Entity Recognition and Linking on Tweets with Random Walks

Zhaochen Guo
Department of Computing Science
University of Alberta
zhaochen@ualberta.ca

Denilson Barbosa
Department of Computing Science
University of Alberta
denilson@ualberta.ca

ABSTRACT

This paper presents our system at the #Microposts2015 NEEL Challenge [4]. The task is to recognize and type mentions from English Microposts, and link them to their corresponding entries in DBpedia 2014. For this task, we developed a method based on a state-of-the-art entity linking system - REL-RW [2], which exploits the entity graph from the knowledge base to compute semantic relatedness between entities, and use it for entity disambiguation. The advantage of the approach is its robustness for various types of documents. We built our system on REL-RW and employed a tweet specific NER component to improve the performance on tweets. The system achieved overall 0.35 F1 on the development dataset from NEEL 2015, while the disambiguation component alone can achieve 0.70 F1.

Keywords

Entity Recognition, Entity Disambiguation, Social Media

1. INTRODUCTION

Microposts such as tweets become popular nowadays. The tweets, though short and simple, can spread information fast and broadly. Events, reviews, news and so on are all posted on Twitter, which make tweets a very valuable resource to support many activities such as political option mining, product development (customer review), or social activism. We need to understand the tweets to make best use of them for such applications. Given the maximum 140 characters limit, there is barely enough useful information in a tweet. Exploiting entities mentioned in tweets can enrich the text with their contexts and semantics in knowledge bases, which is important for a better understanding of tweets. The NEEL task aims to solve this issue by automatically recognizing entities and their types from English tweets, and linking them to their DBpedia 2014 resources. NER and entity linking have been active research subjects. However, most previous works focus on traditional long documents, which do not pose the challenges in tweets, such as the noisy terms, hashtags, retweets, abbreviations, and cyber-slang. Appropriately addressing these problems, and taking advantage of the existing approaches are important.

We developed a NEEL system for the challenge based on a state-of-the-art entity linking approach, and incorporated a tweet specific mention extraction component. Our system takes advantage of the entity graph in the knowledge base, and does not rely on the lexical features in the tweets, which makes it robust on different datasets. In the following sections, we will describe our system and report the experimental results on the challenge benchmarks.

2. OUR APPROACH

2.1 Mention Extraction

As the first component of our system, mention extraction extracts named entities from the given tweets. Our system originally employed the Stanford NER with models trained on the well-formed news documents. However, it cannot handle the short tweets very well. We then used the TwitIE [1] from GATE, a NER tool designed specifically for tweets, to perform the mention extraction in our system.

Compared to the Stanford NER, TwitIE added several improvements. The first is the *Normaliser*. To address unseen tokens and noisy grammars in tweets, TwitIE used a spelling dictionary specific to the tweets to identify and correct spelling variations. The second improvement is a tweet adapted model for the POS tagging. While still employing the Stanford POS Tagger, TwitIE replaced the original model with a new model trained on Twitter datasets which were annotated with the Penn TreeBank with extra tag labels such as retweets, URLs, hashtags and user mentions. With these improvements, TwitIE helps improve the NER performance of our system. Note that we use the types inferred from TwitIE as the types for mentions.

2.2 Candidate Generation

The second component is the candidate generation which selects potential candidates from the knowledge base for mentions in the tweets. Our system utilized an alias dictionary collected from Wikipedia titles, redirect pages, disambiguation pages, and the anchor text in Wikilinks [2], which maps each alias to entities it refers to in Wikipedia.

We simply use exact string matching against the dictionary for the candidate generation. Mentions that do not match any alias in the dictionary are immediately linked to NIL. Otherwise, the mapping entities of the matched alias are selected as candidates. To improve the efficiency, we

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

further prune the candidates by two criteria [2]: *prior probability* which is defined as the probability the alias refers to an entity in the Wikipedia corpus, and *context similarity* which measures the context similarity (cosine similarity) of the mention and the entity. For both criteria, the top 10 ranked candidates are selected and then merged to generate the final candidate list for the given mention.

