Harmonization of German Lexical Resources for Opinion Mining

Thierry Declerck, Hans-Ulrich Krieger
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{declerck|krieger}@dfki.de

Abstract

We present on-going work on the harmonization of existing German lexical resources in the field of opinion and sentiment mining. The input of our harmonization effort consisted in four distinct lexicons of German word forms, encoded either as lemmas or as full forms, marked up with polarity features, at distinct granularity levels. We describe how the lexical resources have been mapped onto each other, generating a unique list of entries, with unified Part-of-Speech information and basic polarity features. Future work will be dedicated to the comparison of the harmonized lexicon with German corpora annotated with polarity information. We are further aiming at both linking the harmonized German lexical resources with similar resources in other languages and publishing the resulting set of lexical data in the context of the Linguistic Linked Open Data cloud.

Keywords: Polarity lexicons, opinion mining, linguistic linked open data

1. Introduction

This paper describes work carried out in the European R&D project TrendMiner (www.trendminer-project.eu) which deals, in part, with the extraction and representation of real time information from dynamic data streams, such as blogs, twitter and newswires. In this context, TrendMiner addresses a use case dealing with the EU-wide tracking of political views, trends, and politician popularity over time. Therefore there is a need for accurate lexical resources for computing polarity associated to relevant entities mentioned in such media. Dealing also with German texts, the project has investigated the re-use of existing German language resources which are marked-up with polarity and sentiment information. For now, we are considering the following resources:

1. A polarity lexicon for German[1] (Clematide & Klenner, 2010), hereafter called "german.lex"
2. GermanPolarityClues[2] (Waltinger, 2010a), hereafter called "GermanPolarityClues.lex"
3. GermanSentiSpin[3] (Waltinger, 2010b)
4. SentiWS[4] (Remus et al., 2010)
5. MLSA: A Multi-layered Reference Corpus for German Sentiment Analysis[5] (Clematide et al., 2012)
6. A collection of nominal phrases and a collection of clauses annotated for polarity, plus their accompanying dependency parses[6] (Klenner et al., 2012)

In this paper we focus on the lexical resources (1-4 in the list just above), describing how those have been mapped onto each other, generating a unique list of entries, with unified Part-of-Speech (PoS) information and basic polarity features. We are currently investigating how this harmonized lexicon can be used for the semi-automatic annotation of German texts with polarity information. We will for this purpose compare the features used in the lexicon and the type of annotation used in the corpora resources listed above (5-6), and will report on this in future updates on our work.

Our work is further aiming both at linking this harmonized language data to similar resources in other languages and their publication in the context of the Linguistic Linked Open Data (LLOD) initiative[7].

2. The considered German Resources for Opinion and Sentiment Analysis

Looking at the different lexical datasets (1-4 in the Introduction section) or at the corpora (5-6 in the same section) we noticed that the developers of those resources make use of different formats for encoding identical or similar language data, while they also describe distinct types of information. Details from two distinct lexical resources are given in Figure 1 and Figure 2 below, showing clearly the heterogeneity of formats and information they encode.

In order to have an as large as possible but unique lexical resource for processing and annotating (German) text with opinion and sentiment information, we needed therefore to go for the harmonization and integration of this set of available lexical data for German.

For the purpose of harmonization we wrote some Perl scripts for mapping the encoded information in the lexicons into hash tables that are being subsequently merged into a unique hash table. Harmonization work consisted in using unique descriptors for opinion/sentiment features and values present in the various sources. So for example the values "+", "p" or "POS" (for positive polarity) or some integers are mapped onto the unique value term "POS" for encoding "positive" polarity/sentiment. The same is done for features: some developers introduce a feature "Polarity weight", while

[1] http://bics.sentimental.li/index.php/downloads/
[2] http://www.ulliwaltinger.de/sentiment/
[3] http://www.ulliwaltinger.de/sentiment/
[4] http://asv.informatik.uni-leipzig.de/download/sentiws.html
[5] http://iggsa.sentimental.li/index.php/downloads/
[6] http://bics.sentimental.li/index.php/downloads/

[7] See http://linguistics.okfn.org/ for more details on the LLOD cloud. In this context, we started a cooperation with the Eurosentiment project (http://eurosentiment.eu/), which is specifically aiming at publishing opinion and sentiment lexical data in the LLOD cloud (Buitelaar et al., 2013).

other have "reduction/gaining factor" and the like. Such features are harmonized to "pol_rank". An example of such an alignment of features, with indication of the provenance, is given in Figure 3.

