Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning

Marieke van Erp
The Network Institute
VU University Amsterdam,
The Netherlands
marieke.van.erp@vu.nl

Giuseppe Rizzo
EURECOM
Sophia Antipolis, France
giuseppe.rizzo@eurecom.fr

Raphaël Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

ABSTRACT

Microposts shared on social platforms instantaneously report facts, opinions or emotions. In these posts, entities are often used but they are continuously changing depending on what is currently trending. In such a scenario, recognising these named entities is a challenging task, for which off-the-shelf approaches are not well equipped. We propose NERD-ML, an approach that unifies the benefits of a crowd entity recognizer through Web entity extractors combined with the linguistic strengths of a machine learning classifier.

Keywords

Named entity recognition, NERD, machine learning

1. INTRODUCTION

Microposts are a highly popular medium to share facts, opinions or emotions. They promise great potential for researchers and companies alike to tap into a vast wealth of a heterogeneous and instantaneous barometer of what is currently trending in the world. However, due to their brief and fleeting nature, microposts provide a challenging playground for text analysis tools that are oftentimes tuned to longer and more stable texts. We present an approach that attempts to leverage this problem by employing an hybrid approach that unifies the benefits of a crowd entity recogniser through Web entity extractors combined with the linguistic strengths of a machine learning classifier.

2. THE NERD-ML SYSTEM

In our approach, we combine a mix of NER systems in order to deal with the brief and fleeting nature of microposts. The three main modules of our approach are: NERD, Ritter et al.'s system, and Stanford NER. NERD [4] is used to spot entities using a variety of Web extractors. The strength of this approach lies in the fact that these systems have access to large knowledge bases of entities such as DBpedia[1] and Freebase[2]. Ritter et al. [3] propose a tailored approach for entity recognition based on a previously annotated Twitter stream; while Stanford NER [1] represents the state of the art in the entity recognition, providing off-the-shelf or customisable NER using a machine learning algorithm. While NERD and Ritter et al.'s approach are used as off-the-shelf extractors, Stanford NER is trained on the MSM training dataset. The outputs of these systems are used as features for NERD-ML's final machine learning module. We have also added extra features based on the token and the micropost format to further aid the system. The generated feature sets can be fed into any machine learning algorithm in order to learn the optimal extractor/feature combination. An overview of our system is shown in Figure 1. In the remainder of this section we explain the components.

[1] http://www.dbpedia.org
[2] http://www.freebase.com

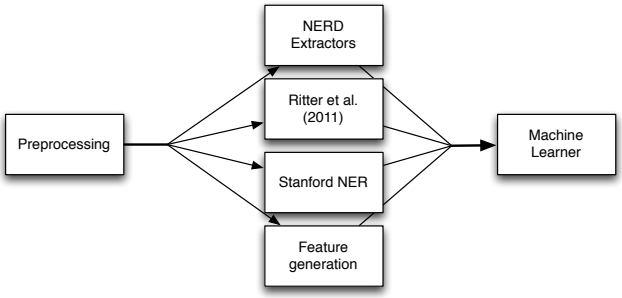

Figure 1: Overview of the NERD-ML System

Preprocessing: In the preprocessing phase, the data is formatted to comply with the input format of our extractors. For ease of use, the dataset is converted to the CoNLL IOB format [5]. Furthermore, posts from the MSM2013 training data are divided randomly over 10 parts in order to a) be able to perform a 10-fold cross-validation experiment and b) comply with NERD filesize limitations.

NERD Extractors: Each of the data parts is sent to the NERD API to retrieve named entities from the following extractors: AlchemyAPI, DBpedia Spotlight (setting: *confidence=0*, *support=0*, *spotter=CoOccurrenceBasedSelector*), Extractiv, Lupedia, OpenCalais, Saplo, TextRazor, Wikimeta, Yahoo and Zemanta (setting: *markup_limit=10*). The NERD ontology consists of 75 classes, which are mapped to the four classes of the MSM2013 challenge.

Ritter et al. 2011: The off-the-shelf approach as described in [3] is taken both as baseline and input for the hybrid classifier. The 10 entity classes are mapped to the four classes of the MSM2013 challenge.

Stanford NER: The Stanford NER system (version 1.2.7) is retrained on the MSM2013 data challenge set, using parameters based on the english.conll.4class.distsim.crf.ser.gz properties file provided with the Stanford distribution. The Stanford results serve as a baseline, as well as input for the hybrid classifier.

