Italian Named Entity Recognizer Participation in NER task @ Evalita 09

Yashar Mehdad[1,2], Vitalie Scurtu[3], Evgeny Stepanov[1]
mehdad@fbk.eu, vitalie.scurtu@liquida.it, stepanov@disi.unitn.it

[1] University of Trento, Trento, Italy
[2] Fondazione Bruno Kessler - FBK-irst, Trento, Italy
[3] RGB s.r.l, Banzai Group, Milan, Italy

Abstract. In this paper, we present our system for Named Entity Recognition (NER), as one of the significantly important preliminary steps prior to main Natural Language Processing tasks, based on Support Vector Machines and feature extraction and selection. The system performed the third best on the task of Italian NER at EVALITA 2009, with an overall F-measure of 81.09, which has less than one percent gap with the best result (82 %).

Keywords: Named Entity Recognition, Support Vector Machines, Feature Extraction and Tuning.

1 Introduction

Named Entity Recognition, in computational linguistics terms refers to the task of identifying the named entities which represent an instance of a name, location, person, organization or geo-political entity. Since many tasks and applications in Natural Language Processing (NLP), such as Information Extraction and Summarization, Information Retrieval, Data Mining and Question Answering, are dependent on Named Entity Recognition, this task is considered as one of the main and important preliminary works in this field. The number of research papers which has been published during last few years is a clear evidence for the significant weight of this job for almost all attempts in the area of Human Language Technology.

Amongst approaches which are applied to solving the problem of NER, statistical methods proved to be effective, fast, and popular. Machine Learning algorithms used in this approach, appeared to be successful, as well as reasonable, providing that a fairly large data set with high quality is available. Results gained using the ML methods, considering the features and data set as two important factors, never disappointed the attempts in this area.

In this paper we describe our system and participation in NER task for Italian. Although the focus of the task is on Italian dataset; however, we claim that our approach in named entity recognition can be easily adaptable to all languages [1]. Concluding the points, we illustrates that our runs was one of the best runs and less than 1 % in F-measure lower than the first ranked participant.

The paper is structured as follows: Section 1 describes the system architecture, its core components, and the workflow. Section 2 presents the resources we have used and the procedures to extract the features with a brief description of the list of features. Section 3 and Section 4 describe the settings we have experimented for our submissions and the experiment results. Finally, in section 5, the conclusion is made and future works are proposed.

2 System Architecture

Our Named Entity Recognition system is based on the YAMCHA classifier machine[4] which was originally created as generic, customizable and open source text chunker, and that can be adapted to the various tag-oriented NLP tasks.

YAMCHA uses Support Vector Machine learning algorithm for classification, at the same time is has a built-in support for multiclass problems (it allows both one-against-one and one-against-all approaches) For the purposes of feature extraction it also allows to manipulate the window-size of the features considered for learning.

For our approach, we have carried out a feature fine tuning study [2] in order to improve the information passed to the classifiers, removing noisy features and incorporating new meaningful features. The final set of the features is shown and described in the next section.

The structure of the system is summarized in two main parts, Feature Extraction and Feature Selection. The Feature Extraction portion was made out of the YAMCHA boundary, the Feature Selection subsequently, selected the set of features based on the experiments and results. At each time, the features extracted, added or deducted to resolute the results and consequently compare the gained results to reach the optimal state. Figure 1 demonstrates the system structure and steps.

For this purpose we tried to choose n number of candidate features for the initialization phase. We selected all possible relevant features for this task which possibly could be extracted easily from the available resources or directly from the data set. Having a large selection of features we tried to find the best features which could stabilize the best performance.

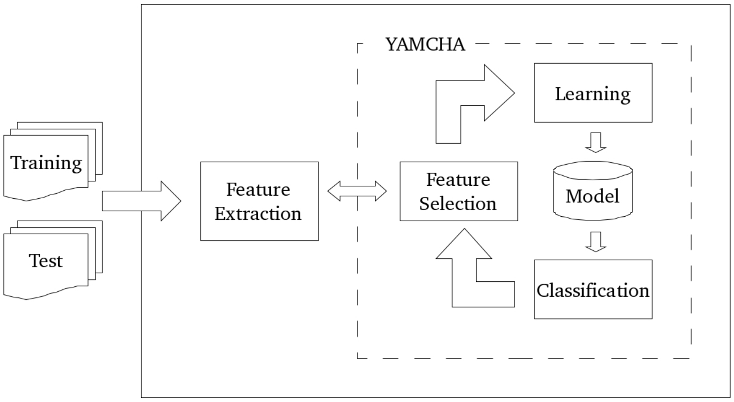

Fig. 1. System architecture.