Format:
Word {NEG|POS|NEU|SHI|INT} PolarityStrength PoS
SHI for Shifters, INT for Intensifiers
INT < 1, e.g. 0.5 is a reduction factor,
> 1, e.g. 2 is a gain factor

Examples of lexical entries:
fehlschlagen NEG=0.7 verben
frisch POS=0.7 adj
Tick NEU=0 nomen
beenden SHI=0 verben
ohne SHI=0 neg
enorm INT=2 adj
viel INT=2 adj
…

Figure 1: Format of polarity information associated with lexical entries in the lexicon "german.lex" (Clematide & Klenner, 2010). For example, the word "beenden" (*to finish*, *to terminate*) is encoded as a verb carrying the polarity feature "Shifter" with the value (PolarityStrength) set to "0".

Format:
Lemmata (\t) Part-of-Speech (\t) PositiveRating (\t) NegativeRating (\t) NeutralRating (\t)
PositiveCorpusProbability (\t) NegativeCorpusProbability (\t) NeutralCorpusProbability

Examples:

illegal	ADJD	0	1	0	0	1	0
Ragen	VVFIN	0	1	0	0.2	0.8	0
Abhilfe	NN	1	0	0	0.2	0.8	0
Sehr	ADV	1	0	0	0.350257	0.388175	0.261568
werden	VVFIN	0	0	1	0.258483		0.493513
gleich	ADJD	0	0	1	0.247549		0.566176

Figure 2: The format of the lexical resource "GermanPolarityClues.lex" (Waltinger, 2010a), together with a few lexical entries. Comparing it to the lexical data in Figure 1, the reader can see the differences in the tagset used ("NN" vs "nomen", etc) and the type of polarity information included. And here, one lexical item can be associated with distinct polarity features, while this is not the case in the example in Figure 1.

3. The Harmonization Strategy

All lexical resources we considered use the lemma of the words to be marked-up with polarity features (while some also include inflected forms). The lemmas of the lexicons have been used as keys in the hash tables resulting from the application of the Perl scripts. This allows to both control the merging procedure and to check the lexical coverage of the harmonized lexicon.

```
"erheblich" => {
        "prov::GermanPC.lex" => {
                "pos::AJ" => {
                        "pol_rank" => "0.15",
                        "pol_val" => "POS",
                },
        },
        "prov::GermanSentiSpin.lex" => {
                "pos::AJ" => {
                        "pol_rank" => "0.0252801",
                        "pol_val" => "POS",
                },
        },
        "prov::GermanSentiWS.lex" => {
                "pos::AJ" => {
                        "pol_rank" => "0.0040",
                        "pol_val" => "POS",
                },
        },
        "prov::german.lex" => {
                "pos::AJ" => {
                        "pol_rank" => "2",
                        "pol_val" => "INT",
                },
        },
}
```

Figure 3: An example of the integration of entries from the 4 different German polarity lexicons, indicating the provenance of each piece of information.

Figure 3 displays one example of the output of our Perl scripts applied to the four original lexicons. In this case, the reader can see that almost all sources agree on the (harmonized) polarity value to be associated to the lemma "erheblich" (*considerable*) , with the source "german.lex" deviating and selecting the value "INT" (for *Intensifier*). Since the "german.lex" source has a more fine-grained set of polarity features (see Figure 1) as the other lexical resources, we decide to select the feature "INT" as being the one to be used in all cases, so that the only remaining difference between the sources is concerned by the values of the "pol_rank" feature. The lexicon is getting thus much more compact and the entry has now the form:

erheblich, AJ, INT, {0.15, 0.0252801, 0.0040, 2}.

For sure, we still have to decide on how to deal with the fact that the value "2" given in the "german.lex" source has another meaning as the figures given in the other sources: it is meant to indicate the level of intensification and not a probability.

Figure 3 exemplifies our strategy for dealing with the different levels of granularity for encoding polarity: we opt for the system being more specific. Some lexicon consider only the values "NEG", "POS" and "NEUT", while "german.lex" include two more values: "SHIFT" and "INT". So that if a lexicon is encoding the German word "nicht" (*not*) as "NEG" and another one as "SHIFT", we choose for the harmonized and integrated lexicon the latter value. We expect from this decision to

generate lexical data to be used in the context of sentiment/opinion detection grammars, since the value "SHIFT" is marking the fact that the polarity value of the word(s) modified by this lexical item will be correspondingly updated. This is clearly the case in the context of negations ("The money is *not* lost."), where the negation word is (usually) shifting the polarity value of the words in its scope to the opposite value. We are currently developing grammars based on this principle, but this topic is outside of the scope of this paper.

