Named Entity Extraction and Linking in #Microposts

Priyanka Sinha
TCS Innovation Lab Kolkata
Indian Institute of Technology Kharagpur
priyanka27.s@tcs.com

Biswanath Barik
TCS Innovation Lab Kolkata
Tata Consultancy Services Limited
biswanath.barik@tcs.com

ABSTRACT
The task of Named Entity Extraction and Linking (NEEL) challange 2015 [5] is considered as two successive tasks : Named Entity Extraction (NEE) from the tweets and Named Entity Linking (NEL) with DBpedia. For NEE task we use CRF++ [1] to create a language model on the given training data. For entity linking, we use DBpedia Spotlight.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

General Terms
Experiment

Keywords
Twitter, Entity, Linking, Social Media, DBpedia

1. INTRODUCTION
Information Extraction (IE) from short messages or microblogs like tweets is an emerging field of research due to its commercial applications like ecommerce, recommendation etc. and social administration like social security. Entity linking (or entity resolution) is one such task which deals with identifying and extracting the Named Entities that belong to the tweets and disambiguating them by linking to the correct reference entities in the knowledge base.

The entity linking problem is well explored on normal text. However, the existing techniques of entity linking do not work well on short messages as the microblogs do not have sufficient context to classify (or disambiguate) the mentions. In this work we have identified the mention by creating an entity recognition model on the given training data and link them to the DBpedia using DBpedia Spotlight.

The rest of the paper is organized as follows: Section 2 describes our proposed approach which includes data preparation and feature selection for named entity recognition model creation and entity linking method. Section 3 describes the setup for web access. The result of our work is discussed in Section 4. Section 5 illustrates the future scope of our work followed by the references.

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.

2. METHODOLOGY
In our approach we have divided the Named Entity Extraction and Linking (NEEL) [5] task into two consecutive subtasks, namely, Named Entity Extraction and Named Entity Linking.

2.1 Named Entity Extraction
The NER task is viewed here as a sequence labeling problem. Given an input tweet, this step aims to identify the word sequences that constitute a Named Entity and classify each such entity into one of the predefined classes. For entity recognition and classification task, we have developed a model on the given training data using Conditional Random Fields (CRFs) which is an undirected graphical model used mainly for sequence labeling.

As we have discussed in the previous section that the context of the tweets is short, sometimes noisy and informal and thus, their syntactic structures are not always comparable to the normal texts. [4] showed that the Part-of-Speech (POS) features of surface tokens, Shallow Parsing (or Chunking) information, Capitalization indicators etc. are useful for improving NE recognition from tweets provided these modules should be trained on twitter data. In this experiment, we have added POS tag information to the training data using Twitter NER[3], used word features and some binary features like punctuations, digits, dots, hashtags, @, capitalization indicators, existence of URLs, underscore, hyphen etc. as features indicating or not indicating NEs for training NE recognition model. We were motivated to use [3] as it allows to tokenize and distinguish between nouns and other punctuations and tweet related artefacts well. We used [1] as it was relatively simple to adapt to our task.

2.1.1 Data Preparation
In the data preparation step, we have identified the word sequences refering to a Named Entity(NE) in the training data using the gold standard. The training data is tokenized, part-of-speech(POS) tagged using Twitter NER[3] and converted into 'BIO' format. For example, the NEs identified in the tweet ID: 100678378755067904, tweet "RT @HadleyFreeman: NOTHING on US news networks about

London riots. Can you imagine the BBC ignoring, say, riots in NYC? #americanewsfail" as follows

```
RT ~ O
@HadleyFreeman @ B-Person
: ~ O
NOTHING N O
on P O
US ^ B-Location
news N O
networks N O
about P O
London ^ B-Event
riots N I-Event
. , O
Can V O
you O O
imagine V O
the D O
BBC ^ B-Organization
ignoring V O
, , O
say V O
, , O
riots N O
in P O
NYC ^ B-Location
? , O
#americanewsfail # O
```

2.1.2 *Feature Selection*

We have experimented with various feature types, various window lengths and their combinations and come up with the following feature set which gave us a good result. We experimented with some context window lengths and 5 gave us good results.

- Contextual (Word) Features: a context window of size five: W_{i-2} W_{i-1} W_i W_{i+1} W_{i+2}
- Part-of-Speech (POS) Features: a context of size five: P_{i-2} P_{i-1} P_i P_{i+1} P_{i+2}
- Word having Capitalization: binary feature
- Word having Punctuation: binary feature
- Is a Digit: binary feature
- Word having a Dot: binary feature
- Word having hashtag: binary feature
- Word having @: binary feature

2.2 Named Entity Linking

For linking, we use the annotations returned by DBpedia Spotlight REST API as the candidates and look for the longest matching surface forms.

We take the output of the NEE task and collect the named entities that are extracted and their categories. To identify correct start position we check for # and @. For each tweet, using the B/I tags we find the longest consecutive entities that make up a single entity. For example, in the tweet above, "London riots" would be treated as a single entity. For each tweet, DBpedia Spotlight REST API is accessed with confidence and support set to 0 with accepted return text in XML. We use the DBpedia Spotlight's annotate endpoint to obtain all the links at once. For each entity returned from DBpedia Spotlight, if the surface form is found to be a substring of any of the entities and if a substring match is found the corresponding URI is returned. For named entities for which no match is found, if it is an existing nil entity then the nil id is returned, else the nil counter is incremented and returned.

3. SETUP

We used perl for transforming the data. We used the CMU Twitter NLP[3] package for generating POS, CRF++[1] package and DBpedia Spotlight[2] REST API.

3.1 Web access

We use JSP to create our REST API, which uses perl which in turn uses curl to connect to DBpedia Spotlight[2] REST endpoints.

4. EVALUATION

The precision for strong link match with the training set itself is 30.49%, recall is 30.29% and f1 is 30.39%. For the tagging of correct entity type the precision with the training set itself is 82.89%, recall 82.35% and f1 82.62%.

The precision for strong link match with the development set is 14.82%, recall is 7.97% and f1 is 10.37%. For the tagging of correct entity type the precision with the training set itself is 41.65%, recall 22.41% and f1 29.14%.

5. FUTURE WORK

As we can see using the CMU POS tagger[3] and CRF[1] discovers the entities well, but the way we do linking needs more work.

6. REFERENCES

[1] Crf++: Yet another crf toolkit.

[2] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[3] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of NAACL*, 2013.

[4] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK, July 2011.

[5] 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.