Bootstrapped Extraction of Index Terms from Normalized User-Generated Content

Piroska Lendvai
Saarland University
Dept. of Computational Linguistics
Saarbrücken, Germany
piroska.r@gmail.com

Thierry Declerck
Saarland University
Dept. of Computational Linguistics
Saarbrücken, Germany
declerck@dfki.de

Abstract

We report on the extraction of key phrases for news events, based on string alignment between social media posts and user-linked web documents. Hashtag normalization is tested for enhancing string similarity, while both token-based tweet similarity and manual event annotations are tested for transferring web links to posts that do not refer to external documents. We are able to identify more terms via web link transfer compared to no link transfer, and obtain syntactically and semantically more complex terms compared to general document-based term extraction.

1 Introduction

Creating the logical representation of a document collection in terms of index terms is a crucial step in information retrieval. The extraction of a meaningful set of index terms for a document collection, instead of making every word (noun) an index term, can profit from human insights. Our goal is to find, collect, and utilize such insights from social media content. Our core assumption is that users who include a reference to an external web document in their social media post are implicitly encoding a relevance signal; this assumption is analogous with utilizing landing page information from click data to classify user intent (see e.g. Joachims (2002)).

We investigate the extraction of key terms for news events, based on string alignment between social media posts and user-linked web documents, as described in Lendvai and Declerck (2015). Manually assigned event annotations are used to transfer the web links to posts that do not refer to external documents, thereby boosting the amount and quality of extracted index terms. Then, token-based tweet similarity is going to be used to the same end.

Hashtag harmonization is supposed to enhance string similarity; in Declerck and Lendvai (2015a) we reported on a hashtag processing approach that we test in our present study as well. Hashtags are normalized, lemmatized and segmented in a data-driven way in a simple offline procedure that generates a gazetteer of hashtag elements. In Declerck and Lendvai (2015b) we developed the basic Linguistic Linked Open Data (LLOD)[1] infrastructure for representing hashtags from social media posts. We explained how the OntoLex model[2] is used both to encode and to enrich the hashtags and their elements by linking them to existing semantic and lexical LOD resources: DBpedia and Wiktionary.

Our goal in the current study is to give a pilot evaluation on application- and data-driven, language-independent approaches for term extraction, comparing the obtained terms with document-based term extraction, and comparing the terms after hashtag harmonization and web link transfer against non-harmonized data and no link transfer. We also report on how term extraction is affected by link transfer based on automatically assigned tweet similarity instead of manual annotations.

2 Hashtag harmonization

Hashtags allow users to classify their social media text, especially Twitter messages, into semantic categories. Those tags are typically named entities such as ”#Ottawa”, terms such as ”#Shooting”, or concatenated phrases such as ”#WearewithCanada”. The relevance of hashtags to identify text topics has been utilized by several approaches. Laniado and Mika (2010) find hashtags to qualify as strong identifiers for Semantic Web applications. However, the

[1] See (Chiarcos et al., 2013) and http://linguistic-lod.org/llod-cloud

[2] OntoLex is a model for the representation of lexicons (and machine readable dictionaries) relative to ontologies. It has been developed in the context of the W3C Ontology-Lexica Community Group, see https://www.w3.org/community/ontolex/.

analysis of lexical variation to identify semantically coincident hashtags has not yet been considered, but is important to identify messages relating to the same topic. Semantic clustering approaches (Pöschko, 2011; D. Antenucci et al., 2011) focus on semantic topics and their relations, but neglect variation within hashtags. This kind of processing is necessary to obtain more precise information on the exact semantics represented by hashtags and identify all related tweets within a dataset.

Hashtags appear in different cases and need to be normalized first. Secondly, hashtags need to be lemmatized to automatically match singular and plural uses of words. Finally, the segmentation of complex hashtags into its individual components is needed if one wants to recognize hashtag paraphrases in related documents. By reducing the (ortho)graphical variation of hashtags, basic string and substring matching across document types is hypothesized to be made more effective.

The corpus we were working on in (Declerck and Lendvai, 2015a) was a UK-Riots corpus established by the Guardian[3]. We are now testing and extending our approach to datasets that have been collected in the context of the Pheme project[4], relating to the events of the Ottawa Shooting and the Gurlitt art collection, as described in (Lendvai and Declerck, 2015). The corpus contains 40,201 tweets (including many retweets) in which we identified 22,825 hashtags.

