Processing and Normalizing 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 ongoing work in linguistic processing of hashtags in Twitter text, with the goal of supplying normalized hashtag content to be used in more complex natural language processing (NLP) tasks. Hashtags represent collectively shared topic designators with considerable surface variation that can hamper semantic interpretation. Our normalization scripts allow for the lexical consolidation and segmentation of hashtags, potentially leading to improved semantic classification.

1 Introduction

The relevance of hashtags used in social media text, and more specifically in Twitter messages, has been recognized after some studies focused on the semantics that can be derived by such text constructs. For example, Laniado & Mika (2010) discussed whether hashtags behave as identifiers for Semantic Web applications. But the authors do not raise the issue of processing the hashtags in order to harmonize them, which would be necessary for gaining information on the specific semantics carried by hashtags.

Our observations are based on a large Twitter corpus dedicated to riots in the UK in the summer of 2011[1].Variants for hashtags that refer to the same topic abound, e.g. "#LondonRiots", "#londonriots", "#RiotsInLondon", "londonriot", "#LONDONRIOTS", and so on. We hypothesize that consolidating variants to a preferred hashtag form would benefit further tasks that draw on semantic similarity, such as the recently organized Semantic Textual Similarity Shared Task on Twitter data[2]. We have implemented a set of scripts in order to normalize the surface forms of hashtags. This includes case normalization, lemmatization and syntactic segmentation. We first describe related work, then our approach, and finally display some of our current results.

2 Related Work and Task Specification

(Pöschko, 2011) focuses on the detection of similar hashtags on the basis of their co-occurrences in a tweet. While the detection of co-occurrences is also present in our pipeline, we are additionally interested in detecting variants in order to reduce the amount of topics that hashtags designate. The expectation is that hashtag variants would all together represent only one topic.

(Antenucci et al., 2011) discuss an algorithm to learn the relationships between the literal content of a tweet and the types of hashtags that describe that content, which is one of our goals as well. Contrary to us, (Antenucci et al., 2011) do not suggest the harmonization (or reduction to a preferred form) of hashtags, but use similarity measurements between hashtags and words, while we implement patterns for explicitly relating variants of hashtags to a preferred form. We will use the results of their study for comparison with our approach.

(Costa et al., 2013) propose an approach that defines meta-hashtags by grouping the most used hashtags and their related hashtags into a meta-class, in order to improve the classification of

[1] This corpus was built on behalf of the newspaper „The Guardian", and its first objective was to gather data for tracking the emergence of rumours in social media. See http://www.theguardian.com/news/datablog/2011/dec/08/twitter-riots-interactive . An example usage of this corpus with NLP approaches for argumentation research is given in (Llewellyn et al., 2014).

[2] http://www.cis.upenn.edu/~xwe/semeval2015pit/

Proceedings of Recent Advances in Natural Language Processing, pages 104–109,
Hissar, Bulgaria, Sep 7–9 2015.

tweets. We aim at a reduced set of hashtag classes as well, but keeping normalization at the surface form level, without reaching a more abstract level. We promote the most frequent, lowercased hashtag variant to be the preferred form.

(Krokos and Samet, 2014) generate hashtags for tweets to be used as identifiers for NLP and Semantic Web applications, for which the preferred hashtag variant that we create would be directly of use.

Closely related work is presented by (Bansal et al., 2015). The authors seek to improve entity linking in tweets via semantic information provided by segmenting and linking entities that are present in a hashtag. Our approach is not limited to entities, but aims at covering the full lexical content of hashtags and targets general NLP scenarios.

Next to the normalization step that we mentioned above, syntactic parsing of hashtags would benefit retrieval tasks. A tweet could be more easily linked to other documents (e.g. documents from other genres that do not include hashtags, such as news articles). The query "#RiotsInLondon" would be less successful than the free-text query "Riots in London" or keyword query "Riots" and "London". Journalists, for example, need to establish verification links between a tweet and other sources in order to corroborate information in user-generated texts. Segmenting the text of the hashtag will allow using the derived components as search terms. The strong semantic and dependency relations between lexical components of the hashtag are typically not taken into account by search engines; this is why we aim to make these explicit.

(Bansal et al., 2015) discusses various algorithms for the segmentation of hashtags, like the Variable Length Sliding Window technique. We plan to investigate the application of this technique is the next steps of our work, but will use a simpler approach in the current study.

