Towards the Harmonization and Segmentation of German Hashtags

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

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

Abstract

We present on-going work on the harmonization and segmentation of German hashtags. Our aim is to reduce the number of variants of hashtags expressing the same content to one harmonized hashtag that can thus serve as a unique “annotation tag” for a large set of tweets.

1 Introduction

When looking at hashtags used in Twitter posts (and probably in all social media) one can observe that one content is often expressed by various hashtags, whereas the degree of variance between the hashtags can heavily differ. Sometimes only the use of lowercase vs. uppercase letters marks the variance, like #EM2016 vs. #em2016. But there are more complex variants, as shown by the use of abbreviations or acronyms, like for example #EURO2016 vs. #europeanchampionship2016 or #Europameisterschaft2016. For the human reader, if she understands both English and German, the three hashtags are clearly related to the soccer event that took place in France in 2016. But for the machine processing of tweets and for supporting queries to them, it might be useful to formally establish this relationship. Both #europeanchampionship2016 and #Europameisterschaft2016 could be marked as a variant of #EURO2016 (or of #euro2016) or vice versa.

While Declerck & Lendvai (2015b) describe a proposal for the formal representation of such hashtag variants, there is, at the best of our knowledge, not yet any implemented method for detecting and marking such hashtag variants in German tweets.

We expect this harmonization step to also improve results of queries addressed to social media, as this has already been suggested in (Berardi et al., 2011).

2 Related Work

Our investigation dealing with the harmonization and segmentation of German hashtag in Twitter posts is influenced by the work applied to hashtags used in English tweets (Declerck & Lendvai, 2015a). Kotsakos et al. (2015) are proposing a very interesting approach to the filtering of meme-hashtags, including German hashtags, but the hashtags harmonization step they implement is limited to lowercasing.

3 Use of Hashtags in German Twitter Texts

An interesting aspect of hashtags in English posts is that they are showing a move to compounding, generating more and more “glued” word constructions, which are not only in use in social media, but are also getting more popular in “classical” text. This is making word decomposition a more and more relevant task for the automated analysis of English.

Now, compounding is an important feature of German and there exist already some segmentation algorithms for the analysis of German text.[1] But we see that the “compounding” mechanisms applied to hashtags are showing relevant differences to the compounding rules applied to the generation of “normal” text. There is therefore a need to develop specific algorithms for the decomposition of German hashtags. And this is even more necessary if one considers the fact that German tweets are making a large use of hashtags, substantially more as in English tweets, as this has been reported in (Weerkamp et al., 2011) on a comparative study of Twitter texts in several languages. This study reveals that 14% of English tweets are tagged by a hashtag, whereas 25% of German tweets include a hashtag. And German tweets used significantly more hashtags

[1] See for example (Henrich & Hinrichs, 2011).

than English ones: 1.9 hashtags per tweet, compared to 1.4 for English tweets.

3.1 Examples of German Hashtags

In this section we present few examples of hashtags we found by just reading some German tweets.

In a first case we are dealing with normal German words, which can simply be lowercased for achieving the intended harmonization:

```
#europameisterschaft, #Europameisterschaft, #EUROPAMEISTERSCHAFT, #EUROPAmeisterschaft
#Europameisterschaft2016,
#europameisterschaft2016
```

But compared to English hashtags, we can see here that we have “classical” compounds within the hashtag, and thus no use of camelCase can further help for segmenting[2]. In this case we will use just standard segmentation algorithms for German.

In a second case, we are dealing with abbreviated hashtags, which can also be lowercased:

```
#EM2016, #em2016
```

While those cases are straightforward candidates for harmonization, it is a bit more challenging to reduce #Europameisterschaft2016 to #em2016, also due to the fact that no camelCase is used (in this case one could relate “E” and “M” to “em”). For this we take advantage of the use of non-classical compound effects, like the addition of digits at the end of the hashtag. An indicator is also given by the fact that those distinct hashtags are sometimes used in the same tweet, although this is more often the case with the use of the #EURO2016 hashtag. In fact the latter hashtag can be used as a pivot over tweets in different languages, but we are not dealing with multilingual issues in this study.

A third case is given by examples like:

```
#StandortDeutschland
#FußballEM2016
```

We assume that the use of camelCase notation in German tweets is really an indicator of non-classical compounding, so that we can segment the hashtag here, and possibly harmonize it with the following sequence: #Fußball-#Europameisterschaft 2016. The last example is an interesting one: grouping two hashtags via a hyphen sign, which seems to be specific to German hashtags.

A fourth case is given by:

```
#Brexit-Befürworter
#Brexit-Votum
```

This case is very similar to #Fußball-#Europameisterschaft with the difference that the word after the hyphen sign is not a hashtag. We can here harmonize to #Brexit – and ultimately to #brexit -- just storing the second word as a modifier.

A fifth and more complicated case is:

```
#warumeuropa
#bestemannschaft
#mussverscrapptwerden
```

Those cases show examples of real chunks or phrases included in one hashtag. It is still not clear if there is any advantage in trying to segment to cases.

