", *Communications of the ACM* 9 (1): 36–45

Fear and Trembling: Annotating Emotions in Czech Holocaust Testimonies

Kateřina Veselovská
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Prague, Czech Republic
E-mail: veselovska@ufal.mff.cuni.cz

Abstract

In this paper, we introduce the Visual History Archive of the USC Shoah Foundation as a multimodal data resource for sentiment analysis in Czech and potentially in all thirty three languages it contains, taking the opportunity of having both physical access to these unique and highly emotional data and the established research group on sentiment analysis at Charles University in Prague. We describe the Czech portion of the archive data and its three-layer transformation present in the Prague DaTabase of Spoken Czech 1.0. Also, we provide a preliminary methodology for sentiment annotation of the multimodal data. Using the recently released Czech subjectivity lexicon, we employ subjectivity detection, i.e. automatic identification of whether a given sentence expresses opinion or states facts, within a treebank in spoken term detection. Moreover, we introduce a new extension of the tree annotation graphical editor TrEd and basic guidelines for annotating emotions in the Czech dependency data.

Keywords: sentiment analysis, multimodal data, visual history archive

1. Introduction

The main resource of the data used in the present contribution is known as the Visual History Archive (VHA) of the USC Shoah Foundation[1]. The archive was founded by Steven Spielberg after releasing the historical drama film "Schindler's List" and it contains almost 52,000 witness testimonies of Holocaust survivors (later extended also with testimonies of survivors of Rwandan, Cambodgian or Armenian genocide) covering the history of entire 20th century. Since it is a very large collection of corpora, filmed interviews are fully accessible only through the access points spread around the world, three of them situated in Europe. MALACH (Multilingual Access to Large Spoken Archives) Centre for Visual History in Prague[2] was officially opened in 2010. On six separate working stations located in the Library of the Faculty of Mathematics and Physics of Charles University, users can search for and view testimonies of interest by using more than 55,000 keywords or a database of 1.1 million names. The testimonies available in the Malach Centre were recorded in 57 countries and in 33 languages, which makes a total amount of about 116,000 hours of video. The Refugee Voices archive provided by the Association of Jewish Refugees complements this collection with additional 150 interviews.

Since survivor testimonies are highly emotional and generally full of very significant affective behaviour like crying, sighing and trembling (but also of positive emotions, such as laughter or weeping with joy), it represents perfect training data for multimodal sentiment analysis and affective speech modelling. The present paper describes the first steps towards multimodal sentiment analysis in Czech.

2. Related Work

The issue of a text-based sentiment analysis has been addressed many times, e.g. in connection with sentiment detection on product reviews (Hu & Liu, 2004), news articles (Balahur et al., 2010) or blogs (Balog et al., 2006). The issues of sentiment analysis in Czech have been tackled by Veselovská, 2012, Veselovská et al., 2012 and Habernal et al., 2013.

Apart from data-driven methods, most of the researchers use the rule-based classifiers along with subjectivity lexicons for the opinion mining task. There is a number of papers dealing with the topic of building subjectivity lexicons for various languages (see e.g. Baklival et al., 2012, De Smedt & Daelemans, 2012, Jijkoun & Hofmann, 2009 or Perez-Rosas et al., 2012). The method for building Czech subjectivity lexicon used in this article is described in detail in Veselovská (2013).

Concerning the affective data for sentiment analysis, one of the most widely used manually annotated corpora is the MPQA corpus (Wiebe et al., 2005). Another manually annotated corpus is the collection of newspaper headlines created during the SemEval 2007 task on affective text (Strapparava & Milhacea, 2007) annotated with the six Eckman emotions (anger, disgust, fear, joy, sadness, surprise) and their polarity orientation (positive, negative). In the present paper, we use the Visual History Archive of USC Shoah Foundation and the Prague DaTabase of Spoken Czech (Hajič et al., 2008)[3] as a data resource for multimodal sentiment analysis in Czech, or more precisely for manual annotation of emotional

[1] Available from *http://sfi.usc.edu/*.
[2] *http://ufal.mff.cuni.cz/cvhm/index-eng.html*
[3] Available from *http://hdl.handle.net/11858/00-097C-0000-0001-4914-D.*

utterances.

Moreover, this work builds upon the research related to multimodal sentiment analysis, i.e. on papers combining different audio-visual features for sentiment detection or combining audio-visual and text features for sentiment analysis, mostly in connection with annotating emotional videos posted on the web (Morency et al., 2011 or Rosas et al., 2013).

3. Data

3.1 Czech Portion of VHA

In the present contribution, we consider only the Czech part of the archive. The Czech language data contain 566 testimonies including the testimonies from the Museum of Romani Culture in Brno which provided the much needed 40 records of genocide and persecution of the Roma (in Czech and Slovak language). Altogether, it suppliesmore than 1,000 hours of video material – the amount of data which is still prohibitive for complete manual annotation (verbatim transcription). The size of the data also posed a challenge for the designers of a retrieval system that works in (or very near to) real time. However, Psutka et al. (2011) employed automatic speech recognition and information retrieval techniques to provide improved access to this large video archive. The resulting system is able to search through the video constituting the Czech portion of the archive and find query word occurrences in the matter of seconds. The phonetic search implemented alongside the search based on lexicon words allows researchers to find even words outside the automatic speech recognition system lexicon such as names, geographic locations or Jewish slang.

3.2 PDTSC1.0

Except for the multimodal form, all the Czech data from the Visual History Archive were transformed into the first version of the Prague DaTabase of Spoken Czech (PDTSC 1.0). The PDTSC has three hierarchical layers and one external base layer (audio), see Figure 1., annotation of the sentence *I think the relationships between the classmates were good.*

Figure 1. Linking the layers

The bottom layer of the corpus (z-layer) contains automatic speech recognition output aligned to audio. It is a simplified token layer which is interlinked with the manual transcription using synchronization points. The second layer (w-layer) is a literal manual transcript, i.e. everything the speaker has said including all slips of the tongue, coughing, laugh etc. The transcription was produced in Transcriber (Baras et al., 2001). The XML-output from Transcriber has been converted into PML (Prague Markup Language, Pajas & Štěpánek, 2009)[4], which is an XML subset customized for multi-layered linguistic annotations.

The actual annotation was performed in the editor MEd[5], an editor of interlinked multi-layered linearly-structured linguistic annotations which is the main annotation tool that is being used for the speech reconstruction annotation (see Figure 2). MEd can handle PML directly, and can work with all of the audio, ASR transcription, manual transcription and the speech-reconstruction annotation at the same time.

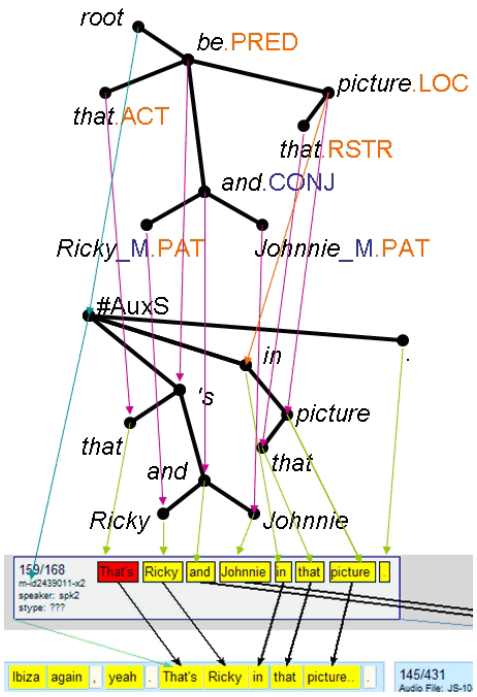


Figure 2. Layers of annotation in MEd

By means of XML references, the transcription is interlinked with the tokens at the bottom z-layer and synchronized with the audio track. The topmost layer (m-layer), called speech reconstruction, is an edited version of the literal transcript. Disfluencies are removed and sentences are smoothed to meet written-text standards. The highest level was further subjected to automatic morphological annotation (tagging, lemmatization) and then the text was automatically parsed by TectoMT (Popel & Žabokrtský, 2010) and transformed into the working version of the Czech treebank of spoken language. For the sentiment annotation task, we take into account also these automatically generated trees in order to detect the opinion target and source.

4. First Step: Using Czech Subjectivity Lexicon

To obtain the first version of the set of evaluative items, i.e. words or phrases inherently bearing a positive or

[4] Available from *http://hdl.handle.net/11858/00-097C-0000-0022-C7F6-3.*
[5] Available from *http://hdl.handle.net/11858/00-097C-0000-0001-48F8-6.*

negative value, in the PDTSC corpus (and, consequently, the Czech part of the Visual History Archive), we have used the Czech subjectivity lexicon[6]: all items present in the lexicon were marked as potentially evaluative. This result was then manually refined in the steps described below. Although Holocaust testimonies, the main source of the PDTSC texts, were supposed to be highly emotional, this step also served as a quick screen determining whether the data can be used at all.

The Czech subjectivity lexicon contains 4,626 evaluative items (1,672 positive and 2,954 negative) together with their part of speech tags, polarity orientation and source information. The core of the Czech subjectivity lexicon has been obtained by automatic translation of a freely available English subjectivity lexicon downloaded from http:// http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/. For translating the data into Czech, we used CzEng 1.0 (Bojar & Žabokrtský, 2006)[7], a parallel corpus containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation. The reliability of the final lexicon was evaluated by comparing it against several previously trained classifiers (see Veselovská et al., 2012).

5. Second Step: Using a New TrEd Extension for Sentiment Annotation

Using the Czech subjectivity lexicon, we have identified potentially evaluative sentences in PDTSC. However, to verify whether the evaluative items were actually used in an evaluative context, it was necessary to review the data manually. For this purpose, we built PML_T_Sentiment, a new extension for TrEd, a tree annotation editor[8]. The extension provides the following GUI supporting the entry and modification of sentiment related information:

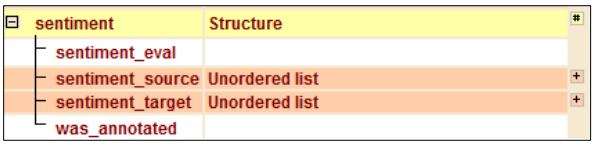


Figure 3. GUI for sentiment annotation

All the polarity items obtained from the subjectivity lexicon and found in the dependency data are highlighted, so that the annotators could easily check one occurrence after another. They are also assigned the primary polarity from the lexicon (using two different colours, green for positive polarity and red for negative polarity). Moreover, the evaluative chunk of the above text is marked with yellow. If the polarity is correct in the given context, the annotator confirms this. If the actual polarity does not correspond with the polarity from the lexicon, it can be altered manually by changing the value of the attribute *sentiment_eval* (attribute concerning the anchor of evaluation). The annotator can choose from various options, depending on the polarity of the given evaluative item: POS for positive, NEG for negative or none when the item is not evaluative at all in this particular context. Once an item was checked/corrected, it is marked both visually and by setting the attribute *was_annotated* to the value of 1.

As for the *sentiment_source*, the assigned value can be either the identifier of the source node in the treebank, or *is_external*, when the source is e.g. the author of the text. This holds also for the *sentiment_target* attribute.

6. Benefits of Sentiment Annotation of Dependency Structures in PDTSC

The annotation described above allows us to effortlessly find the original source and target, which would not be possible within a plain text. Since Czech is a pro-drop language, one needs to employ the additionally generated nodes in order to detect either sources or targets on a deep-syntax layer. Both source and target nodes are clearly marked with the arrows of different colours, which are interlinked with the arrows for coreference. As can be seen in Figure 4, the pink arrow points to the target of evaluation and since it is the substitute node for personal pronoun, it leads through the green arrow to another tree containing the real target.

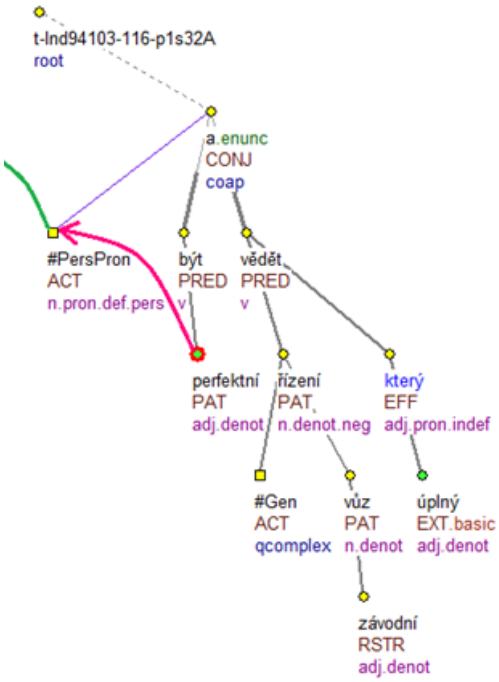


Figure 4. *Byl perfektní a věděl o řízení závodního vozu úplně všechno.*
He was great and he knew everything about racing.

Moreover, it is much easier to assign the target attributes, no matter how far they are from the governing word in the surface structure. In the treebank, one can see the whole dependency subtree immediately. The state-of-the-art research of the evaluative structures has shown that in the basic predicate-argument structure, the source is usually a grammatical subject and the target tends to be

[6] Available from *http://hdl.handle.net/11858/00-097C-0000-0022-FF60-B.*
[7] Available from *http://hdl.handle.net/11858/00-097C-0000-0001-4916-9.*
[8] Available from *http://hdl.handle.net/11858/00-097C-0000-0001-48F7-8.*

in most cases an object (see e.g. Joshi & Penstein-Rosé, 2009 or Qiu et al., 2011). Thus, we can find the sources and targets of evaluative verbs from the Czech subjectivity lexicon by parsing the data.

Another advantage of using the dependency data could be an easy negation detection. In plain text, both sentential and constituent negation in Czech is usually a part of the verb and thus it is difficult to distinguish between the two, i.e. to find the negative scope. This does not hold for the dependency data, where the scope of negation is easily recognizable since it is represented by a separate node. Therefore we can detect the negated items and in consequence switch their polarity (or the polarity of the whole sentence, depending on the negation type).

7. Conclusion and Future Work

We introduced the first steps towards annotation of the Czech portion of Visual History Archive of USC Shoah Foundation – namely a creation of a manually annotated treebank of Czech spoken evaluative sentences based on the multimodal data from Czech Holocaust survivor testimonies. Currently we are undertaking a pilot annotation of a small set of sample sentences to prove the usability of the current TrEd extension and the suitability of the newly provided guidelines for such a task. After that, we would like to run the first round of the sentiment annotation followed by more fine-grained annotation where other sub-attributes, such as *sentiment_type* for different types of emotional statements (e.g. judgement, appraisal, excitement etc.) would take place. After tagging the data, an analysis of the annotation using statistical methods would be applied. In either case, we would like to connect the emotional sentences found in the treebank corpora with the Visual History Archive recordings by spoken term detection provided by Psutka et al. (2011) and investigate the relationship between the linguistic structure and audiovisual component of the data. Moreover, the tagged data will thus be prepared as training data for future sentiment analysis and opinion mining experiments.

8. Acknowledgement

The research described herein has been supported by the by SVV project number 260 140, by the LINDAT/CLARIN project funded by the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2010013, and by the travel funds of the Center for Visual History Malach, funded by Charles University in Prague.

This work has been using language resources developed and stored by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

9. References

Bakliwal, A., Piyush, A. & Varma, V. (2012). Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Poliquen, B. & Belyaeva, J. (2010). Sentiment analysis in the news. In Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010), pp. 2216-2220.

Balog, K., Mishne, G., & De Rijke, M. (2006). Why are they excited?: identifying and explaining spikes in blog mood levels. In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, pp. 207-210.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication, 33(1), pp. 5-22.

Bojar, O. & Žabokrtský, Z. (2006). CzEng: Czech-English Parallel Corpus, Release version 0.5. Prague Bulletin of Mathematical Linguistics, 86. Univerzita Karlova v Praze, ISSN 0032-6585, pp. 59-62.

De Smedt, T. & Daelemans, W. (2012). Vreselijk mooi! (terribly beautiful): A subjec-tivity lexicon for dutch adjectives. In Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).

Habernal, I., Ptáček, T., & Steinberger, J. (2013). Sentiment analysis in Czech social media using supervised machine learning. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 65-74.

Hajič J., Cinková S., Mikulová M., Pajas P., Ptáček J., Toman J. & Urešová Z. (2008). PDTSL: An Annotated Resource For Speech Reconstruction. In Proceedings of the 2008 IEEE Workshop on Spoken Language Technology. IEEE, Goa, India, ISBN 978-1-4244-3472-5, pp. 93-96.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177.

Jijkoun, V. & Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. In Proceeding of EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference.

Joshi, M., & Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

Morency, L. P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces, pp. 169-176.

Pajas, P., & Štěpánek, J. (2009). System for querying syntactically annotated corpora. In Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, ISBN 1-932432-61-2, pp. 33-36.

Perez-Rosas, V., Banea, C. & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012).

Popel M. & Žabokrtský Z. (2010). TectoMT: Modular NLP Framework. In: Lecture Notes in Computer Science, Vol. 6233, Proceedings of the 7th

International Conference on Advances in Natural Language Processing (IceTAL 2010), Springer, Berlin/Heidelberg, ISBN 978-3-642-14769-2, ISSN 0302-9743, pp. 293-304.

Psutka, J., Švec, J., Psutka, J. V., Vaněk, J., Pražák, A., Šmídl, L., & Ircing, P. (2011). System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. EURASIP Journal on Audio, Speech, and Music Processing, 2011(1), pp. 1-11.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. Computational linguistics, 37(1), pp. 9-27.

Rosas, V., Mihalcea, R., & Morency, L. (2013). Utterance-Level Multimodal Sentiment Analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 973–982.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 70-74.

Veselovská, K. (2012). Sentence-level sentiment analysis in Czech. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantic, ACM , New York, NY, USA, ISBN 978-1-4503-0915-8, pp. 65-69.

Veselovská, K., Hajič Jr., J. & Šindlerová, J. (2012). Creating annotated resources for polarity classification in Czech. In Empirical Methods in Natural Language Processing – Proceedings of the Conference on Natural Language Processing 2012, Eigenverlag ÖGAI, Wien, Austria, ISBN 3-85027-005-X, pp. 296-304.

Veselovská, K. (2013). Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification. In Proceedings of SLOVKO, 7th International Conference of NLP, Corpus Linguistics and E-Learning. RAM-Verlag, Lüdenscheid, Germany, ISBN 978-3-942303-18-17, pp. 279-284.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2-3), pp. 165-210.

Using Ambiguous Handwritten Digits to Induce Uncertainty

Heather Pon-Barry

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
ponbarry@asu.edu

Abstract

The lack of ground truth labels is a significant challenge in the field of automatic recognition of emotion and affect. The most common approach to acquiring affect labels is to ask a panel of listeners to rate a corpus of spoken utterances along one or more dimensions of interest. In this paper, we describe a method that uses ambiguous handwritten digits for the purpose of inducing natural *uncertainty*. Using a crowdsourcing approach, we quantify the legibility of each handwritten digit. These images are integrated into visual stimuli that are used in a lab experiment for eliciting spontaneous spoken utterances of varying levels of certainty. While we cannot measure a speaker's actual internal level of certainty, our method generates a novel and interesting approximation for internal certainty.

Keywords: Uncertainty, Methodology for Speech Elicitation, Affect Labels and Ground Truth

1. Introduction

Although significant progress has been made in recent years, the problem of automatically recognizing a person's emotional or cognitive state faces many challenges (Schuller et al., 2011). One of the main challenges is in obtaining ground truth labels for a person's emotional or cognitive state. The most common approach to obtaining labels is to measure perceived emotion, as annotated by one or more human judges. This produces labels that are by definition subjective. We treat them as a gold standard, understanding that the subjectivity makes for a challenging classification problem (Devillers et al., 2005).

In this paper, we present a method for inducing natural *uncertainty* in the context of collecting a corpus of affective speech. We use a crowdsourcing approach to identify a set of ambiguous handwritten digits and to calibrate the difficulty of deciphering each digit. The handwritten digit images are integrated into visual stimuli that are used in a question-answering lab experiment for eliciting spontaneous spoken answers of varying levels of certainty. Details on the speech elicitation, the annotation of uncertainty, and the resulting Harvard Uncertainty Speech Corpus are presented in a separate paper (Pon-Barry et al., 2014).

In previous on recognizing uncertainty, there is little control over how uncertain a person is. To obtain labels for level of certainty, researchers have utilized annotators to label perceived certainty (Litman and Forbes-Riley, 2006). In our past work, we compared perceived level of certainty to speaker self-reported level of certainty. We found that self-reported certainty was often lower (rated as less certain) than perceived certainty (Pon-Barry and Shieber, 2011). In that work, we did not attempt to control the speaker's internal level of certainty. As a result, there was no way to verify whether the perceived certainty or the self-reported certainty was closer to his or her *actual* certainty.

Our interest in improving uncertainty detection is motivated by applications for personalized learning in tutorial dialogue systems, where we are most interested in knowing a student's internal level of certainty. There is evidence indicating that adapting to uncertainty can improve learning, but also that accurately detecting uncertainty is a bottleneck for fully-automated adaptive systems (Forbes-Riley and Litman, 2011). Skilled human tutors can gauge a student's level of certainty and tailor the dialogue appropriately. For example, if a student feels certain but gives an incorrect answer, it may be due to a misconception. Studies of learning in human tutorial dialogue suggest a strong connection between impasses (such as misconceptions) and student learning, to the point of proposing that *cognitive disequilibrium* is a necessary precursor to deep learning (VanLehn et al., 2003; Craig et al., 2004).

We describe in this paper a method for approximating internal certainty based upon crowdsourced judgements of handwritten image legibility. We create speech elicitation stimuli around these images that enable the creation of a speech corpus with three kinds of certainty labels: approximate internal certainty, self-

reported certainty, and perceived certainty (Pon-Barry et al., 2014).

2. Legibility Scores for Handwritten Digits

Here, we discuss our procedure for obtaining the set of handwritten digit images and describe a human computation approach to quantifying the legibility of each image. We make use of the MNIST database of handwritten digit images (LeCun et al., 1998). The database contains 10,000 handwritten digit images from the United States Postal Service.

Our process of selecting handwritten digit images and generating legibility scores has three steps.

1. Identify 400 candidate images (out of all 10,000 images) that may have low legibility.
2. Generate legibility scores for these 400 images via crowdsourcing.
3. Narrow down set of 400 images to identify 50 images with varying legibility scores.

The following sections describe these steps in detail.

2.1. Identify Candidate Images

In the first step, we use an existing support vector machine classifier (Maji and Malik, 2009) to classify all the images in the MNIST database. This classifier outputs a confidence measure along with the most likely label. The 400 images with the lowest confidence measures are used in the crowdsourcing experiment.

2.2. Crowdsourcing Legibility Scores

In the second step, we generate legibility scores for these 400 images by crowdsourcing human labels on Amazon's Mechanical Turk. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks*, or HITs (Paolacci et al., 2010; Mason and Suri, 2011). Our crowdsourcing approach enables each image to be labeled by 100 humans in a short amount of time.

We divide the digit images into twenty sections so that each HIT consists of 20 images. We instruct workers to identify each digit using a drop-down menu. Figure 1 shows a screenshot of the Mechanical Turk HIT. Pon-Barry (2013) includes the full instructions and experiment settings.

We generate a legibility score for each image based on the *entropy* of the human label distribution, a measure

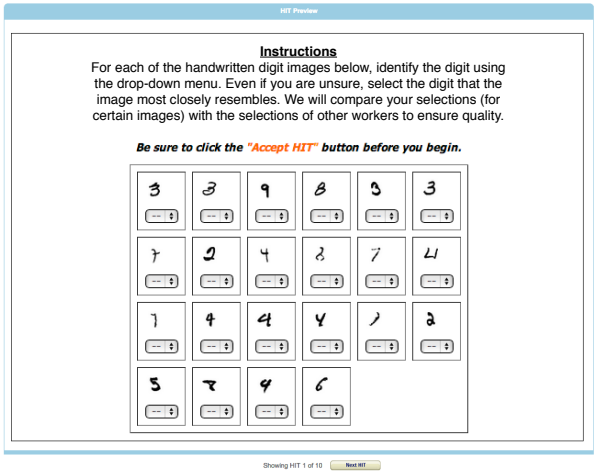


Figure 1: Screenshot of the Mechanical Turk HIT for handwritten digit legibility scores.

of the uncertainty of a random variable X taking on values $x_1, \ldots x_N$ defined by,

$$H(X) = -\sum_{i=1}^{N} P(x_i) log P(x_i) \quad .$$

Using the labels collected on Mechanical Turk, we can compute the maximum likelihood estimate for the probability $P(x_i)$. We take the legibility score to be $1 - H(X)$.

Thus, legibility scores fall in the range [0,1]. A legibility score of 1 (entropy of 0) indicates high legibility (all 100 people choose the same label).

Table 1: Handwritten digits of varying legibility. The individual label frequencies and legibility scores are shown in the columns below each image.

	Crowdsourced Label Frequencies				
Label					
'0'	-	-	-	-	2
'1'	-	-	-	5	34
'2'	-	22	-	-	9
'3'	-	-	-	-	20
'4'	-	-	69	-	4
'5'	100	-	-	-	15
'6'	-	1	31	-	3
'7'	-	77	-	58	5
'8'	-	-	-	-	8
'9'	-	-	-	37	-
Entropy	0.00	0.25	0.27	0.36	0.81
Legibility Score	1.00	0.75	0.73	0.64	0.19

Table 2: The distribution of legibility scores for the 400 images that were classified by human workers on Mechanical Turk.

Legibility Score s	Number of Images
$0.1 < s < 0.2$	1
$0.3 < s < 0.4$	1
$0.4 < s < 0.5$	3
$0.5 < s < 0.6$	2
$0.6 < s < 0.7$	8
$0.7 < s < 0.8$	26
$0.8 < s < 0.9$	33
$0.9 < s < 1$	181
$s = 1$	146

Table 1 shows five digits of varying legibility, the frequencies of the human labels, and the associated entropy values and legibility scores. Table 2 shows the frequency of legibility scores for the 400 images that were classified by workers on Mechanical Turk.

Ensuring Quality. Preventing malicious behavior (e.g., artificial bots designed to complete all the HITs in a batch) is a challenge for researchers collecting data on Mechanical Turk (Ipeirotis et al., 2010; Callison-Burch and Dredze, 2010). We take two measures to ensure worker quality. First, we include a question, such as "What is 4+2?", to verify that the worker is a real person. Second, we include two control images in every HIT. Before paying workers, we verify that they correctly identify the control images.

Experiment Running Time. Our Mechanical Turk experiment was staged in two rounds, with 10 unique HITs per round. Round 1 took 126 hours (about five days) to complete with an average time/HIT of 72 seconds. Round 2 took 33 hours (about one and a half days) to complete, with an average time/HIT of 61 seconds.[1]

2.3. Narrow Down Set of Images

In the final step, we identify 50 images to use in the speech elicitation stimuli based on the entropies of the human-label distributions. We drew uniformly (as uniformly as possible) from the binned range of legibility scores. The resulting set of 50 images is shown in Figure 2. The images are displayed from easiest to hardest (low entropy to high entropy) starting from the top-left and moving left-to-right across the rows.

