# Linking Entities in #Microposts

Romil Bansal, Sandeep Panem, Priya Radhakrishnan,
Manish Gupta, Vasudeva Varma
International Institute of Information Technology, Hyderabad

## ABSTRACT

Social media has emerged to be an important source of information. Entity linking in social media provides an effective way to extract useful information from microposts shared by the users. Entity linking in microposts is a difficult task as they lack sufficient context to disambiguate the entity mentions. In this paper, we do entity linking by first identifying entity mentions and then disambiguating the mentions based on three different features: (1) similarity between the mention and the corresponding Wikipedia entity pages; (2) similarity between the mention and the tweet text with the anchor text strings across multiple webpages, and (3) popularity of the entity on Twitter at the time of disambiguation. The system is tested on the manually annotated dataset provided by Named Entity Extraction and Linking (NEEL) Challenge 2014, and the obtained results are on par with the state-of-the-art methods.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## General Terms

Algorithms, Experimentation

## Keywords

Named Entity Extraction and Linking (NEEL) Challenge, Entity Linking, Entity Disambiguation, Social Media

## 1. INTRODUCTION

Social media networks like Twitter have emerged to be major platforms for sharing information in form of short messages (tweets). Analysis of tweets can be useful for various applications like e-commerce, entertainment, recommendations, etc. Entity linking is the one such analysis task which deals with finding correct referent entities in the knowledge base for various mentions in the tweet. Entity linking in social media is important as it helps in detecting, understanding and tracking information about an entity shared across social media.


Published as part of the #Microposts2014 Workshop proceedings, available online as CEUR Vol-1141 (http://ceur-ws.org/Vol-1141)

#Microposts2014, April 7th, 2014, Seoul, Korea.

Entity linking consists of two different tasks, mention detection and entity disambiguation. Entity linking from general text is a well explored problem. Existing entity linking tools are intended for use over news corpora and similar document-based corpora with relatively long length. But as microposts lack sufficient context, these context-based approaches fail to perform well on microposts.

In this paper we describe our system proposed for the NEEL Challenge 2014 [1]. The proposed system disambiguates the entity mentions in the tweets based on three different measures: (1) Wikipedia's context based measure (§2.2.1); (2) anchor text based measure (§2.2.2); and (3) Twitter popularity based measure (§2.2.3).

The mention detection is done using existing Twitter part-of-speech (POS) taggers [2, 5].

## 2. OUR APPROACH

### 2.1 Mention Detection

Mention detection is the task of finding entity mentions in the given text. We assumed mentions as named entities present inside the tweets. Various approaches for named entity recognition in tweets have been proposed recently [3, 5]. This includes spotting continuous sequence of proper nouns as named entities in the tweet. But sometimes named entities like 'Statue of Liberty', 'Game of Thrones' etc. also includes tokens other than nouns. To detect such mentions, Ritter *et al.* [5] proposed a machine learning based system for named entity detection in tweets. Gimpel *et al.* [2] present yet another approach for POS tagging of tweets. We tried both of these POS taggers to extract proper noun sequences. In our experiments Ritter *et al.*'s tagger gave an accuracy of 77% while Gimpel *et al.*'s tagger gave an accuracy of 92%. So we merged the results from both as shown in Fig. 1. The tweet text is fed to the system and the longest continuous sequences of proper noun tokens detected using the above approach are extracted as the entity mentions from the given tweet. The merged system provided an accuracy of 98% in predicting mentions.

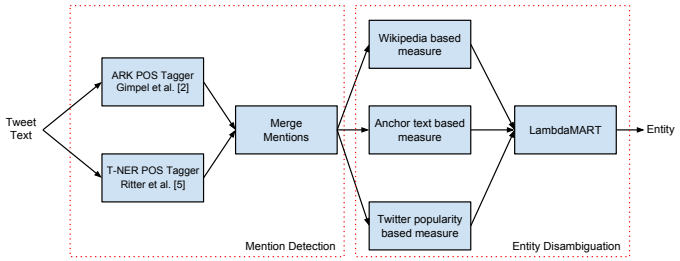


**Figure 1: System Architecture**

## 2.2 Entity Disambiguation

Entity disambiguation is the task of assigning the correct referent entity from the knowledge base to the given mention. We disambiguate the entity mention using three measures as described below. The scores from these three measures are combined using LambdaMART [7] model to arrive at the final disambiguated entity.