3 Harmonizing Hashtags

There are different ways of using hashtags in different languages: Spanish tweets are reported to contain much fewer hashtags than e.g. German tweets[3], while the use of CamelCase[4] notation seems to be much more popular in English tweets than in Spanish tweets[5].

Our first experiment is performed on a subset of the UK Riots corpus, selecting tweets between time stamps *2011-08-08/16:56:58* and *2011-08-08/17:18:53*. The subcorpus comprises 11,898 tweets. In this subcorpus 9,289 tweets are hapaxes. This yields a type-token ratio (TTR) of 78.07[6]. The subcorpus includes 16,716 hashtags tokens[7], but only 1,330 hashtags types, giving us a TTR of 7.95. We have 3,837 hashtags tokens and 188 hashtags types in CamelCase notation, yielding a TTR of 4.89.

Applying the simplest normalisation step – lowercasing all hashtags – leaves 1,156 hashtag types (TTR = 6.91). Lowercasing was applied 6,832 times. The number of matching between lowercased and original hashtags (in lowercase) is 5,921. From this figures we can see that this simple step is already reducing considerably the number of variants.

3.1 #LondonRiots

In order to show the relevance of lowercasing, the distribution of candidate variants of "#londonriots" in our subcorpus is displayed below in Figure 1.

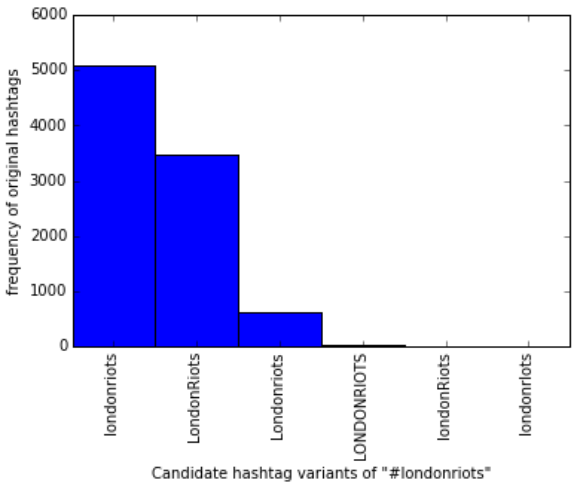

Figure 1: Distribution of candidate variants of #londonriots.

This gives us for those candidate variants a total of 9,218 harmonized hashtags (all original hashtags to be lowercased). We can see that this

[3] See (Weerkamp et al., 2011) on a cross-language study of the use of hashtags.

[4] An example of a CamelCased hashtag is „#LondonRiots". See https://en.wikipedia.org/wiki/CamelCase for more details on CamelCase.

[5] We still have to quantify this observation.

[6] TTR can be important for estimating the amount of noise in the corpus.

[7] The high number of hashtags might be collection-specific, the figures reported in (Weerkamp et al., 2011) for the use of hashtags in English tweets are lower.

harmonization step is considerably increasing the initial number of #londonriots hashtags (5,085). The tag "#londonriots" is promoted as the preferred form of these variants.8 Since the singular forms #londonriot, #LondonRiot (occurring respectively only six and one times in the subcorpus) and others are less frequent, we also replace these with the canonical plural form.

4 Segmenting Hashtags

Most of the English hashtags reflect a trend to use space-free compounding, e.g. "#LondonRiots", similar to e.g. German compounding. We also observe that the order of words in binary compounding follows the NE + N syntactical pattern, "#LondonRiots", while more complex compounding takes place via e.g. N + PP pattern "#RiotsInLondon". The dependency structure of this pattern makes it easier to determine the semantic head of the hashtag. In our example, the semantic head of the compound "#RiotsInLondon" is 'Riots'. Establishing a paraphrase relation to "#LondonRiots" allows us to state that also in the latter case the semantic head is "Riots" (although not being in the first position of the compound).

This approach for detecting paraphrases of compound terms has been investigated first in (Mihaela Vela, 2011). In a large corpus of German texts on financial topics, all detected binary compound words have been segmented. Then a search in the corpus was started in order to find within a small window of words the segments of the original compounds (in the reversed order) separated by either a preposition or a determiner in genitive case. This approach was effectively supporting the building of taxonomic structures from German compounds, since the paraphrases were offering additional semantics on relations between the components, marked by the preposition or by the genitive determiners. We apply a derived version of this approach to the (English) complex hashtags present in our UK Riots corpus.