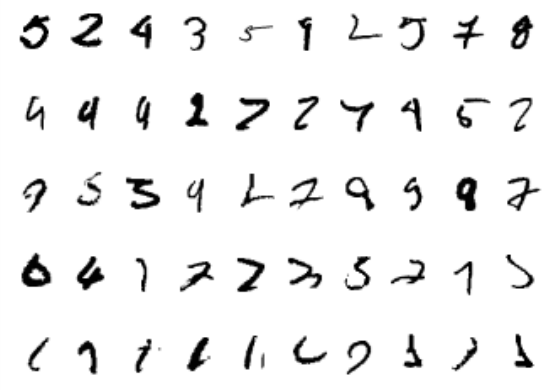


Figure 2: Handwritten digit images of varying legibility, ordered from easiest to hardest.

2.4. Image Ambiguity

When generating legibility scores, we assume that ambiguous images will appear ambiguous to nearly all people. To test this, we conducted a second experiment on Mechanical Turk that asked 100 people whether they found an image to be ambiguous or unambiguous. Figure 3 shows the fraction of people who rated an image as unambiguous versus the image's legibility score. The distribution confirms our hypothesis. Images found unambiguous by a majority of people all have legibility scores in the upper range (greater than 0.75).

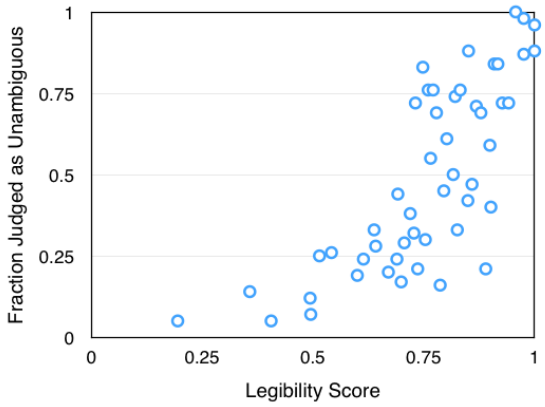


Figure 3: For each image, the fraction of people who judged it to be unambiguous vs. its legibility score.

3. Integrating Images into Stimuli

The materials for eliciting speech are designed so that participants utter a specific digit aloud in the context of answering a question. The handwritten digit images are embedded in an illustration of a train route

[1]The two experiment rounds were identical in all ways except for the images themselves. We speculate that Round 2 took less time than Round 1 due to the time of posting, i.e., weekday vs. weekend.

connecting two U.S. cities. The handwritten digit indicates the train number. An example train route illustration is shown in Figure 4. The handwritten digit on the train was identified as a '7' by 76 people, as a '2' by 22 people, and as a '6' by 2 people.

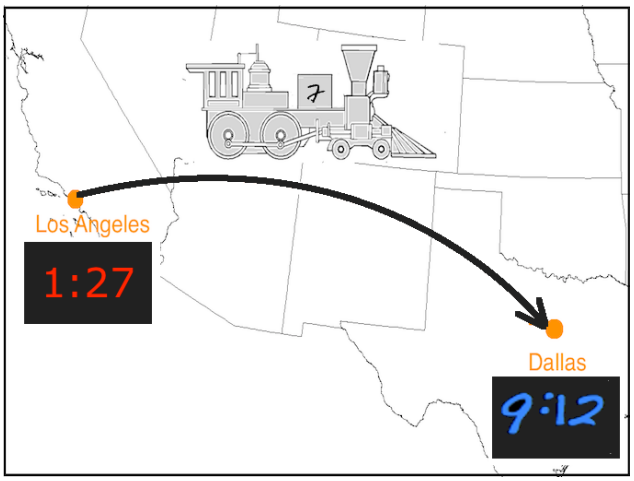


Figure 4: Speech elicitation stimulus integrating an ambiguous handwritten digit indicating the train number.

At the start of the data collection experiment, participants read a task scenario explaining why they are deciphering handwritten train conductor notes and answering questions about them. A question that requires reading the train number is asked and participants respond spontaneously. For example:

Q: Which train leaves Los Angeles and at what time does it leave?

A: Train number seven leaves Los Angeles at 1:27.

Although the question responses are spontaneous, word choice is influenced by a warm-up task where participants are given answers to read aloud. This lets us have indirect influence over the length and lexical content of the utterances, which aids future analysis of utterance-level and word-level prosody.

The key point is that we can assign each image a legibility score, based on the crowdsourced judgements. We assume that when participants are trying to read the digits, their internal certainty is proportional to the image's legibility score. We compare two kinds of certainty labels to these legibility scores: labels from the speaker's perspective and labels from the hearer's perspective. The former, labels from the speaker's perspective, are more strongly correlated with the legibility scores (Pon-Barry et al., 2014).

4. Harvard Uncertainty Speech Corpus

The results of our Mechanical Turk experiment and speech elicitation stimuli are available to the research community through the Dataverse Network.[2] At this site, researchers can also access the level of certainty annotations, acoustic feature vector data, and request access to the audio data. Details on the Harvard Uncertainty Speech Corpus can be found in previous and concurrent published works (Pon-Barry and Shieber, 2011; Pon-Barry et al., 2014).

5. Discussion and Conclusion

In this paper, we introduced a novel method for approximating internal certainty based upon crowdsourced judgements of handwritten image legibility. We collected affective speech in a controlled experiment in a laboratory setting that utilized these images. This allowed us to analyze subtle differences in prosodic expressiveness to better understand individual speaking styles (Pon-Barry and Nelakurthi, 2014). However, there are limitations associated with speech collected in a lab. Integrating these images into new experiments to collect spontaneous affective speech in real-world learning and tutorial environments is an exciting avenue for future research.

This work addresses an issue central to human language technologies and affect recognition: what are the best practices with respect to measuring speaker affect and speaker state? We have presented a method for identifying ambiguous handwritten digits for the purpose of inducing natural uncertainty and we used crowdsourcing to generate a legibility score for each handwritten digit. While crowdsourcing has been used as a way of obtaining labels for a given audio or video segment, we claim that it also has utility in designing stimuli for inducing natural affect. Our work is done in the context of examining uncertainty, though the method is applicable to other forms of affect as well, ones where the source of the affectual state is manipulable.

6. Acknowledgements

This work was supported in part by a National Science Foundation Graduate Research Fellowship and by a Siebel Scholarship. We thank Vanessa Tan and Spencer de Mars for their contributions to the Mechanical Turk experiment, and Nicholas Longenbaugh and Stuart Shieber for their contributions to the development of the Harvard Uncertainty Speech Corpus.

[2] http://dvn.iq.harvard.edu/dvn/dv/ponbarry

7. References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53:1115–1136.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67. ACM.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Diane Litman and Kate Forbes-Riley. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

Subhransu Maji and Jitendra Malik. 2009. Fast and accurate digit classification. Technical Report UCB/EECS-2009-159, EECS Department, University of California, Berkeley.

Winter Mason and Siddharth Suri. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44:1–23.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

Heather Pon-Barry and Arun Reddy Nelakurthi. 2014. Challenges for robust prosody-based affect recognition. In *Proceedings of Speech Prosody*.

Heather Pon-Barry and Stuart M. Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011(251753).

Heather Pon-Barry, Stuart M. Shieber, and Nicholas Longenbaugh. 2014. Eliciting and annotating uncertainty in spoken language. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC)*.

Heather Pon-Barry. 2013. *Inferring Speaker Affect in Spoken Natural Language Communication*. Ph.D. thesis, Harvard University.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53:1062–1087.

Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B. Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249.

Can We Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds

Eshrag Refaee and Verena Rieser
Interaction Lab, Heriot-Watt University,
EH144AS Edinburgh, Untied Kingdom.
eaar1@hw.ac.uk, v.t.rieser@hw.ac.uk

Abstract

Supervised machine learning methods for automatic subjectivity and sentiment analysis (SSA) are problematic when applied to social media, such as Twitter ©, since they do not generalise well to unseen topics. A possible remedy of this problem is to apply distant supervision (DS) approaches, which learn from large amounts of automatically annotated data. In this research, we explore DS for SSA on Arabic Twitter feeds using emoticons as noisy labels. We achieve 95.19% accuracy, which is a 48.57% absolute improvement over our previous fully supervised results. While our results show a significant gain in detecting subjectivity, this approach proves to be difficult for sentiment analysis. An error analysis suggests that the most likely cause for this shortcoming is the unclear facing of emoticons due to the right-to-left direction of the Arabic alphabet.

Keywords: Subjectivity and Sentiment Analysis, Twitter, Arabic, Semi-Autonomous Learning on Big Data, Sarcasm, Cultural Bias

1. Introduction

The growth of social media, especially as a source for analysis, has resulted in a two-fold challenge: managing the costs processing all of that data, as well as developing new ways to make sense of it. In addition, of course, in the small world in which we live, one needs to be able to handle multiple languages and idioms equally well. In this work we explore different approaches to subjectivity and sentiment analysis (SSA) of Arabic tweets. SSA aims to determine the attitude of the user with respect to some topic, e.g. objective or subjective, or the overall contextual polarity of an utterance, e.g. positive or negative. To the best of our knowledge, there is no publicly available large-scale Arabic Twitter © corpus annotated for subjectivity and sentiment analysis. Creating a new data set is costly, and, as we will show in the following, learning from small data sets does not cover the wide scope of topics discussed on Twitter. To the author's knowledge this is the first empirical study of using distant supervision learning for Arabic social networks.

2. Background

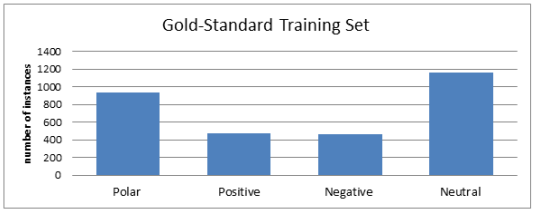
Arabic can be classified with respect to its morphology, syntax, and lexical combinations into three different categories: classic Arabic (CA), modern standard Arabic (MSA), and dialectal Arabic (DA). Users on social networks typically use the latter, i.e. dialectic varieties such as Egyptian Arabic and Gulf Arabic (Al-Sabbagh and Girju, 2012). Dealing with DA creates additional challenges for natural language processing (NLP): Being mainly spoken dialects, they lack standardisation, and are written in free-text (Zaidan and Callison-Burch, 2013). This problem is even more pronounced when moving to the micro-blog domain, such as Twitter (Derczynski et al., 2013). People posting text on social networks tend to use informal writing style, for example by introducing their own abbreviations, as in example (1), or using spelling variations. In addition, tweets may also convey sarcasm, mixed and/or unclear polarity content, as in example (2) taken from our corpus (see Section 3.).

(1)
لول
lol (laugh out loud)

(2)
مصر دلوقتي بقت عَامله زي الفيلم الاَجنبي الغير مترجم، الكل بيتفرج ويترَاجم علي مزَاجه
Egypt now is more like a foreign film without subtitles, so everybody watches and puts their own translation.

Machine learning techniques are in general robust to such variety. Previous work on SSA has used manually annotated gold-standard data sets to analyse which feature sets and models perform best for this task, e.g. (Wilson et al., 2009; Wiebe et al., 1999). Most of this work is in English, but there have been first attempts to apply similar techniques to Arabic, e.g. (Abdul-Mageed et al., 2011; Mourad and Darwish, 2013). While these models work well when tested on limited static data sets, our previous results reveal that these models do not generalise well to new data sets collected at a later point in time due to their limited coverage (Refaee and Rieser, 2014). In addition, while there is a growing interest within the NLP community to build Arabic corpora by harvesting the web, e.g. (Al-Sabbagh and Girju, 2012; Abdul-Mageed and Diab, 2012; Zaidan and Callison-Burch, 2013), these resources have not been publicly released yet and only small amounts of these data are (manually) annotated.

We therefore turn to an approach known as *distant supervision* (DS), as first proposed by (Read, 2005), which uses readily available features, such as emoticons, as noisy labels. This approach has been shown successful for English SSA, e.g. (Go et al., 2009; Suttles and Ide, 2013), and SSA for under-resourced languages, such as Chinese (Yuan and Purver, 2012).



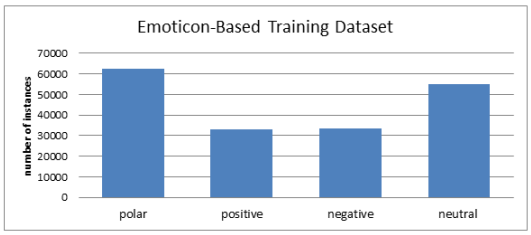


Table 1: Sentiment label distribution of the gold-standard manually annotated and distant supervision training data sets.

3. Arabic Twitter SSA Corpora

We start by collecting corpora at different times over one year to account for the cyclic effects of topic change in social media (Eisenstein, 2013): (1) A gold-standard data-set which we use for evaluation (spring 2013); (2) A data-set for DS using emoticon-based queries (autumn 2013). Table 1 shows the distributions of labels in our data-sets.

We use the Twitter Search API for corpus collection, which allows harvesting a stream of real-time tweets by querying their content. The tweets were collected at different times and days to reduce bias in the distribution of the number of tweets from individual users. In addition, we collected the used-ID of each retrieved tweet. The distribution of tweets per user IDs is 1.12. By setting the language variable to `ar`, all retrieved tweets were restricted to Arabic. The extracted data is cleaned in a pre-processing step, e.g. normalise digits, non-Arabic characters, user-names and links.

3.1. Gold-Standard Dataset

We harvested two data sets at two different time steps, which we label manually. We first harvest a data set of 3,309 multi-dialectal Arabic tweets randomly retrieved over the period from February to March 2013. We use this set as a training set for our fully supervised approach (Refaee and Rieser, 2014). We also manually labelled a subset of 963 tweets of the "emoticon-based" corpus (see Section 3.2.), which we use as an independent held-out test set.

Two native speakers of Arabic were recruited to manually annotate the collected data for subjectivity, i.e. subjective/polar versus objective tweets, and sentiment, where we define sentiment as a positive or negative emotion, opinion, or attitude, following (Wilson et al., 2009). Our gold-standard annotations reached a weighted $\kappa = 0.76$, which indicates reliable annotations (Carletta, 1996).

We annotate the corpus with a rich set of linguistically motivated features using freely available processing tools for Arabic, such as MADA (Nizar Habash and Roth, 2009), see Table 2. For more details please see (Refaee and Rieser, 2014).

3.2. Emoticon-Based Queries

In order to investigate DS approaches to SSA, we also collect a much larger data set of Arabic tweets, where we use emoticons as noisy labels following, e.g. (Read, 2005; Go et al., 2009; Pak and Paroubek, 2010; Suttles and Ide, 2013). We query Twitter API for tweets with variations of positive and negative emoticons to obtain pairs of micro-blog texts (statuses) and using emoticons as author-generated emotion labels. In following (Purver and Battersby, 2012; Yuan and Purver, 2012; Zhang et al., 2011; Suttles and Ide, 2013), we also utilise some sentiment-bearing hash-tags to query emotional tweets.

Examples of hash-tags we queried are: فرح *happiness* and حزن *sadness*. Note that emoticons and hashtags are merely used to collect and build the training set and were replaced by the standard (positive/ negative) labels. In order to collect neutral instances, we query a set of official news accounts, following (Pak and Paroubek, 2010). Examples of the accounts queried are: BBC-Arabic, Al-Jazeera Arabic, SkyNews Arabia, Reuters Arabic, France24-Arabic, and DW Arabic. Using this method, we collected 55,076 neutral instances in total. We then automatically extract the same set of linguistically motivated features as for the gold-standard corpus, see Table 2. After removing re-tweets, duplicates and mixed tweets, the corpus is composed of 120,747 data instances.

Note that this work is the first to investigate distant supervision approaches for Arabic, and as such, no previous automatically labelled data sets are available. The gold-standard data set will shortly be available from the ELRA repository.[1] We also hope to release the automatically labelled data to the community in the near future, where we investigate standardised RDF data schema for linked open SSA data, such as MARL (Westerski, 2011).

4. Classification Experiments Using Distant Supervision

In previous work, we experiment with a fully supervised approach on a hand-labelled data set (Refaee and Rieser, 2014). However, our results reveal that these models do not transfer well to new data sets collected at a later point in time due to their limited coverage. An error analysis confirms that this drop in performance is due to topic-shifts in the Twitter stream. We therefore turn to DS approaches. In this section we empirically evaluate emoticon-based approach to DS.

4.1. Experimental Setup

For classification, we experiment with two alternative problem formulations: Related work has treated SSA as a two-

[1] `http://catalog.elra.info/`

Type	Feature-sets
Morphological	diacritic, aspect, gender, mood, person, part_of_speech (POS), state, voice, has_morphological_analysis.
Syntactic	n-grams of words and POS, lemmas, including bag_of_words (BOW), bag_of_lemmas.
Semantic	has_positive_lexicon, has_negative_lexicon, has_neutral_lexicon, has_negator, has_positive_emoticon, has_negative_emoticon.

Table 2: Annotated Feature-sets

Data-set	majority baseline		fully supervised		emoticon DS: BOW		emoticon DS: BOW+Morph		emoticon DS: BOW+Morph +Semantic.	
	F	Acc.	F	Acc.	F	Acc.	F	Acc.	F	Acc.
polar vs. neutral	0.69	53.0	0.43	46.62	0.94	94.89	**0.95**	**95.19**	0.94	94.28
positive vs. negative	**0.67**	**50.89**	0.41	49.65	0.50	50.29	0.51	51.25	0.49	50.0
positive vs. negative vs. neutral	0.63	46.99	0.28	28.24	**0.70**	**69.67**	0.69	68.43	0.67	67.18

Table 3: 2-level and single-level SSA classification using distant supervision (DS).

stage binary classification process, where the first level distinguishes subjective and objective statements, and the second level then further distinguishes subjectivity into positive and negative sentiment, e.g. (Wiebe et al., 1999; Abdul-Mageed et al., 2011; Abdul-Mageed et al., 2012). Alternatively, the classification can be carried out at as single-level classification (positive, negative, neutral), e.g. (Farra et al., 2010). We experiment with both options. For the first stage of the binary approach, we collapse the positive and negative labels into a "polar" label.

We experiment with a number of machine learning methods and we report the results of the best performing scheme, namely Support Vector Machines (SVMs), where we use the implementation provided by the WEKA data mining package version 3.7.9 (Witten and Frank, 2005). We report the results on two metrics: F-score and accuracy. We use paired t-tests to establish significant differences ($p < .05$). Different to the previous experiments on the gold-standard data (Refaee and Rieser, 2014), we only experiment with a subset of features, which we previously identified as best performing: *Bag-of-Words (BOW) + morphological + semantic* features. Note that, for the DS approach, we exclude the emoticon-based features from the *semantic* feature set. We compare our results against a majority baseline and against a fully supervised approach, i.e. SVMs trained on a manually labelled gold-standard data set using the same feature set. We evaluate the approaches on a separate held-out test set, as described in Section 3.1.

4.2. Emoticon-Based Distant Supervision

In this section, we evaluate the potentials of exploiting training data that is automatically labelled using (noisy) emoticons, see Section 3.2. The results are summarised in Table 3.

Polar vs. Neutral: The results show a significant improvement over the majority baseline, as well as over the classifier trained on the gold-standard data set: We achieve a 95.19% accuracy on the held-out set with BOW and morphological features, which is a 48.57% absolute improvement over our previous fully supervised results. These results indicate that the classifier is able to recognise and distinguish the language used to express neutral/objective utterances from those used to convey personal opinion/attitude. Feature selection, while showing some improvement when adding morphological features, does not have a significant effect on performance.

Positive vs. negative: For sentiment classification, the performance of emoticon-based approach degrades notably to 0.50 F-score (for BOW only), which is significantly better that the fully supervised baseline, but still significantly worse than a simple majority baseline. One possible explanation for this is that the classifier is faced with the naturally harder discrimination task between positive and negative instances. The confusion matrix shows that it's mostly negative instances are misclassified as positive. In Section 4.2.1. we will investigate possible reasons in a detailed error analysis. Again, adding features has no significant effect on performance.

Positive vs. Negative vs. Neutral: When performing three-way SSA on a single level, the SVM outperforms the majority baseline and achieves 0.70 F-score. Again, BOW achieves the highest results. The confusion matrix reveals that detecting the negative class is most problematic, with the lowest recorded precision at 0.55, while the neutral class achieved significantly better precision at 0.96. In this case, adding the semantic features significantly decreases the performance. We hypothesise that this might be the features based on the subjectivity lexicon, which so far only covers MSA. We will address this short-coming in future work.

Feature Selection: In general, our feature selection experiments show no significant impact on performance. However, adding morphological features show a positive trend for improving both, subjectivity and sentiment analysis. This confirms previous results by Abdul-Mageed et al. (2012) for SSA on Arabic tweets using fully supervised learning. Go et al. (2009), in contrast, reports that adding morphological features hurts performance when using emoticon-based DS for SSA on English Twitter feeds. We therefore hypothesise that morphological features are especially useful for Arabic, being morphologically rich language.

Emoticon Label	Predicted label	Manual label	# instances
Positive	Negative	Mixed	8
Negative	Positive	Mixed	10
Positive	Negative	Negative	59
Negative	Positive	Negative	42
Positive	Negative	Neutral	29
Negative	Positive	Neutral	7
Positive	Negative	Positive	62
Negative	Positive	Positive	52
Positive	negative	Sarcastic	8
Negative	Positive	Sarcastic	5
Positive	Negative	Unclear sentiment indicator	19
Negative	Positive	Unclear sentiment indicator	2

Table 4: Results of labelling sarcasm, mixed emotions and unclear sentiment for misclassified instances.

4.2.1. Error Analysis for Emoticon-Based DS

The above results seem to indicate that DS works well for subjectivity analysis (distinguishing neutral vs. polar instances), but proves to be difficult for sentiment analysis (distinguishing positive vs. negative instances). Especially, detecting negative instances seems to be problematic. We conduct an error analysis in order to further investigate the underlying cause. In particular, we investigate the use of sarcasm and the direction of facing of emoticons in right-to-left alphabets.

Use of sarcasm and irony: Using emoticons as labels is naturally noisy, since we cannot know for sure the intended meaning the author wishes to express. This is especially problematic when emoticons are used in a sarcastic way, i.e. their intended meaning is the opposite of the expressed emotion. An example from our data set is: جميل يَا اهلي :(*great job Ahli :(* – referring to a famous football team. Research in psychology shows that up to 31% of the time, emoticons are used sarcastically (Wolf, 2000).

In order to investigate this hypothesis we manually labelled random sample of 303 misclassified instances for neutral, positive, negative, as well as sarcastic, mixed and unclear sentiments, see Table 4. Interestingly, the sarcastic instances represent only 4.29 %, while tweets with mixed (positive and negative) sentiments represent 5.94% of the manually annotated sub-set. In 34.32% of the instances the manual labels have matched the automatic emoticon-based labels. Surprisingly, automatic emoticon-based label contrasts the manual labels in 36.63% of the instances. The rest of the instances were manually annotated as neutral. As such, a large proportion of the misclassified is still unexplained.

Facing of emoticons: We therefore investigate another possible error source following (Mourad and Darwish, 2013), who claim that the right-to-left alphabetic writing of Arabic might result in emoticons being mistakenly interchanged. On some Arabic keyboards, typing ') ' will be producing the opposite ' (' parentheses. The following examples in Table 5 illustrate cases, where we assume that the facing of emoticons might have been interchanged. However, the intended meaning cannot be known for sure.

ID	Original tweet	Translation
1	اكرهك :)	*I hate you :)*
2	تعبت وَانَا اتخيل، ابي شي يصير وَاقع :)	*I'm tired of dreaming, I want something to become true :)*
3	خلَاص مَافي امل :)	*no hope anymore :)*

Table 5: Example of mislabelled tweets

5. Conclusion and Future Work

We address the task of subjectivity and sentiment analysis (SSA) for Arabic Twitter feeds when learning from large data sets. We empirically investigate the performance of an emoticon-based distant supervision (DS) approach on a manually labelled independent test set, in comparison to a fully supervised baseline, trained on a manually labelled gold-standard data set. Our experiments reveal that an emoticon-based DS approach to SSA for Arabic Twitter feeds shows significantly higher performance in accuracy and F-score than a fully supervised approach. Despite providing noisy labels, this approach allows larger amounts of data to be rapidly annotated, and thus can account for the topic shifts observed in social media.

We also find that our emoticon-based DS approach achieves good results of up to 95.19% accuracy for subjectivity analysis, i.e. distinguishing between neutral and polar instances. However, we detect a decrease in performance for sentiment analysis, i.e. distinguishing between negative and positive instances, where the negative instances repeatedly get misclassified as positive.

We conduct a detail error analysis and find that this low performance cannot be attributed to sarcasm, as originally hypothesised, but might be due to the right-to-left alphabet of Arabic, which causes the facing on emoticons to be ambiguous. In future work, we will investigate alternative approaches to DS. One promising direction which we plan to explore is "lexicon-based" DS, i.e. using adjectives from a subjectivity lexicon as noisy labels for DS, following e.g. (Zhang et al., 2011).

Other possible reasons for this drop in performance include cultural specific differences, as well as context-dependent "pragmatic" aspects of opinion (Sayeed, 2013). For example, Hong et al. (2011) observe that in some cultures, such as German, users tend to predominantly post factual/ neutral statements. We will explore if this is also the case for Arabic, as our high performance in detecting neutral instances might suggest.

Acknowledgements

The first author would like to thank the Saudi Arabian government for supporting her with a PhD scholarship.

6. References

Abdul-Mageed, M. and Diab, M. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Abdul-Mageed, M., Diab, M. T., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abdul-Mageed, M., Kuebler, S., and Diab, M. (2012). Samar: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. Association for Computational Linguistics.

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal arabic corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013). Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Farra, N., Challita, E., Assi, R. A., and Hajj, H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1114–1119. IEEE.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Hong, L., Convertino, G., and Chi, E. H. (2011). Language matters in twitter: A large scale study. In *ICWSM*.

Mourad, A. and Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. *WASSA 2013*, page 55.

Nizar Habash, O. R. and Roth, R. (2009). MADA+TOKAN: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Choukri, K. and Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*.

Purver, M. and Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491, Avignon, France, April. Association for Computational Linguistics.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics.

Refaee, E. and Rieser, V. (2014). Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *9th International Conference on Language Resources and Evaluation (LREC'14)*.

Sayeed, A. (2013). An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of NAACL-HLT*, pages 691–696.

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pages 121–136. Springer.

Westerski, A. (2011). MARL ontology. `http://www.gi2mo.org/marl/0.1/ns.html`. Accessed: 2014-03-12.

Wiebe, J. M., Bruce, R. F., and O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3(5):827–833.

Yuan, Z. and Purver, M. (2012). Predicting emotion labels for chinese microblog texts. In *Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD)*, pages 40–47, Bristol, UK, September.

Zaidan, O. F. and Callison-Burch, C. (2013). Arabic dialect identification. *Computational Linguistics*.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2011). Combining lexiconbased and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89.

Detecting Happiness in Italian Tweets: Towards an Evaluation Dataset for Sentiment Analysis in Felicittà

C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, E. Sulis
Dipartimento di Informatica - Università degli Studi di Torino
Corso Svizzera 185 - 10149, Torino (Italy)
{bosco,patti,ruffo,msanguin,sulis}@di.unito.it, {leonardo.allisio,valeria.mussa}@studenti.unito.it

Abstract

This paper focuses on the development of a gold standard corpus for the validation of Felicittà, an online platform which uses Twitter as data source in order to estimate and interactively display the degree of *happiness* in the Italian cities. The ultimate goal is the creation of an Italian reference Twitter dataset for sentiment analysis that can be used in several frameworks aimed at detecting sentiment from big data sources. We will provide an overview of the reference corpus created for evaluating Felicittà, with a special focus on the issues raised from its development, on the inter-annotator agreement discussion and on implications for the further development of the corpus, considering that the assumption that a single right answer exists for each annotated instance cannot be done in several cases in the particular kind of data at issue.

