Leveraging Existing Tools for Named Entity Recognition in Microposts

Fréderic Godin[†], Pedro Debevere[†], Erik Mannens[†],
Wesley De Neve[†*], and Rik Van de Walle[†]

[†] Multimedia Lab, Ghent University - iMinds, Ghent, Belgium
[*] Image and Video Systems Lab, KAIST, Daejeon, South Korea
{frederic.godin,pedro.debevere,erik.mannens,
wesley.deneve,rik.vandewalle}@ugent.be

Abstract. With the increasing popularity of microblogging services, new research challenges arise in the area of text processing. In this paper, we hypothesize that already existing services for Named Entity Recognition (NER), or a combination thereof, perform well on microposts, despite the fact that these NER services have been developed for processing long-form text documents that are well-structured and well-spelled. We test our hypothesis by applying four already existing NER services to the set of microposts of the MSM2013 IE Challenge.

Keywords: microposts, NER, text processing

1 Introduction

Research in the domain of text processing has traditionally focused on analyzing long-form text documents that are well-structured and well-spelled [1]. However, thanks to the high popularity of microblogging sites, research in the domain of text processing is increasingly paying attention to the analysis of microposts as well. Microposts are short-form text fragments that are typically noisy in nature, hereby lacking structure and often containing a substantial amount of slang and misspelled words, frequently in multiple languages. In this paper, we hypothesize that already existing services for Named Entity Recognition (NER), as often used for processing news corpora, perform well on microposts, even without preprocessing, and that future research efforts should regard these NER services as a strong baseline.

2 Evaluation of existing services

Current NER services are tailored to processing long-form text documents that are typically well-structured and well-spelled. Rizzo *et al.* [2] quantitatively evaluated six NER web services on three types of corpora: 5 TED talks, 1000 news articles of the New York Times, and 217 WWW conference abstracts. In this paper, we aim at complementing this evaluation by testing the effectiveness of

these services on a fourth fundamentally different text corpus, namely the microposts of the MSM2013 IE Challenge. Because both Evri and Extractiv are no longer available, we had to limit ourselves to the testing of four services, namely AlchemyAPI[1], DBpedia Spotlight[2], OpenCalais[3], and Zemanta[4].

To test the effectiveness of the aforementioned services, we did not apply any type of preprocessing. Given the MSM2013 IE Challenge guidelines, we evaluated the recognition of four types of entities: persons, locations, organizations, and a set of miscellaneous entities. The miscellaneous category contains the following entities: movies, entertainment award events, political events, programming languages, sporting events, and TV shows.

Given that the services evaluated make use of ontologies that are much more elaborate, we mapped the service ontologies to the four entity types. We evaluated a total of 2813 microposts of the training set. We left out microposts 583 and 781 because OpenCalais could not handle them. Because we used an ontology mapping, our results can differ with other evaluations. We report our results in Table 1.

Table 1. Evaluation of four different services: AlchemyAPI (A), DBpedia Spotlight (S), OpenCalais (O), and Zemanta (Z). For DBpedia Spotlight, we evaluated two configurations: confidence=0.2 and confidence=0.5.

	PER			LOC			ORG			MISC		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
A	81.1%	75.6%	**78.2%**	81.2%	69.0%	**74.6%**	59.5%	50.2%	54.4%	54.2%	5.6%	10.2%
S (0.2)	54.6%	61.0%	57.6%	44.8%	48.1%	46.4%	16.1%	49.7%	24.4%	2.7%	40.7%	5.0%
S (0.5)	87.0%	20.3%	32.9%	54.5%	1.9%	3.7%	19.7%	3.9%	6.5%	5.8%	10.0%	7.3%
O	71.7%	67.2%	69.3%	81.8%	66.1%	73.1%	72.2%	45.5%	**55.8%**	46.2%	23.8%	**31.4%**
Z	91.0%	57.4%	70.4%	83.9%	52.1%	64.3%	71.9%	36.1%	48.1%	37.1%	24.2%	29.3%

Table 2. Evaluation of the Random Forest (RF)-based model for predicting entity types, using 10-fold cross validation. Dependent on the DBpedia Spotlight results obtained, we evaluated two configurations.

