È VANTAGGIOSO USARE UNA COMBINAZIONE DI TAGGER PER L'ASSEGNAZIONE AUTOMATICA DI PARTI DEL DISCORSO?

ARE TWO HEADS BETTER THAN ONE? EXPERIMENTS WITH ITALIAN PART-OF-SPEECH LABELLING

JOHAN BOS · MALVINA NISSIM

SOMMARIO/*ABSTRACT*

Descriviamo come combinare l'output di due sistemi per il *part-of-speech tagging* in italiano. Nonostante ci si aspetti che l'utilizzo di più d'una sorgente di informazione apporti beneficio, i nostri risultati mostrano che il miglioramento nella performance è solo marginale rispetto all'utilizzo di un *tagger* solo. Esperimenti futuri mireranno ad esplorare l'utilizzo di altri tagger e diverse tecniche di combinazione.

There seems to be no obvious way to combine the output of a pair of off-the-shelf POS taggers in order to get improvement over single taggers' accuracy. We combined two well-known retrainable taggers, C&C and TnT, using memory-based learning and tested the resulting tagger on Italian POS data, with respect to two different tagsets. Only for one tagset we observed a slight increase in performance, but the added value is small, and one could spare the effort and use as well a single tagger.

Keywords: POS tagging, memory-based learning

1 Introduction

In theory, two different taggers cooperating on the same task are likely each to make up for deficiencies in the other's methods. Previous work has indeed shown that a semi-supervised technique such as co-training can be used to boost the performance of part-of-speech tagging when only very little annotated data is available [2]. But if large amounts of labelled data *are* available, is it still reasonable to expect the combination of several off-the-shelf retrainable taggers to outperform any of the single taggers?

We tested this hypothesis in the context of EVALITA-2007, a campaign for evaluating Italian natural language processing tools, which included a shared task for part-of-speech (POS) tagging. In what follows we describe the task and material, our method, and the obtained results.

2 Method

2.1 Task and Material

The training data, as provided by the organisers of EVALITA, consisted of data with gold standard parts of speech (POS) for two different tagsets: EAGLES and DISTRIB. The EAGLES tagset is with 32 different tags more fine-grained than the DISTRIB tagset (16 different tags). Each line comprised one token and its POS-tag. Sentence boundaries were not indicated in the training data.

2.2 Data Preparation

Because at least one of the taggers that we used required sentence boundaries, we automatically assigned sentence boundaries in the training (and later, the test data). The method that we used to do this was utterly simple: we considered full stops and question marks as sentence boundaries, and introduced an extra new line after each sentence. We divided the training data into ten folders, split on the basis of sentence boundaries, and used cross validation for the development of our system.

2.3 Training the Taggers

We employed two well-known taggers that can be trained on new annotated data: the TnT tagger [1] and the C&C POS tagger [3]. Both these taggers reach state-of-the-art performance on unseen English newspaper text (ca. 97% per-word accuracy on Section 23 of the Penn Treebank).

TnT is an efficient statistical tagger, which comes with and English and a German model for POS tagging. It incorporates several methods of smoothing and handling unknown words. The C&C tagger follows a Maximum Entropy tagging method, using log-linear probability distributions to model local decisions at each point in the tagging process. Although both taggers are re-trainable for different languages, no additional language-specific features can be included in their standard distributed versions.

Both taggers were retrained on the EVALITA training data for both tagsets. We measured the performance (per-word accuracy) using 10-fold cross validation, observing that TnT outperformed C&C on both tagsets (Table 1).

Table 1: Results (accuracy) of both single taggers and combined on training data using cross-validation.

	C&C	TnT	Combined
DISTRIB	90.97	95.34	95.43
EAGLES	92.78	96.31	96.34

2.4 Combining the Taggers

The general idea was to combine the two taggers by generating a decision tree on their output. Essentially, this would comprise rules of the type: given a token τ, if tagger T-1 outputs α, and T-2 outputs β, then output γ.

To learn the rules, we experimented with a C4.5 decision tree as implemented in the Weka distribution [5], but we obtained better results using the memory-based machine learner Timbl [4]. The setting that yielded the best results included five features: the token itself, the output of T-1 and the output of T-2 for the EAGLES tagset, and the output of T-1 and the output of T-2 for the DISTRIB tagset.

The results on the training data of the combined tagger were marginally better than on the single taggers (see Table 1). Hence, the prospects for a combined tagger using the method that we present here are not quite encouraging.

3 Results

We submitted only one run (the maximum allowed in the EVALITA challenge), and therefore could not experiment much with variations of the settings.

Our expectations regarding performance were confirmed by the official results on the test data (Table 2). Even though the results turned out to be higher than those obtained on training data, they were hardly better than just using the best of the two taggers i.e. TnT.

Table 2: Results (accuracy) on test data of the baseline system, TnT (provided by the organisers), and the system reported in this article (with its official name).

	Baseline	TnT	UniRoma1_Bos_POS
DISTRIB	89.48	95.96	96.21
EAGLES	90.43	96.82	96.76

On the bright side, however, our system convincingly beat the baseline (most frequent tag for known words; absolute most frequent tag for unknown words), and also showed no signs of overfitting the training data.

4 Conclusion

Combining output of different taggers to obtain a better one looks like an attractive idea, but it's not easy to realise in practice. Results probably strongly depend on the individual performances of the taggers. The method described in this paper might work very well for two equally good taggers. But not, as was most likely the case in our experiment, for a good and a very good tagger.

Still, what we presented can only be considered a pilot study, since we only used two taggers and one single combination method. Given that there was a small measureable gain, further experiments might indeed show that combining the output of two fully trained taggers can yield a boost in performance.

REFERENCES

[1] T. Brants. TnT – a statistical part-of-speech tagger. In *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

[2] S. Clark, J. R. Curran, and M. Osborne. Bootstrapping POS taggers using unlabelled data. In *Proc. of CoNLL-2003*, pages 49–55, Edmonton, Canada, 2003.

[3] J. R. Curran and S. Clark. Investigating gis and smoothing for maximum entropy taggers. In *Proc. of EACL-03*, pages 91–98, Budapest, Hungary, 2003.

[4] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. Timbl: Tilburg memory based learner, version 5.0, reference guide. ILK Research Group Technical Report Series 03-10, Tilburg, 2003.

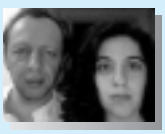
[5] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Diego, 2000.

CONTACTS

JOHAN BOS
Univ. of Rome "La Sapienza" - Email: bos@di.uniroma1.it

MALVINA NISSIM
Univ. of Bologna - Email: malvina.nissim@unibo.it

JOHAN BOS is a computational linguist and currently employed on a "Rientro dei Cervelli" grant at the University of Rome "La Sapienza". He obtained his PhD at the University of the Saarland in Germany, and held a fiveyear post-doc position at the University of Edinburgh, UK. He's president of the ACL Special Interest group in computational semantics (SIGSEM).



MALVINA NISSIM holds a PhD in Linguistics from the University of Pavia. She spent five years at the School of Informatics of the University of Edinburgh as a postdoctoral fellow, and is now assistant professor at the University of Bologna.