Keywords: Sentiment analysis in Twitter, Corpus annotation, Italian

1. Introduction

In the last few years, the linguistic analysis of social media has become a relevant topic of research, and several frameworks for detecting sentiments and opinions in social media have been developed for different application purposes.

One of the possible applications of Sentiment Analysis (SA) is in the social and behavioral sciences field, where SA techniques could contribute to interpret the degree of well-being of a country. The studies concerning life satisfaction have grown substantially since the late 20th Century. New areas of research have arisen, such as the *Subjective Well-Being* (SWB) in Psychology (Diener, 2000) and the *Happiness economics* in Economy, within the debate on alternative measure to Gross Domestic Product (Helliwell et al., 2014). The rise of Big Data and the exponential growth of social media (e.g. Facebook, Twitter) has created vast opportunities and new challenges to the social sciences on this respect. In some pioneering work in this direction, extracting expressed sentiments – typically categorized as positive, negative or neutral – in short messages has been used for several purposes: to detect moods and happiness in a given geographical area from geotagged Tweets (Mitchell et al., 2013), to create a *hate map* based on expressions of homophobia and racism on Twitter[1], to show the correlation with traditional data (Bollen and Mao, 2011) and to measure the well-being of a population (Quercia et al., 2012).

It should also be observed that linguistic analysis of social media has gained in the last few years an increasing relevance in the detection of well-being or happiness (Mihalcea and Liu, 2006). However, various issues should be taken into account in the detection of sentiments and opinions in natural language texts. On the one hand, data on which SA is applied are from texts especially challenging for most Natural Language Processing (NLP) systems. Although, as observed in (Baldwin, 2012), social media texts can also be considered a valuable resource, rather than a foe, by virtue of the richness of non-textual data that can be exploited to enhance the robustness and accuracy of NLP techniques. As a matter of fact, hashtags, emoticons, emojis or links occurring in a post can be used to disambiguate the textual content. On the other hand, training and testing automatic systems requires the availability of several resources that may consist in large datasets of annotated posts or even in lexical databases where affective words are associated with polarity values. But their availability is currently very limited in particular for languages other than English.

In this paper, we would like to contribute to the debate in this area by describing our experience in the development of Felicittà, an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines a lexicon-based approach for SA and visualization techniques in order to provide users with an interactive interface for data exploration (Allisio et al., 2013). (Pianta et al., 2002; Strapparava and Valitutti, 2004).

In particular, we will report the most recent achievements in the development of the platform, especially focusing on the creation of a Twitter dataset for testing the sentiment algorithm in Felicittà. For what concerns the annotation schema and procedure, we rely on the research carried out within the Senti–TUT project (Bosco et al., 2013). The ultimate goal is the creation of an Italian reference corpus that can be used in several frameworks for detecting sentiments from big data sources, such as Twitter.

We will provide an overview of the reference corpus currently developed for Felicittà, by focusing in particular on the issues raised from annotator agreement analysis and their implications for the further development of the corpus.

2. Related Works

For what concerns the resources for SA, for English language, sentiment lexicons (listed in (Nakov et al., 2013)), Twitter datasets and gold standards for the sentiment analysis task on Twitter messages are now available [2], while

[1] `http://users.humboldt.edu/mstephens/hate\/hate_map.html`

[2] See the recent survey and comparison in (Saif et al., 2013).

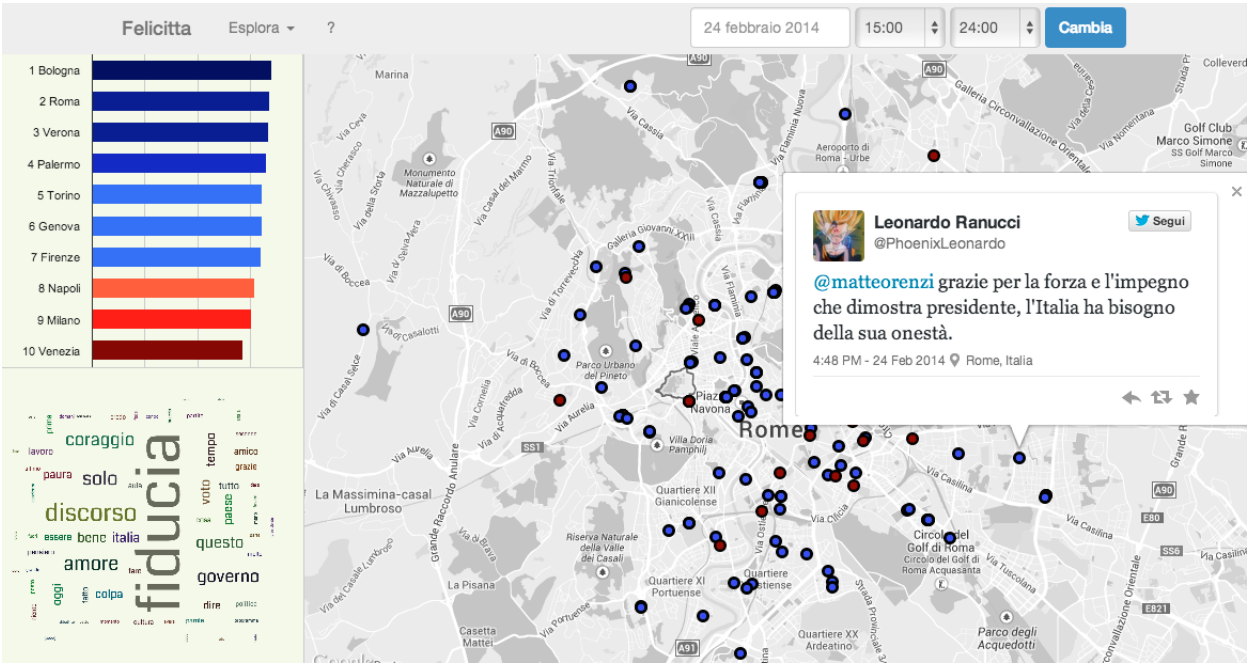


Figure 1: Felicittà: an interactive map displaying Tweets that convey negative or positive polarity and positioned within the area where they have been posted.

for several other languages, like Italian, the availability of such resources is currently very limited. Indeed, several resources are being developed by individual companies for their commercial use in sentiment monitoring services[3], but normally they are not shared nor publicly available.

For what concerns Italian, to the best of our knowledge, Senti–TUT is the first Italian gold corpus developed for Twitter SA (Bosco et al., 2013), which also includes ironic tweets. Irony detection is a hot topic in the SA research community indeed, and in particular the fact that Twitter messages include a high percentage of ironic messages cannot be neglected (González-Ibáñez et al., 2011; Reyes et al., 2013; Davidov et al., 2011; Hao and Veale, 2010). Platforms monitoring the sentiment in Twitter messages experience the phenomenon of wrong polarity classification in ironic messages. Indeed, the presence of ironic devices in a text can work as an unexpected ”polarity reverser” (one says something ”good” to mean something ”bad”, or vice versa), thus undermining systems' accuracy[4].

Recent works (Caselli et al., 2012; Baldoni et al., 2012; Bertola and Patti, 2013) exploited WordNet-Affect (Strapparava and Valitutti, 2004), an affective lexicon which links synsets in the original Princeton WordNet (Fellbaum, 1998) to affects, but, being the affective extension of WordNet domains developed at irst-FBK and aligned with MultiWordNet, WordNet-Affect embeds information on the correlation between English and Italian terms. WordNet-Affect is freely available for research purposes. It is semi-automatically created, based on a manually realized core, and includes 4,787 affective words. Moreover, only very recently a new publicly available lexical resource for Italian has been developed, which is called Sentix (Sentiment Italian Lexicon) (Basile and Nissim, 2013) and is the result of the alignment of several existing lexical and affective resources: WordNet, MultiWordNet (Pianta et al., 2002), BabelNet (Navigli and Ponzetto, 2012) and SentiWordNet (Esuli et al., 2010).

It should also be observed that the development of corpora that can be usefully exploited in this kind of task is in itself very challenging. For other tasks the development of a corpus consists in creating an annotated human ground truth, assuming that for each annotated instance there is a single right answer and that the quality of the annotation can be measured in terms of inter-annotator agreement.

In the development of a corpus for SA this assumption cannot be done, and the disagreement reflects semantic ambiguity in the target instances, thus providing useful information. Under this respect, the annotation of a corpus for SA can be usefully compared to the development of corpora for clinical studies, see e.g. (Xia and Yetisgen-Yldiz, 2012), or those for co-reference where underspecified labels are adopted to cope with the vagueness of data (Versley, 2006). In corpora for SA the reasons for annotator disagreement are also related to the fact that *a*) there are many different ways to linguistically express the same polarity, and *b*) the same linguistic expression may be used for different polarities. This in turn makes context extremely important, for instance in case of humorous and ironic expressions. These factors create, in human understanding, a fairly wide range of possible, plausible interpretations of a post, and as a consequence a disagreement in the annotation.

[3]Think for instance to the affective Italian lexicon used in the social media monitoring platform Blogmeter (`http://www.blogmeter.eu/`), which includes about 10,000 entries (Bolioli et al., 2013).

[4]A pilot subtask concerning irony detection on Italian Tweets will be organised at Evalita: `http://www.di.unito.it/~tutreeb/sentipolc-evalita14/index.html`

3. Felicittà

Felicittà[5] is an online platform for estimating happiness in the Italian cities, which daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets, in order to enable the summarization of SA outcomes and the exploration of Twitter data (Allisio et al., 2013). Interactive maps offered by Felicittà provide users not only with the opportunity to have a comprehensive overview of the SA results about the main Italian cities, but also to zoom-in to a specific region to visualize a fine-grained map of the city or district and the location of the individual sentiment-labeled Tweets (Fig. 1). Interaction possibilities enabled by the platform allow users to tune their view on such huge amount of information and to interactively reduce the inherent complexity, possibly providing a help in the detection of meaningful patterns. Tag clouds highlighting the important words in the Tweets posted in a geographic area are daily generated and visualized together with the sentiment outcomes, with the aim of evoking possible correlations between mood and events.

The heart of the framework is a sentiment analyzer. By exploiting Twitter's APIs, the system collects every day all the Tweets freely downloadable (450,000), geo-located in the main Italian towns, and performs the analysis for each Tweet in order to classify it as positive or negative. This analysis includes, in particular, the application of Freeling[6], a multilingual open source tool for morpho-syntactic analysis, developed at the University of Catalunya (Spain). The grammatical category and lemma of each word is recognized, thus allowing a more efficient association with the lexical item to be searched in the affective lexicon. Finally, the polarity of all the Tweets is aggregated according to their geo-location and the happiness degree of each town and region is evaluated and made available in different visualization modes.

According to a lexicon-based approach, the polarity of each Tweet in Felicittà depends on the affective words detected within it and then found in the affective lexicon, i.e. in WordNet–Affect, that is the resource which most of the recent works for Italian currently exploit, see e.g. (Caselli et al., 2012; Baldoni et al., 2012; Bertola and Patti, 2013).

4. Data annotation for sentiment analysis

In order to validate our approach and to analyze the limits of the sentiment analyzer implemented in Felicittà, we have created a reference corpus including a set of Italian Tweets, called TW-FELICITTA.

4.1. Collection

1,500 Italian Tweets were randomly extracted from those collected by Twitter API, paying attention to avoiding geographic and temporal bias at different level of granularity. As a matter of fact, possible correlations have been observed between sentiment and time of the day or day of the week (weekdays or holidays), or between sentiment and geographical areas in a given time frame due to the occurrence of some special event. Furthermore, we gathered the Tweets for the collection in order to avoid a logical link between a Tweet and the next one, which is a typical situation where two users communicate with each other: this way, it is not possible to infer the discussion topic, unless this is explicitly mentioned; the principle that lies behind this choice is that of preventing both the system and the manual annotator from labeling the Tweets in a different way namely because of such inferred information. We therefore implemented an automatic algorithm for the collection which takes into account such issues.

4.2. Annotation schema

Sentiment annotation was manually performed at the Tweet level. This means that we considered single Tweets as individual documents and annotated them using one of the tags reported in Table 1 and previously applied to the annotation of the Senti–TUT Italian corpus for SA (Bosco et al., 2013)[7].

POS	positive
NEG	negative
NONE	objective (no sentiment expressed)
MIXED	mixed (POS and NEG both)
HUM	ironic
UN	unintelligible

Table 1: Tags annotated in TW-FELICITTA corpus.

The application on TW-FELICITTA has shown the suitability of this schema designed for the annotation also of mixed polarity and ironic expressions, exploiting the MIXED and HUM tags. Indeed, also because the sentiment annotation is performed at the Tweet level, it is often difficult to determine unambiguously the overall polarity of the sentiment expressed in it, especially in presence of irony and mixed sentiment. Ironic Tweets and Tweets containing parts expressing both positive and negative sentiment have recognized to be phenomena that strongly contribute to make the Tweet classification task harder (Nakov et al., 2013). In this context, the classical labels distinguishing only among positive, negative or neutral sentiment may not be sufficient; we thus extended the tag set by including:

- MIXED to mark the presence of more than one sentiment within a Tweet, which can be related to the expression of opinions on different targets or also to a contrast between polarity of the opinion conveyed and expressed mood, see also the gold standard presented in (Saif et al., 2013).
- HUM to mark the intention of the author of the post to express irony, which could be hardly classified as entirely positive or negative;
- UN to mark the difficulty experienced by the annotator due, e.g., to the incompleteness of the message or the absence of a context.

[5] http://felicitta.di.unito.it/

[6] http://nlp.lsi.upc.edu/freeling/

[7] http://www.di.unito.it/~tutreeb/sentiTUT.html

The following examples are applications of the above described labels.

TW-FELICITTA#504 (tagged as POS)
American Horror Story ti amo
#AmericanHorrorStory sei il telefilm che fa la differenza.
(I love you, American Horror Story
#AmericanHorrorStory you're the tv series that makes
the difference.)

TW-FELICITTA#518 (tagged as NEG)
Perche' non riesco a dimenticarla
(Why can't I just forget about her)

TW-FELICITTA#636 (tagged as NONE)
Accadde oggi: 1993: entra in vigore il Trattato di Maastricht, che stabilisce formalmente l'Unione Europea....
(Today in history: 1993: The Maastricht Treaty,
which formally establishes the European Union,
enters into force ...)

TW-FELICITTA#305 (tagged as MIXED)
E' stata una settimana perfetta
Ma questa domenica ha rovinato tutto Ma proprio tutto.
(It was a perfect week. But this Sunday has ruined everything
Absolutely everything)

TW-FELICITTA#683 (tagged as HUM)
RT@lddio:Letta:"I giovani senza lavoro sono l'incubo dell'Italia". Per non essere da meno, anche l'Italia e' l'incubo dei giovani.
(Letta: "Young people out of work are the nightmare of Italy."
Not to be outdone, Italy is the nightmare of young people.)

TW-FELICITTA#771 (tagged as UN)
@Caustica_mente ho detto che sono inconsistenti?
volevo capire i motivi dell'eventuale autogoal.
A leggerti, non e' alfine tale. Bene.
(@Caustica_mente Did I say that they are inconsistent?
I wanted to understand the reasons for an own goal.
By reading you, this is not finally such. All right.

For what concerns the last sample, the English translation was kept ungrammatical on purpose, in order to convey to the non-Italian reader as well the difficulty experienced by the annotator in inferring the meaning of the message.
For what concerns the label HUM, let us notice that, as also pointed out in the literature, there is no agreement on a formal definition of irony, as is the case of most figurative devices. Nonetheless, psychological experiments have given evidence that humans can reliably identify ironic text utterances from an early age in life. These findings provide grounds for developing manually annotated corpora for irony detection. Moreover, the boundaries between irony and other figurative devices, such as sarcasm, satire, or humor, are quite fuzzy (Strapparava et al., 2011). This made us lean on adopting the same approach proposed in Senti-TUT, where no distinction has been drawn among different types of irony.
Notice that, having a distinguished tag for irony do not prevent us to reconsider these Tweets at a later stage, and "force" their classification according to traditional annotation schemes for the SA task, as suggested for instance in (Bosco et al., 2013), where a similar approach has been applied to tackle with the polarity reversing phenomenon due to the presence of irony, and to measure how an automatic traditional sentiment classifier can be wrong. Similarly, identifying Tweets containing mixed sentiment can be useful in order to measure how the phenomenon impacts on the performances of sentiment classifiers[8].
Moreover, having distinguished tags for irony and mixed sentiment can be helpful for a better development of the corpus itself, in order to increase the inter-annotator agreement, since such cases, being typically source of disagreement on the polarity valence, are recognized and labeled apart.

4.3. Annotation process

The annotation process (together with the annotation guidelines) was developed through multiple stages. After a phase where four human annotators (A_1, A_2, A_3, A_4) (native-speakers, different genders, varying ages and background) collectively annotated a small set of data (i.e. 100 Tweets), results on the disagreement were discussed in order to both reach a better agreement on the exploitation of the labels on the entire corpus, and produce a document including annotation guidelines[9] shared by the annotators.
Then, A_1, A_2 and A_3 annotated all the data (i.e. 1,500 Tweets) producing for each Tweet not less than three independent annotations. The inter-annotator agreement has been calculated at this stage according to the Fleiss's Kappa (Fleiss, 1971) andthe measure obtained reached $\kappa = 0.51$. It can be observed that this rate positively compares to that described for the similar task in (Basile and Nissim, 2013), and it is a slightly lower rate with respect to the development of TW-NEWS (Bosco et al., 2013), where only two annotators were involved.
The agreement among the three annotators has been achieved in this step in 46% of cases, corresponding to 695 Tweets. On the remaining 805 Tweets, we can distinguish between Tweets in *hard disagreement*, when three different tags have been annotated, and those in *soft disagreement*, on whose polarity at least two annotators agreed. The former consist of 13% (191 Tweets), while the latter of 41% (614 Tweets) of the entire corpus.
In order to further extend our data set, we discarded the Tweets featured by hard disagreement, but we recovered the agreement on a large portion of those resulting in a soft disagreement after the first annotation step in two ways. First, we applied to this set a 4th independent annotation (by A_4), and we achieved in this way the agreement among three of the four annotations on further 433 Tweets of the

[8] Also in this case it could be interesting to reconsider Tweets tagged as MIXED at a second stage, by classifying them as either (mainly) positive or negative

[9] See: http://www.di.unito.it/˜tutreeb/AnnotationGuidelines.pdf

614 cited above. Second, at a last stage, the four annotators discussed the polarities of the remaining 181 Tweets (i.e. 29%), hypothesizing that the soft disagreement was persisting on them because of annotators' biases or errors. The discussion leaded to an updated version of the guidelines and to the ultimate version of the corpus where further 107 Tweets have been recovered in agreement, thus obtaining two sets: one set of Tweets in agreement composed of around 82% (1,235 Tweets), henceforth indicated as *A-set*, and one of those featured by an unsolvable disagreement composed of around 18% (265 Tweets) of the entire corpus, henceforth indicated as *D-set*.

Therefore we included in the final version of the TW-FELICITTA gold corpus only the 1,235 Tweets on which we achieved the agreement among the annotators, ready to be exploited for training and evaluation purposes. The final tags in the gold corpus are distributed as follows among the Tweets of the *A-set*: around 57% of them were classified as positive (338) or negative (364), 21% is classified as NONE (260), around 14% as HUM, and the remaining as MIXED (3%) or UN (5%), as shown in Figure 2.

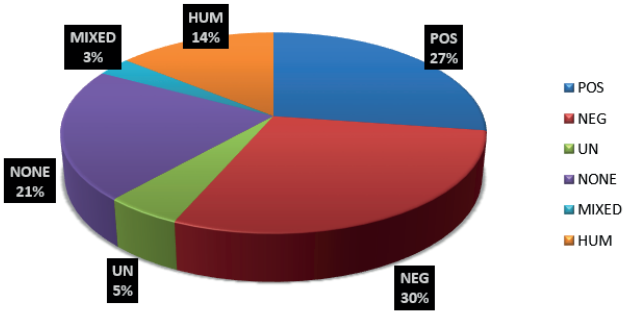


Figure 2: Distribution of the sentiment labels used by the annotators in the gold corpus of Felicittà.

For what concerns instead the remaining cases (around 18%), considered as too ambiguous to be classified according to the detected disagreement, we aim to define a framework to harness the analysis of the disagreement between the human annotators, in order to capture interesting features related to the sentiment and irony detection task. Our preliminary results in the definition of this frame can be seen in section 5.

5. Annotation Analysis

Annotation is an important task for NLP, and the traditional annotation pipeline, including writing detailed guidelines, trained annotators and disagreement calculation, has proved to work well in several projects. Other annotation strategies has been proposed for specific tasks, see e.g. (Xia and Yetisgen-Yldiz, 2012). On the one hand, annotation of polarity for SA is a task featured by specific peculiarities that can be made clear e.g. by observing the distribution of tags and disagreement calculation. On the other hand, the feature of each single corpus should be carefully taken into account and compared with those of other data sets.

For what concerns TW-FELICITTA, we first made a comparison with TW-NEWS (Bosco et al., 2013), a similar Italian corpus that includes Tweets collected in the time frame between October 2012 and February 2013 and that focuses on a specific topic (the past Montis government in Italy). Such comparison shows that in the former there is a meaningfully smaller amount of Tweets with neutral polarity with respect to the other data set we have previously annotated. This can be motivated by the larger frequency of emoticons and *emoji*[10], which are currently often used in social media and supported by smartphones interfaces, as observed also in (Suttles and Ide, 2013), but were very rarely used in 2012, when TW-NEWS has been collected. They are considered by the annotators as useful hints about the polarity of posts, and can also be used by automatic systems for a reliable detection of polarity. This is confirmed by the preliminary analysis performed by the sentiment analyzer implemented in Felicittà.

Second, considering the selection criteria (mentioned above) for the creation of the TW-FELICITTA corpus, there is a high variety in the topics addressed in the Tweets, and their independence with respect to the time frame and geographic area do not allow the annotator to trace back to the original communicative situation. This aspect, as also pointed out in (Basile and Nissim, 2013), together with the wider tag set used in our corpus (w.r.t. the classic annotation schemas for sentiment) and varying annotators' skills (depending, in their turn, on different genders and varying ages and background), is deemed to be a possible source of disagreement.

It should be observed that the final goal of the annotation of a corpus for SA is a consistent annotation rather than a full agreement. If we compare annotation for SA to that performed for other tasks, we can see relevant differences that should be dealt with in different ways with respect to e.g. co-reference annotation (Poesio and Artstein, 2008), where the use of underspecified representations is exploited as a means to cope with the inherent ambiguity of the data to be annotated. By contrast, according to the results of a fine-grained analysis of disagreements (see section 5.1.), for SA the occurrence of genuine ambiguities gives useful hints about what kind of annotation can be more suitable for the task. In particular, observing the features of the task, we investigated some directions of analysis, among which the detection of subjectivity of the sentiment tags according to different measures, and the detection of systematic differences among annotators, devoted to identify the peculiarities of this task.

5.1. Measuring disagreement

For what concerns the detection of the *subjectivity of the sentiment labels* in our annotation scheme, we hypothesized that when a sentiment label is more involved in the occurrence of disagreement, this is because it is more difficult to be annotated, as its meaning is less shared among the annotators and there is a larger range of subjectivity in its interpretation. This phenomenon can be modeled and described according to different perspective and with reference to different portions of the dataset.

[10]Emoji are an alternative for explicit, manual labels, see `http://en.wikipedia.org/wiki/Emoji`.

In order to calculate the subjectivity of each label *L* we propose the following measure: considering all the tags exploited by all the annotators during the annotation process (i.e. 4,936 for the 1,235 Tweets of the *A-set*, and 867 for the 265 Tweets of the *D-set*), we calculated for each *L* the percentage of cases where *L* has been annotated for a Tweet in the *A-set* or for one in the *D-set*. Table 2 shows therefore how much a label has been used in percentage to contribute to the definition of an agreed or disagreed annotation of the Tweets.

label	agreement	disagreement
POS	26.3	14.4
NEG	29.2	17.8
NONE	21.8	23.5
MIXED	3.3	8.8
HUM	11.9	13.0
UN	7.6	22.5

Table 2: A measure of subjectivity of tags annotated in TW-FELICITTA corpus: percentage of Tweets in agreement/disagreement where each label is involved.

It should be observed, in particular, that while POS and NEG labels seem to have a higher reference to the agreement, for UN and MIXED the opposite situation happens, confirming that the annotators are more troubled by the exploitation of the latter tags.

Assuming a perspective oriented to the single annotators and referring to all the annotated tags, as above, we also measured the *subjectiveness* of each *annotator involved in the task* according to the variation in the exploitation of the labels. For each label *L*, starting from the total amount of times when *L* has been annotated, we calculated the average usage of the label. Then we calculated the deviation with respect to the average and we observed how this varies among the annotators. In table 3 the labels are presented from the most to the least used, together with the percentage of positive and negative deviation with respect to the average number of times where they have been annotated.

label	total	average	deviation +	deviation -
NEG	1,592	398	15.32%	14.82%
POS	1,421	355.25	6.68%	5.13%
NONE	1,281	320.25	24.90%	16.31%
HUM	700	175	28.57%	31.42%
UN	569	142	73.94%	35.21%
MIXED	237	59.25	46.83%	80.18%

Table 3: A measure of variation among the exploitation of the labels in TW-FELICITTA corpus.

The deviation is maximum for the tags MIXED and UN, while is meaningfully lower for all the other tags, in particular for POS and NEG, showing that the annotators are more confident in exploiting these latter tags.

Focusing instead the analysis on the *A-set* only, and again assuming a perspective oriented to the single annotators, we can calculate a sort of precision of the annotation done by each of them. We calculated this measure by considering each annotator A as a system whose results should be evaluated against the gold standard represented by our *A-set*. Dividing the amount of Tweets annotated by A with the same tag exploited in the *A-set* over the amount of Tweets included in the *A-set*, we obtained the precision shown by A in the annotation task. The scores for our annotators vary from 0.801 to 0.911, confirming that they can be considered as skilled enough and featured by a limited bias.

On the same set of data, i.e. *A-set*, but focusing on the tags, for each polarity label *L* we calculated the amount of Tweets that contain in their annotation at least one occurrence of *L*, divided by the amount of Tweets whose final annotation has been done with that label. The value of this measure is 1, when *L* is highly precise, that is each time that *L* has been used by some annotator, the final annotation of the Tweet in the released corpus is exactly *L*; it is higher than 1 when *L* is less precise. As reported in table 4, the lower scores are referred for POS and NEG, while the higher for UN and MIXED, which are in effect the labels annotated when the polarity of the Tweet is more ambiguous.

label	precision
POS	1.2
NEG	1.2
NONE	1.5
MIXED	2.0
HUM	1.2
UN	3.5

Table 4: A measure of precision of tags annotated in TW-FELICITTA corpus.

We conclude with some observation on the tag HUM, which we would like to investigate in the future work. If we focus on the *A-set*, we can see that all the Tweets included in it are featured by three or four annotations done with the same tag. If we further limit our observation to the Tweets associated with only three annotations done with the same tag and the fourth different, we see that for more than a quarter of them the fourth annotation is done by the tag HUM.

