Feature Based Approach to Named Entity Recognition and Linking for Tweets

Souvick Ghosh
Jadavpur University
Kolkata, West Bengal 700032
India
souvick.gh@gmail.co

Promita Maitra
Jadavpur University
Kolkata, West Bengal 700032
India
promita.maitra@gmail.com

Dipankar Das
Jadavpur University
Kolkata, West Bengal 700032
India
dipankar.dipnil2005@gmail.com

ABSTRACT

In this paper, we describe our approach for Named Entity rEcognition and Linking Challenge (NEEL) at the #Microposts2016. The task is to automatically recognize entities and their types from English microposts, and link them to corresponding DBpedia 2015 entries. If the resources do not exist, we use NIL identifiers instead. The task is unique as twitter data is informal in nature with non-conformational spellings, random contractions and various other noises. For this task, we developed our system using a hybrid model. We have used various existing named entity recognition (NER) systems and combined them with our classifier to improve the results.

Keywords

Named Entity Extraction; Named Entity Linking; Social Media; DBpedia; Twitter

1. INTRODUCTION

In present day world, the relevance and importance of various social media platforms are immeasurable. Microposts such as tweets are limited in number of characters. However, the conciseness of the text is barely a pointer to its usefulness. From opinion mining during political campaigns to live feeds during sports events, from product reviews to vacation posts, Twitter is almost ubiquitous. Twitter promotes instant communication. Most celebrities use it to form their own digital presence. It also serves as a common forum where people have the capability to rise from obscurity to prominence through sharing of opinions. If we compare microposts to any standard long document such as blog or news articles, there exist a number of differences. Long articles are usually well written. They follow a definite structure, include headings and follow the rules of English grammar. Microposts, on the other hand, are short, noisy and hardly show any adherence to formal grammar. Presence of extraneous characters like hashtags, abbreviations and the lack of structure and context makes it difficult to extract relevant information. Due to this complexity, existing named entity recognition systems (NER) do not perform very well with tweet data. In NEEL challenge [8] of #Microposts2016 [6], we were required to automatically identify the named entities and their types from Twitter data and link them to the corresponding URIs of the DBpedia 2015-04 dataset[1] . Identifying the named entities and linking them to an existing knowledge base enriches the text with more contextual and semantic information. The mentions which could not be linked to any existent DBpedia resource page were recognized as NIL mentions. These mentions were clustered to ensure that the same entity, which does not have a corresponding entry in DBpedia, will be referenced with the same NIL identifier. We have developed three systems for the NEEL challenge, the major difference between the systems being the features used for each run. Our system follows a hybrid approach where Stanford Named Entity Recognition system is used to identify the entity mentions. In the next step, we run ARK Twitter Part-of-Speech Tagger to identify the mentions which are missed formerly. We use our own classifier to detect the type of the mentions. The named entity linking to DBpedia resources is done using Babelfy[2] . It must be noted that we followed a feature-based approach for the NEEL challenge. We also combined the existing tools for Named Entity Recognition and Linking. Each of the existing tools, like the Stanford NER, ARK Part-of-Speech Tagger and Babelfy are state-of-the-art. We explored their strengths and weaknesses in our work.

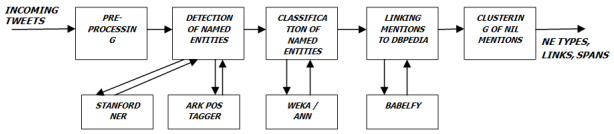

Figure 1: Workflow of the system.

2. OUR SYSTEM

Our system follows four steps in pipeline as shown in Figure 1. Mention detection in two stages, followed by mention type classification, mention linking and NIL clustering.

[1]http://wiki.dbpedia.org/dbpedia-data-set-2015-04
[2]http://babelfy.org/

Published as part of the #Microposts2016 Workshop proceedings, available online as CEUR Vol-1691 (http://ceur-ws.org/Vol-1691)

#Microposts2016, Apr 11th, 2016, Montréal, Canada.

2.1 Preprocessing

From the training data, the mentions referring to the 7 types of entities were extracted to form 7 bags of words. Using the initial words as seeds, the Wikipedia dumps were crawled to expand the set of words. These lists represent potential candidates for Named Entity mentions.

2.2 Detection of Entity Mentions

In this step, the named entity mentions in the given tweets are identified using two different approaches.

