DESCRIZIONE E VALUTAZIONE DEL UNIPISYNTHEMA POS TAGGER

UNIPISYNTHEMA POS TAGGER DESCRIPTION AND EVALUATION

CARLO ALIPRANDI · NICOLA CARMIGNANI · NEDJMA DEHA · PAOLO MANCARELLA · MICHELE RUBINO

SOMMARIO/ *ABSTRACT*

Questo articolo descrive la struttura del *UniPiSynthema* Part-of-Speech (POS) Tagger utilizzato per il task di POS Tagging all'interno di EVALITA 2007.

This paper outlines the structure of the UniPiSynthema *Part-of-Speech (POS) Tagger that we used for the POS Tagging task within the EVALITA 2007 initiative.*

Keywords: Italian POS tagging, statistical POS tagger.

1. Introduction

We present the UniPiSynthema POS tagger, a system for handling inflected languages, namely the Italian language. The system is based on combined statistical and rule based methods and it uses robust lexical resources.

2. Tagger Description

2.1 Structure and Core Methods

The UniPiSynthema POS tagger basic assumption is that contextual information affects the environment where the word has to be tagged. In order to tag the word with the most likely POS it is necessary to have a high-order representation of the context. This assumption has been consolidated into stochastic methods that are based on a second order Markov Model.

We studied and elaborated a complex framework for language modelling in [2], that was used to develop and tune the Lexical Resources underlying the tagger. The UniPiSynthema Language Model has been trained from a large tagged corpus to acquire sentence context.

The Italian POS n-grams, approximated to bigrams and trigrams, have been trained from a large balanced corpus created from newspapers, magazines, documents, commercial letters and emails. The corpus was cleaned, standardised (punctuations, capitalisations) and then parsed using the Italian POS tagger Synthema Lexical Parser (SLP), a rule-based parser (see Lexical Data Base Management System - LDBMS [5]). The result was a corpus tagged with syntactic and morphological information, that was automatically extracted from Italian dictionaries available in LDBMS. We used this corpus as training set for the Language Model.

The core method for POS tagging is based on a rule based lemmatizer, Synthema Lexical Classifier - SLC, available in LDBMS. The Lexical Classifier produces one or more candidate POS and deep morpho-syntactic information for each token, relying on a huge lexicon given by the Lexical Resources.

The Language Model is then applied to the lemmatized token list, in order to solve ambiguities and to assign each token the best POS. Specifically, as detailed in [4], given candidates Part-of-Speech POS_1 ... POS_n, the best POS is assigned by the Language Model.

Due to time constraints, but also to the underlying assumption that our POS tagger has been designed to work on open domains, we used the existing Language Model and the existing Lexical Resources, without adapting it to the Enrolment Data Set given by the EVALITA POS tagging task. Unknown words coming both from the Development set and the Test set were not considered: this can be considered as a disadvantage in our participation.

2.2 Lexical Resources

To grant a high lexical coverage, POS tags for each tokenized word are extracted from tagged lexicon dictionaries. The SLC lemmatizer relies on an Italian dictionary, (43.000 word lemmas - 1.165.000 word forms). We also used specific dictionaries for Geography, Politics, People and Foreign words.

Besides the specific dictionaries, many proper nouns and multiwords are specifically recognized by a multiwords gazetteer, extending our dictionary's tagging capabilities. At run time a multiword recognition grammar searches through the gazetteer and identifies multiword patterns and proper nouns like "D'Alema", "Di Pietro", "New York", "Buenos Aires", etc.

2.3 Unknown Words Handling Methods

We have no specific methods for the unknown words handling. We only have some heuristic rules for determining proper nouns.

3. Results

We are pleased to have achieved the results outlined in Table 1 although our system has not been trained on the EAGLES-like development set. In spite of all UniPiSynthema POS tagger has obtained satisfying results even though it has been designed for on-the-fly tagging rather than the off-line tagging.

Table 1: Results of the EAGLES-like TestSet

	Accuracy	Error Rate
Global Data	88.71	11.29
Unknown Tokens	79.49	20.51

4. Discussion

One of the worst penalties of the UniPiSynthema performance in tagging the Eagles-like TestSet was caused by the recurrent wrong tagging of some words. As an example among others, the word "come" was presented in the Eagles-like set most frequently as an adverb on quite rarely as a conjunction, however UniPiSynthema tagged this word usually like a conjunction because the POS expected by the system has got priority over the words' possible classifications: if the system consider that conjunction, given the sentence context, is very likely, when the user (or the simulated test user) writes a word like "come", that could be an adverb or a conjunction the system will tag the word as the expected conjunction, giving this tag priority over the word's possible classifications.

Many multiwords were also not recognized or treated as bad tokens. This happened because the multiword composition strategy in the Eagles-like TestSet was very different from our own. For a correct tagging, almost all the multiwords in the TestSet should have been inserted in the multiword gazetteer, together with the morpho-syntactic information; due to time reasons such an operation was not possible. Our tagging policies imply the classification of multiwords as single token, without separating each word. For example compound names, such as "D'Alema" or "Di Pietro", are usually tagged as a unique proper noun ("D'Alema" NP) rather than two ("D'" NP and "Alema" NP).

These experiments with the Eagles-like tagset allowed us to find many new research directions for improving our POS tagger: a combination between the POS expected by the system and the most frequent POS for the word to be tagged can improve tagging quality, and an algorithm for automatic multiword recognition and self-updating gazetteer can increase tagger accuracy in dealing with composite expressions and proper nouns. We had a research interest into evaluating our POS tagger on an open domain, thus the EVALITA initiative was interesting. We decided with awareness to apply no specific optimizations for dealing with adaptations like, for example, enrolment, unknown word management and proper noun identification, that could be considered a drawback in accuracy. Nevertheless we consider very positively our global results and the comparison to other systems is a positive source of encouragement to further developments to get even better performances not only in terms of qualitative results.

REFERENCES

[1] C. Aliprandi, D. Barsocchi, F. Fanciulli, P. Mancarella, D. Pupillo, R. Raffaelli and C. Scudellari. AWE, an Innovative Writing Prediction Environment. *Proceedings of the 10th International Conference on Human-Computer Interaction*, pp. 237-238, 2003.

[2] D. Barsocchi. Disabilità, Informatica, Linguistica: un'Istanza del Trinomio. Master's Thesis in Computer Science, Department of Computer Science, University of Pisa, 2002.

[3] N. Carmignani. A Word Prediction System for People with Disabilities based on Part-of-Speech Tagging. Master's Thesis in Computer Science, Department of Computer Science, University of Pisa, 2005.

[4] C. Aliprandi, N. Carmignani and P. Mancarella. An Inflected-Sensitive Letter and Word Prediction System. *Proceedings of the International Conference on Interactive Computer Aided Learning*, 2006.

[5] R. Raffaelli. Lexical Data Base Management System - LDBMS. Synthema Internal Report, Pisa, 2000.

CONTACTS

CARLO ALIPRANDI
Synthema Srl, Via Malasoma 24, Pisa
Email: aliprandi@synthema.it

NICOLA CARMIGNANI, NEDJMA DEHA,
PAOLO MANCARELLA, MICHELE RUBINO
Dipartimento di Informatica - Università di Pisa
Email: {nicola | deha | paolo | rubino}@di.unipi.it

CARLO ALIPRANDI received his M.S. degree in Computer Science in 1992 at the University of Milano. Since 1997 he works with Synthema, an Italian SME leader in Language Intelligence and Text Mining, where he is currently Language and Speech Technology Manager. He is an expert in Natural Language and Speech research, technology and applications. His research is focused on speech subtitling, speech reporting and text entry.