Another aspect we investigated is related to the issue of which tags co-occur more frequently with the tag HUM in the Tweets. Comparing the distribution of the tags on tweets that were labeled as HUM at least by one of the annotators to the overall distribution of the tags (excluding the tweets containing in their annotation a tag HUM), it appears that HUM significantly co-occurs with the UN and MIXED tags. With regard to the co-occurrence of HUM and UN, this result can be explained with the importance of the con-

text and of common ground, which, according to functional psychological models of language use, are often preconditions for understanding if a text is ironic utterance. While with regard to the co-occurrence of HUM and MIXED, in many cases the misinterpretation takes place because a sarcastic expression has been used; as also noted in (Riloff et al., 2013), a common form of sarcasm on Twitter consists of a positive sentiment contrasted with a negative situation, therefore, even though a positive sentiment is expressed in the utterance, the overall perception of the ironic tweet is that it bears a negative polarity. This may lead in annotators that do not recognize the ironic intent (maybe, again, for the absence of a context) to the perception that the Tweet has a mixed polarity.

6. Conclusion and future work

We described a new corpus for SA developed within the context of a platform for the detection of happiness. The development resulted in both a data set for system training and testing (i.e. Tweets on which we achieved the agreement of the annotators), but it also provides the basis for a framework to capture and analyze the nature of the disagreement (i.e. Tweets on which the disagreement reflects semantic ambiguity in the target instances and provides useful information). We propose a new type of ground truth, which is richer in diversity of perspectives and interpretations, and reflects more realistic human knowledge. Moreover, we propose a framework to exploit such diverse human responses to annotation tasks for analyzing and understanding disagreement.

7. References

Allisio, L., Mussa, V., Bosco, C., Patti, V., and Ruffo, G. (2013). Felicittà: Visualizing and estimating happiness in Italian cities from geotagged Tweets. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096, pages 95–106. CEUR-WS.org.

Baldoni, M., Baroglio, C., Patti, V., and Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54.

Baldwin, T. (2012). Social media: friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 58–59.

Basile, V. and Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Bertola, F. and Patti, V. (2013). Emotional responses to artworks in online collections. In *UMAP Workshops*, volume 997 of *CEUR Workshop Proceedings*.

Bolioli, A., Salamino, F., and Porzionato, V. (2013). Social media monitoring in real life with blogmeter platform. In *ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 156–163. CEUR-WS.org.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 44(10):91–94.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Caselli, T., Russo, I., and Rubino, R. (2012). Assigning connotation values to events. In *Proc. of the 8th Language Resources and evaluation Conference, LREC'12*, pages 3082–3089.

Davidov, D., Tsur, O., and Rappoport, A. (2011). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the CONLL'11*, pages 107–116, Portland, Oregon (USA).

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, 55(1):34–43.

Esuli, A., Baccianella, S., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh Language Resources and Evaluation Conference, LREC'10*. ELRA.

Fellbaum, C., editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

González-Ibáñez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hao, Y. and Veale, T. (2010). An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.*, 20(4):635–650, November.

Helliwell, J., Layard, R., and Sachs, J. (2014). *World Happiness Report 2013*. UN Sustainable Development Solutions Network.

Mihalcea, R. and Liu, H. (2006). A corpus-based aproach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs*.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. ACL.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: developing an aligned multilingual database. In *Proc. of Int. Conference on Global WordNet*.

Poesio, M. and Artstein, R. (2008). Inter-coder agreement

for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Quercia, D., Crowcroft, J., Ellis, J., and Capra, L. (2012). Tracking "gross community happiness" from tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work)*, pages 965–968.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proc. of EMNLP*, pages 704–714. ACL.

Saif, H., Fernández, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 9–21. CEUR-WS.org.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA.

Strapparava, C., Stock, O., and Mihalcea, R. (2011). Computational humour. In Cowie, R., Pelachaud, C., and Petta, P., editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 609–634. Springer-Berlin.

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing, CICLing 2013*, volume 7817 of *LNCS*, pages 121–136. Springer.

Versley, Y. (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-)reference. In *Proceedings of ESSLI'06*.

Xia, F. and Yetisgen-Yldiz, M. (2012). Clinical corpus annotation: challenges and strategies. In *LREC 2012 Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*.

Using Tweets for Assigning Sentiments to Regions

Erik Tjong Kim Sang
Meertens Institute
Amsterdam, The Netherlands
erik.tjong.kim.sang@meertens.knaw.nl

Abstract

We derive a sentiment lexicon for Dutch tweets and apply the lexicon for classifying Dutch tweets as positive, negative or neutral. The classifier enables us to test what regions in the Netherlands and Flanders express more positive sentiment on Twitter than others. The results reveal sentiment differences between Flemish and Dutch provinces, and expose municipalities which are a lot more negative than their neighborhood. The results of this study can be used for finding areas with local issues that might be expressed in tweets.

Keywords: Sentiment analysis, Twitter, Dutch

1. Introduction

Measuring sentiment of social media messages is an important application for organizations and individuals that want to track the impact of products, services, events and people on social media. However, the sheer volume of the data stream makes manual impossible for all but small selections of the data. An automatic analysis is difficult because of the ambiguity of language but there is no alternative when large volumes of data need to be processed.

In this paper we describe a two-stage process for identifying positive and negative Dutch tweets. First we create a Dutch sentiment lexicon based on the vocabulary observed in Dutch tweets. Next we use the lexicon for determining if Dutch tweets are positive, negative or neutral.

After this introduction and an overview of the related work, we present our method for deriving a sentiment lexicon from tweets. Next we apply the lexicon in a case study: a comparison of the average sentiment of different regions in The Netherlands and Flanders. The final section of the paper contains some concluding remarks.

2. Related work

The earliest references of work on automatic sentiment analysis are from 2002, for example the work of Pang et al. (2002), who use different machine learning techniques for determining if movie reviews are positive or negative. Application of sentiment analysis to tweets started seven years later, with among others the report of Go et al. (2009) who created a training corpus of 1.6 million positive and negative tweets by using emoticons as noisy labels. This approach has been used by several follow-up works, for example Pak and Paroubek (2010) . Sentiment analysis applied to Dutch tweets was only reported on in Tjong Kim Sang and Bos (2012), who performed a manual sentiment analysis of political Dutch tweets. In 2012, the company Incentro seemed to have developed a sentiment analysis module for Dutch (Incentro, 2012) but its current status is unknown. Sentiment analysis of tweets per region was first covered by Mislove et al. (2010) who studied the average mood of regions of the United States in the course of two days.

3. Sentiment lexicon

We use a lexicon of sentiment words for identifying positive and negative tweets. There are two reasons for favoring this approach over a machine learning approach with a training corpus of positive and negative examples. The first reason is portability: while we can share share a sentiment lexicon created from tweets with the research community, we would not be able to share annotated tweets because of the developer rules of the company Twitter (Twitter, 2011)[1]. The second reason is ease of implementation: our sentiment analysis is part of a parallel tweet search engine implemented on the Hadoop framework (White, 2012). Creating a lexicon-based analyzer required fewer resources than implementing a machine learner on Hadoop.

In order to collect words for the sentiment lexicon, we collected three sets of Dutch tweets, one with tweets that contained smileys – :-) or :) – one with tweets that contained frownies – :-(or :(– and one which did not contain any of the four emoticons. Our assumption is that when we compare the sets, words that express positive sentiment would predominantly be found in the first dataset while words associated with a negative sentiment would be found primarily in the second set. The third set will be used as an approximation of neutral tweets. We used the website twiqs.nl (Tjong Kim Sang and van den Bosch, 2013) for building the two sentiment tweet sets from the available tweets of January 2013 (1,724,642 and 999,685 tweets respectively). The third set was generated from the tweets of 16 January 2013 (2,569,203 tweets).

The search process produced frequency scores for 1,642,659 strings (words, names, punctuation signs and hash tags) from the sentiment tweets. We compared the frequencies of the two datasets with the t-test: $(f_1 - f_2)/\sqrt{f_1 - f_2}$ (Church et al., 1991, page 8), a measure for comparing the usage of a word in different texts. We created two ranked lists of words (strings without punctuation): one of the positive words versus the negative and neutral words and one of the negative words versus the positive and neutral words. We use add 0.5 smoothing to deal

[1]Twitter allows sharing the identification codes of tweets which can be used for retrieving the tweet text from their website. However the retrieval process will fail when tweets have been deleted by Twitter or by the author of the tweet.

Name	Lexicon size	Accuracy
baseline: 4 emoticons	4	87.6%
n-best t-scores, threshold 0	75,000	48.2%
n-best t-scores, threshold 5	4,400	53.2%
n-best t-scores, threshold 10	2,700	53.6%
incremental selection	644	84.2%
incr. sel. with 4 emoticons	100	93.2%
manual selection	338	82.1%

Table 1: Performance of the sentiment lexicon extracted from January 2013 tweets when tested on automatically annotated tweets from July 2013. For the n-best experiments, only the best results per threshold value are shown. The best results have been achieved with incremental selection of words suggested by the t-test combined with the four emoticons of the baseline: :-) :) :-(:(

with zero frequencies.

The t-test does not provide a perfect sentiment ranking: the top of the lists contained some character sequences that accidentally occur in a few positive or negative tweets. Therefore we experimented with frequency thresholds (0, 5 and 10) and removed words from the lexicon that occurred fewer times in either of the sentiment collections. We also tested building the lexicon incrementally, by only adding strings suggested by the t-test to the lexicon if they improved the sentiment performance of the lexicon on the test data.

Next, we devised a method for assigning sentiments to tweets based on the lexicon words. Tweets that do not contain any of the words will be neutral and tweets with words from only one sentiment set will be assigned that particular sentiment. In case a tweet contains both positive and negative words, the majority sentiment can be assigned. In case of a tie, the sentiment of the final sentiment word can be given preference, so that some cases of irony can be handled, like in the tweet: *so happy with with math grade :(*. Sentiment words immediately preceded by any of the words *not* (*niet*) and *no* (*geen*) are interpreted with their opposite sentiment value.

As test data we used a random selection of tweets from July 2013. We manually annotated 500 tweets with at least one of the two smileys, 500 tweets with at least one of the two frownies and 1000 tweets without any of the four emoticons. We selected the first 600 positive, negative and neutral tweets of this set as test set (a total of 1800). We tested different lexicons and measured their accuracy on classifying the tweets in test data with respect to the three sentiment classes positive, negative and neutral. A summary of these experiments can be found in Table 1.

Because of the method we used for selecting the test data, the baseline lexicon with only four emoticons already performed very well (accuracy 87.6%). Using the n words with the best t-scores did not perform as well (best accuracy 53.6%). Restricting the lexicon words to words which appeared at least 5 or 10 times in both positive and negative tweets, was a good idea (best accuracy 53.6% vs 48%). Adding only words to the lexicon which improved their performance on the test data worked very well, both without emoticons (accuracy 84.2%) and with emoticons in the lexicon (best accuracy 93.2%). We also evaluated a manually created sentiment lexicon and found that its performance was between the n-best approaches and the incremental selection methods.

In our experiments, incrementally adding words suggested by the t-test worked best. Words are only added to the lexicon if they improve the performance on the test set. However, this approach amounts to tuning the lexicon to the test data which may lead to performances which cannot be reproduced for other datasets. An inspection of the words in the two lexicons showed that several of the included words did not express sentiment in isolation but they were only added because they appeared in a positive or negative tweet in the test data. For this reason we did not select the lexicons generated with this method but we continued with the manually selected lexicon. This lexicon also has the advantage that it finds more sentiment tweets than the baseline lexicon and the incrementally built lexicons.

4. Measuring sentiment per region

As an application, we measured the average sentiment of regions in The Netherlands and Flanders. The results of this measurement could complement the research on living conditions periodically performed by the Dutch and Belgian government and press. For this purpose we selected the tweets with geolocation information from the period 1 to 31 January 2014 and measured their sentiment using the sentiment lexicon described in the previous section. About 5% of the tweets in the selected time frame contain geolocation information, a total of 2,071,851 tweets.

We started with examining regions. Dutch is spoken in The Netherlands (12 provinces) and Flanders (5 provinces). We made crude map of the 17 provinces and linked the boundaries to longitude and latitude figures from the coordinate system used in the tweet meta data (degrees with decimal part). Next, we determined for every tweet coordinate to which province it belonged using the point-in-polygon algorithm (Sutherland et al., 1974). We found tens of thousands of tweets per province, the lowest number for Belgian Limburg (29,489) and the highest number for South-Holland (313,312). The associated sentiment scores can be found in Figure 1.

The most striking observation that can be made from the map in Figure 1 is the difference between The Netherlands and Flanders. With the exception of the most southern Dutch province Limburg, all Dutch provinces obtain a sentiment score of nine or higher. Meanwhile the maximum score of the Flemish provinces is nine. This is not an isolated feature: we made similar observations for earlier months. There is no obvious reason why people in Flanders would tweet more negatively that people in The Netherlands. Cornips (2014) has suggested that the measured differences might be caused by dialect differences. If people from the southern regions of the map use words for expressing sentiment that are not part of our sentiment lexicon then the sentiment scores measured for their region will be closer to zero than the the scores measured in the northern regions. Since the average sentiment is positive, it will

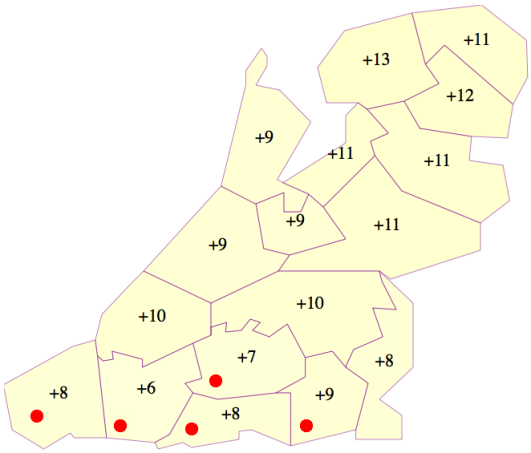


Figure 1: Twitter sentiment scores in the 17 Dutch-speaking provinces of Flanders and The Netherlands measured in January 2014. There is a clear difference between the sentiment scores of the provinces of The Netherlands on one side and the provinces of Flanders (marked with dots) on the other side.

appear that their tweets are less positive while this need not be the case.

Next, we performed sentiment analysis tweets originating from municipalities. The number of municipalities is too large to represent in a easily drawable map so we applied for an official Dutch municipality map from the Dutch mapping registry Kadaster. They offered the digital version of 2012 which was already outdated (417 instead of 403 municipalities). We mapped the tweets to municipalities using the point-in-polygon algorithm. The number of tweets per municipality was lower than for the provinces, with a minimum of 221 for Ouderkerk. For this reason, we computed the average of the sentiments per user (36 for Ouderkerk) rather than per tweet, otherwise one user with many tweets could have a large impact on the sentiment score of a municipality.

The resulting map for January 2014 can be found in Figure 2. Some interesting observations can be made. First, the larger population centers achieve sentiment scores at or below average: neither Amsterdam (+15), Rotterdam (+12), The Hague (+15), Eindhoven (+14), Tilburg (+15), Almere (+13), Breda (+13) nor Nijmegen (+14) does better than average (+15). Utrecht (+17) and Groningen (+16) are the only two of the ten most populated Dutch municipalities that achieve an above-average score.

A second observation is that all five Frisian islands in the north achieve very positive scores with the island of Schiemonnikoog appearing as one of the two most positive municipalities of The Netherlands. The fact that the islands are a popular tourist attraction probably has a positive influence on their mood on Twitter.

A third observation is that there are large differences between some neighboring municipalities. Voerendaal (+3) in the south has a relatively low score but neighboring Gulpen-Wittem (+22) is very positive. Grootegast (+5) in the north also has a relatively low score but it is surrounded by municipalities with scores around +20. Further study of the tweets involved is necessary to see if they mention important local issues that cause discomfort for their inhabitants.

5. Concluding remarks

We have described a sentiment analysis method for Dutch tweets based on a sentiment lexicon automatically derived from tweets. Words in the lexicon have been selected based on a comparison of positive and negative tweets with the t-test (Church et al., 1991). Several versions of the lexicon have been tested. We chose a manually developed lexicon of 338 tweets as the most appropriate for further experiments.

The sentiment lexicon has been used for determining the average sentiment of Dutch-speaking regions: provinces in Flanders and The Netherlands and municipalities in The Netherlands. The province sentiments revealed a surprising difference between Flemish and Dutch regions, most likely caused by differences in tweet vocabularies between the two areas. In the municipality results, we observed neutral busy regions and happy holiday regions. We also found some areas which were much less positive than their neighbors, a possible indication of local problems.

In all cases, one should be cautious in drawing conclusions from the sentiment measurements. They have been performed automatically and contain a certain degree of error. But one should also take into consideration that the demographics of Twitter is different from that of the Dutch-speaking community. This especially true for the users behind the tweets studied: people that freely share their location in their tweets. This group is predominantly male (66%) and over 25 years of age (56%). Only manual study of the tweets themselves can give an insight in why the users are positive or negative.

An obvious followup of this work is to try to find more indicators than the two used in this study (positive/negative), for example, crime, recreation, traffic, pollution, education and politics. A live view of local opinions of these topics would be interesting for policy makers. The main challenge here would be to collect enough tweets to be able to say something meaningful about the topics for all regions. Present day Twitter will probably not be able to satisfy that information need completely but it should prove to be a useful addition to other information sources.

6. References

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

Cornips, L. (2014). Zeur-tweets. *Dagblad De Limburger*, 25 January.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Technical Report, Stanford Digital Library Technologies Project.

Incentro. (2012). Sentiment Analysis for the Dutch Language. `http://www.incentro.com/en/inspiration/blogs/sentiment-analysis-dutch-language`. Web page, accessed 3 February 2014.

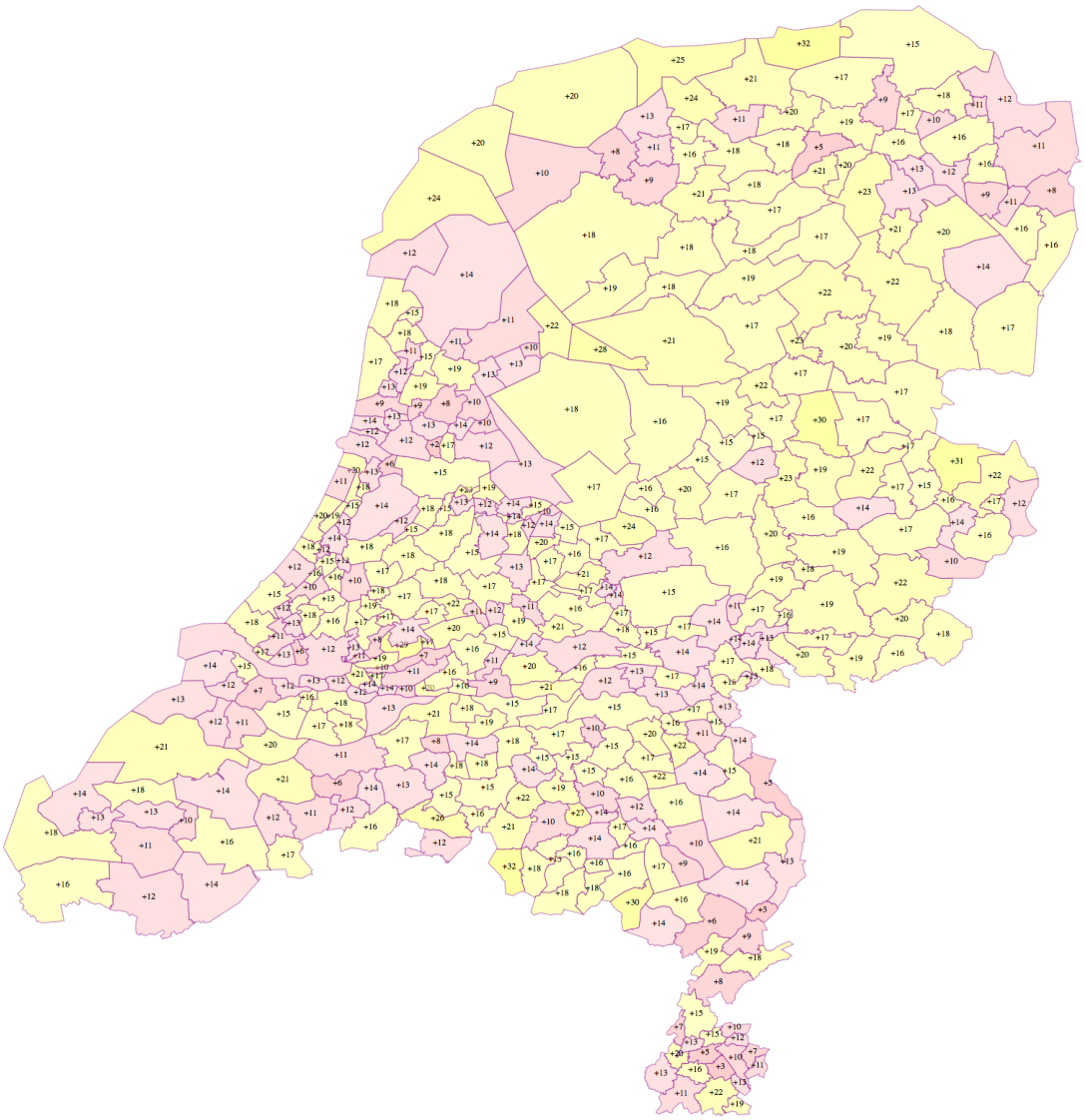


Figure 2: Twitter sentiment scores in the 417 Dutch municipalities measured in January 2014. Municipality areas include both land and water. Areas in yellow achieved an average (+15) or above average score while red areas achieved a score below average. Because of the low number of tweets, sentiment scores have been averaged over users rather than over tweets.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2010). Pulse of thge Nation: U.S. Mood Throughout the Day inferred from Twitter. `http://www.ccs.neu.edu/home/amislove/twittermood/`. Web page, accessed 4 February 2014.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*. European Language Resources Association (ELRA).

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-2002*, pages 79–86. ACL.

Sutherland, I. E., Sproull, R. F., and Schumacker, R. A. (1974). A Characterization of Ten Hidden-Surface Algorithms. *ACM Computing Surveys*, 6(1):1–55.

Tjong Kim Sang, E. and Bos, J. (2012). Predicting the 2011 Dutch Senate Election Results with Twitter. In *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks*, pages 53–60. ACL, Avignon, France.

Tjong Kim Sang, E. and van den Bosch, A. (2013). Dealing with Big Data: the Case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134. ISSN: 2211-4009.

Twitter. (2011). Developer Rules of the Road. `https://dev.twitter.com/terms/api-terms`. Web page, accessed 4 February 2014.

White, T. (2012). *Hadoop: the definite guide*. O'Reilly.

Emotions and Irony per Gender in Facebook

Francisco Rangel[1,2], **Irazú Hernández**[3], **Paolo Rosso**[3], **Antonio Reyes**[4]
[1]Autoritas Consulting, Madrid, Spain
[2]Universitat Politècnica de València, València, Spain
[3]PRHLT Research Center, Universitat Politècnica de València, València, Spain
[4]Laboratorio de Tecnologías Lingüísticas. Instituto Superior de Intérpretes y Traductores, Mexico
francisco.rangel@autoritas.es, {dhernandez1, prosso}@dsic.upv.es, antonioreyes@isit.edu.mx

Abstract

This paper describes a Spanish dataset collected from Facebook that has been labelled with emotions, irony and author's gender. The inter-annotator agreement shows the difficulty and high level of subjectivity of the annotation task, especially with respect to irony. Statistics of the corpus show the relationship among topics, emotions, irony and author's gender. For instance, females used more emotions than males (mainly positive emotions), males were more ironic than females (at least in this dataset), and politics is the topic addressed more ironically and with more *negative* emotions. A social analysis of the results goes beyond the scope of this paper but being the comments mainly about (some of) the Spanish politicians, we are not surprised of the results. The dataset is publicly available for research and social analysis purposes with the name EmIroGeFB at http://ow.ly/uQWEs

Keywords: emotions, irony, gender, Facebook

1. Introduction

Our habits are changing, we are no longer customers searching for products but users looking for new experiences. Social Media even accentuate such changes. The emotional aspect of the life is acquiring a growing importance. Thus, the need of affective processing acquires a new dimension nowadays in order to know what users want and need.

We are interested in social media since we are interested in everyday language and how it reflects basic social, emotional and personal processes. Furthermore, in social media users reflect what they want and need without restrictions and liberty of expression. But there is a lack of annotated resources on affectivity when we talk about social media texts. Even more if we focus on Spanish language, no matter its good penetration in Internet[1].

We focused on Facebook as representative of social media because it is massively used by people, where they express their thoughts freely and without editorial guidelines unlike traditional media like newsletters and with spontaneity unlike blogs. Thus the expected affectivity in such media is very high. Facebook also allows us to obtain demographics such as gender, unlike similar media like Twitter.

The paper describes a dataset collected from Facebook, in Spanish and labelled with emotions, irony and gender of authors. The structure of the paper is as follows. In Section 2 we describe the corpus, how the data was collected and annotated, and the inter-annotator agreement. In Section 3 we analyse the corpus and we present statistics about emotions and irony per gender. In section 4 the distribution of the labelled corpus is described. Finally we draw some conclusions in Section 5.

[1]http://eldiae.es/wp-content/uploads/2012/07/2012 el espanol en el mundo.pdf

2. Corpus description

In this section we describe the corpus, how it was collected, the labelling process and the inter-annotator agreement.

2.1. Dataset collection

Facebook is composed of a hierarchy of objects. Pages are one of the first level objects, as Profiles, Events or Groups. Each Page has an owner who publishes Posts. Posts are second level objects. Posts are written by the owner of the Page and follow the owner's guidelines and thematics. Posts allow other users to participate in the conversation by answering to them with Comments. Comments are third level objects. In Comments people can express what they think about the topic of the Posts but without the guidelines of Pages' owners. For building the dataset, we focused on Comments.

We selected three thematics, with high volume of participation[2], and susceptible of emotional comments, and as representative for each thematic, we selected four of the most well-known pages in such thematics in Spain, as it is shown in Table 1.

We retrieved at least 1,000 posts for each page, and all the comments written in each post. We collected comments with the gender information of their author. We randomly selected 200 comments for each thematic and each gender, balancing the data as shown in Table 2.

Neither selection nor cleaning has been done except for language filtering and for ensuring that comments have some text (e.g. they are not only shared links).

[2]http://www.pewglobal.org/files/2012/12/Pew-Global-Attitudes-Project-Technology-Report-FINAL-December-12-2012.pdf

POLITICS: Four official pages of Spanish political parties

Partido Popular
https://www.facebook.com/pp
Partido Socialista Obrero Español
https://www.facebook.com/psoe
Izquierda Unida
https://www.facebook.com/izquierda.unida
Union por el Progreso y la Democracia
https://www.facebook.com/Union.Progreso.y.Democracia

FOOTBALL: Four official pages of Spanish football clubs

Real Madrid CF
https://www.facebook.com/RealMadrid
FC Barcelona
http://www.facebook.com/fcbarcelona
Valencia Club de Futbol
http://www.facebook.com/vcf1919
Atletico Bilbao
http://www.facebook.com/pages/ATLETICO-BILBAO/103997686354572

CELEBRITIES: Four official pages of Spanish celebrities

Belen Esteban
http://www.facebook.com/BelenEstebanM
Kiko Hernandez
http://www.facebook.com/ElConfesionariodeKiko
David Bisbal
http://www.facebook.com/davidbisbal
Santiago Segura
http://www.facebook.com/pages/Santiago-Segura-Silva/12459228767

Table 1: Official pages selected for collecting data for each thematic

Theme	Gender	Comments
Politics	Male/Female	200/200
Football	Male/Female	200/200
Celebrities	Male/Female	200/200

Table 2: Dataset collected from Spanish Facebook comments

2.2. Labelling emotions, irony and gender

Three independent annotators[3] labelled 1,200 documents with the six basic emotions of the Ekmans theory (Ekman, 1972) (joy, surprise, fear, anger, disgust, sadness), irony and gender.
There are many ways of annotating emotions in texts:

- The emotion profiled by the speaker;
- The emotion produced in the hearer;
- The emotion that is described or expressed.