2.3 Entity Disambiguation

Entity disambiguation is to select the target entity from the candidates of a mention. We use our prior algorithm [2] for this task. The main idea is to represent the semantics of the document (tweet) and candidate entities using a set of related entities in DBpedia for which the weight of each entity is measured by their semantic relatedness with the candidates. We then use the semantic representation to compute the semantic similarity between the candidates and the document. For each mention-entity pair, we measure their prior probability, context similarity, and semantic similarity and linearly combine them together to compute an overall similarity. The candidate with the highest similarity will be selected as the target entity.

The key part of the approach is the semantic representation and relatedness. Knowledge bases such as DBpedia are graphs where entities are connected semantically. We construct an entity graph from the knowledge base and use the connectivity in the graph to measure the semantic relatedness between entities. We use random walks with restart to traverse the graph. Upon convergence, this process results in a probability distribution over the vertices corresponding to the likelihood these vertices are visited. This probability can then be used as an estimation of relatedness between entities in the graph. For each target entity, we restart from that entity in each random walk, generating a personalized probability for the target entity, and use it as the semantic representation. For the semantic representation of the document, we perform the random walk restarting from a set of entities representing the document. Since the true entities of mentions in the documents are not available, we either choose the representative entities from the unambiguous mentions which have only one candidate, or the candidate entities whose weights are approximated by their prior probability. With the representative entities, the semantic representation of the document can then be computed as the probability distribution obtained through the random walk from these entities.

To improve the efficiency, instead of using the entire DBpedia graph, we construct a small entity graph by starting with the set of candidates, and adding all entities adjacent to these candidates in the original graph. This subgraph contains entities semantically related to the candidates and is large enough to compute the semantic representation of entities and the document.

Once obtaining the semantic representation, we measure the semantic similarity between each candidate and the document using the Zero-KL Divergence [3], which is then combined with the prior probability and context similarity to disambiguate candidates.

2.4 NIL Prediction and Clustering

For NIL prediction, mentions are deemed out of a knowledge base (and thus linked to NIL) either when no candidates are available or their similarity with the highest ranked entities is below a threshold. For clustering, we simply group mentions by their name similarity. In the future, we plan to exploit the semantic representation of the tweets to measure their semantic similarity and use it for NIL clustering.

	Precision	Recall	F1
Tagging	0.34	0.22	0.27
Linking	0.35	0.36	0.35
Clustering	0.45	0.29	0.35

Table 1: Results on the development datasets.

3. EXPERIMENTS

We built our system using a 2013 DBpedia dump, including the knowledge base and alias dictionary. Table 3 lists the results of our system on the development dataset. As shown, the performance of the mention extraction (tagging) is very poor, especially the recall. We believe more tuning would improve the performance. Since the novelty of our system is the disambiguation part, we further evaluated the performance of the entity disambiguation component separately (assuming all mentions are correctly recognised), and the system can achieve results of 0.74 precision, 0.66 recall for an F1 of 0.70 on the dataset.

4. CONCLUSION

In this paper, we described a system for the #Microppost2015 NEEL challenge, in which we adopted a tweet specific NER system for mention extraction, and used an entity disambiguation approach that utilized the connectivity of entities in DBpedia to capture the semantics of entities and disambiguate mentions.

Due to time limitation, our system still has much room for improvements. As shown, mention extraction is now the bottleneck of the system and needs further improvement. More features from the tweets could be used to train a better model. For the mention disambiguation, we will explore supervised approaches such as learning to rank to combine the semantic features such as the semantic similarity and lexical features specific to tweets. Also, the semantic representation seems to be valuable for the NIL clustering and worth exploration.

5. REFERENCES

[1] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In *RANLP*. ACL, 2013.

[2] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *CIKM*, pages 499–508, 2014.

[3] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL*, pages 581–589, 2007.

[4] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 44–53, 2015.