2.2.1 Using Stanford Named Entity Recognizer

The Stanford Named Entity Recognizer [3] was used to extract the named entities. It is a CRF classifier implementing linear chain Conditional Random Field. We use the 3 class model to extract the named entities belonging to classes *Location*, *Person* and *Organization*. While the recall was very low, the precision of Stanford NER was quite good.

2.2.2 Using ARK Twitter Part-of-Speech Tagger

The tweets were tokenized and assigned Part of Speech tags using the ARK Twitter Part-of-Speech Tagger [1]. We used the Twitter POS model with 25-tag tagset. The proper nouns (NNP and NNS tagged as ^) and possessive proper nouns (tagged Z) along with hashtags (#) and at-mentions () were extracted as probable candidates for Named Entity mentions. The mentions which were already identified using Stanford NER are not considered for classification step as they are already classified by the tagger itself. The rest of the mentions are classified using our classifier in the next step.

2.3 Classification of Entity Types

In the machine learning software WEKA [2], we use the following features to form a feature set and used the Random Forest classifier to generate a pruned C4.5 Decision Tree for 7-way classification of the named entity mentions âĂŞ Thing, Event, Character, Location, Organization, Person and Product, while providing the identified noun entities from previous steps as input. We checked the accuracy by using various classifiers like NaÃŕve Bayes, k-Nearest Neighbour and Support Vector Machine on training data with a 10-fold cross validation. Random Forest gave the best results.

2.3.1 Features for Run 1

The features used for Run 1 were as follows:
- Length of the mention string
- If the mention is all capitalized
- If the mention contains mixed case
- If the mention contains digits
- If internal period is present in mention string
- If present in list of Persons
- If present in list of Things
- If present in list of Events
- If present in list of Characters
- If present in list of Locations
- If present in list of Organizations
- If present in list of Products

The above-mentioned lists are basically the bag of words produced from the training data in the pre-processing step.

[3]http://nlp.stanford.edu/software/CRF-NER.shtml

2.3.2 Features for Run 2

We made use of various text based features and bag of words in Run 1. In Run 2, we explored various contextual features in addition to the features of Run 1. So we combined ten new features with the previous twelve features for Run 2. The ten additional features used in Run 2 were as follows:
- Context score for Person entity
- Context score for Location entity
- Context score for Character entity
- Context score for Organization entity
- Context score for Event entity
- Context score for Thing entity
- Context score for Product entity
- Frequency of Part-of-speech of mention
- Frequency of previous Part-of-speech
- Frequency of next Part-of-speech

Context score of a particular mention is calculated for a three word window of the mention. For each class, we have the number of occurrences of each word in a three word window. While calculating the feature value, we assign the sum of the frequency of the words forming that fixed-size window as the context score of mention.

2.3.3 Run 3

We wanted to apply a Feed-Forward neural network (also called the back-propagation networks and multilayer perceptron) to our feature set and see how it performs as these kind of Artificial Neural Networks are useful in constructing a function where the complexity of the feature values makes the decision for building such a function by hand almost impossible. We took the same features of Run 2 and employed a feed-forward neural network based regression model with 5 hidden layers.

For the previous two runs, i.e. Run1 and Run2, the tags from Stanford NER were considered as the primary influence over our classifier tags as its accuracy was quite good. For Run 3 however, we omit the Stanford NER influence and let only the neural network model do the tagging to check the efficiency of our classifier.

2.4 Linking Mentions to DBpedia

We used the Babelfy java API service [3] to address the task of entity linking to DBpedia 2015-04 resources. It is a unified, multilingual, graph-based approach to Entity Linking and Word Sense Disambiguation based on a loose identification of candidate meanings coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations [4] . The Babelfy parameters that we tuned according to our preferences are:

setAnnotationType was set to identify both concepts and named entities,

setMatchingType was set to exact matching,

setMultiTokenExpression was on to identify multi-word tokens,

setScoredCandidates was set in a way so that it obtains only top-scored candidate from the disambiguation list.

