Adding Semantics to Microblog Posts (Abstract)*

Edgar Meij
edgar.meij@uva.nl

Wouter Weerkamp
w.weerkamp@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam
Amsterdam

ABSTRACT

In recent years Twitter has become one of the largest online microblogging platforms. Microblogging streams have become invaluable sources for many kinds of analyses, including online reputation management, news and trend detection, and targeted marketing and customer services [4, 9]. Searching and mining microblog streams offers interesting technical challenges, because of the sheer volume of the data, its dynamic nature, the creative language usage, and the length of individual posts [3].

In many microblog search scenarios the goal is to find out what people are saying about concepts such as products, brands, persons, et cetera. Here, it is important to be able to accurately retrieve tweets that are on topic, including all possible naming and other lexical variants. So, it is common to manually construct lengthy keyword queries that hopefully capture all possible variants. We propose an alternative approach, namely to determine what a microblog post is about by automatically identifying *concepts* in them. We take a concept to be any item that has a unique and unambiguous entry in a well-known large-scale knowledge source, Wikipedia.

Little research exists on understanding and modeling the semantics of individual microblog posts. Linking free text to knowledge resources, on the other hand, has received an increasing amount of attention in recent years. Starting from the domain of named entity recognition, current approaches establish links not just to entity types, but to the actual entities themselves [5]. With over 3.5 million articles, Wikipedia has become a rich source of knowledge and a common target for linking; automatic linking approaches using Wikipedia have met with considerable success [2, 6, 8].

Most, if not all, of the linking methods assume that the input text is relatively clean and grammatically correct and that it provides sufficient context for the purposes of identifying concepts. Microblog posts are short, noisy, and full of shorthand and other ungrammatical text and provide very limited context for the words they contain [3]. Hence, it is not obvious that automatic concept detection methods that have been shown to work well on news articles or web pages, perform equally well on microblog posts.

We present a robust method for automatically mapping tweets to Wikipedia articles to facilitate social media mining on a semantic level, involving a two-step method for semantic linking. Our main contributions are: (i) a robust, successful method for linking tweets to Wikipedia articles, based on a combination of high-recall concept ranking and high-precision machine learning, including state-of-the-art machine learning algorithms, (ii) insights into the influence of various features and machine learning algorithms on the task, and (iii) a reusable dataset, with which we aim to facilitate follow-up research.

The goal of the first step of our proposed method is obtain high recall, so we generate a ranked list of candidate concepts for each word n-gram in a tweet. Various methods exist for creating a ranked list of concepts for an n-gram and in our experiments we compare three families of approaches, including lexical matching, language modeling, and other state-of-the-art methods.

In the second step we enhance precision and determine which of the candidate concepts to keep by applying supervised machine learning. Here, each candidate concept is classified as being relevant or not (in the context of the tweet and the user). We use supervised machine learning, that takes as input a set of labeled examples (tweet to concept mappings) and several features of these examples, see [7] for the full list of features. We employ several feature types, each associated with either an n-gram, concept, or their combination. We also include a separate set of Twitter-specific features. The goal of the machine learning algorithm is to learn a function that outputs a relevance status for any new n-gram and concept pair given a feature vector of this new instance. Following Milne and Witten [8], we include a Naive Bayes, Support Vector Machines, and a C4.5 decision tree classifier. Random forests (RFs) are a very efficient alternative to C4.5, since they are (i) relatively insensitive to parameter settings, (ii) resistant to overfitting, and (iii) easily parallelizable. In recent years, gradient boosted regression trees (GBRTs) have been established as the de facto state-of-the-art learning paradigm for web search ranking. It is a point-wise learning to rank algorithm that predicts the relevance score of a result to a query by minimizing a loss function (e.g., the squared loss) using stochastic gradient descent. Finally, since RF is resistant to overfitting and also

*The full version of this paper will appear in *WSDM 2012* [7].

DIR 2012 Ghent, Belgium

often outperforms GBRT, the RF predictions can be used as starting point for GBRT. By doing so, GBRT starts at a point relatively close to the global minimum and is able to further improve the already good predictions.