### *2.2.1 Wikipedia's Context based Measure (M1)*

This measure disambiguates a mention by calculating the frequency of occurrence of the mention in the Wikipedia corpus. Wikipedia's context based measure has been used in various approaches for disambiguating mentions in tweets [4]. We query MediaWiki API[1] with the entity mention. MediaWiki API returns the candidate entities in the ranked order. Each candidate entity is assigned its reciprocal rank as score. Thus, a ranked list of candidate entities with their scores are created using M1.

### *2.2.2 Anchor Text based Measure (M2)*

Google Cross-Wiki Dictionary (GCD) [6] is a string to concept mapping, created using anchor text from various web pages. A concept is an individual Wikipedia article, identified by its URL. The text strings constitute the anchor hypertexts that refer to these concepts. Thus, anchor text strings represent a concept. We query the GCD with a mention along with the tweet text. Based on the similarity to the query string, a ranked list of probable candidate entities are created (which is the ranked list using M2). The ranking criteria is based on Jaccard similarity between the anchor text and the query. So if the mention is highly similar to the anchor text, then the corresponding concept will have a high score.

### *2.2.3 Twitter Popularity based Measure (M3)*

Tweets about entities follow a bursty pattern. Bursty patterns are the bursts of tweets that appear after an event relating to an entity happens. We exploited this fact and tried to measure the number of times the given mention refers to a particular entity on Twitter recently. The mention is queried on Twitter API[2] and the resultant tweets are analyzed. All the tweets along with the mention are then queried on the GCD and the candidate entities are taken. Based on the scores returned using GCD, all the candidate entities are ranked (which is the ranked list using M3). As Twitter popularity based measure captures the people's interests at a particular time, it works well for entity disambiguation on recent tweets. In essence, the methods M2 and M3 are similar but with different inputs. Both use GCD, and produce candidate mentions and score as output. However, M2 takes mention and single tweet text as input whereas M3 takes mention and multiple tweets as input.

We have three rankings available using M1, M2, M3. Now the task is to arrive at the final ranking of the candidate entities by combining the rankings of the three different models. The rankings of different models should be combined such that the overall F1 score is maximized. For this, we use LambdaMART which combines LambdaRank and MART models. LambdaMART creates boosted regression trees for combining the rankings of the three different systems.

# 3. RESULTS AND EVALUATION

The dataset comprises of 2.3K tweets each annotated with the entity mention and its corresponding DBpedia URL. We divided the dataset into the 7:3 (train:test) ratio. Table 1 shows the results obtained using the NEEL Challenge evaluation framework. The best results are obtained when a combination of all the measures were used for disambiguation[3]. A 5-fold cross validation on the dataset gave an average **F1** of **0.52** for **M1+M2+M3**.

**Table 1: Results: M1 represents Wikipedia's Context based Measure (§2.2.1), M2 represents Anchor Text based Measure (§2.2.2) and M3 represents Twitter Popularity based Measure (§2.2.3)**

| Measure | F1-measure |
|---|---|
| M1 | 0.355 |
| M2 | 0.100 |
| M3 | 0.194 |
| M1+M2 | 0.355 |
| M2+M3 | 0.244 |
| M1+M3 | 0.405 |
| M1+M2+M | **0.512** |

# 4. CONCLUSION

For effective entity linking, mention detection in tweets is important. We improve the accuracy of detecting mentions by combining various Twitter POS taggers. We resolve multiple mentions, abbreviations and spell variations of a named entity using the Google Cross-Wiki Dictionary. We also use popularity of an entity on Twitter for improving the disambiguation. Our system performed well with a F1 score of 0.512 on the given dataset.

# 5. REFERENCES


[1] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *Proc., 4th Workshop on Making Sense of Microposts (#Microposts2014)*, pages 54–60, 2014.

[2] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proc. of the* $49^{th}$ *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2 (NAACL-HLT)*, pages 42–47, 2011.

[3] S. Guo, M.-W. Chang, and E. Kıcıman. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In *Proc. of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1020–1030, 2013.

[4] X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei, and Y. Lu. Entity Linking for Tweets. In *Proc. of the* $51^{th}$ *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1304–1311, 2013.

[5] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[6] V. I. Spitkovsky and A. X. Chang. A Cross-Lingual Dictionary for English Wikipedia Concepts. In *Proc. of the* $8^{th}$ *Intl. Conf. on Language Resources and Evaluation (LREC)*, 2012.

[7] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting Boosting for Information Retrieval Measures. *Journal of Information Retrieval*, 13(3):254–270, Jun 2010.


[1] https://www.mediawiki.org/wiki/API:Search

[2] https://dev.twitter.com/docs/api/1.1/get/search/tweets

[3] submitted as Agglutweet_1.tsv