Microblog Entity Linking by Leveraging Extra Posts

Yuhang Guo, Bing Qin*, Ting Liu , Sheng Li
Research Center for Social Computing and Information Retrieval
School of Computer Science and Technology
Harbin Institute of Technology, China
{yhguo, bqin*, tliu, sli}@ir.hit.edu.cn

Abstract

Linking name mentions in microblog posts to a knowledge base, namely microblog entity linking, is useful for text mining tasks on microblog. Entity linking in long text has been well studied in previous works. However few work has focused on short text such as microblog post. Microblog posts are short and noisy. Previous method can extract few features from the post context. In this paper we propose to use extra posts for the microblog entity linking task. Experimental results show that our proposed method significantly improves the linking accuracy over traditional methods by 8.3% and 7.5% respectively.

1 Introduction

Microblogging services (e.g. Twitter) are attracting millions of users to share and exchange their ideas and opinions. Millions of new microblog posts are generated on such open broadcasting platforms every day [1]. Microblog provides a fruitful and instant channel of global information publication and acquisition.

A necessary step for the information acquisition on microblog is to identify which entities a post is about. Such identification can be challenging because the entity mention may be ambiguous. Let's begin with a real post from Twitter.

(1) No excuse for floods tax, says Abbott URL

This post is about an Australia political leader, `Tony Abbot`, and his opinion on flood tax policy. To understand that this post mentions Tony Abbot is not trivial because the name Abbot can refer to many people and organizations. In the Wikipedia page of *Abbott*, there lists more than 20 *Abbott*s, such as baseball player `Jim Abbott`, actor `Bud Abbott` and company `Abbott Laboratories`, etc..

Given a knowledge base (KB) (e.g. Wikipedia), entity linking is the task to identify the referent KB entity of a target name mention in plain text. Most current entity linking techniques are designed for long text such as news/blog articles (Mihalcea and Csomai, 2007; Cucerzan, 2007; Milne and Witten, 2008; Han and Sun, 2011; Zhang et al., 2011; Shen et al., 2012; Kulkarni et al., 2009; Ratinov et al., 2011). Entity linking for microblog posts has not been well studied.

Comparing with news/blog articles, microblog posts are:

short each post contains no more than 140 characters;

fresh the new entity-related content may have not been included in the knowledge base;

informal acronyms and spoken language writing style are common.

Due to these properties, few feature can be extracted from a post. Without enough features, previous entity linking methods may fail. In order to overcome the feature sparseness, we turn to another property of microblog:

*Corresponding author

[1] See http://blog.twitter.com/2011/06/ 200-million-tweets-per-day.html.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 863–868,
Seattle, Washington, USA, 18-21 October 2013.

redundancy For each day, over 340M short messages are posted in twitter. Similar information may be posted in different expressions.

For example, we find the following post,

> *(2) Julia Gillard and Tony Abbott on the flood levy just after 8.30am on @612brisbane!*

The content of post (2) is highly related to post (1). In contrast to the confusing post (1), the text in post (2) explicitly indicates that the *Abbott* here refers to the Australian political leader. This inspires us to bridge the confusing post and the knowledge base with other posts.

In this paper, we approach the microblog entity linking by leveraging extra posts. A straightforward method is to expand the post context with similar posts, which we call Context-Expansion-based Microblog Entity Linking (CEMEL). In this method, we first construct a query with the given post and then search for it in a collection of posts. From the search result, we select the most similar posts for the context expansion. The disambiguation will benefit from the extra posts if, hopefully, they are related to the given post in content and include explicit features for the disambiguation.

Furthermore, we propose a Graph-based Microblog Entity Linking (GMEL) method. In contrast to CEMEL, the extra posts in GMEL are not directly added into the context. Instead, they are represented as nodes in a graph, and weighted by their similarity with the target post. We use an iterative algorithm in this graph to propagate the entity weights through the edges between the post nodes.

We conduct experiments on real microblog data which we harvested from Twitter. Current entity linking corpus, such as the TAC-KBP data (McNamee and Dang, 2009), mainly focuses on long text. And few microblog entity linking corpus is publicly available. In this work, we manually annotated a microblog entity linking corpus. This corpus inherit the target names from TAC-KBP2009. So it is comparable with the TAC-KBP2009 corpus.

Experimental results show that the performance of previous methods drops on microblog posts comparing with on long text. Both of CEMEL and GMEL can significantly improve the performance over baselines, which means that entity linking system on microblog can be improved by leveraging extra posts. The results also show that GMEL outperforms CEMEL significantly.

We summarize our contributions as follows.

- We propose a context-expansion-based and a graph-based method for microblog entity linking by leveraging extra posts.
- We annotate a microblog entity linking corpus which is comparable to an existing long text corpus.
- We show the inefficiency of previous method on the microblog corpus and our method can significantly improve the results.

