Named Entity Recognition for Italian using SVM

Stefan Rigo

Università degli Studi di Trento
stfz@gmx.net

Abstract. This report presents the Named Entity Recognition (NER) system for Italian presented at Evalita 2009. The system uses Support Vector Machines for machine learning and was trained on a large number of static and dynamic features; among those are orthographic and morphological information. Gazetteers have been build extracting information from the Internet. The open source chunker *YamCha* was used for development. While the recognition of person-names and geopolitical entities was satisfactorily, the system performance according to locations and organizations was rather weak. The system ranked on fifth position out of seven participants and performed with an overall F1 measure of 74.98%.

Keywords: Named Entity Recognition, SVM, Evalita 2009.

1 Introduction

A Named Entity Recognition (NER) Task consists in the recognition of nominal entities in a text. For the *Evalita 2009* NER task, a system must identify and classify entities belonging to four categories: persons, organizations, geopolitical names (cities, countries, etc) and locations (addresses, streets, etc). The system described in this paper uses a machine-learning approach based on Support Vector Machine [1, 2]. The open source text chunker *YamCha* was used as development environment [3]. Features were extracted for each word and encoded into vectors in order to learn classification criterions.

2 The System

The available development data was split in a training (66%) and a test (33%) set (used for system-tuning). For each word, the following set of features was extracted:

- the word itself

- lowercased word
- lemma
- stem
- *pos*, the POS tag
- *wnpos*, WordNet tag
- morphological informations:
  - *mood*
  - *tense*
  - *gender*
  - *person*
  - *number*
  - *grade*
  - *case*
  - *class*
- orthographic informations
  - *cap*, first letter capitalized
  - *cmix*, first and third letter capitalized
  - *mixc*, mixed case
  - *upper*, all letters uppercase
  - *lower*, all letters lowercase
  - *abbr*, word is an abbreviation
  - *hyph*, word contains hyphen
  - *digit*, word contains digits
  - *com_and*, word contains ampersand
  - *pCap*, previous word is capitalized
  - *nCap*, next word is capitalized
  - *CapNoN*, current word capitalized, next not
  - *CapNoP*, current word capitalized, previous not
  - *CapP*, current and previous are capitalized
  - *CapS*, previous, current and next are capitalized
  - *CapN*, current and next are capitalized
  - *bCap*, previous and next are capitalized, current not
- *aff*, word affix
- gazetteers
  - *org*, word is in organizations gazetteer
  - *nam*, word is in name gazetteer
  - *loc*, word is in location gazetteer
  - *gpe*, word is in gpe gazetteer

Gazetteers were build extracting information from various Internet-sites. *YamCha* was configured to use the PKI algorithm (2nd degree polynomial kernel). Different window-sizes were used for different subsets of static features; the dynamic features used the tags of the two tokens preceding the current token.

3 Results

The performance of the system on the Evalita 2009 test set is shown in Table 1.

Table 1. Results of the system

	Precision %	**Recall %**	**F1 %**
Overall	81,08	69,73	**74,98**
GPE	84,31	69,12	75,96
LOC	70,69	26,28	38,32
ORG	71,08	52,44	60,36
PER	84,13	82,25	83,18

The system ranked on fifth position out of seven participants, outperforming the baseline by 31%. It can be seen that the system performed poorly especially when classifying location entities and to some degree also for the classification of organization entities, while performance is satisfactorily for the recognition of names and geopolitical entities. It can be argued that the poor performance in classifying location entities must be attributed not only to the used gazetteer but also to the extracted features. Another observed problem belongs to the fact that various tokens are present in different gazetteers. For instance, *Trento* occurs in the Location gazetteer as well as in the GPE, Organization and Names gazetteers.

4 Conclusions

The system described in this report was conceived to be language independent and can thus easily be adapted to other languages (mainly by the use of different gazetteers). According to the results, it can be argued that the location and organization gazetteers are not performing satisfactorily, although they contain most of the tokens belonging to those classes, which are found in the test-set. A possible improvement could be obtained by the collection of predictive words, which could be implemented as a rule for "gazetteer disambiguation" for words found in more than one gazetteer.

References

1. Vapnik, V. N.,: The Nature of Statistical Learning Theory. Springer (1995)
2. Thorsten, J.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of ECML-98, 10th European Conference on Machine Learning. Chemnitz, Germany (1998)
3. YamCha: Yet Another Multipurpose CHunk Annotator, http://chasen.org/~taku/software/yamcha/