[3]Two females and one male

We asked the annotators to use the last approach trying to involve as little as possible issues that are purely personal. Annotators were provided with the information of Figure 1 that was obtained by Greenberg (Greenberg, 2000) on the basis of psychological relationships of emotional states with the six basic emotions of Ekman. It is remarkable that some secondary emotions are shared by more than one primary emotion; for example, *indignation (indignación)* is shared by *anger* and *disgust*, and *fascination (fascinación)* is shared by *joy* and *surprise*. This issue hinders the unique identification of such basic emotions, as it was evidenced in (Ortony and Turner, 1990). Besides, the identification of multiple emotions and the absence of any has been allowed.

ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA
Agradecido	Agresivo	Acomplejado	Aborrecimiento	Extrañeza	Abatido
Alegre	Colérico	Alarmado	Desagrado	Sobresalto	Agobiado
Animado	Crispado	Angustiado	Grima	Susto	Apenado
Calmado	Descontento	Ansioso	Repulsión	Consternación	Confuso
Confiado	Enfadado	Atemorizado	Antipatía	Pasmo	Decepcionado
Contento	Enojado	Aterrado	Aversión	Desconcierto	Deprimido
Dichoso	Excitado	Avergonzado	Repugnancia	Estupor	Desalentado
Encantado	Fastidiado	Confuso	Disgusto	Asombro	Desanimado
Entusiasmado	Furioso	Desesperado	Repudia	Fascinación	Desdichado
Eufórica	Insatisfecho	Desorientado	Repulsa	Admiración	Desmoralizado
Esperanzado	irascible	Horrorizado	Odio	Confusión	Frustrado
Feliz	Malhumorado	Inquieto	Manía	Chasco	Nostálgico
Gozoso	Molesto	inseguro	Rabia	Impresión	Soledad
Satisfecho	Nervioso	Intranquilo	Animadversión	Exclamación	Triste
Tranquilo	Rabioso	Pánico	Nauseabundo	Conmoción	Infeliz
Complacido	Tenso	Preocupado	Indignación	Estupefacción	Desconsolado
Libre	Violento	Temeroso	Enfado		Afligido
Fascinado	Irritado	Tenso	Desprecio		Amargado
Seguro	Indignado	Indeciso	Distanciamiento		Impotente
		Impotencia			

Figure 1: Secondary emotions related to the six basic emotions

Due to the increasing use of irony in social media[4] (Reyes et al., 2013)(Reyes and Rosso, 2012)(Bosco et al., 2013) we labelled each comment also as ironic/not ironic. Irony is a uniquely human mode of communication by which the speaker says something other than what he or she intends (Wallace, 2013). (Grice, 1975) and (Attardo, 2000) consider irony as an *intentional* violation of conversational maxims. We ask annotators for tagging each comment as ironic/not ironic based only on their own concept of irony. No further information or definition was provided.
Texts were also labelled with gender information in order to link this resource to tasks such as Author Profiling at PAN 2013(Rangel et al., 2013). Gender annotation was provided by Facebook, but we ensured the right annotation by manually checking first names and photos of the users.

2.3. Inter-annotator agreement

For emotions annotation we calculated the inter-annotator agreement with the Kappa DS method (Diaz-Rangel, 2013). This metric is based on Fleiss's Kappa but it allows to calculate concordance for more than two annotators (in our case three: A1, A2 and A3) with multiple not

[4]A pilot task on sentiment analysis and irony (in Italian) will be organised at Evalita-2014: http://www.di.unito.it/?tutreeb/sentipolc-evalita14/index.html Another task (in English) should be organised at SemEval-2015.

mutually exclusive categories (in our case six not mutually exclusive: the six basic emotions). Kappa DS is calculated for each couple of annotators and then the average of all of them is calculated to obtain the overall assessment. We show results in Table 3.

	A1	A2	A3	Rest
A1	-	0.0587	0.2738	0.1662
A2	0.0587	-	0.1042	0.0814
A3	0.2738	0.1042	-	0.1890
Total	0.1455			

Table 3: Kappa DS: Inter-annotators agreement for emotions annotation

The average value for Kappa, equal to **0.1455**, shows a low index of agreement according to (Landis and Koch, 1977). But, as it is shown in (Diaz-Rangel, 2013), we have to bear in mind the amount of variables intervening in the evaluation for the right interpretation of such index. We also grouped the nearest emotions, that is, those which share secondary emotions, as we highlighted in Figure 1: *joy / surprise* and *anger / disgust*. Results in terms of inter-annotator agreement are shown in Table 4. In this case, Kappa shows a higher value for the agreement (**0.6016**), what stresses out the need of considering such discordance among annotators when using the dataset for machine learning purposes, especially with respect to *joy/surprise* and *anger/disgust*.

	A1	A2	A3	Rest
A1	-	0.6618	0.5656	0.6137
A2	0.6618	-	0.5773	0.6196
A3	0.5656	0.5773	-	0.5715
Total	0.6016			

Table 4: Kappa DS: Inter-annotators agreement with grouped emotions: joy/surprise and anger/disgust

Respect to irony annotations, we calculate the Fleiss's Kappa to measure the inter-annotator agreement. This method allows multiple annotators (three in our case) and binary variables (ironic or non-ironic). We obtained a Fleiss's Kappa value equal to **0.0989**; i.e. a very low index of agreement. We think this low index is due to the task itself: irony is quite subjective and depends on people, their moods, linguistic and cultural context, etc. to be correctly understood; as well as contextual information that, in our case, was not provided. For instance, we did not provide general definition of irony to set up a common framework of characteristics; or the fact that annotators carried out the task on text which means that information such as facial expressions or tone of voice could not help identifying a text as ironic.

In order to know the agreement rate among annotators in relation to emotions labelled also on ironic comments, we calculated the Kappa DS taking into account only the subset of comments identified as ironic. Results are shown in Table 5

	A1	A2	A3	Rest
A1	-	-0.0854	0.0001	-0.0426
A2	-0.0854	-	-0.1128	-0.0991
A3	0.0001	-0.1128	-	-0.0563
Total	-0.0660			

Table 5: Kappa DS: Inter-annotators agreement of emotional comments labelled with irony

We obtained a negative value of **-0.0660** which means there is no agreement when taking into account emotions and irony. In this respect, no further analysis was performed, only the agreement index among annotators. However, we think there is a close relation between irony and emotions, especially, the kind of emotion triggered by an ironic utterance. This kind of analysis is projected as future work.

3. Corpus analysis

In this section we present basic statistics about the corpus and the annotations for emotions and irony.

3.1. Emotions

As it was shown by the low inter-annotator agreement, emotion labelling is a difficult task. Below some examples are shown in their original language to preserve the meaning. We explain them in English in order to show the difficulty of labelling the basic emotion expressed.

e.g. "guuapaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa ers la mejorrr"

The previous is clearly an encouraging comment to a celebrity. But which is the basic emotion? Using the Figure 1, which word better describe the mood: *euphoria*, *fascination*, *admiration*? Depending on the selection, annotators would doubt between *joy* and *surprise*: Annotator 1 selected *joy* and *surprise*, Annotator 2 selected *joy* and Annotator 3 selected *surprise*.

e.g. "Es una vergüenza, que se financien ellos, que para algo son privadas"

Something similar happens with disgusting comments. In the previous one, the author criticizes some entities for being financed by the government although they are private. Comments like the previous one, may be labelled differently depending on annotators' mood or their own world vision. For instance, Annotator 1 and Anotator 2 labelled it focusing on *disgust* and *anger* while Annotator 3 labelled it focusing only on *anger*.

e.g. "guarda semejante alhaja en una camara acorazada por si os la roban"

In the previous comment, the author is recommending to a celebrity to save something in a safe room to preserve it from thieves. This comment is so ambiguous that the basic emotion labelled by each annotator is different. Annotator 1 reported *no-emotion*, Annotator 2 reported *fear* and Annotator 3 reported *surprise*.

In Table 6 the number of comments annotated for each emotion by annotator is shown. As it was previously mentioned, the difference among annotators is greater in emotions like *joy*, *anger*, *disgust* and *surprise*. It seems that

some annotators (A1 and A3) perceived as *surprise* what others (A2) perceived as *joy*, and similar with *anger* (A2) and *disgust* (A1).

	A1	A2	A3
Joy	255	756	215
Anger	96	265	148
Fear	19	6	7
Disgust	255	78	166
Surprise	626	140	460
Sadness	165	72	83
None	97	42	160

Table 6: Number of comments per emotion and annotator

We finally selected those annotations in which, at least, two out of three annotators agreed with. In Table 7 the number and percentage of comments per emotion is given. For example *joy* has a higher value (338) than that was obtained by two annotators (A1=255; A3=215). Something similar happens to other emotions. This means that the perception of *joy* or *surprise* is quite subjective.

	Total	%
Joy	338	28.17
Anger	151	12.58
Fear	3	0.25
Disgust	129	10.75
Surprise	390	32.50
Sadness	76	6.33
None	262	21.83

Table 7: Number and percentage of comments per emotion

In Table 8 the distribution of emotions labelled per gender is shown. Results seem to be quite balanced, no matter there are less comments without emotions for females (18 vs. 37), or more positive/neutral emotions like joy (194 vs. 144) or surprise (215 vs. 175).

	Male	Female
Joy	144	194
Anger	79	72
Fear	2	1
Disgust	66	63
Surprise	175	215
Sadness	37	39
None	37	18

Table 8: Emotions per gender

In Table 9 the distribution of emotions per topic is shown. As it was expected, politics is the most negative perceived topic with higher values for anger, disgust and sadness emotions, and also with lower values for non-emotional comments. Football and celebrities have similar values for joy and surprise, but celebrities have higher values for disgust. Maybe this is due to the fact that people write in celebrities' pages for supporting or criticizing them, depending on the affinity to them.

	Politics	Football	Celebrities
Joy	50	153	135
Anger	114	10	27
Fear	2	1	0
Disgust	79	7	43
Surprise	53	180	157
Sadness	52	9	15
None	9	23	23

Table 9: Emotions per topic

3.2. Irony

Some examples of ironic comments are shown below. Comments are shown in their original language in order to preserve their ironic sense. We provide an English explanation based on our own interpretation, in order to show the difficulty of the task.

e.g. "Pitbul es cultura, no ves que te enseña a contar? aunque sea sólo hasta 3"

In the previous comment, the authors criticises the singer for including in his lyrics "one, two, three...". The author says that this is culture because listening such singer, anyone can count. At least, until number three. The author expresses a positive comment using a remark in order to emphasizing his negative opinion about this singer. In this comment, two of three annotators agreed.

e.g. "Que viva, pero muy lejos!"

In this comment, the author expresses his intention of being far from someone, mentioning at first a positive desire and finally showing his real intention. In this comment all the annotators agreed.

e.g. "Pobres, en el fondo producis ternura...que triste tiene que ser haber votado al PP."

In this comment, the author expresses shame towards people. The author uses this remark in order to show his despise about people's judgement for choosing the current politician party. The author expresses a negative comment in order to show his real intention. In this comment two of the three annotators agreed.

e.g. "Eres muy injusto y quiero que sepas que la infanta cuando se fue a vivir a su nueva vivienda recien reformada y a pesar de ser mucho mas pequeña que la zarzuela se mudo convencida de que era una VPO o no?....."

In the previous comment, the author says that the the Spanish King's daughter moves to a new residence. She says that the Spanish King's daughter is convinced that this new house is a kind of state subsidy housing because it is smaller than Zarzuela's Palace, the Residence of the Spanish royal family. The author expresses a positive remark about someone's judgement including comparisons in order to emphasize the utterance's ironic sense. In this comment all the annotators agreed.

e.g. "Yo soy presunta ciudadana española y digo esto porque no estoy segura de si realmente lo soy o si vivo en una realidad paralela donde nuestro presi es más inútil que una neurona de Paris Hilton."

In the last comment, the author alludes the possibility of living in a parallel reality because her country is governed for someone useless than a Paris Hilton's neuron. The author compares two remarks in the same comment, in order to emphasizing her real intention to show disagreement with government of her country. In this comment all the annotators agreed.

In Table 10 the number of ironic comments labelled by each annotator is shown. The percentage of comments labelled with irony is very low, although one annotator labelled a higher number of comments than the rest.

Annotator	Comments	%
A1	52	4.33
A2	189	15.75
A3	48	4.00

Table 10: Number of comments with irony per annotator

We determined as ironic only those comments that were annotated as ironic by at least two annotators. As can be seen in Table 11 only 42 comments fits this criteria.

	Total	%
Ironic	42	3.62
Non-ironic	1158	96.37

Table 11: Number and percentage of ironic and non-ironic comments

In Table 12 is shown the number of ironic comments per gender and topic. We can see that males used irony more than females and politics is the topic with most ironic comments.

	Female	Male	Total
Football	1	3	4
Politics	11	16	27
Celebrities	3	8	12
Total	15	27	42

Table 12: Ironic comments per gender and topic

Finally, in Table 13 we show the number of comments per emotion, in which, at least, two of three annotators agree.

Emotion	Ironic comments
Joy	8
Anger	4
Fear	0
Disgust	6
Surprise	6
Sadness	0
None	3

Table 13: Number of ironic comments per emotion

4. Corpus distribution

In its data use policy[5] Facebook says: "Because Pages are public, information you share with a Page is public information. This means, for example, that if you post a comment on a Page, that comment may be used by the Page owner off Facebook, and anyone can see it.". We collected comments from public pages thus the data collected is public and can be seen by anyone.

For distributing the collection we use a XML file with the structure described in Table 14.

```
<dataset>
    <comments count="1200">
        <comment ID="FACEBOOK_COMMENT_ID"
                 gender="male|female"
                 topic="POLITICS|FOOTBALL|CELEBRITIES">
           <annotator1>
                <joy>true/false</joy>
               <surprise>true/false</surprise>
                <sadness>true/false</sadness>
                <anger>true/false</anger>
                <disgust>true/false</disgust>
                <fear>true/false</fear>
                <no-emotion>true/false</no-emotion>
                <irony>true/false</irony>
            </annotator1>
            <annotator2>
                ...
            </annotator2>
            <annotator3>
                ...
            </annotator3>
        </comment>
        ...
    </comments>
</dataset>
```

Table 14: XML structure of distributed data

Each Facebook comment is identified by an unique ID with the form:

`pageID_postID_commentID`

For example:

`208701145825784_582486558447239_1966964`

For downloading contents a Facebook token is needed. It may be generated at Facebook Developers website [6]. With the Facebook comment ID and the generated Facebook token, content is available through Facebook Graph [7]

Result is provided in JSON format with the structure described in Table 15.

[5] `https://www.facebook.com/full_data_use_policy`
[6] https://developers.facebook.com/tools/explorer/
[7] https://graph.facebook.com/COMMENTID?access_token=TOKEN

```
{
   "id": "208701145825784_582486558447239_1966964",
   "from": {
      "name": "COMMENTER NAME",
      "id": "COMMENTER ID"
   },
   "message": "COMMENT CONTENTS",
   "can_remove": [false|true],
   "created_time": "DATETIME",
   "like_count": NUMERIC,
   "user_likes": [false|true]
}
```

Table 15: JSON format of Facebook response

The described dataset is available at http://ow.ly/uQWEs with the name EmIroGeFB.

5. Conclusions

In this paper we describe a Spanish dataset collected from Facebook that has been labelled with emotions, irony and author's gender. Such dataset was manually labelled considering four layers: the six basic emotions described in Ekman's theory, the absence of emotions, irony, and finally, gender. To our knowledge, this is the first attempt to link the gender of an author with emotions and irony.

In order to evaluate the annotation of the dataset, we carried out a Kappa-DS analysis of concordance for emotions. To this respect, we shown that there is low concordance due to some emotions such as *joy/surprise* and *anger/disgust* that are very close to each others. In the case of irony, we carried out a Fleiss's Kappa analysis, resulting in a very low concordance. This shows how subjective irony is. A Kappa-DS analysis of concordance was carried out with cases of comments labelled both with irony and emotions. No agreement was found among annotators. The main reason is the high level of subjectivity when annotating the texts. This negative result may suggest that people express irony independently of the emotions they feel. This issue will be investigated further in the future.

The statistics show that (at least in this dataset): i) females tend to use more words related to emotions than males, mainly positive emotions; ii) males tend to be more ironic than females; or iii) the category *politics* is the one with more negative emotions and irony than other the rest of categories. Being the comments mainly about (some of) the Spanish politicians, we are not surprised of the results.

Finally, this dataset was used in (Rangel and Rosso, 2013) for automatic identification of emotions in text, and also for gender identification, showing to be a valuable resource for research in social media in Spanish.

6. Acknowledgements

The work of the first author was partially funded by Autoritas Consulting SA and by Ministerio de Economía de España under grant ECOPORTUNITY IPT-2012-1220-430000. The National Council for Science and Technology (CONACyT-Mexico) has funded the research work of the second author (218109/313683 grant). The work of the third author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

7. References

Attardo, S. (2000). Irony as relevant innappropriateness. In *J. Pragmat*.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. In *IEEE Intelligent Systems 28(2)*, pages 55–63.

Diaz-Rangel, I. (2013). Detección de afectividad en texto en español basada en el contexto lingístico para síntesis de voz. In *Tesis Doctoral. Instituto Politécnico Nacional*, México.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In *Symposium on Motivation*, pages 207–283, Nebraska.

Greenberg, L. (2000). Emociones: Una guía interna. In *Descleé De Brouwer*, Bilbao.

Grice, H. (1975). Logic and conversation, syntax and semantics. In *Academic Press*, pages 41–58.

Landis, R. and Koch, G. (1977). The measurement of observer agreement for categorical data. In *Biometrics (35)*, pages 159–174.

Ortony, A. and Turner, T. (1990). What's basic about basic emotions? In *Psychological Review (97)*, pages 315–331.

Rangel, F. and Rosso, P. (2013). On the identification of emotions in facebook comments. In *ESSEM Workshop on Emotion and Sentiment in Social and Expressive Media, AIxIA, CEUR-WS.org, vol. 1096*, pages 34–46, Turin, Italy.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. (2013). Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013*, Valencia, Spain.

Reyes, A. and Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customers reviews. In *Journal on Decision Support Systems, vol. 53, issue 4*, pages 754–760.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. In *Language Resources and Evaluation, vol. 47, issue 1*, pages 239–268, Turin, Italy.

Wallace, B. (2013). Computational irony: A survey and new perspectives. In *Artificial Intelligence Review*, pages 1–17.

On-line Annotation System and New Corpora for Fine-Grained Sentiment Analysis of Text

Ekaterina Volkova[1,2] **and Betty J. Mohler**[1]
[1]Human Perception, Cognition and Action,
Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Graduate School of Neural & Behavioural Sciences, International Max Planck Research School,
University of Tübingen, Tübingen, Germany
{ekaterina.volkova, betty.mohler}@tuebingen.mpg.de

Abstract

We present a new on-line annotation system that allows participants to perform manual sentiment analysis of coherent texts for emotions, as well as indicate the intensity of the emotion and the emphasis in each phrase. We have developed the following set of emotion categories: *amusement, anger, contempt, despair, disgust, excitement, fear, hope, joy, neutral, pride, relief, sadness, shame, surprise*. This set greatly expands the boundaries of the often used basic emotion categories and is balanced for positive and negative emotions. Using this new annotation tool and its predecessor version, we have collected two corpora of fairy tale texts manually annotated for emotions on the utterance level. One corpus encompasses 72 texts in German, each annotated by two participants. The other corpus is a work in progress and contains three fairytale texts, each annotated by seven participants. The inter-annotator agreement in both corpora is "fair". Although annotation conflict resolution strategies can be developed for merging several annotations into one, we suggest that for manual SA, the researchers should aim at recruiting more annotators and use the consensus method for retrieving an annotation based on the opinion of the majority.

Keywords: sentiment analysis, corpora, manual text annotation

1. Introduction

The amount of work carried out in the field of sentiment analysis (SA) during the last decades is impressive, with projects spanning across a variety of methods and domains, cf. reviews by Pang and Lee (2008) and Liu (2010). Most SA systems are implemented for specific goals. The final application field ranges from extracting appraisal expressions (Whitelaw et al., 2005) to opinion mining of customer feedback (Lee et al., 2008). However, simulation of human emotion perception of text is rarely discussed even in recent reviews on possible applications of SA (Karlgren et al., 2012). An automatic SA system that could simulate emotion perception of text would be useful in many areas of human-computer interaction, including but not limited to first and/or second language acquisition, social training, intelligent social agents and research in cognitive psychology. Such a system has to operate on relatively small linguistic units (sentences, clauses, phrases) and use a rich set of emotion labels. By now, the initial foundations for simulating human motional perception of text have already been laid (Liu et al., 2003; Strapparava and Mihalcea, 2008; Alm, 2008; Aman and Szpakowicz, 2008; Gill et al., 2008; Neviarouskaya et al., 2010).

For many manual annotation tasks, e.g., parts-of-speech tagging or syntactic functions labelling, the annotators can be provided with a detailed manual for solving most of the uncertainties during the annotation process. In manual SA, however, if the researchers want to capture the intuitive, natural perception of text, a detailed set of rules often developed to improve consistency comes at the cost of a decrease in the annotation quality with regard to individual annotator's preferences. Human emotion perception and expression is complex and varies by the individual, thus it is important to provide annotators with a rich set of emotion categories in order to allow them to express their emotion perception of text in the finest detail. It is important to let the annotators work with linguistic units smaller than a sentence, since it is very probable that the locus of an emotion instance can be narrowed down to a short utterance. The interface of the annotation system, e.g., web-based tool or a stand-alone application, should be intuitive for the user.

We present a new on-line annotation system that allows participants to perform manual SA of coherent texts for emotions. From the researcher's perspective, this new annotation tool allows the management of content for annotation, supervision over annotators' progress and export the resulting data. For the annotators the system is meant to be easy to use and requires minimum instruction from the researcher. The annotation system is open for both researchers and annotators, is web-based and does not require any specific software. It allows the annotators to encode their emotion perception of text with the help of a rich set of emotions, indicate the intensity for each emotion instance and mark the word that bears most of the emotional charge.

We demonstrate the performance of the proposed annotation scheme and its basic principles by collecting two corpora of texts: a German corpus annotated by two people, and a smaller corpus of English texts with seven annotators for each story. The approaches share crucial features, e.g., a large set of emotions used as annotation labels, and smaller than sentence, utterance-based annotation units; but differ in the implementation details. We show that consensus method can be successfully used for annotation merge as it results in a final annotation rich with emotions, suitable for establishing a *gold standard* annotation.

Category (German)	mean	stdev	min	max
Approval (Zustimmung)	2.3	1.00	1	4
Comfort (Zufriedenheit)	2.7	1.55	1	5
Compassion (Mitgefühl)	2.5	1.36	1	5
Hope (Hoffnung)	3.5	1.28	1	5
Interest (Interesse)	2.2	1.17	1	5
Joy (Freude)	4.6	0.66	3	5
Surprise (Überraschung)	1.9	0.94	1	3
Anger (Ärger)	-2.9	0.70	-4	-2
Anxiety (Unruhe)	-1.6	1.20	-4	-1
Despair (Verzweiflung)	-4.2	1.08	-5	-2
Disgust (Ekel)	-3.4	1.69	-5	-1
Fear (Angst)	-3.2	1.54	-5	-1
Hatred (Hass)	-4.8	0.40	-5	-4
Sadness (Trauer)	-2.8	1.08	-4	-1
Neutral (Neutral)	0	0	0	0

Table 1: Emotion categories used in the German corpus annotation. The categories in italics were later excluded from the corpus during the two annotations merge as a result of conflict resolution between the two annotators. Polarity of the German emotion categories (columns 2-5) was determined by ten participants during a pilot experiment.

2. German Corpus of Brother Grimm Fairytales

In this section we report on a corpus of texts in German annotated for fifteen emotion categories by two native speakers. We chose the Brother Grimm's fairy tales as the main annotation material. The texts were chosen based on their genre, for in spite of the references to the depths human psyche and national traditions that were shown in works of Propp (1968) and Von Franz (1996), folk fairy tales are uncomplicated in the plot-line and the characters' personalities. Due to this relative simplicity of the content, we expected the participants' emotional reactions to be more coherent than to other texts of fictional literature. While it may seem that fairy tale texts are generally more emotional than texts of other genres, this impression comes from the density of emotion instances rather than from the intensity of each individual instance (Mohammad, 2011).

The initial available collection of Brother Grimm fairy tales contained more than 300 texts from which we chose 139 texts that had length between 1400 and 4500 word tokens (lower and upper quartiles of the texts length distribution respectively). However, many of the 139 texts were written down in dialectal language or were very close versions of one and the same fairy tale. In the end we were left with 72 fairy tale texts 2500 word tokens long on average, written in Standard German. No two texts chosen for the annotation belonged to the same story although some universal plot lines, or functions (Propp, 1968) were common for several different stories. A list of fifteen emotion categories was employed for text annotation (see Table 1). Their polarity was determined during a pilot experiment (Volkova et al., 2010).

2.1. Text Delimitation into Annotation Units

Despite the well-established recognition of the fact that semantic content of single words and morphological units influence the emotion information contained in an utterance (Polanyi and Zaenen, 2006; Neviarouskaya et al., 2009), classification units in most SA approaches range from whole documents to sentences (Liu et al., 2003; Alm, 2008), but seldom go to a more fine-grained level. In our approach we have chosen a short phrase to be the basic annotation unit, henceforth referred to as utterance. This ensures that the annotators have more freedom in expressing their perception of text — they can annotate a sentence with several emotion categories or mark only one part of the sentence leaving some parts unmarked (*neutral*). The decision to split sentences into utterances before the annotation process was motivated by the fact that participants of a previous annotation study did it naturally (Volkova et al. 2010). Therefore the 72 texts were split into utterances before the annotators worked with them.

The delimitation was done also for practical reasons: to simplify and speed up the annotation process and to limit the disagreement on the placement of annotation unit borders. The sentence delimitation was performed automatically and was based on syntactic sentence structure. It relied on the requirement for the final utterances to be relatively short — the typical size of annotations units was empirically estimated as four to seven words long (Volkova et al., 2010). Before the delimitation, each text was processed with the Berkley Constituent Parser (Petrov and Klein, 2007) via the on-line service provided by the WebLicht Project (Hinrichs et al., 2010). Each parsed sentence was searched top to bottom for constituents that had seven to four word tokens as descendants. If the current constituent had more token leaves than required, the procedure was repeated with its immediate children. Additionally, post-processing filters were added to ensure that utterances never included punctuation marks.

2.2. Text Annotation Procedure

After the texts were split into utterances, they were ready to be marked for emotions with the help of desktop software we have developed for this purpose (see Figure 1).

