COMBINING NAMED ENTITY RECOGNITION TOOLS

Štefan DLUGOLINSKÝ, Peter KRAMMER, Marek CIGLAN,
Michal LACLAVÍK, Ladislav HLUCHÝ

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
{upsysdlu,upsypkra,upsymaci,laclavik.ui,upsylhlu}@savba.sk

Abstract. In this paper, we describe our successful submission to the #MSM2013 IE Challenge organized under WWW2013 conference. The challenge was aimed at concept extraction from microposts restricted to four classes; i.e. PER, ORG, LOC and MISC. Our approach was based on a combination of several existing NER tools, which used different classification methods. In prior evaluation of these tools, we have observed that some of the tools performed better on different entity types than other tools. In addition, different tools produced diverse results, which brought a higher recall when simply combined than that of the best individual tool. But as expected, the precision went significantly down. We proposed a more advanced method of combination involving machine-learning techniques and trained several classification models. Evaluation results showed that several classification models have achieved better results than the best individual extractors.

1. Introduction

A significant growth of social media interaction can be observed in recent years. The easiest and probably the most popular way of interaction on the Web is through microposts – short text messages posted on the Web. Notorious examples of microposts include tweets, Facebook statuses, comments, Google+ posts, Instagram photos. Microposts analysis has a big potential and knowledge hidden in microposts can be used in wide range of domains. The most important task in order to analyze and make sense of microposts is the Named Entity Recognition (NER). NER in microposts is a challenging problem because of a limited size of a single micropost, prevalence of term ambiguity, noisy content or multilingualism.

There was a challenge organized to foster research into novel, more accurate concept extraction for Micropost data. It was the Making Sense of Microposts Workshop

(#MSM2013) Concept Extraction Challenge, hosted in conjunction with the 2013 World Wide Web conference (WWW'13). The challenge required participants to build semi-automated systems to identify concepts within Microposts and extract matching entity types for each concept; i.e. PER, LOC, ORG and MISC. Out of a total of 22 complete submissions 13 were accepted for presentation at the workshop [3]. Our team's submission Annotowatch was also accepted. In this paper, we give a short overview of our approach taken in the challenge.

2. Evaluation of NER Tools for Combining

In our evaluation of several chosen NER tools; i.e. ANNIE, Apache OpenNLP, Illinois Named Entity Tagger, Illinois Wikifier, Open Calais, Stanford Named Entity Tagger and Wikipedia Miner we observed that the tools return diverse results [2]. But when properly combined, this might lead to a new composite named entity (NE) recognizer that performs better than any individual classifier on its own. For example, a simple composite recognizer might incorporate the best performing tool for each NE class. The diversity of the results can be seen in Figure 1. If we make a union of all the entities recognized by the evaluated tools, we get a very high recall, but with a drawback of a very low precision. Machine learning techniques such as decision trees can help us to choose appropriate technique in a specific context and thus improve the precision, while keeping the recall relatively high.

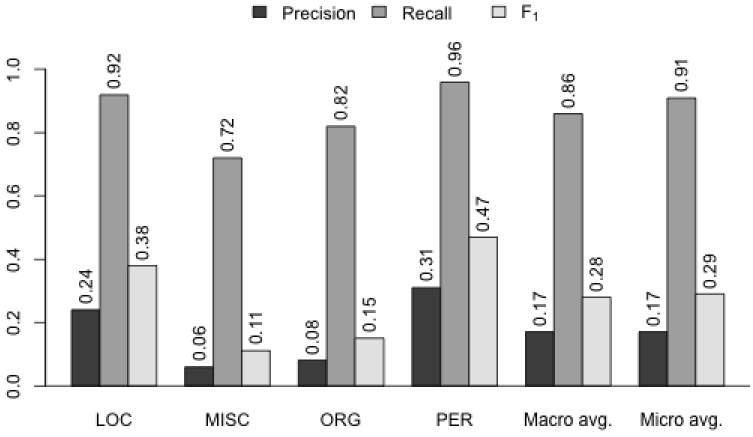

Figure 1. Performance of unified tools.

