Mapping Microblog Posts to Encyclopedia Articles*

Uta Lösch and David Müller
Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany
uta.loesch@kit.edu,
david.mueller@student.kit.edu

Abstract: Microblog posts may contain so-called hashtags that mark keywords or topics. These hashtags are used in an ad-hoc fashion; their meaning is implicitly defined via their use, which makes understanding and querying hashtags difficult. In this paper we devise a method for annotating microblog posts which contain hashtags with related encyclopedia entities. Thus, users have the means to quickly grasp the meaning of a hashtag, and find starting points for further exploration of the hashtags' context. We implement our method based on Twitter and an existing system for linking content to Wikipedia.

1 Introduction

Microblogging services allow users to publish short messages online. Twitter[1] is the largest dedicated microblogging site; as of March 2011, its users create an average of 140 millions posts a day[2]. Given a 140-character limit for posts, Twitter users invented shortcuts for certain expressions. So-called hashtags (starting with a # sign followed by a keyword) are frequently used to associate posts with a specific topic, place, person or event. For example, posts covering current events in Lybia are tagged with $\#Libya$.

Hashtags are useful for organising microblogging messages and for highlighting topics of a specific message. Thus, it becomes possible to search for messages on a topic by searching for the hashtag used for the topic. Microblogging platforms support this search for specific hashtags. However, as hashtags are implicitly defined via their use, the meaning of a hashtag is often unclear. Services such as tagdef[3] provide means to provide a definition of a hashtag, but these services rely on user input and thus only cover a subset of all hashtags. In addition, since hashtags are not related to other structured information sources, querying is limited to direct keyword search.

In this paper we propose a method for associating hashtags with encyclopedia entities. We envision a system which automatically finds entities equivalent to a hashtag. However, this

*The work presented in this paper has been supported by the European Community's Seventh Framework Programme FP7/2007-2013 (PlanetData, Grant 257641) and by the German Research Foundation (DFG) in scope of the project Multipla (Grant 38457858

[1] http://twitter.com/
[2] http://blog.twitter.com/2011/03/numbers.html
[3] http://www.tagdef.com/

proves impossible for many hashtags, as by far not all hashtags represent entities [LM10]. In our system we therefore aim at finding related entities not necessarily equivalent ones. However, if a hashtag can be associated with an entity with high confidence, the entity can be seen as descriptive for the hashtag. In our system we use Wikipedia as entity source which covers a wide range of topics. Furthermore, automatic tools for annotating text with Wikipedia entities are readily available, for example Milne and Witten's Wikifier [MW08]. Wikipedia entities are widely linked to external data sources via DBpedia [BLK+09]. Thus, in our system we enable combined querying of DBpedia and Twitter.

Matching entities in microposts is a difficult problem [AAG+10] as microblog posts are very short and contain little information. Our approach uses search results for a given hashtag as input to the entity matching component and returns messages using the hashtag including a set of related Wikipedia entities. The matched entities not only help to understand the search terms, but also serve as a starting point for refining the search or searching for further information related to the search result.

Our contributions thus are an approach for finding entities which are related to search results on microblogging platforms and an implementation of our method based on Twitter as microblogging platform and on Wikipedia as source of entities.

The remainder of this paper is organised as follows: an overview of the approach is given in Section 2, its implementation is presented in Section 3. We discuss related work in Section 4 before concluding in Section 5.

2 Method Overview

The goal of our method is, given a search query q as input, to provide the user with an RDF[4] document describing the search result R, i.e. the most recent messages that match the query, the authors of these messages and the most relevant entities E for this result. The architecture of our system is shown in Figure 1. In the following we will illustrate our approach based on a search for the hashtag $\#Libya$, a hashtag which is frequently used to describe posts related to the country of Libya.

In a first step, the most recent search results R for the query are obtained. In the example of searching for $\#Libya$ a document containing the most recent Twitter posts whose content matched the search string $\#Libya$ and their meta information will be returned. An example[5] for a single retrieved post is the following:

```
"Arab League calls for no-fly zone in #Libya http://t.co/ZsLNIWa"
```

Additional information such as who posted the message (Twitter user `SenJohnMcCain`) and where and when the message was posted are returned[6] is also returned.