There are also cases of use of incompatible polarity feature values for an entry in different lexical resources. Figure 4 displays an example of the use of incompatible polarity values for the lemma of the noun "Erhalt" (*receipt, reception, acceptance, preservation*). One lexicon is giving the value "POS", while the other is opting for the value "NEG".

```
"erhalt" => {
        "prov::GermanSentiSpin.lex" => {
                "pos::N" => {
                        "pol_rank" => "0.0182394",
                        "pol_val" => "NEG",
                },
        },
        "prov::german.lex" => {
                "pos::N" => {
                        "pol_rank" => "0.7",
                        "pol_val" => "POS",
                },
        },
}
```

Figure 4: Example of a lemma with contrary polarity values in distinct sources. The noun "Erhalt" (*receipt, reception, acceptance, preservation*) is considered to have a positive polarity in one case and a negative one in the other case

Our strategy to deal with this case consists in checking if one of the sources is giving the same value to syntactic variants of one entry. For the example in Figure 4, we are considering the corresponding adjectival and verbal forms of the noun "Erhalt". The assumption (to be still verified) is that all syntactic variants of a word should bear a compatible polarity value. So that if the lexical source "GermanSentiSpin.lex" is associating the value "NEG" to the noun "Erhalt", we would expect this source to do the same for the corresponding verbal form "erhalten" or the adjectival variant "erhaltbar". Looking at the corresponding harmonized entries in Figure 5, the reader can see that this is not the case: The lexical source is marking up both the verbal and adjectival entries as having a positive polarity value, leading us to the assumption that the source is not reliable on this particular example. Since other lexical sources are marking the entries as carrying a positive polarity value in the cases displayed in Figure 5 and the source "german.lex" is also associating a positive polarity value to the noun "Erhalt" (Figure 4), our approach consists in overwriting the negative polarity value given by the source "GermanSentiSpin.lex" in Figure 4, leading to a unified value.

While this approach still has to be validated for all entries, we didn't find at first sight an example invalidating our approach for this type of contrary polarity value included in distinct lexical sources.

```
"erhaltbar" => {
        "prov::GermanSentiSpin.lex" => {
                "pos::V" => {
                        "pol_rank" => "0.0210568",
                        "pol_val" => "POS",
                },
        },
},

"erhalten" => {
        "prov::GermanPC.lex" => {
                "pos::V" => {
                        "pol_rank" => "0.277778",
                        "pol_val" => "POS",
                },
        },
        "prov::GermanSentiSpin.lex" => {
                "pos::V" => {
                        "pol_rank" => "0.00794075",
                        "pol_val" => "POS",
                },
        },
},
```

Figure 5: Example of lemmas related to the noun "Erhalt" in Figure 4. The adjective "erhaltbar" (*perceivable, maintainable, preservable, available*) and the verb "erhalten" (*to perceive, to maintain*) are associated by the source GermanSentiSpin.lex with a positive polarity value although the same source is marking the noun "Erhalt" with a negative polarity value.

4. First Results of the Harmonization and Merging of German Polarity Lexicons

The four German polarity/sentiment lexicons have been mapped onto a corresponding hash table by Perl scripts. The resulting hash tables have been merged, whereas the harmonization procedures described in the preceding section have been applied. The total of entries in provenance from the four lexicons is 116970, where the largest part is coming from the GermanSentiSpin lexicon (95572 entries). This lexicon includes a lot of compound words (the lexicon is in fact a translation from other sources). Our harmonization approach allows reducing the size of the lexicon to 97162 entries. Actual work consists in performing decomposition on the entries of the GermanSentiSpin lexicon, and so to reduce to a larger extent the resulting harmonized lexicon (many entries of the actual harmonized lexicon are containing the lemma originating only from of the GermanSentiSpin lexicon). Looking at the result when we consider only "german.lex" and "GermanPolarityClues.lex", we have following figures:

