Named Entity Recognition through Redundancy Driven Classifiers

Roberto Zanoli, Emanuele Pianta, and Claudio Giuliano

FBK-irst,
via Sommarive 18, I-38123 Povo (TN), Italy
{zanoli, pianta, giuliano}@fbk.eu

Abstract. We present Typhoon, a classifier combination system for Named Entity Recognition (NER), in which two different classifiers are combined to exploit *Data Redundancy* and *Patterns* extracted from a large text corpus. *Data Redundancy* is attained when the same entity occurs in different places in documents, whereas P*atterns* are 2-grams, 3-grams, 4-grams and 5-grams preceding, and following entities in documents. The system consists of two classifiers in cascade, but it is possible to use a single classifier making the system faster (100 times faster, with a speed rate of about 20,000 tokens/sec); whereas the second classifier in the cascade can be used when more accuracy is needed. Moreover the system can use additional features such as that given by using a Text Classifier able to recognize the category to which the story belongs. The system performed the best on the task of Italian NER at EVALITA 2009, with an F_1 of 0.82.

Keywords: NER, Named Entity Recognition, Entity Detection, Entity Detection and Recognition.

1 Introduction

This paper investigates the combination of two different classifiers on the task of Named Entity Recognition (NER). NER is a subtask of Information Extraction which aims to classify words in text into predefined categories. Examples of named entities are person names, location and organization names, date and time indications, etc. Spurred on by the Message Understanding Conferences (MUC), a considerable amount of work has been done in last years on the Named Entity Recognition. The most representative machine-learning approaches used in NER are Hidden Markov Model (HMM) [1], Maximum Entropy [2], Support Vector Machines (SVMs), and Conditional Random Fields (CRFs) [3, 4], whereas attempts have been made to use global information (e.g., the same named entity occurring in different sentences of the same document) [5, 2, 8]. Drawing from our participation at Evalita 2007 [6] and at ACE08[1], we built Typhoon, a system for Named Entity Recognition in which two different classifiers are combined in a cascade to exploit *Data Redundancy* and *Patterns* extracted from a large text corpus. *Data Redundancy* is attained when the same entity occurs in different places in documents, whereas *Patterns* are 2-grams, 3-grams, 4-grams and 5-grams preceding, and

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

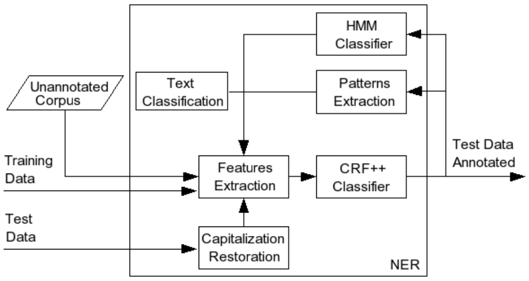
following recognized entities in documents. The rest of the paper is organized as follows: Section 2 gives an introduction to the Evalita 2009 NER task, Section 3 describes the system architecture, Section 4 describes the experiments, and Section 5 analyses the results we obtained.

2 The task description

In the Named Entity Recognition task at Evalita 2009, systems are required to recognize the Named Entities occurring in a text by identifying their extension and type, i.e. Person (PER), Organization (ORG), Geo-Political Entity (GPE) and Location (LOC). Both training data and test data are part of the Italian Content Annotation Bank (ICAB), developed in the context of the Ontotext Project[2]. Training data consists of 525 news stories taken from the local newspaper L'Adige for a total of around 180,000 words, whereas the test data consists of about 86,000 words. Both the datasets were provided with Part of Speech (PoS) annotation, while the training data also contains the information about the category (Attualita', Cronaca, Cultura, Economia and Sport) to which the story belongs. The use of external resources was allowed and systems were evaluated in terms of Precision (Pr), Recall (Re) and F_1 measure by using the CoNLL 2002 scorer.

3 The system architecture

The system consists of two different classifiers in cascade: the CRF++[3] classifier that is a customizable and open source implementation of Conditional Random Fields (CRFs), and the HMM classifier based on disambig[4], an implementation of the Hidden Markov Models. In the first phase the CRF++ trained on the training data is used to identify Named Entities in the large collection of documents and classify them into the predefinited four categories: Person, Organization, Geo-Political Entity and Location. In the second phase the HMM classifier, trained on the data produced by the previous phase, recognises entities both in the training data and in test data; furthermore in this phase a number of patterns (about 90,000) are extracted : 2-grams, 3-grams, 4-grams and 5-grams preceding and following each recognized entity in the collection of documents. Finally the CRF++ exploits the data of the second phase as features to perform the final annotation.



[2] http://tcc.fbk.eu/projects/ontotext/

[3] http://crfpp.sourceforge.net

[4] http://www.speech.sri.com/projects/srilm/

The system can use additional features such as that given by the category of the document. As this information was only available for the documents in the training data, we trained a Text Classifier, based on Support Vector Machines (SVMs), to predict the topic of the test documents; tf-idf was used to weight the words and, in addition, chi-squared was used for feature selection. Typhoon was built not only to take part in official evaluation campaigns but also to be employed in real world applications, such as the LiveMemories project (http://www.livememories.org), which aims at scaling up content extraction techniques towards very large scale extraction from multimedia sources. In these type of applications high classification speed is as important as the annotation accuracy: the system architecture we proposed allows the system to work with a single classifier (CRF), making the system faster (100 times faster, with a speed rate of about 20,000 tokens/sec); whereas the second classifier (HMM) will be used in combination to the first one when more accuracy is needed. In addition a system for Capitalization restoration has been implemented when the capitalization information is not available (e.g., speech transcriptions).