First, we process hashtags that feature CamelCase notation: "RiotsInLondon" is segmented in 'Riots', 'In', and 'London'. We perform part-of-speech (POS) tagging[9] on the segments to check if they are part of the English vocabulary. This step is done in order to validate the segmentation. For our example, the NLTK default tagger delivers:

```
[('Riots', 'NNS'), ('In',
'IN'), ('London', 'NNP')]
```

And in this case, we are also lucky that no ambiguity is present in this tagged example, but the main purpose of tagging the resulting segments is to verify that there are no unknown words among the segments.

On the top of the results of the tagger, we are applying our SCHUG constituency and dependency parser[10]:

```
<NP
    TYPE="gen/5-attach_en/16"
    STRUCT="1_23_25"
    STRING="Riots In London"
    NP_HEAD_STEM="riot"
    NP_HEAD="Riots"
    NP_RULE="NP-PP"
    NP_MOD="[In London]">
    <TOKEN ORD="1" POS="1"
    TC="22"    STTS_POS="NNS"
    STEM="riot">Riots</TOKEN>
    <TOKEN STEM="in"
    STTS_POS="IN" TC="21"
    POS="23"
    ORD="2">In</TOKEN>
    <TOKEN ORD="3" POS="1"
    TC="22" STEM="London"
    STTS_POS="NNP">London</TO
    KEN>
</NP>
```

The main information for us is in this result the fact that the word "Riots" has been identified as the head of the segment, and "London" as part of the modifier.

Following strategies consisting in extracting semantic relations from dependency structures[11], we can infer that the sequence "RiotsInLondon" is a subclass of the class "Riots". Or that "RiotsInLondon" are an instance of "Riots". Another possibility consists in stating that the class "Riots" is equipped with a property "hasLocation".

While we are not now generating an ontological structure out of those segmented hashtags, we

[8] We are working on encoding all information in the emerging W3C standard 'Ontolex', cf. https://www.w3.org/community/ontolex/

[9] We use for this the part-of-speech tagger included in NLTK. See section 1 of http://www.nltk.org/book/ch05.html

[10] See (Thierry Declerck, 2002) for more details.

[11] See (Buitelaar et. al, 2004) and (Mihaela Vela, 2011).

are linking those with existing semantic resources in the Linked Open Data cloud[12]. We applied for this the sparqlwrapper module for Python[13], an example of which given just below:

```
from SPARQLWrapper import
SPARQLWrapper, JSON

sparql = SPARQLWrap-
per("http://dbpedia.org/sparql")

sparql.setQuery("""
PREFIX rdfs:
<http://www.w3.org/2000/01/rdf-
schema#>
PREFIX owl:
<http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl:
<http://dbpedia.org/ontology/>
PREFIX dct:
<http://purl.org/dc/terms/>
   SELECT ?var
   WHERE
{<http://dbpedia.org/resource/Riot>
dct:subject  ?var }
   """)
sparql.setReturnFormat(JSON)
results = sparql.query().convert()
for result in re-
sults["results"]["bindings"]:
   print(result["label"]["value"])
```

In the example above the reader can see that we linked the hashtags “#riots”, “#Riots”, “#riot” and “#Riot” to a DBpedia entry, which is named “Riot”. Since we segmented hashtags like “#LondonRiots” or “#TottenhamRiots”, we can also link their “Riots” segments to this DBpedia entry, while the segments “London” and “Tottenham” can be linked to the corresponding DBpedia entries. At the end, we can compute the information that we are dealing with riots in UK cities.

The process is the same for both hashtag types "#LondonRiots" and “#RiotsInLondon”, since we established that both hashtags are paraphrases of each other[14]. The process of segmentation helps gaining evidence that the main topic of the corpus is riots; while specific locations can be designated by specific hashtags, e.g. “#hackneyriots”. Next, the components extracted from camel notation hashtags are used for supporting the segmentation of similar hashtags that are not written in camel case notation. For example “#londonriots” could be segmented into “london” and “riots” (as a reminder, the hashtag “#londonriots” is occurring 5,085 times in the corpus used in our experiment), or “#riotpolice” into “riot” and “police” (this hashtag in this form occurring only twice, and also twice in the CamelCase form).

