UNIALI: CREAZIONE DI UN SISTEMA PER IL RICONOSCIMENTO DI ENTITA' NOMINATE PER L'ITALIANO INDIPENDENTE DALLE RISORSE

UNIALI: BUILDING A RESOURCE INDEPENDENT SYSTEM FOR ITALIAN NAMED ENTITY RECOGNITION

ZORNITSA KOZAREVA · ANDRÉS MONTOYO

SOMMARIO/*ABSTRACT*

Per il task di riconoscimento di entita' nominate di EVALITA, abbiamo sviluppato un sistema indipendente dalla lingua e dalle risorse usate, basato su tecniche di apprendimento automatico.

For the Italian Named Entity Recognition challenge, we have developed a resource and language independent system, which is based on machine learning techniques. The features capture the lexical and the orthographic information of the words. Our system ranked on ninth position with performance of 66.59%.

Keywords: named entity recognition, language and resource independent

1 Introduction

The Named Entity Recognition (NER) task consists of the identification and classification of name entities in a text. In the Italian NER challenge EVALITA, the NER system must classify the detected names into four categories: person, organization, geopolitical names such as cities, countries, capitals, and locations such as streets, avenues, addresses.

There are two types of NER approaches: a rule-based and a machine-learning one. For the construction of a rule-based systems, we need to integrate a lot of knowledge, plenty of hand-written rules and large amounts of gazetteers. While for the machine-learning systems, we can encode a series of features into vectors, learn a classification criteria and separate automatically the NE candidates into classes.

Among the authors of our Italian NER system, there are no Italian speaking members. Therefore, the creation of rules or the usage of Italian grammar would have been difficult. For this reason, we decided to develop a machine learning system.

2 System Architecture

Our Italian NE recognition system identifies and classifies the NEs in one step. Some of the characteristics are integrated from the previously developed Spanish NER system of Kozareva et. al, [1]. The difference with the Italian system is that the current features are generated on sentence level and that we did not use any triggers or pos-tagging information. Our feature sets is:

- $cont$: word forms of w_0, w_{-1}, w_{+1}
- cap: w_0, w_{+1}, w_{-1} whole in capitals
- low: w_0, w_{+1}, w_{-1} whole in lowercase
- cl: w_0, w_{+1}, w_{-1} start in capital, followed by lowercase
- cd: w_0, w_{+1}, w_{-1} start in capital, contain dot
- cn: w_0, w_{+1}, w_{-1} start in capital, contain digit
- cs: w_0, w_{+1}, w_{-1} start in capital, contain special symbol
- lc: w_0, w_{+1}, w_{-1} start in lowercase, contain capital
- ld: w_0, w_{+1}, w_{-1} start in lowercase, contain dot
- ln: w_0, w_{+1}, w_{-1} start in lowercase, contain digit
- ls: w_0, w_{+1}, w_{-1} start in lowercase, contain special symbol
- d: w_0, w_{+1}, w_{-1} are digits
- dx: w_0, w_{+1}, w_{-1} start in digit, contain other characters
- $bigrams$: $w_{-1}w_0$ and w_0w_{+1}
- enc: encode capitalization with C, X, L, where C-capital, L-lowercased, X-other
- $start$: w_0 is at the beginning of a sentence
- end: w_0 is at the end of a sentence
- end: binary attribute for whether the sentence has more than 3 words or not
- wDP: w_0, w_{+1}, w_{-1} in dictionary of person names
- wDL: w_0, w_{+1}, w_{-1} in dictionary of location names

The generation of the gazetteer lists is done on the basis of a graph-based and pattern matching algorithms [2]. The Italian text collections from which the names are gathered come from CLEF[1]. The number of extracted locations is 8227 and the number of person names is 1084.

The machine learning algorithm which was used in the experiments is the memory based nearest neighbor algorithm TiMBL [3]. During the development of our approach, we have conducted a five-cross validation in order to establish the definitive attributes for our approach. The performance of the different training sets varied from 61 to 76%.

3 Evaluation with the EVALITA data set

The performance of our system with the EVALITA test data sets is shown in Table 1. The system ranked on ninth position from eleven participants and outperformed the baseline with 25%. The 66.59% performance can be improved with around 10% when organization gazetteer lists are incorporated. Our claim is supported by the high performance of the person and location names for which we had gazetteer lists. With this additional information, the person names reached 78.66% f-score and the locations reached 72.60%.

class	P.	R.	F.
GPE	70.50	74.84	**72.60**
LOC	40.59	56.56	47.26
ORG	42.77	54.21	47.81
PER	76.36	81.11	**78.66**
total	62.73	70.95	66.59

Table 1: Performance of the UniAli NER system

Some of the occurred errors concern the identification of the tag sequence and the length of the NEs. For instance, the expression “Opera universitaria di Trento” should be identified as organization, meanwhile our system determines “Opera” as organization and “Trento” as geopolitical entity. Similar error comes with “provincia di Trento” where only Trento is detected as geopolitical entity and the other two words are omitted. In order to improve the detection, trigger words which reveal the external properties of the entities should be collected.

[1] www.clef-campaign.org/

4 Conclusions

For the participation in the Italian NER challenge, we have developed a language and resource independent system. The advantages of the system are its ease to adapt to other domains or languages. During the experimental evaluation, the best performance was obtained for the location and person categories, therefore in the future we will focus toward the automatic collection of organization names. The participation in this challenge served as a confirmation about the performance of our system which was previously evaluated with the Spanish, English and Portuguese languages. The obtained performance of the system is 66.59% which is with 25% better than the baseline. We can claim that the obtained results are promising given the usage of the low level information.

5 Acknowledgements

This research has been funded by QALL-ME number FP6 IST-033860 and TEXT-MESS number TIN2006-15265-C06-01.

REFERENCES

[1] Z. Kozareva, O. Ferrandez, A. Montoyo, R. Muñoz, A. Suárez and J. Gómez. Combining data-driven systems for improving Named Entity Recognition. *Journal of Data and Knowledge Engineering*, volume 61, 3, pp. 449–466, 2007.

[2] Z. Kozareva. Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists. *Proceedings of the EACL*, 2006.

[3] W. Daelemans, J. Zavrel, K. Sloot and A. Bosch”, TiMBL: Tilburg Memory-Based Learner.” *ILK 03-10*,2003

CONTACTS

ZORNITSA KOZAREVA, ANDRÉS MONTOYO
DLSI, Universidad de Alicante
03080 Alicante, Spain
Email: {zkozareva | montoyo}@dlsi.ua.es

ZORNITSA KOZAREVA is a PhD student in Computational Linguistics at the University of Alicante since November 2004. Her research interests are Information Extraction, Machine Learning and Textual Entailment.

ANDRES MONTOYO is a full-time professor at the University of Alicante. His research interests are Information Extraction, Word Sense Disambiguation and Textual Entailment. He is responsible for several projects and edited three books.