Normalization We normalize hashtags by lowercasing all letters. On Twitter, typographical errors and misspellings are common. We used the string similarity measure implemented in the Python module *difflib*[5] to detect basic spelling mistakes such as "#shotting". In order to avoid valid words to be corrected as misspellings, e.g. 'from' and 'form', the strings are matched to the unix words list[6]. If one of the strings is not in the list, the change is made.

Lemmatization Variation in hashtags also originates from suffixation, a frequent suffix is the plural sign. While there might be some semantic difference due to the use of plural or singular , it is worth reducing the plural in hashtags to the singular in order to be able to compare and to link hashtags to documents external to the Twitter sphere. We use a straightforward approach: comparing words ending with an '-s' to the unix word list. If the word ending in 's' is present in this list, like for example the word 'news', no action is taken. Otherwise we perform lemmatization. We are currently evaluating if this approach is accurate for hashtags compared to a proper lemmatizer adapted to user-generated content. We assume that this step will be needed in any case for languages with a richer morphology as English.[7].

Segmentation Deriving components from segmented hashtags as search terms presumably facilitates the automatic linking of tweets to documents from other genres, which do not contain hashtags, such as news articles. We use a simple approach to segmentation, starting from hashtags that use camel notation (see Declerck and Lendvai, 2015a), e.g. '#OttawaShooting', yielding the segments 'ottawa' and 'shooting', which will in turn be utilized to segment its casing-variant '#ottawashooting'. In our corpus we have 1,363 occurrences of 'OttawaShooting' and 232 occurrences of 'ottawashooting', whereas '#shooting' is used only 18 times as a standalone string. Hashtag segmentation is able to impact hashtag distribution, resulting in e.g. 1,611 occurrences of '#shooting', enabling better term relevance metrics.

Our simple approach to segmentation includes the risk to generate arbitrary segments (e.g. 'Wearewith')[8]. The unix words list can again be put to use for checking the validity of the components resulting from segmentation. Additionally, we apply queries to named entities resources in the LOD for validating such components.[9] These validation procedures make the harmonization of hashtags to be considered as an offline procedure, generating specialized gazetteers. We are investigating whether rules or patterns are possible to be extracted from the results of the current experiments, to be reused on incoming tweet streams for online processing.

[3] http://www.theguardian.com/news/datablog/2011/dec/08/twitter-riots-interactive

[4] http://www.pheme.eu/

[5] https://pymotw.com/3/difflib/

[6] https://en.wikipedia.org/wiki/Words_\%28Unix\%29

[7] See for example the work by (Horbach et al., 2014) on improving the performance of PoS taggers applied to German Computer mediated Communication

[8] We are grateful to an anonymous reviewer pointing out this issue.

[9] The querying procedure, implemented on Python, is described in details in (Declerck and Lendvai, 2015b).

3 Tweet-to-Document Linking

Very recently, creating systems for Semantic Textual Similarity judgements on Twitter data has been a Shared Task in the Natural Language Processing community (Xu et al, 2015). Given two sentences, the participating systems needed to determine a numerical score between 0 (no relation) and 1 (semantic equivalence) to indicate semantic similarity on the hand-annotated Twitter Paraphrase Corpus. The sentences were linguistically preprocessed by tokenization, part-of-speech and named entity tagging. The system outputs are compared by Pearson correlation with human scores: the best systems reach above 0.80 Pearson correlation scores on well-formed texts. The organizers stress that "while the best performed systems are supervised, the best unsupervised system still outperforms some supervised systems and the state-of-the-art unsupervised baseline."

In Lendvai and Declerck (2015) we proposed a cross-media (CM) linking algorithm in the PHEME project to connect User-Generated Content (UGC) to topically relevant information in complementary media, which we use in the current study as well. Each tweet in our datasets is manually annotated for an event. E.g. the tweet '`RT @SWRinfo: Das Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius #gurlitt an.`' is assigned the event *'The Bern Museum will accept the Gurlitt collection'*, while '`NORAD increases number of planes on higher alert status ready to respond if necessary, official says. http://t.co/qsAnGNqBEw #OttawaShooting`' is assigned the event *'NORAD on high-alert posture'*, etc.