The segmentation step can provide information about the number of semantic units located in the hashtagged text; this has been found in previous studies[15] indicative about the level of factuality expressed by a hashtag. Below we see two examples of segmented hashtags. The counts represent the position and frequency of the components of a compound hashtag.

```
'#LondonRiots' => {
      '0' => {'London' => 3461},
      '1' => { 'Riots' => 3461},
      'freq' => 3461},

'#SouthernFairiesCantHandleTheirWineGums' => {
  '0' => { 'Southern' = 1  },
  '1' => { 'Fairies' => 1 },
  '2' => { 'Cant' => 1 },
  '3' => {'Handle' => 1 },
  '4' => {'Their' => 1 },
  '5' => {'Wine' => 1 },
  '6' => {'Gums' => 1 },
  'freq' => 1
  },
```

#LondonRiots occurs 3,461 times. We then just add this frequency to the components of the compound. We can add this figure to the number of occurrences of the single hashtags “#Riots” and “#riots” (originally with a total of 1,096 occurrences, now with a total of 9778 occurrences), giving more evidence that a major topic of the corpus is “riots”. This evidence is increasing still when we consider the cases of “#HackneyRiots” and the like. The increase of frequency of the from our algorithm partly generated hashtag candidates “#riots” and “#Riots” is show in **Figure 2** and **Figure 3**. Looking at the values for “#riots” we see a dramatic increase of frequency for the terms “riots” and “Riots”, but also significant changes for the names of locations.

Additionally, we observed that both of the components of “#londonriots” and similar are often co-occurring in tweets in a non- or only a

[12] See http://lod-cloud.net/ for more details.
[13] See https://rdflib.github.io/sparqlwrapper/
[14] As mentioned earlier, we adapted for this a method used for German, consisting in searching for paraphrases of compounds. See (Mihaela Vela, 2011).

[15] Kotsakos et al. (2014) suggest that the length of a hashtagged text as one of the features that help in differentiating meme tweets from event reporting tweets: the longer the hashtagged text, the higher the probability that the tweet is a meme.

partly hashtagged form (like "London #riots"). This fact can give supplementary evidence that the segmentation of the hashtags was well motivated.

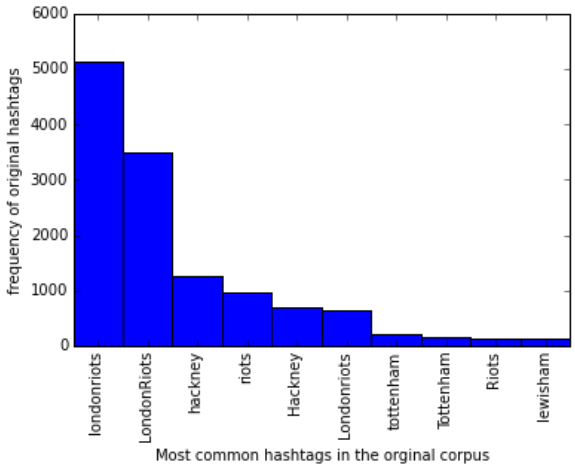

Figure 2: Most common hashtags in the original corpus.

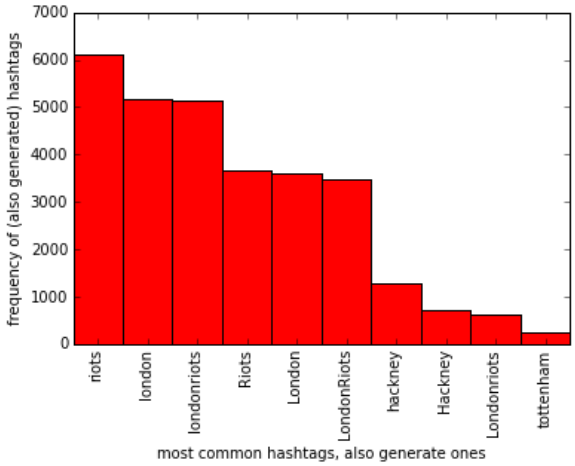

Figure 3: Increased frequency of certain hashtags, after segmentation of complex hashtags.

Related to this, we displayed above the example of the segmentation of the longer hashtag string "SouthernFairiesCantHandleTheirWineGums". We observed that none of the components are co-occurring in any relevant way in the tweets of our corpus. This can lead to the classification of such hashtags as spam or as not factual. This would be in accordance with the findings by (Kotsakos et al. 2014), stating that longer hashtags tend to not represent facts.