2 Task defination

The microblog entity linking task is that, for a name mention in a microblog post, the system is to find the referent entity of the name in a knowledge base, or return a NIL mark if the entity is absence from the knowledge base. This definition is close to the entity linking task in the TAC-KBP evaluation (Ji and Grishman, 2011) except for the context of the target name is microblog post whereas in TAC-KBP the context is news article or web log.

Several related tasks have been studied on microblog posts. In Meij et al. (2012)'s work, they link a post, rather than a name mention in the post, to relevant Wikipedia concepts. Guo et al. (2013a) and Liu et al. (2013) define entity linking as to first detect all the mentions in a post and then link the mentions to the knowledge base. In contrast, our definition (as well as the TAC-KBP definition) focuses on a concerned name mention across different posts/documents.

3 Method

A typical entity linking system can be broken down into two steps:

candidate generation This step narrows down the candidate entity range from any entity in the world to a limited set.

candidate ranking This step ranks the candidates and output the top ranked entity as the result.

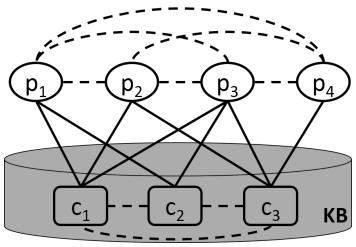

Figure 1: An example of the GMEL graph. $p_1 \dots p_4$ are post nodes and $c_1 \dots c_3$ are candidate entity nodes. Each post node is connected to the corresponding candidate nodes from the knowledge base. The edges between the nodes are weighted by the similarity between them.

In this paper, we use the candidate generation method described in Guo et al.(2013). For the candidate ranking, we use a Vector Space Model (VSM) and a Learning to Rank (LTR) as baselines. VSM is an unsupervised method and LTR is a supervised method. Both of them have achieved the state-of-the-art performances in the TAC-KBP evaluations.

The major challenge in microblog entity linking is the lack of context in the post. An ideal solution is to expand the context with the posts which contain the same entity. However, automatically judging whether a name mention in two documents refers to the same entity, namely cross document co-reference, is not trivial. Here our solution is to rank the posts by their possibility of co-reference to the target one and select the most possible co-referent posts for the expansion.

CEMEL is based on the assumption that, given a name and two posts where the name is mentioned, the higher similarity between the posts the higher possibility of their co-reference and that the co-referent posts may contains useful features for the disambiguation. However, two literally similar posts may not be co-referent. If such non co-referent post is expanded to the context, noises may be included.

Take the following post as an example.

(3) AG Abbott says that bullets have crossed the border from Mexico to Texas at least four times. URL

This post is similar to post (1) because they both contains "says" and "URL". But the Abbott in post (3) refers to the Texas Attorney General Greg Abbott. In this mean, the expanded context in post (3) could mislead the disambiguation for post (1). Such noise can be controlled by setting a strict number of posts to expand the context or weighting the contribution of this post to the target one.

Our CEMEL method consists of the following steps: First we construct a query with the terms from the target post. Second we search for the query in a microblog post collection using a common information retrieval model such as the vector space model. Note that here we limit the searched posts must contain the target name mention. Then we expand the target post with top N similar posts and use a typical entity linking method (such as VSM and LTR) with the expanded context.

Figure 1 illustrates the graph of GMEL. Each node of this graph represents an candidate entity (e.g. $c_1 \dots c_3$) or a post of the given target name (e.g. $p_1 \dots p_4$) In this graph, each node represents an entity or a post of the given target name. Between each pair of post nodes, each pair of entity nodes and each post node and its candidate entity nodes, there is an edge. The edge is weighted by the similarity between the two linked nodes. Entity nodes are labeled by themselves and candidate nodes are initialized as unlabeled nodes. For the edges between post node pairs and entity node pairs, we use cosine similarity. For the edges between a post node and its candidate entity nodes, we use the score given by traditional entity linking methods. We use an iterative algorithm on this graph to propagate the labels from the entity nodes to the post nodes. We adapt Label Propagation (LP) (Zhu and Ghahramani, 2002) and Modified Adsorption (MAD) (Talukdar and Pereira, 2010) for the iteration over the graph.

4 Experiment

4.1 Data Annotation

Till now, few microblog entity linking data is publicly available. In this work, we manually annotate a data set on microblog posts[2]. We collect 15.6 million microblog posts in Twitter dated from January 23 to February 8, 2011. In order to compare with existing entity linking on long text, we select a subset of target names from TAC-KBP2009 and inherit the knowledge base in the TAC-KBP evaluation. The

[2]We published this data so that researchers can reproduce our results.