In order to obtain manual annotations (both for training and evaluation), we have asked two volunteers to manually annotate 562 tweets, each containing 36.5 terms on average. They were presented with an annotation interface with which they could search through Wikipedia articles. The annotation guidelines specified that the annotator should identify concepts contained in, meant by, or relevant to the tweet. They could also indicate that an entire tweet was either ambiguous (where multiple target concepts exist) or erroneous (when no relevant concept could be assigned). Out of the 562 tweets, 419 were labeled as not being in either of these two categories and kept for further analysis. For these, the annotators identified 2.17 concepts per tweet on average. In order to facilitate follow-up research, we make all annotations and derived data available.[1]

The first research question we address is: What is the performance of state-of-the-art approaches for linking text to Wikipedia in the context of microblog posts? We approach the task of linking tweets to concepts as a ranking problem; given a tweet, the goal of a system implementing a solution to this problem is to return a ranked list of concepts meant by or contained in it, where a higher rank indicates a higher degree of relevance of the concept to the tweet. The best performing method puts the most relevant concepts towards the top of the ranking. Our second research question concerns a comparison of methods for the initial concept ranking step; we consider lexical matching, language modeling, and other state-of-the-art baselines and compare their effectiveness. Our third research question concerns the second, precision-enhancing step. We approach this as a machine learning problem and consider a broad set of features, some of which have been proposed previously in the literature on semantic linking, some newly introduced. In addition to multiple features, we also consider multiple machine learning algorithms and examine which of these are most effective for our problem. Finally, we examine the relative effectiveness of the precision-enhancing step on top of different initial concept ranking methods.

We employ three sets of baselines to which we compare our approach and to which we apply supervised machine learning. They include: (i) lexical matching of the n-grams with the concepts, (ii) a language modeling baseline, and (iii) a set of other methods, including the one proposed by Milne and Witten [8], which represents the state-of-the-art in automatic linking approaches. We use the algorithm and best-performing settings as described in [8], trained on our version of Wikipedia. We also include a novel service provided by DBpedia, called DBpedia Spotlight. The third baseline in this set (Tagme) is provided by Ferragina and Scaiella [1]. Our last concept ranking method in this set corresponds to the *COMMONNESS* feature, detailed in [7]. This method scores each concept based on the relative frequency with which the n-gram is used as an anchor text for that particular concept. Applying this concept ranking method achieves the highest scores of all baselines on all metrics. This relatively simple approach is able to retrieve over 75% of the relevant concepts and place the first relevant concept around rank 1.4 on average. We exclude OpenCalais from this set, since this webservice only recognizes entity types without performing any kind of disambiguation. Moreover, the precise algorithmic details are not made public.

Our experiments show that the second, machine learning step—in particular using random forests or gradient boosted regression trees—can significantly improve over the baseline. Especially in terms of precision, placing the first relevant concept around rank 1.3 on average. It is even able to improve when the concept ranking performance is already strong on its own. In sum, we find that the iterative machine learning methods obtain the best improvements overall and that they are able to significantly improve precision in all cases.

[1]See http://ilps.science.uva.nl/resources/wsdm2012-adding-semantics-to-microblog-posts/.

Acknowledgments

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, 727.-011.005, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti and by the ESF Research Network Program ELIAS.

References

[1] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM '10*, 2010.

[2] J. He, M. de Rijke, M. Sevenster, R. van Ommering, and Y. Qian. Generating links to background knowledge: A case study using narrative radiology reports. In *CIKM '11*, 2011.

[3] G. Inches, M. J. Carman, and F. Crestani. Statistics of online user-generated short documents. In *ECIR '10*, 2010.

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10*, 2010.

[5] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *ACL: HLT '11*, 2011.

[6] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Mapping queries to the Linking Open Data cloud: A case study using DBpedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):418 – 433, 2011.

[7] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

[8] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM '08*, 2008.

[9] E. Tsagkias, M. de Rijke, and W. Weerkamp. Linking online news and social media. In *Fourth ACM Web Search and Data Mining (WSDM)*, Hong Kong, 2011.