Twitter hashtags: Joint Translation and Clustering

Simon Carter
ISLA
University of Amsterdam
Science Park 904,
1098 XH Amsterdam
The Netherlands
s.c.carter@uva.nl

Manos Tsagkias
ISLA
University of Amsterdam
Science Park 904,
1098 XH Amsterdam
The Netherlands
e.tsagkias@uva.nl

Wouter Weerkamp
ISLA
University of Amsterdam
Science Park 904,
1098 XH Amsterdam
The Netherlands
w.weerkamp@uva.nl

ABSTRACT

The popularity of microblogging platforms, such as Twitter, renders them valuable real-time information resources for tracking various aspects of worldwide events, e.g., earthquakes, political elections, etc. Such events are usually characterized in microblog posts via the use of hashtags (#). As microbloggers come from different backgrounds, and express themselves in different languages, we witness different "translations" of hashtags which, however, are about the same event. Language-dependent variants of hashtags can possibly lead to issues in content-analysis. In this paper, we propose a method for translating hashtags, which builds on methods from information retrieval. The method introduced is source and target language independent. Our method is desirable, either instead of, or complimentary, to the direct translation of the hashtag for three reasons. First we return a list of hashtags on the same topic, which takes into account the plurality and variability of hashtags used by microbloggers for assigning posts to a topic. Second, our framework accounts for the problem that microbloggers in different languages will refer to the same topic using different tokens. Finally, our method does not require special preprocessing of hashtags, reducing barriers to real-world implementation. We present proof-of-concept results for the given Spanish hashtag *#33mineros*.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis, machine translation*

General Terms

Algorithms

Keywords

Twitter, hashtag, translation, clustering

1. INTRODUCTION

Microblogging platforms such as Twitter have become important real-time information resources [5], with a broad range of uses and applications [1, 7, 10]. A particular feature of Twitter are its hashtags: a short-hand convention adopted by microblog users to manually assign their posts to a wider corpus of messages on the same topic. They are denoted with the # symbol preceding a short string, often a name or abbreviation. In this way, hashtags are simple way to make the wide variety of published microblog material searchable [3], and they serve to give accurate and timely statistics about trending topics of posts on the platform [9]. As users contribute content from different geographical regions and language, we witness a multitude of hashtags variants, which refer to the same topic. These variants can introduce bias in the statistics which, ultimately, may not reflect the true underlying global topic distributions.

To illustrate this, let us look at a specific example. On August 5, 2010, 33 men were trapped 700 meters below ground in a mining accident at the San José copper-gold mine in Chile. The reaction of people was recorded in status updates originating from all around the world. While Spanish speaking people were referring to this topic using hashtags such as #33mineros, #rescatemineros, #fuerzamineros, English speaking people used different hashtags: #chilemine, #chileanminers, #minerescue.

The example above shows two main issues regarding the use of hashtags in different languages: (1) tags, or part of them, are literal translations, and, more interesting, (2) the specificity of the hashtag is different per location. As to (1) , people could use "simple" translations, going from #33mineros to #33miners. As to (2) , we observe that people in Chile refer to this event with quite specific tags, whereas English tags also include general terms like #Chile. From a non-Chilean perspective, the mining accident is "the only thing" happening in Chile at that moment.

A standard approach to hashtag translation would be to use statistical machine translation (SMT) methods, where a SMT system is asked to translate a hashtag h from a source language s to a target language t. We identify two potential problems with this approach: (a) hashtags may not be proper words, raising sparsity issues, and (b) the direct translation may not map to a commonly used equivalent in the target language. To this end, we believe that SMT methods are not sufficient for hashtag translation.

In this work, we propose a method that is robust, in that it can deal with the two issues raised before. It builds on the translation of a hashtag "profile", and uses the translation to retrieve microblog posts in the target language. From this set, we can extract hashtags that refer to the same topic as the original hashtalk.

2. RELATED WORK

The use of tags has been examined for standard natural language processing tasks, such as supervised sentiment classification [1] and sarcasm detection [2], event detection [10, 11], and information diffusion [7, 9]. However, as well as using them as features in wider systems, academics have recently focused on hashtag specific problems, including ranking them with relation to their interestingness [12], retrieving hashtags [3], and finally [4] look at the problem of tag recommendation.

The framework for translating hashtags proposed in this paper is based upon the previous work on extracting translations from noisy parallel corpora [8]. The difference here is that we do not aim to create a bitext or translation directory, but find translations specifically for hashtags that refer to the same topic as the source tag. Our task can therefore be viewed as topic matching between across languages.

3. METHOD

In Algorithm 1 we present the framework we use to generate translation candidates for a specific foreign hashtag h in H, the set of hashtags we are interested in. In line 3 we retrieve a large set of tweets that contain the foreign hashtag we are interested in (for example, using the Twitter API). Given the retrieved set of tweets T containing the hashtag h, we translate each of them (for example, using a translation engine such as MOSES [6]) into the target language (line 5).