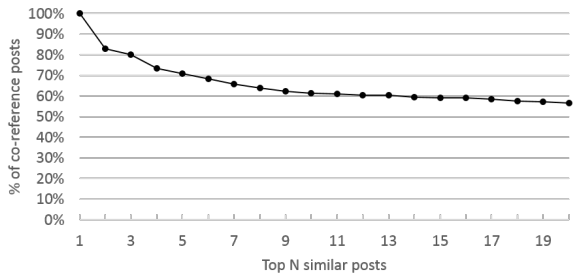

Figure 2: Percentage of the co-reference posts in the top N similar posts

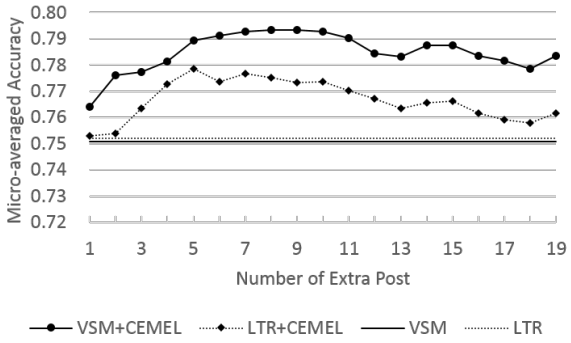

Figure 3: Impact of expansion post number in CEMEL

TAC-KBP2009 data set includes 513 target names. We search for all the target names in the post collection and get 26,643 matches. We randomly sample 120 posts for each of the top 30 most frequently matched target names and filter out non-English and overly short (i.e. less than 3 words) posts. Then we get 2,258 posts for 25 target names and manually link the target name mentions in the posts to the TAC-KBP knowledge base.

In order to evaluate the assumption in CEMEL: similar posts tend to co-reference, we randomly select 10 posts for 5 target names respectively and search for the posts in the post collection. From the search result of each of the 50 posts, we select the top 20 posts and manually annotate if they co-reference with the query post.

4.2 Settings

We generate candidates with the method described in (Guo et al., 2013b) and use Vector Space Model (VSM) (Varma et al., 2009) and Learning to Rank (LTR) (Zheng et al., 2010) as the ranking model. We

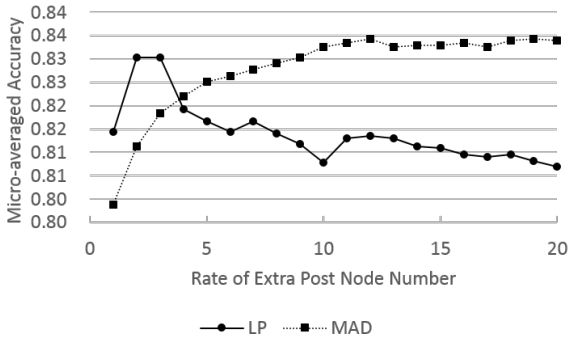

Figure 4: Accuracy of GMEL with different rate of extra post nodes

use Lucene and ListNet with default settings for the VSM and LTR implementation respectively. We use bigram feature for VSM and the feature set of (Chen et al., 2011) for LTR. LTR is evaluated with 10-fold cross validation. Given a target name, the GMEL graph includes all the evaluation posts as well as a set of extra post nodes searched from the post collection with the query of the target name. We filter out determiners, interjections, punctuations, emoticons, discourse markers and URLs in the posts with a twitter part-of-speech tagger (Owoputi et al., 2013). The similarity between a post and its candidate entities is set with the score given by VSM or LTR and the similarity between other nodes is set with the corresponding cosine similarity. We employ junto[3] with default settings for the iterative algorithm implementation .

4.3 Results

Figure 2 shows the relationship between similarity and co-reference. From this figure we can see that the percentage decreases with the growth of N. When the N is up to 10, about 60% of the similar posts co-reference with the query post and the decrease speed slows down. The Pearson correlation coefficient between the percentage and the number of top N is -0.843, which shows a significant correlation between the two variables (with p-value 0.01 under t-test).

Figure 3 shows the impact of the extra post number for the context expansion in CEMEL. We can see that the accuracies of VSM and LTR are improved