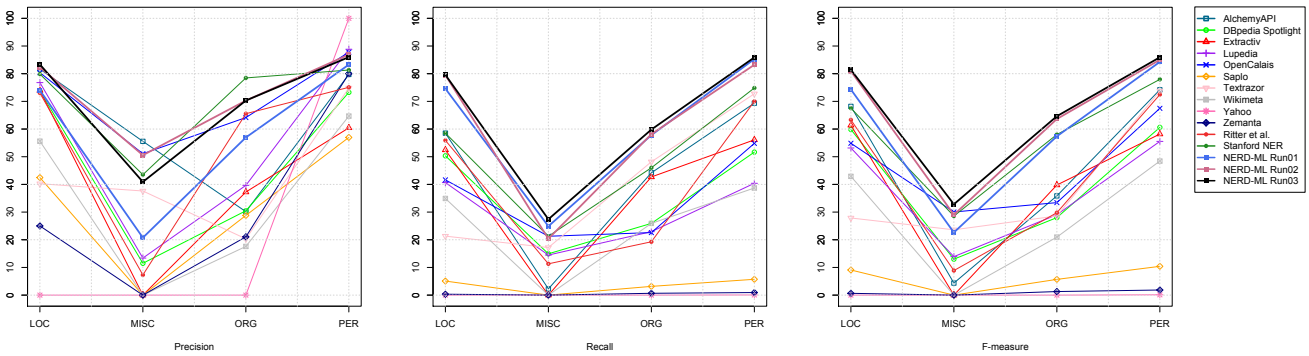

Figure 2: Results of individual and combined extractors in 10-fold cross validation experiments

Feature Generation: To aid the classifier in making sense of the structure of the microposts, we added 8 additional features to the dataset inspired by the features described in [3]. We implemented the following features: capitalisation information (initial capital, allcaps, proportion of tokens capitals in the micropost), prefix (first three letters of the token), suffix (last three letters of the token), whether the token is at the beginning or end of the micropost, and part-of-speech token using the TwitterNLP tool and POS-tagset from [2].

NERD-ML: The output generated by the NERD extractors, Ritter et al.'s system, Stanford NER system and the added features are used to create feature vectors. The feature vectors serve as input to a machine learning algorithm in order to find combinations of features and extractor outputs that improve the scores of the individual extractors. We experimented with several different algorithms and machine learning settings using WEKA-3.6.9[3].

3. RESULTS

In Figure 2, the results of the individual NER components and the hybrid NERD-ML system are presented. The first run is a baseline run that includes the full feature set. The second run only includes the extractors and no extra features. The third run uses a smaller feature set that was compiled through automatic feature selection. The settings of the three runs of the hybrid NERD-ML system are:

Run 1: All features, k-NN, k=1, Euclidean distance, 10-fold cross validation

Run 2: AlchemyAPI, DBpedia Spotlight, Extractiv, Lupedia, OpenCalais, Saplo, Yahoo, Textrazor, Wikimeta, and Zemanta, Stanford NER, Ritter et al., SMO, standard parameters, 10-fold cross validation

Run 3: POS, Initial Capital, Suffix, Proportion of Capitals, AlchemyAPI, DBpedia Spotlight, Extractiv, Opencalais, Textrazor,Wikimeta, Stanford NER, Ritter et al., SMO, standard parameters, 10-fold cross validation

Results are computed using the conlleval script and plotted using R. All settings and scripts are publicly available[4].

[3] `http://www.cs.waikato.ac.nz/ml/weka/`
[4] `https://github.com/giusepperizzo/nerdml`

4. CONCLUSIONS

Extracting named entities from microposts is a difficult task due to the ever-changing nature of the data, breadth of topics discussed and linguistic inconsistencies it contains. Our experiments with NERD-ML show that the combination of different NER systems outperforms off-the-shelf approaches, as well as the customised Stanford approach. Our results indicate that an hybrid system may be better equipped to deal with the task of identifying entities in microposts, but care must be taken in combining features and extractor outputs.

Acknowledgments

This work was partially supported by the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911).

5. REFERENCES

[1] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *43nd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, USA, June 2005.

[2] 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 *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, Atlanta, GA, USA, June 2013.

[3] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Empirical Methods in Natural Language Processing (EMNLP'11)*, Edinburgh, UK, July 2011.

[4] G. Rizzo and R. Troncy. NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *13th Conference of the European Chapter of the Association for computational Linguistics (EACL'12)*, Avignon, France, April 2012.

[5] E. F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Conference on Computational Natural Language Learning (CoNLL'02)*, Taipei, Taiwan, Aug-Sept 2002.