- Nr of entries in german.lex (lexicon1): # 8714
- Nr of entries in GermanPolarityClues.lex (lexicon2) : # 9231
- After merging
  - # intersection = 3014
  - # entries only in lexicon1 = 5700
  - # entries only in lexicon2 = 6217
  - # entries in harmonized lexicon = 14931

5. Actual Work

We noticed that the existing German corpora annotated with polarity information do not include lemmas in the encoding of the terminal elements. We advocate that it would be very useful to link corpora data and lexicons for sentiment analysis. We will therefore map the content of the terminal nodes in the corpora onto lemma included in the harmonized polarity lexicon. For all the resources encoding also full forms (at various level of granularity) we will harmonize and reduce those to inflectional paradigms, which we have at our disposal for example in the NooJ computational lexicon[8]. So for example the German word "Aggressor" is encoded in NooJ this way: "aggressor,N+FLX=HERR+Hum" (thus specifying that the lemma "aggressor" is a noun being inflected like the noun "Herr", and with the semantic being set to "Human"). In a next step, we will abstract over this particular encoding of inflectional information. We will also map the different tagsets used in the distinct lexicons to the STTS tagset[9]. Additionally to this mapping, we will check for the use of the ISOcat registry[10], which is now also available in the Linked Data framework, for allowing cross-lingual comparisons.

We can add basic semantics to all the entries of the mentioned German resources (see the example for the NooJ lexicon entry "Aggressor" above). But we think that linking to semantic data in the Linked Data framework will help in getting a more widely recognized semantic organization of the data (linking for example to the semantic categories associated to German Wiktionary entries, now available in the Linked Data framework, as can be seen at http://dbpedia.org/Wiktionary) or to BabelNet (which has been very recently released in the Linked Data format, see http://lcl.uniroma1.it/babelnet/). An example of encoding the polarity of lexical entries in RDF and *lemon* is discussed in (Buitelaar et al., 2013). We are currently working on linking the German harmonized lexical resources to the opinion ontology developed in the context of the TrendMiner project[11] and which is based on the work by (Westerski et. al, 2013), adding some specific constructs for dealing with the goal of the project, for example allowing a large scale population of opinion ontology elements.

[8] See http://www.nooj4nlp.net/pages/nooj.html for more details.

[9] See http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/stts.asc. We are also aware that a new version of this de facto standard for German language data is on the way.

[10] See http://www.isocat.org/ for more details

[11] www.trendminer-project.eu. The ontologies of the project are available under www.dfki.de/lt/onto.

6. Conclusion

In this paper we have described on-going work on the harmonization and integration of different German lexical resources for opinion/sentiment analysis. We have presented also first quantitative results. The harmonized German lexical resources will be made freely available for research purposes, and can be used for setting up gold standards in the field of opinion mining. Actual work is dedicated in linking the harmonized lexical resources to knowledge objects encoded in the form of ontologies or similar, in order to support their publication in the Linked (Open) Data framework, more specifically in the context of the Linguistic Linked (Open) Data initiative.

7. Acknowledgements

The research described in this paper has been co-financed by the European Commission, in the context of the FP7 ICT project TRENDMINER, under contract number 287863. We would like to thank all the authors of the resources mentioned and discussed in this paper for their authorization to use their data.

8. References

Buitelaar, P., Arcan, M., Iglesias, C., Sánchez, F. and Strapparava, C. (2013). Linguistic Linked Data for Sentiment Analysis. *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL 2013)*, Pisa, 23 September 2013, 1-9.

Clematide, S. and Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*; Held in conjunction to ECAI 2010 Portugal, Lisbon, Portugal, 17 August 2010 - 17 August 2010, 7-13.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U. and Wiegand, M. (2012). MLSA -- A Multi-layered Reference Corpus for German Sentiment Analysis. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, May 2012.

Klenner, M., Clematide, S., Petrakis, S. and Luder, M. (2012). Compositional syntax-based phrase-level polarity annotation for German. *Proceedings of the 10th International Workshop on Treebanks and Linguistic Theories (TLT 2012),* Heidelberg, 06 January 2012 - 07 January 2012,

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. and Wummer, T. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation 46, no. 4* (2012): 701-719.

Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Ressources and Evaluation (LREC'10)*.

Waltinger, U. (2010a). GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Waltinger, U. (2010b). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features. *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10).*

Westerski, A. and Sánchez-Rada, J.F. (2013). *Marl Ontology Specification, V1.0 May 2013*. Available at http://www.gsi.dit.upm.es/ontologies/marl