In a second (optional) step hyperlinks posted in messages of the fetched feed are followed, content of the referenced website is fetched. The idea is to use the websites' content as

[4] `http://www.w3.org/RDF/`
[5] `http://twitter.com/SenJohnMcCain/statuses/46619460060196865`
[6] The original Twitter message posted by SenJohnMcCain was not geo-tagged

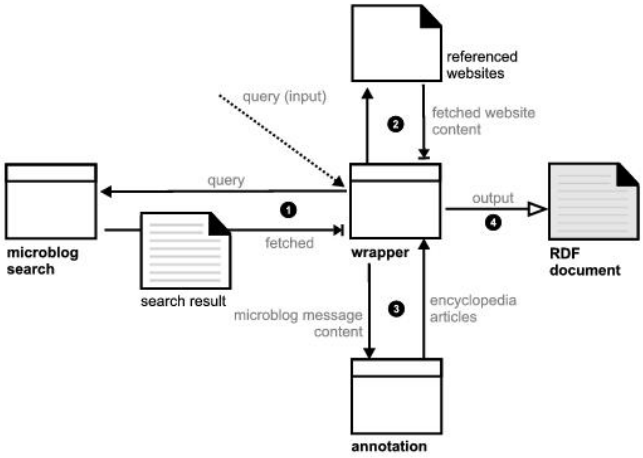

Figure 1: System Architecture

additional signal when searching for entities (as proposed by Cilenk et al [CAS11]). In our example, the message contains a link to `http://t.co/ZsLNIWa`, which redirects to an article on the situation in Libya published on `http://reuters.com/`. Dereferencing this URL and grabbing the text found at the link returns the following text:

```
"Arab League calls for Libya no-fly zone-state TV - CAIRO, March 12
(Reuters) - The Arab League on Saturday called on the U.N. Security
Council to impose a no-fly zone on Libya, Egyptian state television
reported, a decision that would give a regional seal of approval that
NATO has said is needed for any military action."
```

In a third step entities matching the search result are retrieved. The content of the Twitter messages that matched the search query and, if available, the content of referenced websites is merged to a single input string and used as input for the annotation. As result a list of matching articles of the English Wikipedia is obtained. In our example, we find entities which are related to Libya, like Libya, Gaddafi or the Arab_League.

Finally, an RDF document representing the search results and the annotations is generated. In our example, the output RDF document contains the following triples to reference the DBPedia mappings:

```
<http://km.aifb.kit.edu/services/twittersearchwrap/searchwrap
?q=\%23lybia>
     rdfs:seeAlso <http://dbpedia.org/resource/Lybia> ,
        <http://dbpedia.org/resource/Gaddafi> ,
        <http://dbpedia.org/resource/Arab\_League> .
```

Note that these entities are associated with the whole search result and not with individual messages. The individual messages are however also described in the system output. The description for the example post results in the following triples:

```
<http://km.aifb.kit.edu/services/twitterwrap/statuses/show/
```

```
46619460060196865\#id>
   foaf:page <http://twitter.com/SenJohnMcCain/statuses/
      46619460060196865>;
   dc:date "2011-03-12T17:12:00Z";
   geo:lat "49.010239" ;
   geo:long "8.411879" ;
   foaf:maker <http://km.aifb.kit.edu/services/twitterwrap/users/
      show?screen\_name=SenJohnMcCain\#id> ;
   dc:description "Arab League calls for no-fly zone in \#Libya
      http://\newline t.co/ZsLNIWa" .
```

3 Implementation

We provide an implementation of our approach in the Twitter Search Wrapper[7] which is based on a set of existing components. Twitter search results are obtained by using the Twitter search API[8], a RESTful API which delivers search results in JSON. The result is a feed containing the 100 most recently published Twitter messages, which are publicly visible and match the query, and their authors. The messages themselves are described by their content, publishing date, authors and optionally a geographic location. In the optional step of dereferencing URLs posted in messages contained in the result set only the first 50 words of the referenced sites are used (to keep input size manageable in the next steps). The content retrieved from external websites is beautified for our purpose removing HTML-tags and scripting elements using the CyberNeko Java HTML Parser Library[9]. The Wikifier [MW08] is used for obtaining content annotations. This tool takes a text as input and returns a set of Wikipedia entities which are relevant for the analysed text. It represents a state-of-the-art annotation tool and was chosen as it allows for finding entities not from some tool-specific knowledge base but from DBpedia, which is one of the most widely used data sources on the Linked Data Web. The Wikifier is available in several languages, four of which (English, German, Spanish, French) can be used in our system. If the language of the analysis is not specified beforehand, the language of each message in the search result is guessed using JLangDetect[10]. The system generates and outputs an RDF document in RDF/XML Syntax containing a description of the retrieved messages matching the search query for the given query and references to the DBpedia articles that have been returned by the Wikifier. The annotations are added to the search result using `rdf:seeAlso` links to DBpedia entities [BLK+09] which are obtained by a direct transformation of the Wikipedia entities' URLs. DBpedia has been chosen as these entities allow for a more coherent presentation of the content in the context of RDF.