Finally we computed the frequency of usage of each word in different compound hashtags. We display below the example for "Riots". In this representation we also provide for information on the position of the components of the compounds: the word Riots is in the first position within the compound hashtag "#RiotsAffectOthers" ("0 =>"), etc. The representation provides thus contextual information of the word "Riots" when used in distinct hashtags.

```
'Riots' => {
        '0' => {
              '#RiotsAffectOthers' => 1
             },
        '1' => {
              '#BirminghamRiots' => 2,
              '#CroydonRiots' => 1,
              '#EnfieldRiots' => 1,
              '#HackneyRiots' => 9,
              '#LondonRiots' => 3461,
              '#StopRiots' => 1,
              '#StopRiotsInLondon' => 1,
              '#TottenhamRiots' => 9
             },
        '2' => {
            '#Hackney#LondonRiots'=> 1,
              '#NorthLondonRiots' => 1,
              '#StopTheRiots' => 1
             },
        'freq' => 3489
       },
```

5 Current Work

We are currently investigating if our approach can help in concrete applications. In one scenario, hashtag normalization is used to preprocess tweets in a tweet-vs-document similarity task. Similarity is computed by means of string alignment (across a tweet and each of the sentences of a document), and we hypothesize that hashtag normalization would allow for more matching.

In a second application we are aiming at improving the output of cluster algorithms applied to our data. In a preprocessing phase we normalized hashtags and we could already observe that the behavior of the used clustering algorithm (included in the NLTK package) was sensitive to this kind of lexical variation.

Finally, we started to investigate if and how Textual Entailment can be applied to social media text. We are using for this the Excitement Open Platform (EOP)[16]. Since one algorithm deployed in EOP is making strong use of detection of paraphrases, in order to support the system in recognizing similar statements, it is important to either add unifying semantic information to the text segments under entailment judgement and/or

[16] See http://hltfbk.github.io/Excitement-Open-Platform for more details.

to apply methods for reducing the lexical variety of the text segments (supporting the detection of longer matching segments between two text snippets). Our work on the segmentation and harmonization of hashtags is the first step for the investigation on the use of TE for Twitter text. An evaluation of our approach is currently on the way and will be reported soon.

Acknowledgments

The work described in this paper is partly funded by the European FP7 project PHEME ("Computing Veracity across Media, Languages, and Social Networks"), under agreement no: 611233.

References

D. Antenucci, G. Handy, A. Modi, and M. Tinkerhess. 2011. Classification of tweets via clustering of hashtags". EECS 545 FINAL PROJECT, FALL 2011.

Piyush Bansal, Romil Bansal and Vasudeva Varma. 2015. Towards Deep Semantic Analysis of Hashtags. In Proceedings of the 37th European Conference on Information Retrieval *(*ECIR 2015*)*, Vienna, Austria.

Paul Buitelaar, Daniel Olejnik, Mihaela Hutanu, Alexander Schutz, Thierry Declerck, Michael Sintek. (2004). Towards Ontology Engineering Based on Linguistic Analysis. Proceedings of International Conference on Language Resources and Evaluation

Joana Costa, Catarina Silva, Mário Antunes and Bernardete Ribeiro. 2013. Defining Semantic Meta-Hashtags for Twitter Classification. In Proceedings of the 11th International Conference on Adaptive and Natural Computing Algorithms, ICANNGA 2013, Lausanne, Switzerland, April 4-6, 2013. Proceedings

Thierry Declerck. (2002). A set of tools for integrating linguistic and non-linguistic information. In Proceedings of SAAKM (ECAI Workshop).

Genevieve Gorrell, Johann Petrak, Kalina Bontcheva 2015. LOD-based Disambiguation of Named Entities in @tweets through Context #enrichment. In *Proceedings of ESWC 2015*, Portoroz, Slovenia.

Dimitrios Kotsakos, Panos Sakkos, Ioannis Katakis and Dimitrios Gunopulos, 2014. "#tag: Meme or Event?" In *Proceedings of ASONAM 2014*, Beijing, China

Eric Krokos Hanan Samet. 2014. A Look into Twitter Hashtag Discovery and Generation. In *Proceedings of the 7th ACM SIGSPATIAL Workshop on Location-Based Social Networks (LBSN'14)*, Dallas, TX, November 2014

David Laniado and Peter Mika. 2010. Making sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference*. Shanghai, China, November 2010.

Clare Llewellyn, Claire Grover, Jon Oberlander and Ewan Klein. 2014. Re-using an Argument Corpus to Aid in the Curation of Social Media. In *Proceedings of the 9th Language Resources and Evaluation Conference*, 26-31 May, Reykjavik, Iceland

Jan Pöschko. 2011. Exploring Twitter Hashtags. CoRR abs/1111.6553.

Mihaela Vela, Extraction of Ontology Schema Components from Financial News. (2011). PhD Thesis, Saarbrücken

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, Pages 1031-1040

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