Two untrained young adult German native speakers, a female and a male, took part in the annotation project. They were shown how to use the software and were then asked to rely on their intuition. Importantly, our annotators were asked to imagine that they were narrating the stories to a child or children. This motivation placed them into a social scenario, where they could reflect on which emotions would be appropriate for expression at any moment of the story. Otherwise, the annotators were not explicitly instructed to use any other annotating strategies (e.g., to keep the emotion annotation balanced for emotions, to use all the available categories, to leave few annotation units *neutral*). The participants were presented with the full text of a story and were free to select either one utterance or several consecutive utterances and mark the choice for an emotion category. In Figure 1 the utterance *"die Königin gebar ein Mädchen" (the queen gave birth to a daughter)* is selected as the annotator chooses an emotion category from the drop

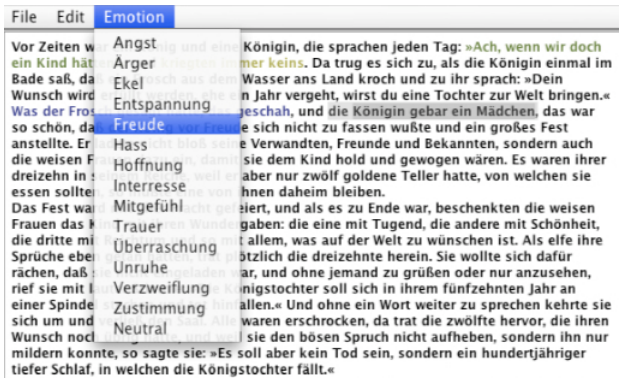


Figure 1: Screenshot of the desktop annotation software. An utterance is selected and the annotator chooses a suitable emotion category from the list. Previously marked utterances a shown in colour in the text.

down list. A few utterances that were already marked for different emotion categories and are shown in colour in the text. To avoid association of colours with emotion categories, the colours were assigned randomly to each emotion category for every annotation session.

The multiple annotation sessions were approximately each two hours long, but there was no specific time limit set and the annotators could work on each text for as long as they needed. Typically, two or three texts (several hundred utterances) were annotated in one session. The participants were encouraged to take breaks between the stories. The order in which the texts were to be annotated was randomised for each participant and the annotators worked individually on each story without consulting each other.

Additionally, we compiled a common lexicon for all the 72 fairy tale texts and asked our participants to mark each word for its inherent polarity (*positive* or *negative*) or leave it *neutral*. Namely, the annotators were asked to estimate each word's potential to change the overall emotional colouring of a short sentence or a phrase towards negative or positive emotions. The size of the lexicon, filtered for about 260 functional words like prepositions, pronouns, articles and conjunctions, contains more than 5500 lemmas. The lexicon was randomised and then split into three parts, approximately 1800 words each. Each part was given to the two annotators in randomised order and was annotated in a separate session.

2.3. Annotation Merge and Conflict Resolution

Annotations of one and the same text are bound to contain conflicts when compared to each other. A conflict or a tie in this context is a situation when an annotation unit received an equal count for two of more labels. These conflicts have to be resolved should a merged annotation be required, e.g., for a machine learning architecture that does not support multiple labels for training and testing items. Thus, it is necessary to develop appropriate strategies for conflict resolution and annotation merge, in order to create a joined annotation that takes both original annotations into account. In order to study and potentially simulate human emotional perception of text, it is important to preserve coherency of the original texts. Thus it is not acceptable to simply discard the conflict items.

In the annotations of the German corpus we observed several distinct tendencies in the conflicts. The analysis of the original annotations shows that one annotator was more sensitive to the emotions expressed in the text and has left less text *neutral* when compared to the other annotator, which caused conflicts between *neutral* and *non-neutral* categories during the annotation merge (by *non-neutral* here and henceforth we denote any other emotion category but the *neutral* category). The other main type of conflict also supports the observation that the annotators had different perception of the texts and/or different annotation strategies, as the first annotator often chose more context dependent and subtle categories where the second annotator preferred more basic emotion categories, e.g. *despair* or *anxiety* instead of *anger* and *fear*.

Assuming that the researchers are interested in a corpus rich in various emotion instances, we suggest a few measures that can be useful for conflict resolution during the annotation merge. First, if one annotator categorised the utterance as *neutral* and the other as *non-neutral*, the *non-neutral* category should be accepted for the merged annotation. Second, if one annotator stably used one category where the other annotator employed a range of categories, prefer the annotation instances that provide more variability. Finally, depending on the set of emotion categories, some categories can be substituted for their closest equivalents, if the former are barely represented in the corpus. The closest equivalents can be established by different methods, e.g. by analysing the conflicts between the original annotations and establishing which emotions frequently form conflict pairs. Naturally, the above strategies should be used with care and the annotation items surrounding the conflict item should also be taken into account. An alternative solution would be of course to hire a third annotator specifically for the tie-breaking task.

2.4. Results

The resulting Cohen's kappa (Cohen, 1960) inter-annotator agreement rate for all 72 texts between the two annotators is 0.21 ("fair") and the observed agreement is 0.48 ("moderate"). Most of the conflicts between the two original annotations were between *neutral* and a *non-neutral* emotion category. This observation is supported by the fact that if items with the *neutral – non-neutral* conflict are taken out of the IAA measurement, the observed agreement rises to 0.93. Similar conflict distribution holds for the lexicon annotation: the observed inter-annotator agreement of word lists was 0.48. In 88% of the conflict cases, the disagreement was between the *neutral* and the *non-neutral* labels, not between *positive* and *negative*.

To create the merged annotation we used the conflict resolution strategies described in the previous section which resulted in resolving most of the ties and brought the observed agreement to 0.94. The rest of the conflicts (6% of the corpus) were solved manually by the first author through analysing the surrounding context for both annotations.

Four emotion categories, namely *approval (Zustimmung), compassion (Mitgefühl), disgust (Ekel)*, and *hatred (Hass)*

Emotion category	Frequency
Anger (Ärger)	6%
Anxiety (Unruhe)	8%
Comfort (Zufriedenheit)	7%
Despair (Verzweiflung)	2%
Fear (Angst)	3%
Hope (Hoffnung)	3%
Interest (Interesse)	15%
Joy (Freude)	3%
Neutral (Neutral)	37%
Sadness (Trauer)	5%
Surprise (Überraschung)	11%

Table 2: Categories distribution in the German corpus for the merged annotation.

	Sentences with ...	
	... # categories >1	... # non-neutral cat. >1
A1	2972 (42%)	445 (15%)
A2	5108 (72%)	1618 (32%)

Table 3: Category counts across sentences for two annotators. Annotator 1 (A1) was less prone to mark several emotion categories in one sentence than annotator 2 (A2).

were excluded as the result of conflict resolution procedure, as their frequency of each category in the corpus was under 1%. These categories were substituted with *comfort (Zustimmung), sadness (Trauer), anger (Ärger)* respectively (both *disgust*, and *hatred* were substituted with *anger*). The resulting distribution of emotions in the final corpus is shown in Table 2.

The two annotations for the lexicon were merged in the following way: when a word was annotated as inherently *non-neutral* (*positive* or *negative*) by one annotator, their label was chosen over the *neutral* one. Those words that were annotated with conflicting *non-neutral* labels were neutralised. The final distribution of polarity labels in the lexicon is as follows: 2639 (48%) *neutral* items, 1380 (25%) *positive* items, and 1518 (27%) *negative* items.

The final German corpus contains 7061 sentences. Table 3 shows that a great portion of sentences was annotated using at least two categories (including the *neutral* category), and a significant portion of those were annotated for two or more *non-neutral* emotion categories. This shows that short utterances are suitable for this kind of sentiment annotation.

3. English Corpus of Andrew Lang Fairytales

The textual base for the second annotation project comes from fairy tales written down by Andrew Lang in a collection of twelve books, published between 1889 and 1910. The full collection is comprises of 437 fairy tales and is truly unique (cf. Lobo and de Matos (2010)). While Andrew Lang did not record any of the fairytales from oral sources, like Brothers Grimm did, he collected previously recorded fairy tales from various cultures and languages and translated many of them into English for the first time. The texts come from Africa, India, Japan, China, Russia and many European cultures. Such broad spectrum of cultures is valuable for researchers since, should many of the texts be annotated by multiple people of various origins, one could gain insight into both cross-cultural differences and shared properties of emotion expression and perception through simple texts. It goes without saying that some emotion nuances and cultural norms can be obscured as the result of the translation into a different language. However, it is of great advantage that all the texts were collected and edited mainly by one single person, ensuring consistent style. At the same time Andrew Lang's language is eloquent and poetic, its rich vocabulary being yet another benefit for the researchers in SA.

Over four hundred texts is a large corpus and its annotation is currently an ongoing project. At this early stage we have three stories, each annotated by seven participants. The stories ("Blue Beard", "Jack my Hedgehog" and "Twelve Brothers") vary in their origins, plots, emotionality and characters but also share some features like elements of magic in the narration and a general happy ending. In the rest of the section we describe the text delimitation and annotation procedure for this corpus since both aspects have undergone important changes in comparison to the German corpus. We also show the results of acquiring annotations remotely with the help of our new on-line annotation tool.

3.1. Text Delimitation into Annotation Units

In order to delimit full texts into utterances we use the Festival TTS (Taylor et al., 1998) for the English corpus. The Festival Speech Synthesis System is a general full text to speech synthesis system as well as an environment for development and research of speech synthesis techniques. Festival is designed to support multiple languages, and comes with support for English, Welsh, and Spanish. Voice packages exist for several other languages, but unfortunately German is not one of them. Due to this we could not use Festival TTS for text delimitation for the German corpus annotation project.

The *phrasify* module of Festival TTS predicts for each word three possible break scenarios: no break, a short phrase-level break, or a long sentence-level break. Originally, this information is used later during speech synthesis for pause generation. Although there may seem no direct connection between utterances in speech and annotation units for SA, our participants found that the texts delimited in this way were easy to work with. After all, the resulting utterances represent prosodic and semantic structure of each sentence and do not interfere with the flow of the story. The delimited texts were then submitted to our new on-line annotation system[1] described in the next section.

3.2. Annotation Process: the on-line annotation tool

In German corpus, the annotators marked the texts for emotions with the help of desktop software we had developed for the purpose. While intuitive to use, it was not ideal in several ways, the major one being that a potential annotator had to install the software on their computer if they wanted

[1] www.epetals.org

to annotate the texts remotely. We thus developed an on-line system that solved the remote annotation problem — a new user can register in the system, receive new annotation tasks and submit finished annotations without the need of installing new software. From the researcher's perspective, the new annotation tool allows management of content for annotation by adding pre-formatted files to the system. They can register new annotators, assign them new texts, send reminders and monitor their work. A log file is kept for the annotation progress making it easy to track the time spent by each user for the annotations and estimate their work routine. Finished annotations can be downloaded and used for further research needs.

Based on previous research in (non-)verbal emotion expression (Izard, 1971; Bänziger and Scherer, 2007; Neviarouskaya et al., 2010), we have developed a new set of emotion categories for English annotation (see Table 4). Although the new set overlaps with the one used in the German corpus, a few emotions are new: *amusement, excitement, pride, relief, contempt* and *shame*. During the annotation process, the categories were displayed next to each utterance in a drop-down list, sorted by polarity (Figure 2). This list greatly expands the boundaries of the often used basic emotion categories suggested by Ekman (1992): *anger, disgust, fear, joy, sadness, surprise*; and unlike the latter is balanced for positive and negative emotions.

We also introduced several new features to the annotation procedure that were absent in the desktop software used for the German corpus annotation. Whereas previously the annotator was instructed to indicate only the emotion category for each utterance, the new on-line annotation tool additionally requires to indicate the intensity of the emotion on a five-point Likert scale and the emphasis in each utterance. Moreover, there are new restrictions that aim to ensure better quality of annotation, possibly at some time and efficiency cost. The annotators of the German corpus were not asked to explicitly mark utterances that they considered to be neutral. They were also free to select several consecutive utterances and mark them for one single category in one action. Partially it was done to speed up the annotation process and avoid fatigue. Since the annotations of the German corpus were performed in the laboratory and with only two annotators, the collaboration between the annotators and the researchers was strong enough to provide environment for responsible behaviour on the part of the annotator. It is harder to ensure high quality of annotations when the participants work remotely. Thus we had to make the *neutral* category arbitrary for active assignment, along with other emotion categories.

When the user was finished, they could submit the annotated text. Their work was then automatically checked for any missing information, e.g. an utterance that did not receive an emotion label, an intensity left unassigned or a missing emphasis. The system keeps pointing out missing information until all information has been provided, after which the researcher receives a message that the annotation assignment has been completed and is ready for download.

Using the latest version of our system, we have collected a smaller size corpus of three unabridged English fairy-tale texts from the Andrew Lang books collection, each of which was analysed by seven annotators. The annotators varied in their cultural origin but had a strong command of the English language. All annotators were young adults (mean age = 25.4 y.o, sd=5.7), three of the annotators were female. Similar to the two annotators of German corpus, our annotators were asked to imagine they were telling the stories to a child or children and to mark down appropriate expression of emotions.

Positive	Negative
Amusement	Anger
Excitement	*Contempt*
Hope	Despair
Joy	Disgust
Pride	Fear
Relief	Sadness
Surprise	*Shame*
Neutral	

Table 4: Emotion categories used in the English corpus and in the new on-line annotation tool in general. The categories in italics were not used in the German corpus.

Figure 2: Screenshot of the on-line annotation tool. The text, split to utterances is presented in its original order on the left side. Each utterance is assigned an emotion category. The annotator also has to indicate the intensity of the emotion on a 5-point scale and the word that bears most emphasis (highlighted in green).

3.3. Results

At the moment the annotated part of corpus contains 1170 annotation units and each unit has received seven annotation labels. Fleiss' kappa (Fleiss, 1971) IAA rate for the seven raters is "fair" (0.25). Fifteen categories is a large set for untrained annotators, and is probably the primary reason for "fair" levels of inter-annotator agreement in this corpus. The disagreements among annotators however were mostly observed within the same polarity (e.g., *sadness'* vs. *dispaire* or *amusement* vs. *joy*). In this case

the observers agreed on the polarity of the item but disagreed on the specific labels. To demonstrate this we organised non-neutral emotion categories according to their polarity (*positve: amusement, excitement, hope, joy, pride, relief, surprise* and *negative: anger, contempt, despair, disgust, fear, sadness, shame*. Fleiss' kappa IAA performed on just three categories (*positive, negative* and *neutral*) is 0.39. The other major source of disagreement was the *neutral* category vs. any of the *non-neutral* categories — in all but one of the 1170 annotation units there was always at least one annotator who marked the unit as *neutral*.

Like in the German corpus, many sentences contained several instances of emotions which shows that during the annotation the participants naturally operated on units smaller than sentence, even though they were not prohibited to mark all utterances within one sentence with the same category. Unsurprisingly, the number of utterances in a sentence strongly correlates with the number of distinct categories assigned to the sentence ($r = 0.70$). The three annotated texts contain 326 sentences (direct speech instances were counted as separate sentences), since each sentence was annotated by seven participants, the total number of annotated sentences is 2282. In 1188 of those (52% of the corpus) at least one non-neutral emotion category is present in the sentence along with the neutral category. In 709 sentences (31% of the corpus) more than two non-neutral emotions were marked in one sentence.

The overall distribution of emotion categories in the English corpus is shown in Table 5, columns 2 and 3. In order to compile maximally objective corpus, we suggested to use multiple annotators and to build final merged annotation on the modal value of response distribution for each annotation unit. Modal value in this context is simply the emotion category that has been assigned to the current annotation unit by majority of the annotators. Should the distribution of assigned labels have no unique modal value, the annotation unit belongs to the tie case. In our English corpus, the resulting merged annotations use the whole spectrum of fifteen emotions. The proportion of *neutral* category is higher by 12.16% in the merged annotation compared to the the full response range (Table 5, columns 4 and 5). Percentage of *anger, joy* and *sadness* is also marginally higher in the merged annotation than in the original annotations.

	All responses		Modal values	
Category	Count	%	Count	%
Amusement	200	2.44	4	0.41
Anger	374	4.57	58	5.93
Contempt	245	2.99	18	1.84
Despair	424	5.18	49	5.01
Disgust	146	1.78	9	0.92
Excitement	779	9.51	58	5.93
Fear	513	6.26	52	5.32
Hope	585	7.14	68	6.95
Joy	548	6.69	70	7.16
Neutral	2848	34.77	459	46.93
Pride	332	4.05	25	2.56
Relief	172	2.10	11	1.12
Sadness	441	5.38	54	5.52
Shame	76	0.93	3	0.31
Surprise	507	6.19	40	4.09
Total	8190	100	978	100

Table 5: Distribution of emotion categories in the English annotation corpus. Second and third column show distribution for all responses; fourth and fifth column show distribution across the modal values for all annotation times when a unique modal value was present (84% of all annotation items).

4. Discussion

We presented two manual SA projects: one large corpus of 72 texts in German language and one smaller corpus of annotations in English. Both corpora are available upon request and the English corpus is a work in progress. The annotation approaches differ between the two corpora in some aspects (number of participants, annotation software, emotion categories), but share crucial points of minimum training for the annotators, a rich set of emotions and smaller than a sentence, utterance based annotation units. The results show fair to moderate agreement, typical for the complex task of sentiment annotation, and are encouraging for using modal value to obtain more objective annotation from multiple users.

As Bayerl (2011) discusses, high inter-annotator agreement rates are important but hard to achieve and the success depends on many factors, e.g. the number of annotators, their expertise level, and the complexity of the annotation scheme. Manual SA of coherent texts is subject to more variability in annotators choices, because people can perceive one and the same story in different, yet perfectly valid ways. This fact is supported by relatively low inter-annotator agreement in several other studies (Alm and Sproat, 2005; Neviarouskaya et al., 2010).

Both annotation projects shed light onto the main the principles of good practice for sentiment annotation methods, as well as the challenges and limitations. Tasks like part-of-speech analysis or semantic relations annotation require a carefully written manual and a set of annotation rules. Manual sentiment analysis on the contrary can hardly benefit from extra restrictions and instructions since this endangers the naturalness of the resulting annotation, as the user is usually asked to mark down their intuitive emotional perception of text.

Nevertheless, a few guidelines can help the user to understand the annotation process better. Besides the understandably necessary instruction in annotation tool operation, the researchers need to state clearly whether the user is supposed to mark down their emotional reaction to the text or rather the emotions directly expressed in the text. The distinction can be illustrated by the following example. In a fairy tale, an evil character can express *joy* over a good character's misfortune, which can trigger *disgust* or *disapproval* in the reader. Which emotion, *joy* or *disgust*, the annotator should mark in that particular segment depends primarily on the task. If the task is to annotate the text with emotions one would express when reading the text to an audience or acting it out, then *joy* would most possibly be a better choice since it represents the currently active char-

acter. If the task is, however, to note down emotions that a person should feel while reflecting on the story, *disgust*, *disapproval* or even *fear* would be more appropriate.

SA traditionally used various annotation schemes and label sets. We argue that a rich set of categories is necessary in order to understand human emotion perception and to build an automatic SA system capable of simulating human emotion perception of text. This gives the annotators the possibility to express their perception of text on a very fine-grained level. However, this approach is not without a challenge - the more categories we add, the more complex and time consuming the task gets for the participants. Moreover, using more annotation categories naturally results in more disagreement between annotators. Individual properties of the emotions in the label set are of crucial importance as well. Some basic categories, like *joy* and *sandess*, are likely to be used more frequently than other more subtle categories (e.g., *contempt* or *shame*). So far we have made sure that the non-neutral emotion categories used for annotation form equal polarity groups (*negative* vs. *positive*). Of course, polarity is not the only aspect across which various emotions vary. In the English corpus annotation we added intensity measure to the annotation scheme.

The distribution of the emotions in the resulting annotation corpus depends not only on the emotion category set, but also on the textual material and the annotators' emotional state and character in general (Volkova et al., 2010; Mohammad, 2011). An annotator, as we have already seen in both studies presented in this paper, can be sensitive to emotions and their shades, or can use predominantly *neutral* category. Modal approach with multiple annotators shows fewer conflicts than when only two annotators are employed. Using the modal value, the merged corpus contains only 16.4% annotation units that need conflict resolution. In the German corpus this number is as high as 52% (100% minus 48% of the observed agreement). The conflict cases in English corpus can be dealt with in the same manner as in the German corpus. Note that at this point of time we have only seven annotators for each text. It is desirable to have as many annotators as there are emotion categories so that in each annotation unit (utterance, sentence, etc.) each category has an equal chance of being chosen. We have also shown that for this task annotation units smaller than sentence are optimal. Both corpora reported in this paper show that people tend to mark a part of a sentence for a non-neutral emotion instead of selecting the full sentence and often mark several different *non-neutral* emotions in one sentence.

One major challenge is quality control of the annotations, especially when the annotators are working remotely. Although fairy tale texts are relatively simple to understand and to interpret at least on the surface level, the task is still time consuming and demands reflection and concentration from the annotator. One of the benefits of our on-line annotation tools is that the participants are free to define their own work routine, since every new annotation action is instantly saved in the system. Thus, it is not necessary to annotate a whole story in one session and it is possible to come back and change previously assigned values. However, such a setup also allows participants to have long breaks between the annotation sessions when working on one story, switch between stories if several are assigned to them and select random categories in an attempt to finish the annotation more quickly. Our system keeps a log record of every annotation action and thus makes it possible for the researcher to analyse annotator behaviour, such as time spent annotating each story, number of sessions and their distribution across time, category distribution and so forth. The system allows the researcher to send messages with reminders and requests to the annotators.

One of the major questions in the SA in general is the one of *"what should be considered the gold standard?"*. Emotion perception is inherent to human nature and thus one might get an illusion that a sentiment annotation task is easy and intuitive to perform. However, multiple studies show how much variability there is in human annotated text for emotion categories. Who can be considered as "professional" annotator is a question difficult to answer. For the specific task of sentiment annotation of fairy tales, theatre plays, screen scripts and other literature of short dynamic nature, professional actors might be a good source of good quality annotations, since they have experience at mediating emotions from text to the public. It is still questionable however if specific training for sentiment annotation of text apart from annotation tool operation, general guidelines, and motivation can be of benefit in this situation. The researcher may wish to use a very short emotional text, e.g., a piece of well-known literature with previously established emotions as acceptable annotation labels and use the text for training a new annotator as well as to test their ability to cope with the task. Apart from such training of an annotator, modal values of response distribution for each annotation item, should multiple annotators be available, can give a stable emotion description of the text approaching the definition of the *gold standard*. Alternatively, the annotation of a participant whose annotation has the highest agreement rate with the merged annotation can be considered as the *gold standard* as well.

5. Conclusion

In this paper we have presented two new corpora of texts manually annotated with a rich set of emotion categories. One corpus is a finished collection of 72 Brother Grimm fairy tales in German. Two native speakers annotated each of the texts with a set of fifteen emotion categories using predefined utterances as annotation units. The resulting agreement is "fair" and the corpus is available upon request (both the original annotations from each of the two annotations as well as the merged version resolved for conflicts). This project confirmed some aspects of the proposed scheme to be useful, e.g., small annotation units and a large set of annotation labels. However, the results also revealed several problems that we have attempted to remedy in the English corpus annotation.

The result of the second project is primarily the new on-line annotation system, open for the scientific community. The tool allows the annotators to work remotely and for the researchers to recruit annotators from all over the world. The annotations are saved online at each new annotation action and the progress is easy to track both for the researcher and

the annotator. In addition to assigning an emotion category to an annotation unit, we added the functionality of recording perceived emotion intensity and the most emotionally charged word. The annotation of the new corpus of texts is still in progress, currently only a few texts have been annotated by seven participants each. The preliminary results show fair inter-annotator agreement and moderate IAA when participants' annotations were analysed on polarity levels and not emotion categories. Most importantly however, we demonstrate that recruiting multiple annotators and then drawing modal value from response distribution for each annotation unit is a method suitable for acquiring an annotation of high quality and rich in emotions.

6. References

Alm, C. O. and Sproat, R. (2005). Emotional sequencing and development in fairy tales. In *Affective Computing and Intelligent Interaction*, pages 668–674. Springer.

Alm, C. O. (2008). *Affect in text and speech*. Ph.D. thesis, Graduate College of the University of Illinois at Urbana-Champaign.

Aman, S. and Szpakowicz, S. (2008). Using roget's thesaurus for fine-grained emotion recognition. In *IJCNLP*, pages 312–318.

Bänziger, T. and Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *Affective computing and intelligent interaction*, pages 476–487. Springer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. (2008). The language of emotion in short blog texts. In *CSCW*, volume 8, pages 299–302.

Hinrichs, E., Hinrichs, M., and Zastrow, T. (2010). Weblicht: Web-based lrt services for german. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.

Izard, C. E. (1971). *The face of emotion.* Appleton-Century-Crofts.

Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., and Hamfors, O. (2012). Usefulness of sentiment analysis. *Advances in Information Retrieval*, pages 426–435.

Lee, D., Jeong, O.-R., and Lee, S.-g. (2008). Opinion mining of customer feedback data on the web. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 230–235. ACM.

Liu, H., Lieberman, H., and Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 627–666.

Lobo, P. V. and de Matos, D. M. (2010). Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Semantically distinct verb classes involved in sentiment analysis. In *IADIS AC (1)*, pages 27–35.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2010). User study on affectim, an avatar-based instant messaging system employing rule-based affect sensing from text. *International journal of human-computer studies*, 68(7):432–450.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *HLT-NAACL*, pages 404–411.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

Propp, V. I. A. (1968). *Morphology of the Folktale*. University of Texas Press.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *the 2008 ACM symposium*, pages 1556–1560, New York, New York, USA. ACM Press.

Taylor, P., Black, A. W., and Caley, R. (1998). The architecture of the festival speech synthesis system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia*, pages 147–151. International Speech Communication Association.

Volkova, E. P., Mohler, B. J., Meurers, D., Gerdemann, D., and Bülthoff, H. H. (2010). Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106. Association for Computational Linguistics.

Von Franz, M. L. (1996). *The interpretation of fairy tales*. Boston: Shambhala.

Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM.

Correlating Document Sentiment Scores with Web-Sourced Emotional Response Polls for a More Realistic Measure of Sentiment Performance

Elizabeth Baran
Lexalytics, Inc.
320 Congress Street
Boston, MA 02210
E-mail: elizabeth.baran@lexalytics.com

Abstract

We explore a novel approach for evaluating document-level sentiment analysis that measures correlation between sentiment scores from a text analysis engine called *Salience* and emotional responses from naïve web polls. We conduct tests in both Italian and Chinese by leveraging web-sourced news data that ask readers to rate how they feel after reading an article. We correlate vote distributions for each of the polled emotions with scores produced by our engine and show that these correlations mimic our intuitions about which emotions are positive, negative, or ambiguous. For Italian, correlation steadily increases in magnitude from -.17 to -.32 for negative emotions and .17 to .32 for positive emotions as more soundly-annotated phrases are added to the sentiment phrase dictionary that underpins *Salience*. For Chinese we saw positive correlations between our engine and emotions that were intuitively positive and negative correlations with those that were intuitively negative. The emotions *shocked, bored*, *funny*, and *moved* did not show clear positive or negative correlations but provided interesting insights into potential data for sentiment analysis that extend beyond polarity identification.

Keywords: sentiment evaluation, web-sourced corpora, multilingual corpora

1. Introduction

A sentiment-tagged corpus is an essential part of creating and testing a system that can detect sentiment in text. However, generating corpora of this type has often been expensive and time consuming (Hsueh, Melville, & Sindhwani, 2009). One of the most prominent corpora of this sort, the Multi-Perspective Question Answering corpus (MPQA) (Yu & Hatzivassiloglou, 2003) is testament to this. Other more efficient methods of generating sentiment-tagged corpora have included the extraction of content from reviews (Pang, Lee, & Vaithyanathan, 2002) and from user-generated free text (Asmi & Ishaya, 2012; Boiy & Moens, 2008) as well as crowdsourcing to anonymous non-expert annotators (Sheng, Provost, & Ipeirotis, 2008).