3 Classifying Features

In the feature extraction phase a set of features, initially extracted and selected to be implemented in the training phase. In each step, number of features, which were estimated to be effective in the experiments, extracted and applied in the

[4] Available at: `http://chasen.org/~taku/software/yamcha/`

system, in addition amongst all features extracted, each iteration, a selection took place to approximate the optimal set of features.

The following features, in abstract, is the final set of extracted features which estimated to optimize the performance of the system with the small training data.

- Part of Speech tags (provided with data)
- Prefixes and suffixes (1, 2, 3, or 4 characters from the beginning and end of each token)
- Orthographic information and stop-word list
- Collocation bigrams (36,000 bigrams from Italian newspapers ranked by MI value [3])
- Gazetteers for proper nouns, geographical locations, companies and organizations[3, 4].
- Frequency and normalized frequency of the token
- Morphological features using [5]
- Token lemmas and stems using [5, 6]

Each of these features was extracted for the current, preceding and following words. We refer to these features as static, while dynamic features were decided dynamically during tagging.

4 Experiments

A set of experiments was conducted on the EVALITA09 NER data set. Both development data and test data are part of the Italian Content Annotation Bank (I-CAB)[7], developed in the context of the Ontotext Project.

As it was mentioned in the guidelines [8], the development data, consisting of 525 news stories taken from the local newspaper "L'Adige", include development and test data distributed for the Named Entity Recognition task at Evalita 2007 (about 180,000 words). The news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004); and they are grouped into five categories: News Stories, Cultural News, Economic News, Sports News and Local News. Test data, on the other hand, consist of completely new data.

We have used all the features mentioned in section 2, while the window size for static features (all extracted feature) and dynamic features (class labels) were tuned for each experiment. Different window sizes were tried and, finally, the best was selected: for the static features -2 to 2 window (two feature before and after the current token); the same -2 to 2 window size was chosen for the dynamic feature as well. The dynamic features window size would take into account a window of -i to i from the class labels as features. In another word, the class labels are used as features during the learning and classification.

5 Results

Our best result was scored third in NER task in EVALITA09 and it was less than percent in overall F-measure performance from the the system ranked as first. In the first submitted run, we used all the features mentioned above except the morphological features. In the second submitted run we have used morphological features as well. Table 1 illustrates our results for the participation in EVALITA 09.

Table 1. Results of the submitted runs.

RUN1					**RUN2**			
Category	**Results**				**Category**	**Results**		
	Pr	Re	F1			Pr	Re	F1
All	83.05	78.86	80.90		All	83.20	79.08	**81.09**
GPE	84.57	85.83	85.19		GPE	85.29	85.21	85.25
LOC	73.12	43.59	54.62		LOC	71.91	41.03	52.24
ORG	72.33	66.72	69.41		ORG	72.13	67.26	69.61
PER	88.30	84.40	86.30		PER	88.41	85.03	86.69

Results show that morphological analysis improved our performance about 1 % in the overall F-measure. Moreover, it can be observed that the system performed very good in recognizing the PER and GPE categories, while the lowest performance is over LOC category which is mainly because of the ambiguity of location names and GPE category. Another reason is the category distribution in the training data: ORG and LOC categories are fewer in number.

In the future, we plan to use different feature sets – increase feature number by adding some new features such as extra morphological information, which intuitively could help in improving the model. Moreover, because of the limited availability of NLP tools for Italian, some features, which could be informative, were omitted. In order to fill up this gap, on of the future plans is to experiment with the syntactic and semantic features.

6 Conclusion

In this paper we described our system developed for Named Entity Recognition task of EVALITA 09. With our sets of features, using SVM learning algorithm, we could gain a very good and promising results for Italian NER. The results indicate that morphological features were helpful in improving the system. In the future, more detailed feature and error analysis will be performed.

References

1. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition In: Proceedings of CoNLL-2003 (2003)
2. Ferrández, Ó., Toral, A., Muñoz, R.: Fine tuning features and post-processing rules to improve named entity recognition. In: Proceedings of NLDB, pp. 176–185 (2006)
3. Pianta, E., Zanoli, R.: Exploiting svm for italian named entity recognition. In: Proceedings of EVALITA 2007 (2007)
4. Kozareva, Z.: Bootstrapping named entity recognition with automatically generated gazetteer lists. In: EACL '06: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 15–21. Association for Computational Linguistics (2006)
5. Pianta, E., Girardi, C., Zanoli, R.: The textpro tool suite. In: Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference (2008)
6. Porter, M.F.: An algorithm for suffix stripping (1997)
7. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: I-CAB: the italian content annotation bank. In: Proceedings of LREC (2006)
8. Speranza, M.: Named entity recognition task guidelines for participants, http://evalita.fbk.eu (2009)