3. Combining the Tools

We have defined a baseline NE recognizer according to the evaluation. The baseline NE recognizer was built as a combination of the best NE recognizers for each target entity class. It included OpenCalais for LOC, MISC and ORG named entity types and Illinois NET for PER named entity type. Our goal was to overcome the performance of the baseline NE recognizer with a model produced by machine learning approach. We have examined several machine learning techniques and tried them for combining the NE recognizers. The point of our approach was in describing how particular tools performed on different entity types compared to the response of other tools and a manual annotation. This relation defined a training vector for the machine-learning step. Then a set of training vectors was generated from training dataset and used for model training. To get an idea of our candidate models performance, we have trained them on an 80% split of training dataset cleaned from duplicate records and have evaluated them on the remaining 20% split. The best performing NER models were random forest and decision tree based on C4.5 algorithm. NER models produced by these algorithms have achieved performance superior to that of underlaying NE recognizers as well as the baseline recognizer. An example of annotated micropost by one of our best performing model DTJ48 M13[1] and by underlaying tools is depicted in Figure 2. More detailed information is available in [2].

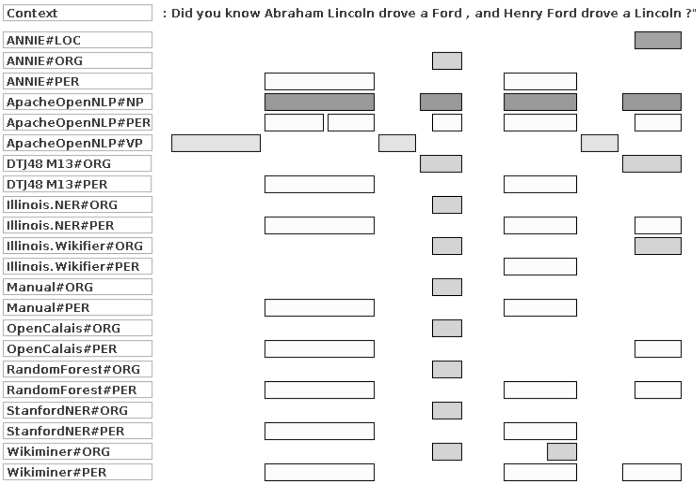

Figure 2. An example of annotated micropost.

4. Conclusions

We have submitted three runs in the challenge based on the best performing model evaluated on the 20% split of training dataset. They were different in configuration of training algorithm and post-processing of annotations. The challenge committee has evaluated one of our runs as the second best in F1 among the final 13 accepted submissions, as it was the best in recall and the second best in precision [3].

Acknowledgement: This work was supported by projects VEGA 2/0185/13, VENIS FP7-284984, TRADICE APVV-0208-10 and ITMS: 26240220072.

[1] This model was trained after the challenge deadline

References

to other papers publishing the results that are summarized here

[1] Dlugolinský, Š., Krammer, P., Ciglan, M., Laclavík, M.: MSM2013 IE Challenge: Annotowatch. In A. E. C. Basave, M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, Making Sense of Microposts (#MSM2013) Concept Extraction Challenge, pages 21–26, May 2013.

[2] Dlugolinsky, S., Ciglan, M., Laclavik, M.: Evaluation of named entity recognition tools on microposts. In Intelligent Engineering Systems (INES), 2013 IEEE 17th International Conference on, pages 197–202, 2013.

Other references

[3] Basave, A. E. C., Rowe, M., Stankovic, M., Dadzie, A.-S. editors: Proceedings, Concept Extraction Challenge at the 3rd Workshop on Making Sense of Microposts (#MSM2013): Big things come in small packages, Rio de Janeiro, Brazil, 13 May 2013, May 2013.