[3]See https://github.com/parthatalukdar/junto

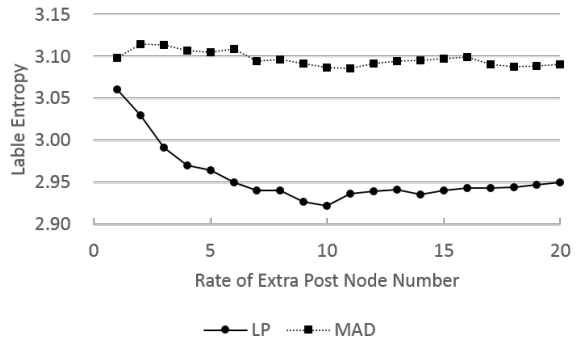

Figure 5: Label entropy of GMEL with different rate of extra post nodes

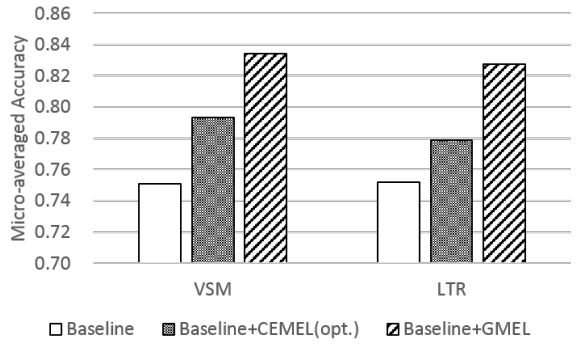

Figure 6: Accuracy of the systems

by CEMEL. The improvements peak with 5-10 extra posts. Then more extra posts will pull down the accuracy.

Figure 4 shows the accuracy of GMEL. The x-axis is the rate of the extra post number over the evaluation post number. We can see that the accuracy of MAD increases with the number of extra post nodes at first and then turns to be stable. The accuracy of LP increases at first and drops when more extra posts are added into the graph.

Figure 5 shows the information entropy of the labels in LP and MAD. The curves show that the prediction of LP tends to converge into a small number of labels. This is because LP prefers smoothing labelings over the graph (Talukdar and Pereira, 2010).

We also evaluate our baselines on TAC-KBP2009 data set (LTR is trained on TAC-KBP2010 data set). The accuracy of VSM and LTR are 0.8338 and 0.8372 respectively, which are comparable with the state-of-the-art result (Hachey et al., 2013).

Figure 6 shows the performances of the systems on the microblog data. We set the optimal expansion post number of CEMEL and use MAD algorithm for GMEL with all searched extra post nodes. From this figure we can see that the results of VSM and LTR baselines are comparable and both of them are significantly lower than that on TAC-KBP2009 data. CEMEL improves the VSM and LTR baselines by 4.3% and 2.7% respectively. GMEL improves VSM and LTR by 8.3% and 7.5% respectively. The results of GMEL are also significantly better than CEMEL. All of the improvements are significant under Z-test with $p < 0.05$.

5 Conclusion

In this paper we approach microblog entity linking by leveraging extra posts. We propose a context-expansion-based and a graph-based method. Experimental results on our data set show that the performance of traditional method drops on the microblog data. The graph-based method outperforms the context-expansion-based method and both of them significantly improve the accuracy of traditional methods. In the graph-based method the modified adsorption algorithm performs better than the label propagation algorithm.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) via grant 61273321, 61073126, 61133012 and the National 863 Leading Technology Research Project via grant 2012AA011102. We would like to thank to Wanxiang Che, Ruiji Fu, Yanyan Zhao, Wei Song and several anonymous reviewers for their constructive comments and suggestions.

References

Zheng Chen, Suzanne Tamang, Adam Lee, and Heng Ji. 2011. A toolkit for knowledge base population. In *SIGIR*, pages 1267–1268.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013a. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030, Atlanta, Georgia, June. Association for Computational Linguistics.

Yuhang Guo, Bing Qin, Yuqin Li, Ting Liu, and Sheng Li. 2013b. Improving candidate generation for entity linking. In Elisabeth Mtais, Farid Meziane, Mohamad Saraee, Vijayan Sugumaran, and Sunil Vadera, editors, *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*, pages 225–236. Springer Berlin Heidelberg.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with wikipedia. *Artificial Intelligence*, 194(0):130 – 150. ¡ce:title¿Artificial Intelligence, Wikipedia and Semi-Structured Resources¡/ce:title¿.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, pages 945–954, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 457–466, New York, NY, USA. ACM.

Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, Furu Wei, and Yi Lu. 2013. Entity linking for tweets. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

P. McNamee and H.T. Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Proceedings of the Second Text Analysis Conference (TAC2009)*.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. 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. ACM.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

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

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL2013*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA, June. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2012. Linden: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 449–458, New York, NY, USA. ACM.

Partha Pratim Talukdar and Fernando Pereira. 2010. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1473–1481, Uppsala, Sweden, July. Association for Computational Linguistics.

Vasudeva Varma, Vijay Bharat, Sudheer Kovelamudi, Praveen Bysani, Santosh GSK, Kiran Kumar N, Kranthi Reddy, Karuna Kumar, and Nitin Maganti. 2009. Iiit hyderabad at tac 2009. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA, November.

Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. 2011. Entity linking with effective acronym expansion, instance selection, and topic modeling. In Toby Walsh, editor, *IJCAI 2011*, pages 1909–1914.

Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Learning to link entities with knowledge base. In *NAACL2010*, pages 483–491, Los Angeles, California, June. Association for Computational Linguistics.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.