Sentiment analysis is traditionally based on one of two techniques: machine learning classification (Pang, Lee, & Vaithyanathan, 2002) or phrase-based dictionary look-up (Dragut, Yu, Sistla, & Meng, 2010). During evaluation, both of these methods for the most part hinge on the assumption that scoring sentiment is a binary task in which something is either tagged correctly or incorrectly. For sentiment at the document level, this often means tagging a positive document as positive and a negative document as negative (Abbasi, Hsinchun, & Salem, 2008).

Even after achieving high levels of inter-annotator agreement, an annotated corpus inevitably becomes tailored to a particular world view that can feel too constrained when dealing with language that expresses human emotion (Liu, 2010). If our goal is to consider how well our system will fare in the real world, we are forced to consider the inherent subjectivity of human sentiment and to what extent we can constrain the evaluation task and still produce realistic measures of performance. For instance, do we consider something negative that makes us feel negative emotions? Is text that refers to something negative, even though conveyed factually, considered negative or neutral (Balahur et. al, 2013)? Some have already begun to question current evaluation methodologies and have proposed other possible evaluation techniques that take into account emotional response (Devitt & Ahmad, 2008).

We explore a method for evaluating document-level sentiment polarity that uses the results of naïve emotional response polls found on two different news sites in two different languages, Italian and Mandarin Chinese. We use Lexalytics' sentiment analysis engine called *Salience*[1], which uses a sentiment phrase dictionary to analyse document-level sentiment. We look at how the engine's document-level sentiment scores correlate with the sentiment polled data. Finally, we show how these correlations change with additions or modifications to our sentiment phrase dictionaries.

The rest of the paper is organized as follows. In Section 2 we outline our proposed method for evaluating document-level sentiment. In Section 3 we discuss our experiment setup for both Italian and Chinese. In Section 4 we give examples of articles from our data set. In Section 5 we describe our phrase annotation process. In Section 6 we show the results of our experiments with our new evaluation methodology. Lastly, we conclude in Section 7 and discuss future work in Section 8.

2. Our Method

Our method for evaluating document-level sentiment observes correlation trends between sentiment scores

[1] http://lexalytics.com/technical-info/sentiment-analysis-measuring-emotional-tone

from our sentiment analysis engine and emotional response data from a naïve web poll.

We perform our experiments on data in Italian and Chinese. For both of these languages, we were able to find news sources that polled their readers on how they felt after reading an article. The articles and votes or vote distributions are publicly available and can be used as a crowd-sourced data set for sentiment evaluation. Therefore there is minimal effort and cost required in creating a sentiment test corpus. We use the *Salience* sentiment analysis engine[2] loaded with the Italian and Chinese data directories to help iteratively build sentiment phrase dictionaries for each respective language, and then to evaluate their accuracy for document polarity detection.

Salience scores document sentiment on a scale from -1.0 to 1.0, where -1.0 is negative and 1.0 is positive. We attempt to show that the addition or modification of sentiment phrases to the phrase dictionaries powering Salience can improve correlations between positive and negative sentiment scores and positive and negative emotion types, respectively, from our data set.

3. Experiment Setup

We extracted articles from the naïve web sources discussed in the previous section, and were therefore able to compile a sentiment annotated corpus that was free from task-related bias. This was our sentiment test corpus. To build are sentiment phrase dictionaries for each language, we used two different approaches that we will describe in the following subsections.

3.1 Italian Setup

We gathered 2,543 articles between November 2013 and March 2014, from *Corriere*[3], a popular Italian news site. *Corriere* asks readers to rate how they feel after reading an article, and gives them the options of *Indignato* (angry), *Triste* (sad), *Preoccupato* (worried), *Divertito* (Amused/Entertained), or *Soddisfatto* (satisfied). We extract articles and their emotional response vote distributions to serve as our document-level sentiment corpus in Italian.

3.2 Chinese Setup

For Chinese we gathered 770 articles from the site *Huanqiu*[4] between November 2013 and March 2014, which, like Italian, polls readers on how they feel after reading an article. However, *Huanqiu* provides a a greater number and variety of emotions, namely 震惊 (shocked), 愤怒 (angry), 悲伤 (sad), 感动 (moved), 喜悦 (joyful), 幸福 (happy), 无聊 (bored), and 可笑 (funny). While our Italian source provides the distribution of votes, this Chinese source provides the actual number of votes per emotion, which we then convert into percentages. We incorporate only articles that have a minimum of 10 votes into our corpus.

[2] Lexalytics Salience Version 5.1.1.7443

[3] www.corriere.it

[4] www.huanqiu.com

4. Cross-linguistic Comparison of Emotional Response Data

Before we began our experiments, we needed to gain a deeper understanding of what emotional responses actually mean in each language and how they might correspond to *positive* and *negative* sentiment polarity metrics. We sought to understand whether similar events could evoke the same emotional response in participants of different languages and culture groups. In other words, is sentiment universal and to what extent can we leverage sentiment data from one language to another? Answering this question far exceeds the scope of this paper, but we believe this data set can at least work as a starting point for this inquiry.

The emotion types in our Italian data set differed significantly with the types in our Chinese news source. *Corriere* polls readers on five emotions, all of which we can reasonably categorize as either *positive* or *negative*. *Huanqiu* provides a broader and more diverse option set of emotions for readers to choose from, some of which cannot be intuitively placed into a category of *positive* or *negative*.

In order to better understand each emotion, we take a sample of articles from our data set, where each article represents a majority distribution of votes for a particular emotional response. We make parallels between emotion types across the two languages where possible, but otherwise list them separately. We choose articles that exhibit greater than a 90% distributional vote for the emotion of focus, unless none exist. In that case we lower this threshold to 60%. We look at the titles of the example articles supplied for each article to observe how emotional responses may differ cross-culturally. By analyzing the data at this level of granularity, we hope to enumerate the variables present in sentiment analysis from a cross-linguistic perspective.

4.1 Negative Sentiment

4.1.1. Indignato/愤怒 (Angry)

The emotions *indignato* and 愤怒 both translate to "angry" in English. The following Italian articles evoked anger in Italian readers:

> Mercoledì sciopero dei mezzi pubblici Sospesa l'Area C, resta attiva la Ztl (Wednesday, public strikes, Area C suspended, Restricted Traffic Zone remains active)
>
> Raucedine, ecco quali sono le cause (Hoarseness, here are the causes)
>
> Utero in affitto, coppia di Iseo condannata ma le altre no (Surrogacy, couple in Iseo condemned but the others not)

Titles of articles that evoked anger for Chinese readers included the following.

> 马民航局：对于航班监测结果军方可能有所隐瞒 (Malaysian Civil Aviation Authority: With regards to the results of the flight monitoring, the military may have concealed things)

中石油再陷质量门 对柴油掺水超标 40 倍仍未回应 (China Petroleum has again fallen into a quality scandal, with regards to exceeding the standard of water mixed with oil by a factor of 40, they have yet to respond.)

新疆人民痛恨所谓民族精英 美却赞暴徒是斗士 (Xinjiang citizens hate the so-called "ethnic elites"; the US however praises rioters as "[brave] fighters")

When we compare the content that evoked anger in Italian versus Chinese we notice several salient points. Firstly, the source of this anger appears to be universal even if it is not relatable. Strikes that cause traffic delays, sickness, lawful injustices, lack of government response during times of catastrophe, being cheated, political agendas that go against our own – all of these situations are scenarios that might cause a human to feel anger and are themes of in our example content. From this perspective, the particular content of the event may not matter as much as the words used to describe the event. Words like "hate", "conceal", "condemned", all have negative implications and can be used in a variety of contexts. In fact, these are often the typical words found in sentiment dictionaries.

The second observation we can make is that although we are able to empathize with someone from a very different culture and situation than our own, there is still a considerable amount of background knowledge and understanding required to do that. For example, it is likely that a common emotional response to the Malaysian Airlines situation in the West might be less anger and more sadness, or even neutrality, merely because the event has less of an effect on the people and families in the West. Contrast this with the fact that the majority of the passengers on the Malaysian Airlines flight were Chinese Nationals; the Chinese response is understandably more elevated.

Similarly, Americans may also be angry after reading the particular article on Xinjiang but probably more so because they disagree with the tone and bias towards Chinese sentiments. The notion of 新疆独立 (Xinjiang Independence) has very different connotations for Chinese readers than it does for American readers.

We will explore this idea of culturally relevant sentiment content further as we look at more articles in different emotional categories.

4.1.2. Triste/悲伤 (Sad)

Both our Italian and Chinese data sets have an emotion type that equates to "sad" in English. The following articles evoked a sad response.

Nonna investita da un Porsche (Grandmother hit by a Porsche)

Ultraleggero precipita sul Monte Conero, morto il pilota (Ultralight aircraft crashes on Mount Conero, pilot is dead)

"玉兔"号月夜后未被唤醒 NASA 官微悼念 (Jade Rabbit could not be awoken after lunar night; NASA's official Weibo mourns)

香港 TVB 荣誉主席邵逸夫今晨逝世 享年 107 岁 (Hong Kong's honorary TVB president Run Run Shaw passed away this morning at the age of 107.)

Feelings of sadness seem to be strongly associated with extreme loss or death in both Italian and Chinese. Still cultural bias is apparent as the metaphorical death of a Chinese moon rover evokes similar feelings of sadness amongst the Chinese.

4.1.3. Preoccupato (Worried)

The word *preoccupato* or "worried", in English, existed in our Italian data but not in Chinese. Examples of articles with high concentrations of "worried" votes are referenced in the titles below.

La Grecia trema, terremoto di magnitudo 6 Scossa avvertita anche nel Sud Italia (Greece trembles, earthquake of magnitude 6, shocks also felt in Southern Italy)

Sprechi alimentari: ancora troppo il cibo buttato via dagli italiani (Food waste: still too much food thrown out by the Italians)

From these articles, we see that "worry" seems to indicate fear and uncertainty. Both of these articles indicate typical emotional responses given the event stimuli, but specific details are more relevant to Italian culture. Other cultures may feel more neutral with regards to both of these topics. In the first case for example, sentiment may be largely dependent on the reader's physical location in relation to site of the earthquake.

4.2. Positive Sentiment

4.2.1. 喜悦 (Joyful)

In Chinese, 喜悦 can mean "excitement" or "joyfulness". There is no direct equivalent in the Italian data set.

玉兔活着就有希望！外媒过早宣布死讯忙改口 (There is hope that the Jade Rabbit is still alive! Foreign media prematurely announced loss of communication and is now correcting their previous statements)

日华媒：属马的安倍晋三势必将在马年下马 (Nikka Media: Shinzo Abe, born in the year of the horse, is bound to "dismount from the horse" in the year of the Horse [2014].)

What the Chinese consider to be "joyful" has strong cultural significance in the articles above. Jade Rabbit, the Chinese lunar rover, is a symbol of a technological achievement that establishes China as a viable international competitor in space exploration. It is a source of pride for the Chinese but might not evoke the same response in readers from other countries.

The second article is infused with cultural bias. There are very strong negative sentiments towards the leader of Japan in China currently, and the act of him leaving office is seen as a positive event for the Chinese. Furthermore, framing this within the context of the Chinese zodiac, which is a ubiquitous component of Chinese culture, serves to fuel Chinese nationalistic sentiments and in this case, against Abe. Sentiment towards a political leader

could differ dramatically across cultures.

4.2.2. Divertito (Amused/Entertained)

The following are example titles of "amusing" articles.

> Winter Marathon, si parte Sfida fra piloti sulle Dolomiti (Winter Marathon, let's get started; a challenge between drivers in the Dolomites)
>
> L'evoluzione dei centri commerciali: da energivori a virtuosi del green (The evolution of shopping malls: from energy-gorgers to green virtuosos)

As we notice from these articles, a source of amusement in Italian can be the announcement of a fun event, such as the Winter Marathon or improvements to existing infrastructures. In this case, we see that Italy is a culture that values energy efficiency.

4.2.3. Soddisfatto (Satisfied)

Satisfaction might entail happiness, but it is its own emotion. Therefore we have put it in its own category separate from the Chinese notions of "joyful" and "happy".

> Social street, la carica delle donne intraprendenti Così su Fb il vicino di casa diventa una risorsa (Social Street, the office of entrepreneurial women. Thus, on Facebook, the neighbor becomes a resource)
>
> «Abbado mi ha suggerito Pereira per la Scala» (Abbado suggested Pereira to me for the Scala).
>
> I biglietti metro e bus valgono 15 minuti in più (Metro and bus tickets are valid 15 minutes longer)

The first example is particularly revealing about cultural values. The concept of an "entrepreneurial woman" adheres to the Western mentality of female empowerment – a hot topic in our modern age. This value is not shared cross-culturally and is a strong case for why a noun phrase such as "entrepreneurial woman" does evoke sentiment. Furthermore, it evokes positive sentiment in a culture that shows strong media-based trends to empower women. In other cultures, the opposite may be true.

The second example alludes to the naming of a new official to head the famous opera house in Milan. The implication here is that Pereira is recognized as a good choice. In this case, we might say that pre-disposed sentiments towards specific people can have an effect on the overall sentiment of an article.

In the last article, getting more for your money and increased convenience are both sources of satisfaction.

4.2.4. 幸福 (Happy)

In English, 幸福 is translated as "happiness", but it also carries the nuances of "good fortune" and "fulfilment". It represents a deeper sort of happiness than 喜悦 which we translated previously as "joyful". An example of a "happy" article in Chinese is the following.

> 香港渔民疑捞获稀世巨型沉香木 或价值过亿 (Hong Kong fisherman finds rare giant Agarwood – valued at over 100 million [HKD])

Finding a rare species evokes a more profound sense of happiness. We might argue that scientific discoveries or innovations are recognized on a more global scale since they carry significance to humans as a species.

Interestingly, of the set of articles that we evaluated for Chinese, only one showed a majority vote in favour of this emotion and only at 77.4%. In contrast, there were 130 articles that demonstrated greater than a 90% distribution of votes for 喜悦 (joyful). This could be a demonstration of the depth of 幸福 as an emotion. It is more difficult to feel 幸福 as opposed to the transient feelings of excitement or joy implied by 喜悦.

These two emotions share a semantic overlap with each other and, to some extent, force the reader to make a choice between the two. It could be the case that the readers favour 喜悦 when offered the choice but might be content to vote 幸福 when not.

4.3 Other Sentiments

The polarities of the following emotions were ambiguous and were therefore listed separately from the *positive* and *negative* categories.

4.3.1. 可笑 (Funny)

可笑 is composed of the characters meaning "can" (as a modal) and "to laugh". In contrast, the verb "divertire" in Italian, means "to amuse" or "to entertain". The word "divertito" is just the past participle of this verb.

There is semantic overlap between these two words, however the Chinese word places explicit reference on the act of laughing. The connotation is therefore that we find something funny or ridiculous.

Articles that were considered funny in Chinese were.

> 全国人大二次会议第三次全会听取和审议两高报告 (The Second Session of the Third Plenary Meeting of the National's People's Congress listen and examine two high-level reports)
>
> 英学者称秦始皇兵马俑创作灵感源于古希腊雕塑 (English scholars say the creation of the Terracotta Warriors was inspired by ancient Greek sculptures.)

These articles prove why "funny" can be such a difficult emotion to classify as positive or negative. The first article is merely a summary of a recent government meeting. The article is fairly neutral but the majority of the people who voted found the article to be humorous, making it clear that there is a cultural knowledge here that is not at all present in the words themselves. Furthermore, the perspective of the people who find this content to be humorous contrasts sharply with the perspective of the government that clearly intends for it to be taken seriously.

The second article exemplifies a strong nationalistic sentiment even though the emotional response is quite ironic. The implication here is that from a Chinese perspective, the proposal of one of Chinese history's most impressive feats as stemming from another culture's inspirational work is just preposterous and therefore funny. The Greeks on the other hand might find this

article more satisfying.
"Funny" articles, as evidenced by the examples above, could be a doorway into a better understanding and identification of sarcasm and irony in text.

4.3.2. 无聊 (Bored)

Articles rated "boring" by Chinese readers seemed to be culturally insignificant, strange, or a critique of some event.

> 英男子网上拍卖"又老又懒"女友 获 50 人竞价 (English man auctions off "old and lazy" girlfriend online and receives 50 bids)

> 央视春晚节目单出炉：语言类仅 5 个 成老歌演唱会 (CCTV Spring Festival Program List released: only 5 speaking type [acts], [event] becomes oldies music concert)

The first article is a strange story about an Englishman which bears no global significance. The second article discusses the official list of programs for the yearly Spring Festival show that is broadcast live. The article points out that there are few speaking-type acts and that the majority of the program is dedicated to the performance of classical songs. The impression here is that readers would prefer more speaking or scripted acts.

4.3.3. 震惊 – Shocked

The following were considered "shocking" articles.

> 尼日利亚一餐馆售卖人肉 菜单上有"烤人头"(A Nigerian restaurant sells human flesh – the menu has "roasted human head")

> 澳两岁男童成滑板高手 穿纸尿裤滑滑板视频爆红 (Two-year old Australian becomes a skateboarding master – Video of him wearing diapers and skateboarding trends)

4.3.4. 感动 - Moved

Finally, "moving" articles tend to emphasize triumph or goodwill during times of hardship. These situations are often spotted with both negative and positive elements making it difficult to place them in a definitive category.

> 玉兔探月日记：月球之旅，没有遗憾(86%) (Jade Rabbit's Lunar Exploration Diary: [my] journey on the moon, no regrets)

> 夫妻 18 载护林情痴大山 愿倾余生栽种"桃源"(A couple spends 18 years protecting the forest, loving the mountains –willing to spend the rest of their lives planting in "paradise")

5. Sentiment Phrase Annotation

We utilize Lexalytics' POS-taggers for Italian and Chinese, as well as the possible-sentiment-phrase method to retrieve phrases that match certain POS patterns for both Italian and Chinese.
For Italian we built a dictionary from scratch. We gathered 91,474 possible sentiment phrases using the above method call and distributed them to five native Italian speakers. The annotators were provided with guidelines instructing them on how to annotate each phrase. They could choose from "very negative", "negative", "has negative undertones", "neutral", "has positive undertones", "positive", "very positive", or "odd grammar/not meaningful". These labels corresponded to phrase scores of -0.9, -0.6, -0.3, 0.0, 0.3, 0.6, 0.9, and "null" respectively. This annotation procedure in total produced 28,314 polar phrases.
In order to increase phrase coverage, we took our initial set of annotated polar phrases and extracted single words that appeared consistently under the same polarity regardless of context. This allowed us to extract general and very polar sentiment words that might otherwise have been missed. For example, if our dictionary contained "very happy", "happy person", and "very sad person", which are positive, positive, and negative respectively, we should be able to deduce that "happy" is most likely positive. After extracting these single words, we again gave them to annotators to weed out any incorrect polarity tags. These final words were added to the phrase dictionary. In the end, our Italian phrase dictionary consisted of a total of 32,470 *positive* and *negative* phrases.
The procedure for Chinese differed since *Salience* already shipped with a sentiment phrase dictionary. Instead, we had two annotators annotate these existing phrases to help refine past sentiment scores. We extracted another 5,151 phrases for annotation using the possible-sentiment-phrase method on data from the news domain. We gave these new phrases to three annotators to score according to the same 7-tier sentiment model that was used for Italian. After annotation, 2,007 new polar phrases were added to the dictionary.

6. Sentiment Tests

For both Italian and Chinese, we observed the relationship between predicted improvements to sentiment dictionaries and the correlation of the Salience document-level sentiment score with emotional response vote distributions per article. We used the following formula to measure correlation:

$$Correlation(X,Y) = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}}$$

Here, X is the set of vote proportions, where $0 \le x \le 1$, across all articles for a single emotion. Y is the corresponding set of document-level sentiment scores from *Salience* for each article, where $-1 \le y \le 1$.

6.2 Italian

We tested Italian in seven rounds, where each round marked an increase in size of the sentiment dictionary. Phrases were added in equal amounts and at random to the sentiment dictionary at each round. We performed document-level sentiment analysis using the *Salience Engine* and each of these dictionaries to obtain a sentiment score for each article at each round. The correlation between Salience scores and the percentage of

votes per emotion was calculated. Following our discussions in Section 4, we also measured the sum of the distribution of votes for the positive emotions of *amused* and *satisfied* for a separate *positive* correlation, and the negative emotions of *angry*, *sad*, and *worried* for a *negative* correlation. This mimics the polarity annotation framework that we typically see for document-level sentiment. The correlation between Salience and the polled emotional response distributions for Italian are shown in Table 1.

Iter	HSD Size	Emotion Types					Positive Emotions	Negative Emotions
		Angry	Sad	Worried	Amused	Satisfied		
1	4045	-0.088	-0.126	-0.040	0.068	0.137	0.177	-0.177
2	8090	-0.113	-0.148	-0.072	0.077	0.186	0.229	-0.230
3	12135	-0.129	-0.169	-0.081	0.079	0.217	0.261	-0.262
4	16180	-0.136	-0.186	-0.090	0.102	0.223	0.282	-0.282
5	20225	-0.135	-0.195	-0.072	0.089	0.227	0.277	-0.278
6	24269	-0.146	-0.191	-0.084	0.085	0.243	0.290	-0.291
7	28314	-0.150	-0.194	-0.096	0.088	0.254	0.303	-0.303
8	32470	-0.148	-0.230	-0.098	0.073	0.286	0.325	-0.325

Table 1: Correlation measures between Salience document sentiment scores and emotional response types as the size of the hand-scored sentiment phrase dictionary increases.

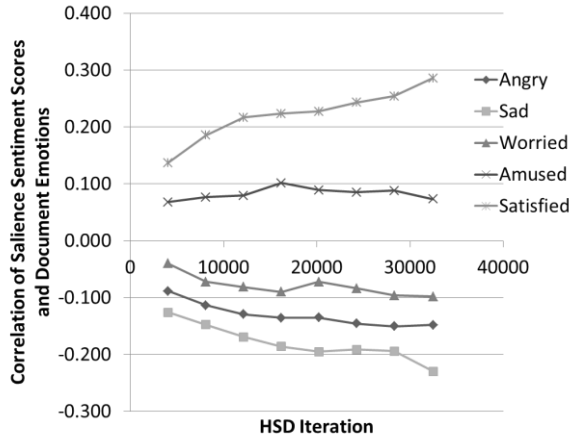


Figure 1: Correlation of each emotional response type to the Salience sentiment scores as the size of the sentiment dictionary increases.

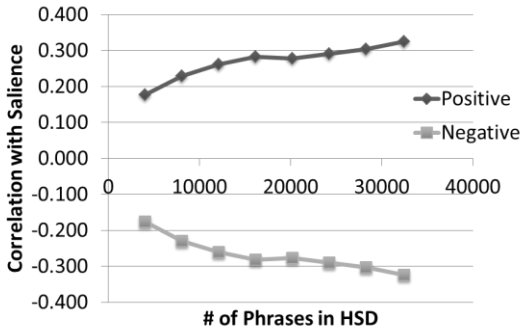


Figure 2: Correlation of positive and negative groupings of polled emotions to the salience sentiment score as the size of the sentiment dictionary increases.

In Figure 1, we see the clear distinction between positive and negative emotions as evidenced by the positive and negative correlations, respectively. All emotions except for "amused" show clear increases in magnitude in their predicted directions.

When the emotions are grouped according to *positive* and *negative* polarity as in Figure 2, the increase in magnitude is much more predictable. Going from a phrase dictionary that is less than 5,000 phrases to one that is a little less than 35,000 phrases we see the correlation with Salience increase in magnitude by 83%, from .177 to .325 for *positive* sentiments, and -.177 to -.325 for *negative* sentiments.

6.3 Chinese

We evaluated Chinese in three rounds. First, we tested the original dictionary. Second, we did a round of pruning to clean the current dictionary and improve phrase scoring. And in the third round we added some more phrases to the dictionary. From round to round, the size of the sentiment dictionary did not undergo significant changes in size as the Italian dictionary had. The correlations are illustrated as bars in Figure 3.

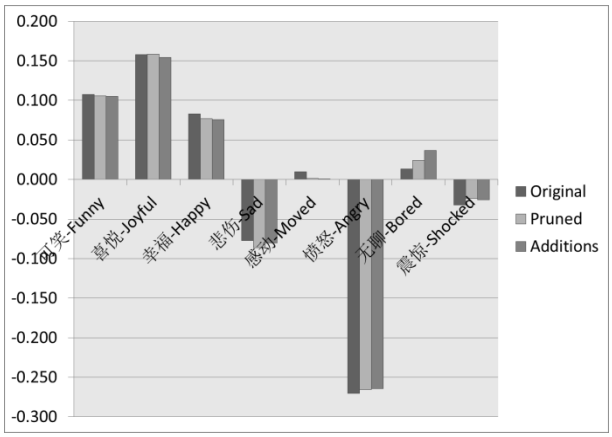


Figure 3: Correlations between Salience and each emotion type for Chinese.

With only a small addition of phrases to the sentiment dictionary for Chinese, the *change* in correlation by round is less informative than its Italian counterpart. Furthermore, it was difficult to categorize the emotions that existed for polling in Chinese into absolute groups of *positive* and *negative* as we showed in 4.2. Given these factors, we did not attempt to group the Chinese emotions into *positive* and *negative* groups as we did in Italian.

In Figure 3, we see that more intuitively positive emotions, i.e. *funny*, *joyful*, and *happy*, show clear positive correlations with document sentiment scores from our engine. Conversely, more intuitively negative emotions, i.e. *sad*, *angry*, showed negative correlations with document sentiment scores from our engine. *Bored*, *shocked*, and *moved* all showed very low correlations. It is not clear whether these emotions are evoked by positive or negative sentiments and therefore these were also the emotions that posed the biggest challenge when attempting to categorize the individual emotions into polar groups.

The scores for our system reflect the ambiguity or lack thereof with respect to each of these emotions.

7. Conclusion

We demonstrate a novel way to measure the performance of systems that identify document-level sentiment by measuring the correlation between *Salience* document-level sentiment scores and third-party naïve sentiment data extracted from web polls. We argue that since sentiment is inherently subjective and opinions vary across individuals and cultures, measuring sentiment should reflect this irregularity, instead of adhering to absolute, binary measures of right and wrong. We measure correlation against size increases and modifications to the phrase dictionary that underlies our system.

For Italian, we show a positive correlation between Salience scores and the polled emotions that were *positive* and a negative correlation between Salience scores and the polled emotions that were *negative*. Furthermore, with the addition of phrases to our sentiment dictionary, this correlation increases in magnitude.

In Chinese, we show that emotions identified as having polarity follow our intuitions about positive and negative emotions, and mimic this polarity through corresponding directional correlations. However, we see that emotions such as *bored*, *shocked*, and *moved* show minimal correlation with our engines document sentiment scores and may not be suitable for a polarity task.

By looking at sources in both Italian and Chinese and showing predictable polarity correlations for both, we show that this method of evaluation works cross-linguistically. For Italian, we saw significant improvements to correlation scores with the addition of sentiment phrases to the underlying sentiment dictionary.

8. Future Work

We engage in a cross-cultural and cross-linguistic comparison of web content that has been annotated with emotional responses from naïve readers. Although the emotional responses show adherences to foundational and universal principles of human emotion, the type or polarity of the emotion can differ immensely with regards to certain contextual factors. In our data, we have seen these factors to be current geopolitical situations, cultural relevance to the reader, value systems, and even geospatial proximity.

If we continue to pursue sentiment in terms of negative and positive polarity, more analysis should be carried out to discover emotions that can act as polarity beacons. For example, we show evidence that "angry" correlates with negative sentiment, and "happy" correlates with positive sentiment. If we gather documents that have higher vote distributions for these two emotions, we can construct a polar data set quickly and efficiently.

