UTILIZZO DI SVM PER IL RICONOSCIMENTO DI NAMED ENTITY IN ITALIANO

EXPLOITING SVM FOR ITALIAN NAMED ENTITY RECOGNITION

EMANUELE PIANTA · ROBERTO ZANOLI

SOMMARIO/ *ABSTRACT*

In questo articolo presentiamo EntityPro, un sistema per il Named Entity Recognition (NER) basato su Support Vector Machines. EntityPro è addestrato considerando sia features statiche che dinamiche. Il sistema, testato su EVALITA 2007, ha ottenuto una misura F_1 di 82.14% (miglior sistema per il NER dell'italiano).

We present EntityPro, a system for Named Entity Recognition (NER) based on Support Vector Machines. EntityPro was trained with a large number of both static and dynamic features. The system performed the best on the task of Italian NER at EVALITA 2007, with an F_1 measure of 82.14.

Keywords: Named Entity Recognition, SVM

1. Introduction

Named Entity Recognition (NER) is a subtask of Information Extraction which aims to locate and classify words in text into predefined categories such as the names of persons, organizations, locations, time expressions, etc.

The most frequently applied techniques for this task are based on machine learning: Hidden Markov Models, Maximum Entropy Models, Support Vector Machines (SVMs). SVMs were introduced in Text Categorization by T. Joachims [2] and subsequently used for many other NLP task, as they scale up well to high feature dimension.

SVMs are now among the most popular machine learning techniques, and a number of implementations and development environment are available for them, such as YamCha [3], an open source text chunker that can be easily adapted to other NLP tasks. YamCha allows for handling both static and dynamic features, and for defining a number of parameters such as window-size, parsing-direction (forward/backward) and algorithm of multi-class problems (pair wise/one vs rest).

We used YamCha to build EntityPro, a system for recognition of Italian Named Entities, exploiting a rich set of linguistic features such as the Part of Speech, and the occurrence in proper nouns gazetteers. EntityPro is part of TextPro a suite of modular NLP tools developed at FBK-irst.

The EntityPro tagger has recently been trained on the EVALITA development set in which named entities are represented with IOB2 format. The data contains entities of four types: Geo-Political entity (GPE), Location (LOC), Organization (ORG) and Person (PER). We assume that named entities are not-recursive and not-overlapping. In this rest of the paper we provide further details on the feature space that we used, and the results we obtained.

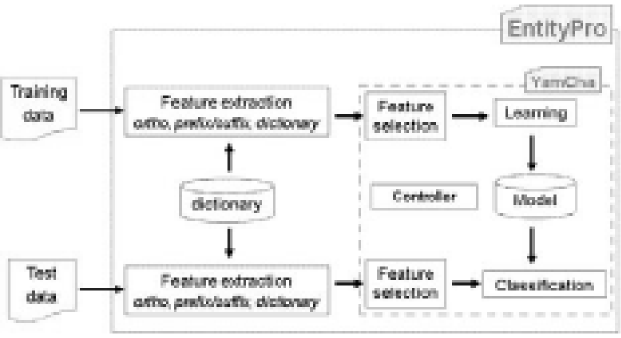


Figure 1: EntityPro's architecture

2. EVALITA NER task

Both development test data are part of the Named Entity task of EVALITA 2007. Other external resources are allowed. EntityPro was configured splitting the development set randomly into two parts: a data set for training (92,241 tokens) and a data set for tuning the system (40,348 tokens). The resulting best configuration was tested on the test set.

For each running word rich set of features (18) are extracted: the word itself, both unchanged and lower-cased; its Part of Speech, as produced by TagPro [1]; prefixes and suffixes (1, 2, 3, or 4 characters at the start/end of the word); orthographic information (e.g.

capitalization and hyphenation), collocation bigrams (36,000 bigrams from Italian newspapers ranked by Mutual Information value); gazetteers of generic proper nouns extracted from the Italian phone-book and from Wikipedia (154,000 proper names), from various sites about Italian and Trentino's cities, (12,000), Italian and American stock market (5,000 organizations) and Wikipedia geographical locations (3,200); moreover a list of 4,000 proper nouns extracted from a sport newspaper (*Gazzetta dello Sport*, year 2004).

Each of these features was extracted for the current, previous and following words. We refer to these features as static, as opposed to dynamic features, which are decided dynamically during tagging. For the latter, we used the tag of the three tokens preceding the current token. YamCha was set to work with the PKI algorithm with 2nd degree of polynomial kernel and one vs. rest as method for solving multi-class problems.

3. Results

Entity Pro (FBKirst_Zanoli_NER_r2) scored as the best system in the Italian Named Entity Recognition task, at EVALITA 2007 (evaluation based on exact match).

Table 1: EntityPro with external resources

Category	Pr	Re	F_1
All	83.41	80.91	82.14
GPE	84.80	86.30	85.54
LOC	77.78	68.85	73.04
ORG	68.84	60.26	64.27
PER	91.62	92.63	92.12

Table 2: EntityPro without external resources

Category	Pr	Re	F_1
All	75.79	72.43	74.07
GPE	78.56	76.51	77.53
LOC	81.08	49.18	61.22
ORG	57.09	52.28	54.58
PER	85.71	85.50	85.60

Table1 gives the results of the best run on the test set in terms of Precision (Pr), Recall (Re) and F1 measure. Table 2 reports the same measures when the system does not exploit external resources.

4. Discussion

F_1 values for both PER and GPE categories appear rather good, comparing well with those obtain in CONLL 2003 for English. Recognition of LOCs and ORGs seems more problematic. We suspect that the number of LOCs examples in the corpus is insufficient for the learning algorithm, and ORGs appear to be highly ambiguous.

EntityPro without external resources (Table 2) performed better than all other systems in EVALITA 2007. This confirms that SVMs perform as state of the art machine learning algorithms. Adding proper noun gazetteers produced a clear improvement in the overall system performance (+8% F_1, see Table 1; 31% error reduction). The EVALITA development set contains a high number of news related to sport events of year 2004. The specific list of proper nouns extracted from 2004 sport newspaper enhanced the performance of the system of 2 points of F_1 measure (18% error reduction) on the PERson category.

Acknowledgments

This work has been partially supported by the three-year project ONTOTEXT, funded by the Provincia Autonoma di Trento.

REFERENCES

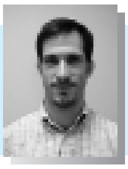
[1] E. Pianta, R. Zanoli. TagPro: A system for Italian PoS tagging based on SVM. EVALITA 2007 Workshop, Roma.

[2] J. Thorsten. Text categorization with support vector machines: learning with many relevant features. Proc. of ECML-98, 10th European Conference on Machine Learning, 1998.

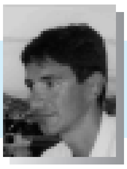
[3] http://chasen.org/~taku/software/yamcha/

CONTACTS

EMANUELE PIANTA, ROBERTO ZANOLI
FBK-irst, via Sommarive, 18, 38050 Povo (Trento)
Email: {pianta | zanoli}@itc.it



EMANUELE PIANTA is researcher at FBK-irst, Trento. His research interests include development of multilingual resources (e.g. MultiWordNet), basic linguistic processors for Italian and English, parsing, and information extraction.



ROBERTO ZANOLI is a research technician at the Cognitive and Communication Technologies division of the FBK-irst. His research interests include information extraction and machine learning.