[7] Online at `http://km.aifb.kit.edu/services/twittersearchwrap/`

[8] `http://dev.twitter.com/doc/get/search`

[9] `http://sourceforge.net/projects/nekohtml/`

[10] `http://www.jroller.com/melix/tags/jlangdetect` using the European Parliament Proceedings Parallel Corpus (`http://www.statmt.org/europarl/`) as training data

4 Related Work

Bringing Microblogging and Semantic Web technologies together, has been proposed several times. The SemanticTweet[11] service is a tool which generates a FOAF file describing one's network of followers and friends on Twitter. Thus, an automatic way of mapping the Twitter social network to semantic data is made available. The transformation is syntactic and does not include links to topics. Laniado and Mika [LM10] have studied how hashtags are used on Twitter. In their evaluation they showed that about 50% of the hashtags can be mapped to entities in Freebase. Passant et al. [PHBB08] have proposed a data model which allows for the association of URIs with users, microblogs and microposts. To this end, SIOC and FOAF vocabularies are used and extended. Additionally, so-called semantic hashtags, i.e. the use of URIs as hashtags (e.g. *#geo:Paris_France*), are proposed. It becomes thus possible to link microposts to entities on the Linked Data Web. In contrast to this approach we do not require users to change their behavior by requiring them to use another type of hashtags. Instead, our method finds DBpedia entities which are related to the query automatically. Softic et al [SEM+10] have developed a framework and system for mining data from social networks. The system has been instantiated to analyse data from Twitter. Data from one or several users is collected by the system and transformed to semantic data using the system by Passant et al. [PHBB08]. Enriching the result data using entities from other Linked Open Data vocabularies like DBpedia[12] or GeoNames[13] is proposed, but no actual method for this annotation is provided.

Relating semantic concepts with Twitter messages has been attempted by various authors: Stankovic et al. [SRL10] propose to use the Zemanta[14] keyword extraction API to detect topics of single Twitter messages. Instead of mining single messages which provide little signal to the extraction algorithm we use an aggregation of messages as input to the extraction component. Similar to our approach, Wagner [Wag10] proposes to use aggregations of social awareness streams for knowledge extraction. As such, our approach is closely related. Kinsella et al. [KPB10] use hyperlinks in posts to derive a richer representation of those posts. We too use external data to augment the set of Twitter messages supplied to the Wikifier system in a variant of our method. The challenge task in the WePS-3 Online Reputation Management Task[AAG+10] requires systems to detect mentions of company names in Twitter messages. The corpus contains evaluation results solicited via Mechanical Turk. In contrast to the WePS-3 task, we associate generic hashtags to Wikipedia/DBpedia concepts.

5 Conclusion and Future Work

We have presented an approach for annotating Twitter hashtags with DBpedia entities and provide an implementation of the method. The results obtained by our system serve as a

[11] http://semantictweet.com/
[12] http://dbpedia.org/
[13] http://www.geonames.org/
[14] http://developer.zemanta.com/

starting point for exploration of topics related to a specific search on Twitter.

The next step in our work will be an evaluation of the relevance of the found entities for the search term as well as of the stability of the found entities over time (we assume that if an entity is descriptive for a given search, then it should be annotated in the search result at any time that the search is executed).

References

[AAG+10] Enrique Amigó, Javier Artiles, Julio Gonzalo, Damiano Spina, Bing Liu, and Adolfo Corujo. WePS3 Evaluation Campaign: Overview of the On-line Reputation Management Task. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[BLK+09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 2009.

[CAS11] Ilknur Celik, Fabian Abel, and Patrick Siehndel. Towards a Framework for Adaptive Faceted Search on Twitter. In *Proceedings of the International Workshop on Dynamic and Adaptive Hypertexts: Generic Frameworks, Approaches and Techniques*, June 2011.

[KPB10] Sheila Kinsella, Alexandre Passant, and John G. Breslin. Using hyperlinks to enrich message board content with linked data. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 1:1–1:9. ACM, 2010.

[LM10] David Laniado and Peter Mika. Making sense of Twitter. In *9th International Semantic Web Conference (ISWC2010)*, November 2010.

[MW08] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518. ACM, 2008.

[PHBB08] Alexandre Passant, Tuukka Hastrup, Uldis Bojars, and John Breslin. Microblogging: A Semantic and Distributed Approach. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, 2008.

[SEM+10] Selver Softic, Martin Ebner, Herbert Mühlburger, Thomas Altmann, and Behnam Taraghi. @twitter Mining#Microblogs Using #Semantic Technologies. In *Proceedings of the 6th Workshop on Semantic Web Applications and Perspectives (SWAP 2010)*, 2010.

[SRL10] Milan Stankovic, Matthew Rowe, and Philippe Laublet. Mapping Tweets to Conference Talks: A Goldmine for Semantics. In *Procedings of the Third Social Data on the Web Workshop*, 2010.

[Wag10] Claudia Wagner. Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams. In *Proceedings of the 7th Extended Semantic Web Conference – Doctoral Consortium*, pages 493–497, 2010.

ISBN 978-3-88579-286-4