CIRGDISCO at RepLab2012 Filtering Task: A Two-Pass Approach for Company Name Disambiguation in Tweets

Arjumand Younus[1,2], Colm O'Riordan[1], and Gabriella Pasi[2]

[1] Computational Intelligence Research Group, National University of Ireland Galway, Ireland
[2] Information Retrieval Lab,Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
arjumand.younus@nuigalway.ie, colm.oriordan@nuigalway.ie, pasi@disco.unimib.it

Abstract. Using Twitter as an effective marketing tool has become a gold mine for companies interested in their online reputation. A quite significant research challenge related to the above issue is to disambiguate tweets with respect to company names. In fact, finding if a particular tweet is relevant or irrelevant to a company is an important task not satisfactorily solved yet; to address this issue in this paper we propose a Wikipedia-based two-pass algorithm. The experimental evaluations demonstrate the effectiveness of the proposed approach.

1 Introduction

Twitter[1] is a microblogging site which currently[2] ranks 8^{th} world wide in total traffic according to **Alexa**[3]. This huge popularity has turned Twitter into an effective marketing platform with almost all the major companies maintaining Twitter accounts. Moreover, Twitter users often express their opinions about companies via 140-character long Twitter messages called tweets [8] [10]. Companies are highly interested in monitoring their online reputation; this, however involves the significant challenge of disambiguating company names in text. The task becomes even more challenging in tweets due to huge noise, short length and lack of context for company name disambiguation [6] [7]. This paper describes our experience in dealing with some of these challenges in the context of RepLab2012 filtering task where we are given a set of companies and for each company a set of tweets, which contain some tweets relevant to the company and some irrelevant ones.

Our approach consists in a two-pass filtering algorithm that makes use of Wikipedia as an external knowledge resource in the first pass, and a concept

[1] http://twitter.com
[2] Data as of July 31, 2012
[3] http://www.alexa.com/siteinfo/twitter.com

term score propagation mechanism in the second pass. The first step is precision-oriented, where the aim is to keep the noisy tweets to a minimum. The second step enhances the recall via a score propagation technique, and by making use of other sources of evidence. Our technique shows high accuracy over the RepLab2012 dataset.

The rest of the paper is organized as follows: section 2 presents a more detailed description of the considered problem. Section 3 describes the proposed methodology in detail. Section 4 gives presents the experimental evaluation of our method, and Section 5 summarizes the related work. Finally section 6 concludes the paper.

2 Problem Statement

In this section we give a brief overview of the RepLab2012 filtering task. We were given company names and a set of tweets obtained via issuing a query (i.e. the name of the company) to Twitter Search[4]. Due to the issue of noise in Twitter, the tweets obtained via the query may or may not be relevant to the company. Our task addresses the problem of distinguishing the tweets relevant from those non-relevant to a company.

3 Methodology

This section describes the proposed filtering method in detail. We first explain how we use Wikipedia as an external knowledge resource: we use only portions of a company's Wikipedia page, as only some of the portions are meaningful for filtering purposes. Next, we explain the two steps of our algorithm.

3.1 Wikipedia as External Knowledge Resource

As Meij et al. point out in [7], a simple matching between tweets and Wikipedia texts would produce a significant amount of irrelevant and noisy concept terms. The authors further mention that such noise can be eliminated on either the Wikipedia or the textual side. On the basis of the intuition that the Wikipedia page of a company contains significant information about the company in certain portions of the Wikipedia article (i.e. concept terms exist in some portions of text), we perform this cleaning on the Wikipedia side as follows:

- we use the information within the Wikipedia infoboxes.
- we use the information within the Wikipedia category box.
- we parse the paragraphs that in the Wikipedia text are followed by application of POS tagging to these paragraphs. After the application of POS tagging, we extract unigrams, bigrams and trigrams for the significant POS tags [1].

[4] http://twitter.com/search

Finally, the concept terms extracted from the various portions of Wikipedia are split into single terms. These are then used for matching against the terms in the tweets for the task of filtering.

3.2 First Pass

After collecting all the extracted concept terms from Wikipedia we order them by their specificity in the Wikipedia article thereby giving a score to each term. We then check for the occurrence of these concept terms in the tweets, and the number of occurrences per term is multiplied by the score of that particular concept term to constitute a score for the tweet. Tweets that have a score above a certain threshold are considered to be relevant. The intuition behind the use of Wikipedia concept terms for the first phase is to keep the precision as high as possible, and to get relevant tweets only.