The rest of the parameters were kept to their default value. The named entities identified by both Babelfy and ARK Tagger were allowed to the linking stage. Initially, we provided the original tweet texts as input to Babelfy. We observed that the number of named entities and concepts recognized and linked solely by Babelfy service was quite low. The named entity recognition suffered because of the

Table 1: Summary of Experimental Results

	Precision	Recall	F1
Run1			
Strong Mention Match	0.729	0.626	0.674
Strong Typed Mention Match	0.301	0.259	0.278
Strong Link Match	0.586	0.161	0.252
Mention ceaf	0.699	0.600	0.646
Run2			
Strong Mention Match	0.729	0.626	0.674
Strong Typed Mention Match	0.144	0.124	0.133
Strong Link Match	0.586	0.161	0.252
Mention ceaf	0.699	0.600	0.646
Run3			
Strong Mention Match	0.729	0.626	0.674
Strong Typed Mention Match	0.411	0.353	0.380
Strong Link Match	0.586	0.161	0.252
Mention ceaf	0.699	0.600	0.646

noisy nature of tweet text. However, the accuracy of the linked resources was satisfactory. So, we modified our system by altering the tweets slightly. We removed the # and and considered only the alphabets from an already recognized named entity (tagged by the ARK tagger). After successfully linking such named entities, we searched for more entities which were syntactically similar to the previously known entities. We linked these new entities to corresponding DBpedia resources and also obtained the disambiguation scores.

2.5 Clustering of NIL Mentions

The entities which could not be linked to any existing DBpedia resource are supposed to have NIL identifiers so that each NIL may be reused if there are multiple mentions in the text which represent the same (s/similar/identical) entity. We have considered only a spelling based approach here to calculate the similarity between entities. Two unlinked entities are taken to be similar if one of them contains the other (letter only). In that case, the new entity is assigned the same NIL identifier as that of the previous one.

3. RESULTS

We evaluated our approach on the development set consisting of 100 tweets made available by the organizers. In Table 1 we have reported on the official metrics for entity detection, tagging, clustering and linking. The precision, recall and f-scores for the above-mentioned three runs show that the Run 3 produces best results for the task with f-score 0.674, 0.380, 0.252 and 0.646 in the categories Strong Mention Match, Strong Typed Mention Match, Strong Link Match and Mention Ceaf respectively.

While all the Runs yield same score in other categories, in Strong Typed Mention Match, we observe better result for our feed-forward neural network model. Our systems for the three different runs only differ in entity type classification module while all other subtasks follow the same system in all three cases. This results in same result in the last two categories which were mainly the evaluation metrics for linking and nil clustering.

4. CONCLUSION

In this paper, we have described our approach for the #Microposts2016 Named Entity rEcognition and Linking (NEEL) challenge. We have developed a hybrid system using the existing Named Entity Recognizer systems and Twitter-specific Part-of-Speech Taggers in conjunction with the classifier developed by us. The Named Entity Linking was done mainly by using Babelfy, which performs as a multilingual encyclopedic dictionary and a semantic network. The performance of our system suffered because of certain restrictions in time. The classification module was slightly biased and the accuracy of classification suffered as result of that. Identifying and selecting better features would have improved results. Also a disambiguation module to treat overlapping classes would have been useful. The accuracy of the linking would also improve by taking a semantic similarity approach using synonym sets for the mentions or context word overlapping from the sets while NIL clustering.

5. REFERENCES

[1] K. Gimpel, N. Schneider, B. OĆonnor, 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 *ACL 2011*, pages 42–47, 2011.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11:231–244, 2009.

[3] A. Moro, F. Cecconi, and R. Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *13th International Semantic Web Conference, Posters and Demonstrations (ISWC 2014)*, pages 25–28, Riva del Garda, Italy, 2014.

[4] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, pages 231–244, 2014.

[5] D. Nadeau. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD thesis, Ottawa-Carleton Institute for Computer Science, School of Information Technology and Engineering, 2007.

[6] D. Preoţiuc-Pietro, D. Radovanović, A. E. Cano-Basave, K. Weller, and A.-S. Dadzie, editors. *Proceedings, 6th Workshop on Making Sense of Microposts (#Microposts2016): Big things come in small packages, Montréal, Canada, 11th of Apr 2016*, 2016.

[7] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP 2011*, pages 1524–1534, 2011.

[8] G. Rizzo, M. van Erp, J. Plu, and R. Troncy. Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. In Preoţiuc-Pietro et al. [6], pages 50–59.

[9] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*, pages 1361–1370, Mumbai, 2012.