For each URL-containing tweet within each event, a tweet-to-document similarity calculation cycle is run between tweets that link an external web document, and the linked web document. Similarity is evaluated in terms of the Longest Common Subsequence (LCS) metric. LCS returns a similarity value between 0 (lowest) and 1 (highest) based on the longest shared n-gram for each text pair, without the need for predefined n-gram length and contiguity of tokens (cf. Lin (2004)).[10]

3.1 LCS terms extraction

We use LCS to collect the top-5 scored longest common token subsequences identified for a linked document, based on a series of LCS computations producing LCSs between one, but sometimes more, tweets linking this document and each sentence of the document. No linguistic knowledge is used, except for stopword filtering by the NLTK toolkit[11]. Then the LCS cycle is applied to the same document set but paired with tweets that did *not* link external documents, based on the hand-labeled events. We are able to extract more, and lexically different phrases due to the link transfer.[12] For example, for the web document with the headlines *"Swiss museum accepts part of Nazi art trove with 'sorrow' — World news — The Guardian"* the extracted top terms based on tweets linking to this document are: 'swiss museum accepts part nazi art trove', 'nazi art', 'swiss museum', 'part nazi', 'nazi', whereas the extracted top terms based on tweets *not linking any document* but being annotated with the same event as the tweets referring to this document, are 'kunstmuseum bern cornelius gurlitt', 'fine accept collection', 'museum art', 'kunstmuseum bern cornelius gurlitt', 'kunstmuseum bern gurlitt', exemplifying that the Gurlitt dataset holds multilingual data, since we obtain terms not only in English, but in German as well.

3.2 Term extraction evaluation

3.2.1 No transfer to URL-less tweets

We are able to grow the set of extracted unique terms significantly if we perform the web link transfer step, when compared to not performing this step: from 110 to 186 in Gurlitt, and from 171 to 320 in Ottawa. The obtained term sets are highly complementary: about 70-90% of the phrases extracted from URL-less tweets are unseen in the phrase set extracted from URL-ed tweets.

3.2.2 Transfer based on automatically grouped tweets

We have compared the results of our LCS approach to experimental results where instead of using tweet clusters based on manual event annotations, we create tweet clusters by computing tweet similarity between each tweet and a centroid tweet for each event (designated by the phrase used in the manual event annotation), via a LCS similarity threshold. Inspired by Bosma and Callison-Burch (2007) who use an entailment threshold value of 0.75 for detecting paraphrases, we obtained our LCS similarity

[10] For details please see (Lendvai and Declerck, 2015).

[11] http://www.nltk.org/index.html

[12] For more details we again refer to (Lendvai and Declerck, 2015).

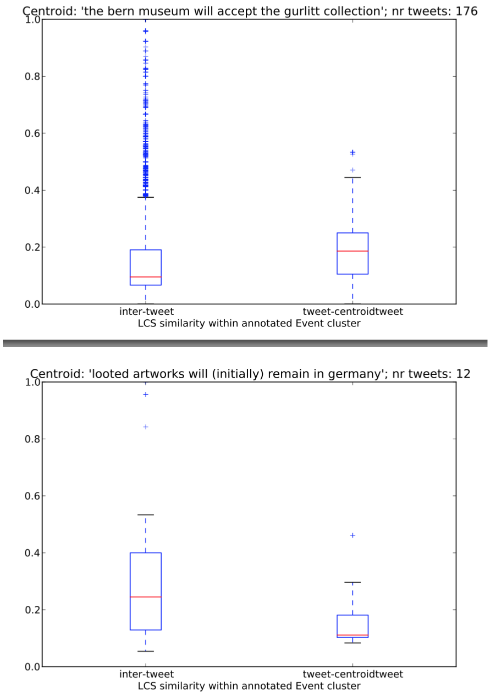

Figure 1: Tweet similarity distribution in terms of LCS values for two events from the Gurlitt dataset: tweet-tweet similarities within an event cluster, as well as centroid tweet - tweet similarities are plotted.

threshold t empirically by averaging the third quartile of LCS value distributions relating to an event over all events in a dataset ($t > 0.22$). Figure 1 illustrates tweet similarity distribution in terms of LCS values for two events from the Gurlitt dataset. We computed LCS values both in an intra-tweet way (i.e., LCS for all pairs of tweets within a tweet event cluster, the size of which is indicated in the upper right corner), and in the centroid-tweet way (i.e., LCS for all centroid-tweet pairs within the event cluster). Since Gurlitt is a multilingual set, the LCS scores often have a very wide distribution, also indicated by the large number of outliers in the plot.