3.3 Second Pass

The second pass makes use of the idea of concept term score propagation in order to discover more tweets relevant to a particular company i.e. to increase the recall. The score propagation technique is based on the intuition that terms co-located with significant concept terms may have some relevance to that concept. The scores for concept terms in a relevant tweet obtained from the first pass are redistributed among co-located terms. This in turn gives some score to the non-concept terms, and by using the scores of these non-concept terms we perform a second computation to obtain the scores of tweets. Moreover, this phase uses more sources of evidence, these are:

- POS tag of the company name occurring within the tweets
- URL occurring within the tweets
- Twitter username occurring within the tweets
- Hashtag occurring within the tweets

The score propagation technique as well as the extra sources of evidence mentioned above enable us to extract more tweets relevant to the company, thus increasing the recall. We show the results of the evaluation metrics in the next section.

4 Experimental Evaluations

As mentioned in section 2 the task comprises binary classification and hence we report the effectiveness of our algorithm through the standard evaluation metrics of precision and recall. We were given a very small trial dataset (six companies) and a considerably larger test dataset (31 companies). We report our results for the test dataset. Table 1 shows the precision and recall figures after the application of the first pass, and after the application of the second pass of our algorithm.

	First Pass	Second Pass
Precision	0.84827	0.81129
Recall	0.16307	0.76229

Table 1. Precision and Recall Scores for Two Passes of the Algorithm

As Table 1 shows the first pass yields a high precision but an extremely low recall. The application of the second pass increases the recall by a large degree, while not overly reducing the precision. The significantly large increase in recall proves the effectiveness of the score-propagation technique combined with the use of multiple sources of evidence.

The RepLab2012 filtering task used the measures of Reliability and Sensitivity for evaluation purposes, these are described in detail in [3]. Table 2 presents a snapshot of the official results for the Filtering task of RepLab 2012, where CIRGDISCO is the name of our team.

Table 2. Results of Monitoring task of RepLab 2012

Team	Accuracy Filtering	R Filtering	S Filtering	F(R,S) Filtering
Daedalus_2	0.72	0.24	0.43	0.26
Daedalus_3	0.70	0.24	0.42	0.25
Daedalus_1	0.72	0.22	0.40	0.25
CIRGDISCO	0.70	0.22	0.34	0.23

Table 2 shows that our algorithm performed competitively. It is the second best among the submitted systems and fourth best among the submitted runs. We only managed to submit a single run due to the shortage of time.

5 Related Work

There has been an increasing interest in research on applying natural language processing techniques to tweets over the past few years. We provide an overview of some of these works in this section.

Named entity recognition in tweets has been an area of active research over the past few years with a focus on semi-supervised learning techniques [5] [9]. Given the lack of context in tweets for the task of named entity recognition, Liu et al. [5] use a large training set and apply a KNN classifier in combination with CRF based labeler whereas Ritter et al. [9] employ Labeled LDA over a FreeBase corpus.

Understanding and analysing the content of tweets is another significant research direction where the goal is to extract keyphrases from the content of tweets [4] [11]. Hong and Davidson [4] proposed several schemes to train standard LDA, and the Author-Topic LDA models for topic discovery over Twitter data. Zhao et al. [11] extracted and ranked topical key phrases on Twitter through the use of topic models [12] followed by topical PageRank.

Semantic enrichment of microblog posts has also been studied with the aim of determining what a tweet is about [2] [7]. Abel et al. make use of news articles to contextualize tweets [2] while Meij et al. [7] provide a fine semantic enrichment of tweets through matching with Wikipedia.

6 Conclusion

We proposed a two-pass algorithm for company name disambiguation in tweets. Our algorithm makes use of Wikipedia as a primary knowledge resource in the first pass of the algorithm, and the tweets are matched across Wikipedia terms. The matched terms are then used for score propagation in the second pass of the algorithm that also makes use of multiple sources of evidence. Our algorithm showed competitive performance and demonstrates the effectiveness of the techniques employed in the two passes. As a future work, we aim to refine the score propagation technique of the second pass by taking into account better features for effective scores computation.

References

1. http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ESWC'11, pages 375–389, Berlin, Heidelberg, 2011. Springer-Verlag.
3. E. Amigo, J. Gonzalo, and F. Verdejo. *Reliability and Sensitivity: Generic Evaluation Measures for Document Organization Tasks*. UNED, Madrid, Spain, 2012. Technical Report.
4. L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.
5. X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
6. K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 362–367, Berlin, Heidelberg, 2011. Springer-Verlag.
7. E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 563–572, New York, NY, USA, 2012. ACM.
8. A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

9. A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
10. S. R. Yerva, Z. Miklós, and K. Aberer. What have fruits to do with technology?: the case of orange, blackberry and apple. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 48:1–48:10, New York, NY, USA, 2011. ACM.
11. W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 379–388, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
12. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.