Algorithm 1 Hashtag translation algorithm.

```
 1: $\vec{w} \leftarrow 0$
 2: for h in H do
 3:     $T \leftarrow RETRIEVE(h)$
 4:     for t in T do
 5:         $TO \leftarrow TO + TRANS(t)$
 6:     end for
 7:     $q \leftarrow CREATEQ(TO)$
 8:     $D \leftarrow RETRIEVE(q)$
 9:     $w_h \leftarrow RANK(D)$
10: end for
11: return $\vec{w}$
```

The $CREATEQ$ function in line 7 takes as input the translated tweets and returns a query in the target language. For this, we order terms in the translated text using their TF-IDF. The advantage of using TF-IDF is that it ranks terms highly that (1) occur often in the translated text, but (2) do not occur too often in "regular" text. We take the top-N highest rank terms $t_1, t_2, t_3 ... t_N$ and create a query consisting of these terms. We issue the query against the microblogging platform in line 8, and extract the hashtags from the returned tweets D. Finally, we rank these hashtags in line 9 according to our ranking function $RANK$. For now, our ranking is based solely on frequency, but more advanced methods are possible here (e.g., discounting very common hashtags).

4. EXPERIMENTS AND RESULTS

As a proof-of-concept, we test our framework using as input the Spanish hashtag #33mineros (see Section 1). We use Topsy[1] to retrieve tweets on this topic. We then translate these tweets using the Google API, and feed the translated text to our $CREATEQ$ function (defined in Algorithm 1). The result of this function is a list of terms that we can use as query: these terms are listed in Table 1 (left). The resulting top scoring hashtags related to tweets returned by this query are displayed in Table 1 on the right.

Query terms	Hashtags
rescue	#Chile
rescued	#chilemine
miners	#fuerzamineros
miner	#CNN
trapped	#miners
capsule	#Chilean
Chilean	#chileanminers
	#chileminers
	#rescue
	#minerescue

Table 1: (Left) English terms used as query. (Right) Top 10 hashtag translations for #33mineros.

From the resulting hashtags we conclude that almost all of them are suitable as translation of the original hashtag (#33mineros). Only the #CNN tag is not relevant, given that this tag is put to all CNN tweets. As mentioned before, more intelligent ranking of hashtags (using collection statistics) can take care of this. The tag #Chile ranks highest, which shows the issue raised in the introduction regarding specificity and locality. As we can see from some example tweets in Table 2, English tweets do often use this hashtag to refer to the Chilean mining incident. This example illustrates why a direct translation method could lead to unhelpful translations, as hashtags used for referencing global topics in one language do not match those used in an other.

#Chile rescue worker reaches bottom of emergency shaft to cheers of trapped miners
1st #Chile rescue worker Manuel Gonzalez sent down in capsule, with presidential sendoff, for 3rd test
Installation of the rescue capsule in #Chile has begun - NBC News
Awaiting rescue of 9th trapped Chilean miner. #CNN #chilemine #minerescue

Table 2: Tweets returned using the query generated by our method.

5. CONCLUSIONS

We have presented a framework that learns the corresponding target side hashtags for a given foreign hashtag. As opposed to directly translating the hashtag, we rely on words that occur alongside it to extract target language hashtags that belong to the same topic, and are thus translations of the original source tag.

We have demonstrated how this framework improves upon the a direct translation of the hashtag in three ways. First, our framework returns a list of hashtags on the same topic, which takes into account the plurality and variability of hashtags on the same topics. Second, our framework accounts for the problem that microbloggers in different languages will refer to the same topic using different tokens. Finally, our method does not require special preprocessing of hashtags, such as word-segmentation, spelling correction and token normalisation.

For future work, we plan on conducting a full analysis of the components defined in our framework, and presenting detailed empirical findings.

[1] http://topsy.com

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.061.815.

References

[1] D. Davidov, O. Tsur, and A. Rappoprt. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceeding of the 23rd international conference on Computational Linguistics (COLING)*, 2010.

[2] D. Davidov, O. Tsur, and A. Rappoprt. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceeding of Computational Natural Language Learning (CoNLL)*, 2010.

[3] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, 2010.

[4] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth. Towards tagging and categorization for micro-blogs. In *21st National Conference on Artificial Intelligence and Cognitive Science (AICS)*, 2010.

[5] G. Golovchinsky and M. Efron. Making sense of twitter search. In *Proceedings of CHI 2010 Workshop on Microblogging: What and How Can We Learn From It?*, 2010.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 2007.

[7] J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how twitter is used to spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[8] D. S. Munteanu and D. Marcu. Improving machine translation performance by exploiting non-parallel corpora. 2005.

[9] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the Conference on the World Wide Web*, 2011.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web (WWW 2010)*, pages 851–860, 2010.

[11] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI 2010)*, pages 1079–1088, 2010.

[12] J. Weng, E.-p. Lim, Q. He, and C. Wing-ki Leung. What do people want in microblogs? measuring interestingness of hashtags in twitter. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, 2010.