	PER			LOC			ORG			MISC		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
RF (0.2)	78.4%	86.3%	**82.2%**	80.9%	71.1%	**75.7%**	62.8%	58.1%	**60.4%**	62.0%	38.3%	**47.4%**
RF (0.5)	75.0%	89.5%	81.6%	81.7%	68.2%	74.3%	71.9%	50.6%	59.4%	62.2%	30.0%	40.5%

[1] http://www.alchemyapi.com/
[2] http://dbpedia.org/spotlight/
[3] http://www.opencalais.com/
[4] http://www.zemanta.com/

As highlighted in bold, AlchemyAPI outperforms the other three services in identifying persons and is a close first in recognizing locations. On the other hand, OpenCalais performs best in recognizing organizations and MISC entities. Although Zemanta never wins, this service is characterized by a high precision. DBpedia Spotlight performs poorly because it returns an extensive list of possible entity types that often adhere to all four categories, instead of returning a single entity type.

When zooming in on the individual results, we can notice that AlchemyAPI performs bad in recognizing exotic names, small villages and buildings (e.g., St. Georges Mill), and recognizing abbreviations of organizations (e.g., DFID and UKGov). Furthermore, AlchemyAPI performs poorly in recognizing well-known events and TV shows such as "Super Bowl" and "Baywatch". Zemanta suffers from similar problems. However, Zemanta performs worse than AlchemyAPI because it is more dependent on the usage of capital letters (e.g., Uruguay - uruguay and URUGUAY). We can observe similar behavior for OpenCalais and AlchemyAPI, for recognizing locations and organizations. OpenCalais is also capable of recognizing well-known events like the Super Bowl. When the confidence is set high (0.5), a lot of well-known entities cannot be recognized by DBpedia Spotlight, such as "Katy Perry". When the confidence is set low (0.2), "Katy Perry" is recognized but a lot of noise is recognized as a person too (e.g., love, follow, guy).

3 Combining existing services

To further improve the results of NER on the training set, we combined the outputs of the different services. E.g., one can imagine that it is more plausible that a word is an entity when multiple services claim this with high confidence than when only one service claims this with low confidence. For each of the recognized entities, we constructed a feature vector and classified it using the technique of Random Forest. The goal was to predict one of the four entity types. For each service, our feature vector contained an element referencing one of the four challenge entity types, the original entity type according to the service ontology used, and a confidence and/or relevance value. In the case of DBpedia Spotlight, we omitted the original entity type element because this element was too sparse. We created a negative set by making use of the entities that were recognized by the services, but that were not in the training set.

We evaluated our set of feature vectors by means of the Weka toolkit. We applied 10-fold cross validation. We made use of two sets: the first set contained the DBpedia Spotlight results when querying this service with a confidence of 0.2, whereas the second set contained the DBpedia Spotlight results when querying this service with a confidence of 0.5. We applied Random Forest with 20 trees and four attributes per tree. We report the results of our evaluation in Table 2.

We highlighted the best results of our Random Forest-based fusion approach in bold for categorizing entity types. When we make use of the entities recognized by DBpedia Spotlight with a low confidence as part of the feature vector, the

use of Random Forest leads to better results than when making use of high-confidence DBpedia Spotlight results. Applying Random Forest on noisy data with low precision and recall values yields significant improvements. Especially in the MISC category where we obtained an improvement of almost 7%. (Note: The result in Table 1 and 2 cannot be compared directly because the evaluation was conducted in a different way. In Table 1, this was on a word-by-word basis. In Table 2, this was on a entity type-by-type basis.)

The next step is to make use of this categorization approach to decide whether we should trust the combined result of the different services for recognizing a certain named entity type. The final evaluation of the proposed algorithm is part of the Making Sense of Micropost Challenge 2013 and was conducted on the test set. The results were presented at the workshop itself and were therefore not available yet at the time of writing.

4 Conclusions

In this paper, we have shown that existing NER services can recognize named entities in microposts with high F_1 values, especially when aiming at the recognition of persons and locations. In addition, we have demonstrated how the results of several services can be combined with the goal of achieving a higher precision. We can conclude that already existing NER services make for a strong baseline when aiming at the design and testing of new NER algorithms for microposts.

5 Acknowledgments

The research activities in this paper were funded by Ghent University, iMinds, the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union.

References

1. M. W. Berry and J. Kogan, editors. *Text Mining: Applications and Theory.* Wiley, Chichester, UK, 2010.
2. G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW 2012, 5th Workshop on Linked Data on the Web*, 2012.