4 Experiments and Results

Typhoon was tuned splitting the training data into two parts: a data set for training (132,589 tokens) and a data set for finding the best system configuration (79,889 tokens). Feature selection plays a crucial role in this task; we used a large set of features: the word itself, both unchanged and lowercased, the *Data Redundancy* as well the extracted *Patterns*. Table 1 lists the top 5 4-grams patterns extracted from the large collection of documents.

Table 1. Top 5 4-grams patterns for the three categories GPE, ORG, PER.

GPE	ORG	PER
vigili del fuoco di X	, il presidente della X	, a cura di X
e‘ un comune della X	il segretario generale della X	X (nella foto)
e‘ una citta‘ dell’ X	in collaborazione con l’ X	dopo la morte di X
stati e territori dell’ X	situato nel dipartimento dell’ X	per la regia di X
situato nella provincia di X	l’ ex presidente del X	X , ex calciatore e

In addition the system used the Part of Speech (PoS) produced by TagPro [7], prefixes and suffixes (1, 2, 3, or 4 characters at the start/end of the word); orthographic information (capitalization and hyphenation), gazetteers of generic proper nouns extracted from the Italian phone-book and from Wikipedia, from various sites about Italian and Trentino’s cities, Italian and American stock market and Wikipedia geographical locations. Each of these features was extracted for the current, previous and following word, whereas for both the classifiers we used the default system configuration. Table 2 shows the results obtained on the development set: the baseline was produced considering the tokens only as features, whereas other experiments were made to evaluate the importance of each type of feature in comparison to the baseline.

Table 2. Results on the development set

	Pr	Re	F_1
baseline	83.56	36.70	51.0
Data Redundancy	84.40	75.23	79.55
PoS	73.19	64.69	68.68
Patterns	84.27	38.80	53.14
Topic	83.65	38.08	52.33
All features	84.94	80.28	82.54

Results were reported in terms of Precision, Recall and F_1. As said in Section 3, the system allows to be used as a combination of the two classifier or as a single classifier when higher classification speed is needed. Table 3 refers to the accuracy of the two classifiers when they work separately.

Table 3. CRF, HMM accuracy

	Pr	Re	F_1
HMM	67.15	73.42	70.14
CRF	81.72	77.57	79.59

Typhoon performed as the best system in the Italian Named Entity Recognition task at EVALITA 2009 (evaluation based on exact match). Table 4 gives the official results on the test set.

Table 4. Official results

Category	Pr	Re	F_1
GPE	86.12	84.16	85.13
LOC	72.09	39.74	51.24
ORG	71.71	69.43	70.56
PER	90.29	86.42	88.31
Overall	84.07	80.02	82.00

5 Discussion and Conclusion

Starting from the Message Understanding Conferences, a considerable amount of work has been done in last years on the Named Entity Recognition and attempts have been made to use global information: generally named entities occurring in the same document. Differently from most of those systems, Typhoon can exploit information given by named entities occurring both in the same and different documents of a large collection, by means of Data Redundancy and Patterns. Data Redundancy and Patterns were first introduced in conjunction with the system that participated in the EVALITA 2009 NER task [Zanoli et al., to appear]. *Data Redundancy* resulted one of the most important feature, whereas *Patterns* contributed to the best system performance, even if not as well as we expected; deciding which patterns (number and type) should be used is not an easy task. We are confident enough about their importance in this task, so in the next future we are going to further investigate how to exploit them (e.g. sorting them by using different techniques, using lemmas too). Another strength is that Typhoon is designed as a combination of two classifiers: the CRF and the HMM classifier; we showed that the two classifiers can take advantages from each other, whereas a single classifier could work on its own when a higher classification speed is needed. Typhoon performed as the best system at Evalita 2009, and when tested on the Evalita 2007 data set, it shows an accuracy higher than the system we presented at Evalita 2007 (+0.4 of F_1); moreover Typhoon was configured with few experiments (e.g. for both the classifiers we used the default system configuration), that is differently from the Evalita 2007 system that took advantages from more than one hundred of experiments and from specific list of proper nouns about sport events of the year 2004 extracted according to entities in the development set. In the next future we will try using different system configurations to obtain better results, furthermore a web service version of the system is under construction.

Acknowledgments. This work has been partially supported by the project LiveMemories, funded by the Provincia Autonoma of Trento.

References

1. Bikel, D.M., Schwartz, R.L., Weischedel, R.M.: An Algorithm that Learns What's in a Name. Machine Learning, vol. 34, issue 1-3, pp. 211–231 (1999)
2. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition. Ph.D. thesis, Computer Science Department, New York University (1999)
3. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of ICML, pp. 282–289 (2001)
4. McCallum, A., Li, W.: Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In: Proceedings of CoNLL, pp. 188–191 (2003)
5. Mikheev, A., Grover, C., Moens, M.: Description of the LTG system used for MUC-7. In: Proceedings of the Seventh Message Understanding Conference (1998)
6. Pianta, E., Zanoli, R.: Exploiting SVM for Italian Named Entity Recognition. Intelligenza Artificiale, Special Issue on NLP Tools for Italian, vol. IV, issue 2 (2007)
7. Pianta, E., Zanoli, R.: TagPro: A system for Italian PoS tagging based on SVM. Intelligenza Artificiale, Special Issue on NLP Tools for Italian, vol. IV, issue 2 (2007)
8. Leong Chieu, H., Tou Ng, H.: Named entity recognition: a maximum entropy approach using global information. In: Proceedings of the 19th international conference on Computational linguistics, pp. 1–7 (2002)