TAGPRO: UN SISTEMA PER IL POS-TAGGING DELL'ITALIANO BASATO SU SVM

TAGPRO: A SYSTEM FOR ITALIAN POS TAGGING BASED ON SVM

EMANUELE PIANTA · ROBERTO ZANOLI

SOMMARIO/ *ABSTRACT*

In questo articolo presentiamo TagPro, una sistema per il PoS-tagging basato su Support Vector Machines. TagPro usa un insieme ricco di feature, tra cui l'analisi morfologica, ed è risultato il miglior sistema per il PoS-tagging dell'italiano in EVALITA 2007 (accuratezza di 98.04 sul tagset EAGLES; 97.68 sul tagset DISTRIB).

We present TagPro, a system for PoS-tagging based on Support Vector Machine. TagPro exploits a rich set of features, including morphological analysis. It scored as the best system in the Italian Pos Tagging task at EVALITA 2007, with an accuracy of 98.04 on the EAGLES tagset and of 97.68 on the DISTRIB tagset.

Keywords: PoS tagging, SVM, morphological analysis.

1. Introduction

Part of speech tagging is the problem of determining the correct parts of speech of a sequence of words.

The most frequently applied approaches for this task are based on machine learning: Hidden Markov Models [1], Maximum Entropy taggers [5], Transformation-based learning, Memory Based learning [2], Decision Trees [3] and Support Vector Machines (SVMs) [4]. SVMs are among the most widely used techniques, and various implementations are available. As argued by T. Joachims [6], one of advantages of SVMs is that dimensionality reduction is usually not needed, as they are robust to overfitting and scale up well to high feature dimensions.

We used YAMCHA, an SVM-based machine learning environment [8], to build TagPro, a PoS-tagging system exploiting a rich set of linguistic features, such as morphological analysis and proper name gazetteers. TagPro is part of TextPro, a suite of NLP tools developed at FBK-irst, which includes MorphoPro, a morphological analyzer that provides the morphological analysis exploited by TagPro.

TagPro was trained on the EVALITA development set, using the standard EAGLES tagset and a new, structurally different, tagset (DISTRIB). In the rest of the paper we give further details on SVMs, the feature space that we used, and the results we obtained.

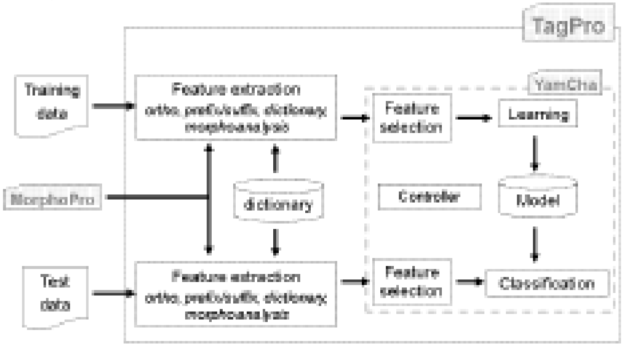
TagPro
Training data
Feature extraction ortho, prefix/suffix, dictionary, morphoanalysis
Feature selection
YamCha
Learning
MorphoPro
dictionary
Controller
Model
Test data
Feature extraction ortho, prefix/suffix, dictionary, morphoanalysis
Feature selection
Classification

Figure 1: TagPro's architecture

2. SVM and YamCha

Support Vector Machines are based on the Structural Risk Minimization strategy [7], which aims at finding a hypothesis H for which we can guarantee the lowest true error, that is the probability that H will make an error on an unseen and randomly selected test example.

YamCha is a generic, customizable, and open source text chunker that can be adapted to a number of other NLP tasks. It 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).

3. EVALITA PoS-tagging

The EVALITA 2007 evaluation campaign provided development and test data extracted from the CORIS/CODIS corpus. External resources were allowed for training of the systems.

TagPro was configured splitting the development set randomly into two parts: a data set for training (100,000 tokens) and a data set to set for tuning (50,000 tokens). The resulting configuration was then tested on the test set.

For each word a rich set of features (38) are extracted: the word itself (both unchanged and lower-cased); morphological features produced by MorphoPro; prefixes and suffixes (2, 3, 4 or 5 characters at the start/end of the word); orthographic information (e.g. capitalization, hyphenation); and occurrence in gazetteers of proper nouns (154,000 proper names, 12,000 cities, 5,000 organizations and 3,200 locations).

Each of these features is extracted for the current word, and for the previous and following words. We refer to these features as static features, as opposed to dynamic features, which are decided dynamically during tagging. For the latter, we used the tag of the two tokens preceding the current token. Moreover, 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. The same system configuration was used for both tagsets and no specific method was applied to classify unknown words.

4. Results

We evaluated our approach on the evaluation set corpus by exploiting the EVALITA scoring software.

Table 1: TagPro results (FBKirst ZANOLI_POS)

TagSet	TAccuracy	UTAccuracy
EAGLES	98.04	95.02
DISTRIB	97.68	94.65

The performance is given both in terms of Tagging accuracy (TAccuracy) and Unknown Words Tagging Accuracy (UTAccuracy). The first is defined as the number of correct PoS tag assignment divided by the total number of tokens. The second as the Tagging Accuracy related to unknown words. TagPro ranked as the best system in the Italian PoS tagging task, at the EVALITA 2007 evaluation campaign (FBKirst ZANOLI_POS).

5. Discussion

We have presented an approach to PoS-tagging for Italian that uses SVMs as learning algorithm. We used all available features without any feature reduction and no specific method was applied so as to classify unknown words. Results confirm that SVMs can deal with a big number of features (38), without incurring in overfitting.

The linguistic features that contributed mostly to the final performance of the system, if compared with a baseline exploiting a 3-word window (86.70, EAGLES) are affixes and orthographic information (+8.56 over baseline), morphological analysis (2.13 improvement over affixes). Gazetteers instead do not contribute any further significant improvement over affixes and morphology (0.03) .

REFERENCES

[1] T. Brants. TnT - A Statistical Part-of-Speech Tagger. In *Proc. of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA, 2000.

[2] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. MBT: A Memory-Based Part of Speech Tagger-Generator. In *Proc. of the 4th Workshop on Very Large Corpora*, 1996.

[3] L. Marquez, and H. Rodriguez. Part-of-Speech Tagging Using Decision Trees. In *Proc. of ECML*, 1998.

[4] T. Nakagawa, T. Kudoh, and Y. Matsumoto, Unknown word guessing and Part-of-Speech tagging using support vector machines, in: *Proc. of the 6th Natural Language Processing Pacific Rim Symposium*, 2001, pp. 325-331.

[5] A. Ratnaparkhi. A Maximum Entropy Part-Of-Speech Tagger. In *Proc. of the Empirical Methods in Natural Language Processing Conference*, May 17-18, 1996.

[6] T. Joachims, Text categorization with support vector machines: learning with many relevant features. In *Proc. of ECML,* 1998, pp. 137-142.

[7] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] 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 FBK-irst, Trento. His research interests include information extraction and machine learning.