4 Our Approach for Segmenting German Hashtags

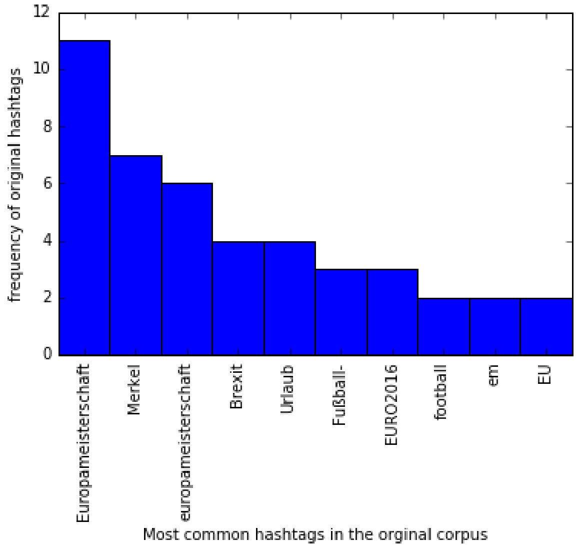
At the current stage of our investigation we have been implementing solutions for the four first cases mentioned in the preceding section. The data basis for our first experiment consists of 164 hashtag tokens just copied from some Twitter threads. The main topics were the 2016 European Championship in soccer[3] and the Brexit[4]. For now, we want to test some few algorithms on this small data set. In a next step we will apply and evaluate the algorithms to larger corpora.

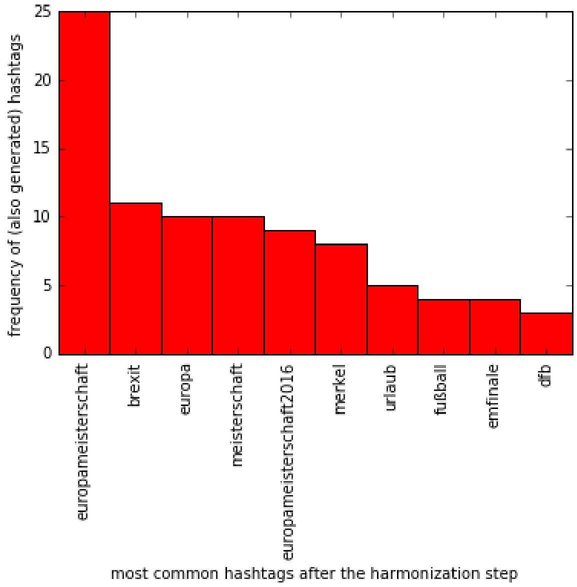
Below we display two charts showing first the 10 most frequent hashtags in our small data set before any segmentation and harmonization steps. The second chart shows the frequency of hashtags or words that result from the application of the current version of our segmentation and harmonization process to our data set.

[2] The intensive use of camelCase notation in English hashtags was a feature helping to segment those in the study reported in (Declerck & Lendvai, 2015a).

[3] https://en.wikipedia.org/wiki/UEFA_Euro_2016

[4] https://en.wikipedia.org/wiki/Brexit





We observe some significant changes in the ranking, so that in the second chart the term "brexit" is now the second in frequency (we had in the original data set both #Brexit" and "#brexit" as standalone hashtags but also compounds like "#Brexit-Befürworter" or "BrexitVote").

Interesting is also the emergence of the term "meisterschaft", which was not appearing as a standalone hashtag in the original collection. But as in this case they were some examples of camelCase notation, the term "meisterschaft" has been extracted by the algorithms.

We also observe the rise of frequency for the harmonized hashtag "europameisterschaft". This is not only resulting from the addition of the frequency of the uppercase term "Europameisterschaft", but also to an acronym resolution step linking "em" to "europameisterschaft" (and "EM" to "Europameisteschaft"). As a consequence, the hashtags "#em" and "#EM" are deleted from the ranking list, and other topics are now visible in this list. It is for now not clear which hashtag should be selected as the harmonized one: we expect the application scenarios to specify this point.

We are currently analyzing those first results, while we immediately see that we have to mark the context or the domain in which "europa" or "meisterschaft" are occurring. The same is valid for "brexit". This aspect is relevant for queries: we aim at suggesting this context to the users submitting the queries.

We are currently working on processing also the hashtags containing numerical and other special symbols and extending the investigation to a larger selection of hashtags, also in the full context of the tweets they are occurring in.

Acknowledgments

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

Reference

Giacomo Berardi, Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2011. ISTI @ TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking. Proceedings of the 20th Text Retrieval Conference (TREC 2011), Gaithersburg, US, 2011.

Thierry Declerck, Piroska Lendvai. 2015a. Processing and Normalizing Hashtags. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitko (eds.): Proceedings of RANLP 2015, Pages 104-110, Hissar, Bulgaria.

Thierry Declerck, Piroska 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.

Henrich, V. & E. Hinrichs (2011). Determining Immediate Constituents of Compounds in GermaNet. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria. pp. 420-426.

Dimitrios Kotsakos, Panos Sakkos, Ioannis Katakis, Dimitrios Gunopulos. 2015. Language agnostic meme-filtering for hashtag-based social network analysis. Social Network Analysis and Mining 5 (1), 1-14

Wouter Weerkamp, Simon Carter and Manos Tsagkias. 2011. How People use Twitter in Different Languages. In: Proceedings of Web Science.