The approach is rather crude and on the current toy datasets achieves a event-based-mean precision of 1.0 for Gurlitt and 0.32 for Ottawa, while a event-based-mean recall of 0.67 for Gurlitt and 0.78 for Ottawa. With this approach, we get much less URL-less tweets (Gurlitt: 16 vs 43, Ottawa:117 vs 182), but this seems to have an impact only on the Gurlitt multilingual dataset on the amount of extracted unique phrases from URL-less tweets (Gurlitt: 64 vs 93, Ottawa: 178 vs 197). Importantly, the quality and semantics of the extracted phrases for both datasets remain in line with those based on link transfer via hand-labeled events.

3.2.3 Frequency-based term extraction

We extracted a document-based term set from all tokens in the fetched documents that were automatically classified as nouns; part-of-speech information was obtained from the NLTK platform. These sets seem semantically more general than the terms obtained by the LCS approach (e.g. 'ausstellung', 'sammlung', 'suisse', i.e., 'exhibition', 'collection', 'switzerland') and are also smaller in size: 75 unique terms from all documents linked from the Gurlitt set, obtained in a top-5-per-document cycle to simulate the LCS procedure, and 116 for Ottawa. The obtained term set consists of single tokens only, while the average phrase length using the LCS approach is 3.65 for Gurlitt and 3.13for Ottawa.

4 Results and Conclusion

Our approach, based on longest common subsequence computation, uses human input for extracting semantically meaningful terms of flexible length. We link tweets to authoritative web documents, and create lexical descriptors extracted from tweets aligned with documents. The method is language-independent and unsupervised. The indexing terms can be used in their multi-word form or could be tokenized further. Hashtag normalization has currently no significant impact on our toy-sized datasets, and has been tested on German data for the first time. Scaling up from our current pilot setup, we plan to report on qualitative and quantitative results with enhanced html parsing and cross-media, cross-lingual text linking in forthcoming studies.

Acknowledgment

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233).

References

D. Antenucci, G. Handy, A. Modi, and M. Tinerhess (2011). Classification of tweets via clustering of hashtags. EECS 545 Final Project, 545:1-11.

W. Bosma and C. Callison-Burch (2007). Paraphrase substitution for recognizing textual entailment. In: Evaluation of Multilingual and Multi-modal Information Retrieval (pp. 502-509). Springer Berlin Heidelberg.

S. Bird, E. Klein, and E. Loper (2009). Natural Language Processing with Python, O'Reilly Media.

C. Chiarcos, P. Cimiano, T. Declerck, J.P. McCrae (2014). Linguistic Linked Open Data (LLOD) - Introduction and Overview in: Christian Chiarcos, Philipp Cimiano, Thierry Declerck, John P. McCrae (eds.): 2nd Workshop on Linked Data in Linguistics, Pages i-xi, Pisa, Italy, CEURS, 2013

T. Declerck and P. Lendvai (2015a). Processing and Normalizing Hashtags. in: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.): Proceedings of RANLP 2015, Pages 104-110, Hissar, Bulgaria, INCOMA Ltd, Shoumen, BULGARIA, 9/2015

T. Declerck and P. Lendvai (2015b). Towards the Representation of Hashtags in Linguistic Linked Open Data Format. in: Piek Vossen, German Rigau, Petya Osenova, Kiril Simov (eds.): Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, Hissar, Bulgaria, INCOMA Ltd, Shoumen, BULGARIA, 9/2015

A. Horbach, D. Steffen, S. Thater and M. Pinkal (2014). Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of the 12th edition of the Konvens conference (Konvens 2014).

T. Joachims (2002). Optimizing Search Engines Using Clickthrough Data. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).

D. Laniado and P. Mika (2010). Making sense of Twitter. In Proceedings of the Semantic Web Conference – ISWC 2010, pages 470-485.

P. Lendvai and T. Declerck (2015). Similarity-Based Cross-Media Retrieval for Events. In: Ralph Bergmann, Sebastian Görg, Gilbert Müller (eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, Trier, Germany, CEURS, 10/2015

Chin-Yew Lin (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8.

J. Pöschko (2011). Exploring twitter hashtags. arXiv preprint arXiv:1111.6553, 2011.

Xu, Wei, Chris Callison-Burch, and William B. Dolan. (2015). SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).