We can also use this type of emotional response data to extend the capability of our sentiment analysis engines past polarity to more fine-grained levels of sentiment. We found articles that were "funny", "shocking", or "boring" difficult to categorize in terms of polarity, but they could prove interesting for other sorts of sentiment tasks such as the detection of irony (Reyes & Rosso, 2012).

Finally, after analyzing emotional responses from readers in two very different languages, it is difficult to ignore the amount of cultural bias that exists in the data and even more difficult to factor it out of the sentiment equation completely. The question of perspective comes to the forefront when we brainstorm ways to improve current sentiment analysis techniques. For example, if we are building sentiment analysis for Chinese in mainland China, we may want to consider incorporating more phrases or features that embody the culture, even if they are biased in relation to other cultures. Building sentiment analysis engines using knowledge and data from the language and culture for which it is built may be crucial in unveiling truly accurate and complete understandings of sentiment.

9. References

Abbasi, A., Hsinchun, C., & Salem, A. (2008). Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Trans. Inf. Syst.*, 1-34.

Asmi, A., & Ishaya, T. (2012). A Framework for Automated Corpus Generation for Semantic Sentiment Analysis. *World Congress on Engineering Vol I.* London, U.K.: WCE 2012.

Balahur, A., Steinberger, R., Kabadjov, M. A., Zavarella, V., Van der Goot, E., Halkia, M., et al. (2013). Sentiment Analysis in the News. *CoRR*.

Boiy, E., & Moens, M.-F. (2008). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*.

Brew, A., Greene, D., & Cunnigham, P. (2010). Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. *Procs of PAIS*, (pp. 1-11).

Das, A., & Bandyopadhyay, S. (2011). Dr Sentiment Knows Everything. *ACL-HLT 2011 System Demonstrations* (pp. 50-55). Portland, Oregon, USA: Association for Computational Linguistics.

Devitt, A., & Ahmad, K. (2008). Sentiment Analysis and the Use of Extrinsic datasets in Evaluation. *Sixth International conference on Language Resources and Evaluation* (pp. 1063-1066). Marrakech, Morocco: European Language Resources Association (ELRA).

Dragut, E., Yu, C., Sistla, P., & Meng, W. (2010). Construction of a Sentimental Word Dictionary. *CIKM '10.* Toronto, Ontario, Canada: ACM.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing: a Study of Annotation Selection Criteria. *NACCL HLT Worksop on Active Learning for Natural Language Processing*, (pp. 27-35). Boulder, Colorado.

Koncz, P., & Paralic, J. (2013). Active Learning Enhanced Document Annotation for Sentiment Analysis. *CD-ARES 2013, LNCS 8127* (pp. 345-353). International Federation for Information Processing 2013.

Liu, B. (2010). Sentiment Analysis: A Multi-Faceted Problem. *IEEE Intelligent Systems.*

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity. *ACL*, (pp. 271-278).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: snetiment classification using machine learning techniques. *Empirical Methods in Natural Language Processing* (pp. 79-86). Philadelphia, PA, USA: Association for Computational Linguistics.

Reyes, A., & Rosso, P. (2012). Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. *Journal on Decision Support Systems*, 754-760.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. *Fourteenth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. Las Vegas, Nevada, USA.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating fracts from opinions and identifying the polarity of opinion sentences. *Empirical Methods in Natural Language Processing (EMNLP 2003)* (pp. 129-136). Sapporo, Japan: Association for Computional Linguistics.

Modelling user's attitudinal reactions to the agent utterances: focus on the verbal content

Caroline Langlet, Chloé Clavel
Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI
CNRS LTCI Télécom ParisTech 46 rue Barrault F-75634 Paris Cedex 13
caroline.langlet@telecom-paristech.fr, chloe.clavel@telecom-paristech.fr

Abstract

With the view to develop a module for the detection of the user's expressions of *attitude* in a human-agent interaction, the present paper proposes to go beyond the classical positive *vs.* negative distinction used in sentiment analysis and provides a model of user's *attitudes* in verbal content, as defined in (Martin and White, 2005). The model considers the interaction context by modelling the link between the user's attitude to the previous agent's utterance. The model is here confronted with the SEMAINE corpus. We provide firstly an overall analysis of the annotation results in terms of labelled user and agent's schemas and, secondly, an in-depth analysis of the relation between the agent's schemas and the user's schemas. The analysis of these annotations shows that user's attitudes and their previous agent's utterances have properties in common. Most of the user's attitudes linked to agent's utterance expressing an attitude have the same polarity. Moreover, a quarter of the targets appraised by the user refer to a target previously appraised by the agent.

Keywords: Sentiment analysis, virtual agent, human-machine interaction

1. Introduction

One of the key scientific challenges of the research field of embodied conversational agents (ECA) is to improve the interaction with human users by giving to the agent the capability of integrating user's sentiments and opinions. Most of the proposed solutions takes into account the acoustic features (Schuller et al., 2011) or facial expressions and the verbal content of the user is more and more integrated, but partially exploited, given the recent advances in sentiment analysis.

The research field of sentiment analysis and opinion mining proposes a bank of methods dedicated to detect opinions and sentiments in written texts such as the ones provided by social networks (Pang and Lee, 2008). These methods differ by their applicative goals, the theoretical frameworks to which they refer and the terminology used (affects, sentiments, feelings, opinions, evaluations). While some of them were designed to classify texts and only focus on the valence (positive *vs.* negative) of sentiments, other methods, such as (Neviarouskaya et al., 2010), aim to go beyond and propose a fine-grained analysis of these phenomena. These methods rely on more complex frameworks such as the Martin and White's model (Martin and White, 2005) – described in Section 2 –, which provides a classification of attitudes as they are expressed in English.

The development of a module for the detection of user's sentiment in a human-agent interaction requires to tackle various scientific issues (Clavel et al., 2013): the use of a relevant theoretical framework – as in the opinion mining approaches –, the integration of interaction context, the integration of the multimodal context for face-to-face interactions and the processing time of sentiment detection. The present paper proposes to tackle two of these issues by providing a model of user's *attitudes* in verbal content grounded on the Martin and White's model (Martin and White, 2005) and dealing with the interaction context (Section 3). The user's attitudes are confronted with the agent's speech by linking it to the previous agent's utterance. With this model, we aim to figure out whether the agent's speech can trigger or constrain an expression of attitude, its target or its polarity and to obtain first information about syntactic and semantic features of *attitude* expressions.

The final aim of this work is to design a module for the detection of user's *attitude* which will use the information given by the system about agent's speech and will be grounded on linguistic rules considering semantic and syntactic clues.

Such a model is also especially suited for further studies on alignment in interactions (Campano et al., 2014), where the user's alignment to the agent at the attitudinal level can be investigated as a cue of engagement.

The model is here confronted with the SEMAINE corpus (McKeown et al., 2011) (Section 4) which has been manually labelled according to the annotation schema derived from this model. The obtained annotations are thus analysed in the same section.

2. Background: theoretical frameworks of sentiment modelling

The major part of detection systems refer to the psychological dimensional model from Osgood (Osgood et al., 1975) by focusing on the valence axis. Other approaches, as (Wiebe et al., 2005) or (Breck et al., 2007), refer to the Private State Theory, which defines mental states as involving opinions, beliefs, judgements, appraisals and affects. Our model is grounded on Martin and White (2005), which has already proven its worth in (Neviarouskaya et al., 2010), (Bloom et al., 2007) and (Whitelaw C. and Argamon, 2005). It provides a complex framework, for describing how *attitudes* are expressed in English, and go beyond the classical positive *vs.* negative distinction. This model involves three sub-systems :

- the sub-system of *Attitude* refers to the emotional reactions and the evaluations of behaviors or things. Three

kinds of *attitude* are defined : the affects, which are concerned with emotional reactions, the judgements, which relate to evaluations toward people's behaviours according to normative principles, and the appreciations, which deal with evaluations toward semiotic and natural phenomena. The authors specify that an *attitude* has a source, the person evaluating or experiencing, and a target, the entity which is evaluated or triggered an affect.

- the sub-system of *Engagement* concerns the inter-subjective dimension and how the speaker deals with the potential other positions on the topic. Example

- the sub-system of *Graduation* describes how the degree of an evaluation can be adjusted.

In order to simplify our annotation model, we gather appreciations and judgements into the same main category of "evaluation". They share several linguistic properties and same patterns can express both of them (Bednarek, 2009).

3. Annotation model of user's attitudinal expressions in interaction

The proposed annotation model aims to identify the particularities of the user's attitudinal expressions in interaction. It integrates the verbal content of both agent's utterances and user's utterances, in order to model the link between the user's attitudinal expressions to the agent utterances. Specific labels are defined for both the agent and the user.

Illocutionary acts of agent's utterances In order to model the potential influence of the agent's speech over the user's expressions of attitude, we label the agent's utterances regarding the illocutionary acts that they perform. We refer to Searle's classification (Searle, 1976) which includes five categories:

- the *representative acts* : their purpose is to commit the speaker to something's being the case, to the truth of the expressed proposition. Example : *It's raining* ;

- the *directive acts* which attempt to get the hearer to do something. Example *I order you to leave* ;

- the *commissive acts*, which commit the speaker to some future course of action. Example *I promise to pay you the money*;

- the *expressive acts*, which express the speaker's psychological state about a state of affairs. Example *I apologize for stepping on your toe* ;

- the *declaration acts*, whose the successfull performance provides that the propositional content corresponds to the world. Example : *You're fired*.

For labelling each agent's utterance, we use the *agent utterance unit* to which we add a feature specifying the type of the illocutionary act. One agent's speech turn can contain several utterances performing different illocutionary acts. For example, in the sentence "well things will normally get better, can you eh think of something that might make you happy", we identify two *agent utterance units*: "well things will normally get better" is a representative illocutionary act, and "can you eh think of something that might make you happy" is a directive illocutionary act.

Specific features of *attitudes* Our model considers both the agent and the user's expressions of attitude. An *attitude* expression comprise three components which we need to label : the linguistic mark referring to the attitude, the source and the target. Information about the *attitude* type and its polarity has to be also specified.

- The *attitude* type (affect or evaluation) and the polarity (positive or negative) are specified by a feature-set associated to the user's schema – described below – and the agent utterance unit. It should be specified that when the agent does not refer to an *attitude* these features receive the *none* value.

- When the user and the agent's expressions of *attitude* have a target and a source expressed, we use the *target unit* and the *source unit*. The *target unit* deals with the phrase referring to the entity, the process or the behaviour evaluated or trigger the attitude, and the *source unit* has to do with the phrase referring to the source of the evaluation or the emoter of the affect. In order to check the influence of the agent's speech over the user's attitudes, the user's target was relied to the agent's target when it was referring to the same entity or one of its sub-aspects.

- The *attitude mark* is only labelled for the user's *attitude* and not for the agent's one. Regarding the agent, the feature specifying the *attitude* type is enough to specify if his utterance conveys an attitude. However, since our further detection module will have to focus on the user's expressions of attitude, we need to retrieve information about its linguistic mark. The *attitude mark unit* concerns, at the phrase level, both linguistic marks referring to an *attitude* and modifiers which can shift, intensify or diminish its semantic value or its valence. For example, in a sentence as "I dont really like my work", "dont really like" is tagged as an *attitude mark*.

Linking the user's *attitude* to the agent's previous speech turn Finally, the *agent utterance units* and the *source* and *target units*, to which it may be linked, compose an **agent schema**. Similarly, the *attitude mark unit* and the source and target units, to which it may be linked, compose a **user schema**. Each *user schema* is linked to the previous *agent schema* by a simple relation notifying this precedence.

Topic segmentation A same conversation can comprise different topics. Here, we define the topic as a sequence, which takes place across several agent and user speech turns. A main topic can include sub-topics, which can refer to its sub-aspects. For example, in a conversation where speakers talk about "christmas" and the gifts they received, "christmas" is tagged as the topic, and "gift" as the sub-topic. It may be important to check whether the user's *attitudes* are linked to the ongoing topic of the conversation.

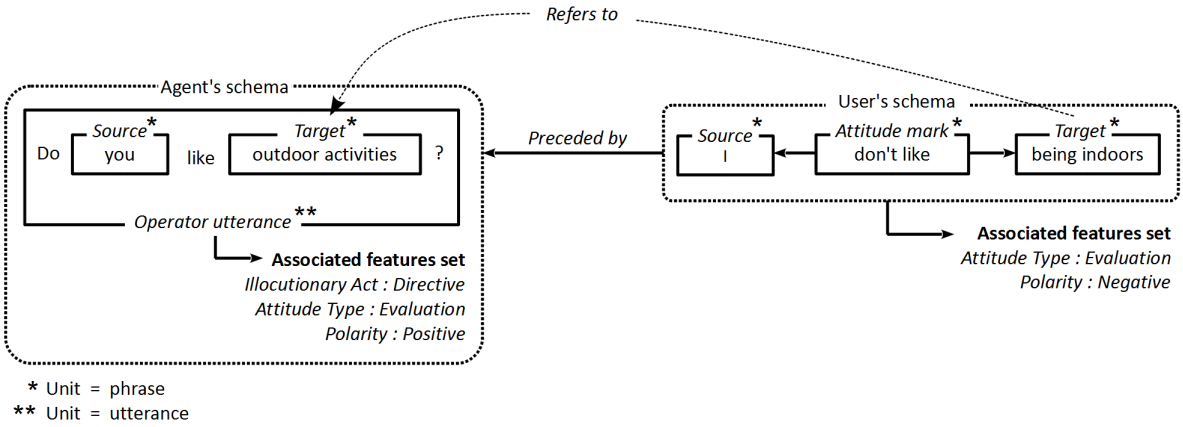


Figure 1: Annotation of two specific utterances. The noun "activities" here opens a new topic

For this purpose, two units are dedicated to the topic labelling, the *topic unit* and the *sub-topic unit*. Even if the topic is sequential and crosses several speech turns, we label the first occurrence of the typical word of the topic or sub-topic. A dedicated relation links a *sub-topic* to its main *topic*. A feature was added to both *topic* and *sub-topic* units, which specify if the topic or sub-topic has been started by the user or the agent. When a target was referring to a topic or a sub-topic, we link them with specific relation. As a summary, Figure 1 presents how two specific sentences are labelled by referring to our model. Since the agent's utterance have an interrogative form, it is labelled as a directive illocutionary act. As explained by Searle (1976), questions are species of directives since they are attempts by the speaker to get the the hearer to answer, i. e. to perform a speech act. The propositional content of this question concerns an evaluation expressed by "like". The source of the evaluation does not match with the speaker – i. e. the agent – but with the hearer – i. e. the user. The agent asks to the user to express himself about an positive evaluation regarding to a target already chosen, "outdoor activities". Regarding the user's sentence, "don't like" is labelled as an *attitude* mark – and "i" as is source. The target of the user's *attitude*"being indoors" is linked to the target of the agent's *attitude*: even if they are kind of antonyms, an ontological reference exists between them.

4. Labelling user's *attitude* in a human-agent (operator) interaction: the SEMAINE Corpus

The model is here confronted with the SEMAINE corpus (McKeown et al., 2011). As a preliminary study, one annotator labelled the corpus, but we plan to provide a second annotation with other annotator. This corpus comprises 65 manually-transcribed sessions where a human user interacts with a human operator acting the role of a virtual agent. These interactions are based on a scenario involving four agent characters : Poppy, happy and outgoing, Prudence, sensible and level-headed, Spike, angry and confrontational and Obadiah, depressive and gloomy. Agent's utterances are constrained by a script (however, some deviations to the script occur in the database) with the aim to push the user toward the character played's state. 15 sessions were labelled according to the previously described annotation model. Sixteen sections, four sessions for each characters, were labelled. Thirteen different users are involved in these different sessions. Regarding the agents, there are four different actors playing the role of Poppy, three for Prudence, three for Obadiah and three for Spike. Finally, for all sessions, there are five different actors.

As an annotation tool, we use the *Glozz Plateform* (Widlöcher and Mathet, 2012). By using *Glozz*, we can locate, identify and describe linguistic phenomena in textual documents.

Overall analysis The sessions labelled have got variable number of speech turns (132 for the longer session, 38 for the smaller). In the entire corpus, the users and the agents have got a number of speech turns almost similar: 559 for the users and 579 the agent. Over the 450 labelled agent schemas (0.77 per speech turn), 46% express a directive act, 42% an expressive act and 12% representative. No declaration and commissive speech act was founded. This is probably due to the nature of the scenario on which is grounded the Semaine corpus: a narrative conversation where the user is pushed to talk about this life. As shown in Figure 2, the distribution of illocutionary acts is the same regarding the agent's identity with the exception of Obadiah. The Obadiah sessions show that the expressive acts hold a majority and the expressive acts occur more often than in the other agents sessions.

As explained above, the operator's utterances can express attitudes. In our corpus, 339 operator's schemas contain expressions of *attitude*: 187 affects and 152 evaluations. With regard to the agent's identity, the expressions of affect are more numerous than the evaluation ones, excepting for the Prudence sessions (see Figure 3). This is probably due to the Prudence's personnality that the operator have to play : a sensible and level-headed person who expresses evaluations about the user's behaviour and asks the user to express attitudes about specific things.

The annotation model allows us to give an insight also into the user's expressions of *attitude* (238 labelled user's

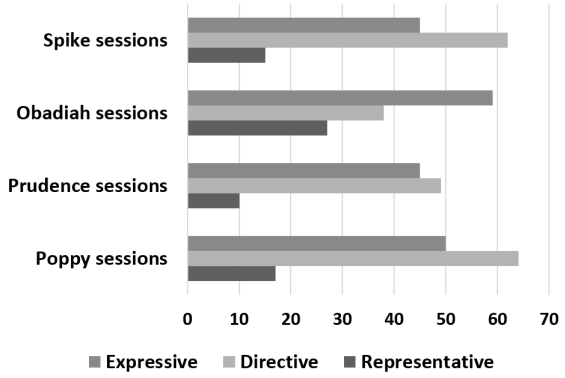


Figure 2: Distribution of illocutionary acts according to the agent's identity

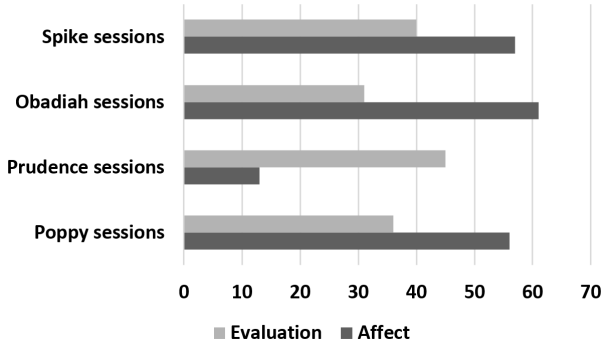


Figure 3: Distribution of affects and evaluations according to the agent's identity

schemas). In particular, we show that expressions of evaluation are more frequent than expressions of affect, but it is not a clear majority : 44% are affects and 56% evaluations.

Illocutionary acts and user's attitude Both the directive and expressive acts have a significant number of occurrences in the entire labelled corpus. Nevertheless, regarding the specific relation linking the user's attitudes to the agent's utterances, the directive illocutionary acts prevail : 57% of user's attitudes are linked with an agent's directive act. It seems that most of the directive acts labelled in the corpus have an interrogating form : the agent asks the user to tell about something. Thus, as requests, these agent's utterances may easily involve an attitudinal reaction from the user. Moreover, some of them are explicit requests, from agent to the user, to express himself about attitudes.

Polarity accordance The user and the agent attitudes are studied according to the type of agent played by the operator (see Table 1). As expected, Poppy and Prudence sessions express more positive attitudes, whereas Spike essentially expresses attitudes with a negative polarity. The distribution is more balanced concerning Spike. Moreover, with regard to the relation between user schemas and agent schemas, the polarity of user's attitudes is mostly the same as the polarity expressed by the agent. The polarity of 71% of the user schema linked to an agent schema containing an *attitude* matches with the polarity of the agent's attitude. Furthermore, this polarity accordance occurs in most of the sessions (see Figure 4).

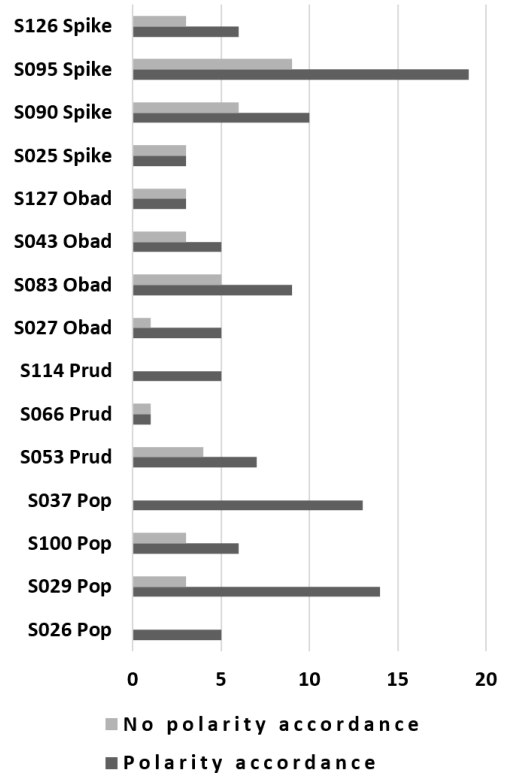


Figure 4: For each session, number of user's attitudes, linked to an agent's attitude, which have the same polarity

Sessions	Agent positive attitudes	Agent negative attitudes	User positive attitudes	User negative attitudes
Poppy sessions	92%	8%	83%	17%
Prudence sessions	96%	4%	67%	33%
Spike sessions	20%	80%	40%	60%
Obadiah sessions	41%	59%	52%	48%

Table 1: Polarity of user and agent's *attitudes* according the agent identity

Target Among the attitudes labelled by the user's schemas (238 in the entire corpus), 172 have a target. When the user schema is linked to an agent schema containing an attitude, the user target can refer to the agent target : one relation in our model notifies its reference. In the entire labelled corpus, a quarter of the target appraised by the user refer to a target appraised by the agent (27% of them). These results show that the user does not always choose what object he will appraise. This phenomena needs to be

considered in order to design a sentiment analysis module : with regard to the agent target - which will be known by the system - it will may be easier to process the potential user's attitudinal reaction and find its target.

Topic In the entire labelled corpus, 61 lexical units are labelled as topic and 51 as sub-topic – an average number of 4 topics and 3.4 sub-topics by session. Among these topics and sub-topics, 57% are started by the agent, 30% by the user, and 13% of them arise due to a kind of collaboration between the agent and the user. For example, in the session 26, Poppy asks to the user, "where is the best wake up you ever had?". The user answers "in a tent in kilimanjaro". Here, the agent's question opens a new topic but not completely defines it. It is the user's answer which chooses the new topic, but by following the indications given by the agent's question. When the user's targets are not linked to an agent's target, they may be linked to a topic or a sub-topic. In the entire labelled corpus, out of a total of 172, 28 user's targets are linked to a topic and 47 to a sub-topic. 32% of these topics and sub-topics are started by an agent, 40% by a user and 28% result form the collaboration - described above - between the agent and the user. Thus, as for the targets, the user's expressions of *attitude* are grounded on elements which are arised in the conversation through the agent's speech.

5. Discussion and tracks to design a module for detection of user's attitudes

As shown by the framework of our model and its different units and features, we aim to describe the expressions of *attitude* in a compositional way, i. e. we consider their meanings as built by the sum of the meaning of their constituents. The categories of our model provide the first semantic values to describe this meaning and that of the agent's utterances. This compositional representation will be useful to build our attitudes detection module.

First, the semantic values regarding the agent's utterances (illocutionary acts, attitudes, etc.) will allow us to process the user's ones. Since there is some semantic accordance between the agent's utterances and user's attitudes, a semantic characterisation of each agent's utterance could be used to anticipate the possible following user's expressions of *attitude* and to improve their analysis when they occur. This characterisation could be implemented as a simplified semantic feature set and used as an input of the module. For instance, the feature set associated to the agent's sentence "Do you like outdoor activities" (see Figure 1) indicates that the agent's utterance conveys the user to express an attitude. The user's expected attitude can thus be modelled in the module, which can check whether its source and its target are the same as the ones in the agent's utterance by using linguistic rules grounded on syntactic and semantic clues. If no accordance is founded, a more complex analysis can be done.

Second, some refinements which will help the future detection module can be done in our model. Regarding the agent's utterances, as shown in Section 4, the expressive and the directive illocutionary acts have a large number of occurrences in the corpus. Thus, these categories could be refined. For example, two sub-categories could be linked to the directive illocutionary act category: question and suggestion. Such sub-categories could give more accurate information about the meaning of the agent's utterance, which will be helpful to improve the performance of our detection module. For instance, if the feature set introduced above could specify that the sentence has an interrogative form, this can limit the number of likely user's sentences to consider and allows the module to use a more specific linguistic rule to process it. Regarding the user's attitudes, other features could be added too. For example, with regard to the graduation dimension, we need to distinguish different semantic values among the valence modifiers or shifters : a negation will not have to be analysed in the same way as an intensifier. Moreover, it is important to provide – as an output – information about how graduating is the attitude expressed. Indeed, it could be interesting that the agent do not perform in the same way attitudes like "I like outdoor activities" and "I like very much outdoor activities". Finally, the model needs to also deal with multimodality issue: in order to ensure this, the user's attitudes could be also linked to the agent's non-verbal signal.

6. Conclusion and further work

This paper proposes a model of user *attitudes* in verbal content. In order to go beyond the classical positive *vs.* negative distinction, this model examines some features as the source, the target or the *attitude* type and deals with interaction by integrating information about the illocutionary acts and the relations between user and agent units. This model – confronted with the SEMAINE corpus – shows that these features are relevant. The user's attitudes have properties in common with the agent's attitudes, like the polarity and, less, the target. In further work, as explained in Section 5, the simplified semantic representation provided by the model will be refined. By doing so, our model will be a strong foundation for designing our detection model.

7. Acknowledgment

The authors thank Catherine Pelachaud for valuable insights and suggestions. This work has been supported by the european collaborative project TARDIS and the SMART Labex project[1].

8. References

Bednarek, M. (2009). Language patterns and attitudes. *Functions of Language*, pages 165–192, 16/2.

Bloom, K., Garg, N., and S., A. (2007). Extracting appraisal expressions. *HLT-NAACL*, pages 165–192, April.

Breck, E., Choi, Y., and Cardie, C. (2007). Identifying expressions of opinion in context. In Sangal S., M. H. and K., B. R., editors, *International Joint Conference On Artifical Intelligence*, pages 2683–2688, San Francisco, CA. Morgan KoffMann Publishers.

Campano, S., Durand, J., and Clavel, C. (2014). Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Language Resources and Evaluation Conference*. to appear, May.

[1] http://www.smart-labex.fr/

Clavel, C., Pelachaud, C., and Ochs, M. (2013). User's sentiment analysis in face-to-face human-agent interactions prospects. In *Workshop on Affective Social Signal Computing, Satellite of Interspeech*. Association for Computational Linguistics, August.

Martin, J. R. and White, P. R. (2005). *The Language of Evaluation. Appraisal in English*. Macmillan Basingstoke, London and New York.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, Jan-March.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 806–814, Stroudsburg, PA, USA. Association for Computational Linguistics.

Osgood, C., Mai, W. H., and S., M. M. (1975). *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, Urbana.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, November.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, 5(01):1–23.

Whitelaw C., Garg, N. and Argamon, S. (2005). Using appraisal taxonomies for sentiment analysis. *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, April.

Widlöcher, A. and Mathet, Y. (2012). The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, pages 171–180, Paris, France, September.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotation expressions of opinion and emotions in language. *Language Resources and Evaluation*, pages 165–210, Vol. 39/2-3.