LA MASSIMA ENTROPIA PER IL PART OF SPEECH TAGGING DELL'ITALIANO

MAXIMUM ENTROPY FOR ITALIAN POS TAGGING

FELICE DELL'ORLETTA · MARIA FEDERICO · ALESSANDRO LENCI · SIMONETTA MONTEMAGNI · VITO PIRRELLI

SOMMARIO/ *ABSTRACT*

L'articolo illustra le prestazioni del ILC-UniPi MaxEnt PoS Tagger in Evalita 2007.

The report contains a description of the ILC-UniPi MaxEnt PoS Tagger performance in Evalita 2007.

Keywords: Maximum Entropy, PoS tagging.

1. System description

The ILC-UniPi Tagger is a combination of two Maximum Entropy PoS taggers [2,3], Tagger-A and Tagger-B, operating on the output of MAGIC, an Italian rule-based morphological parser [1], equipped with a general-purpose lexicon of about 100.000 entries. For each word form in the text, MAGIC yields one or more parses, each consisting of a word lemma, its part-of-speech (PoS) plus a bunch of morpho-syntactic features (e.g. gender, number, person, mood, tense, etc). Tagger-A deals with morphologically ambiguous word forms only, i.e. those forms that are assigned more than one morphological parse by MAGIC. On the other hand, Tagger-B is specifically designed to deal with "out of vocabulary" words that MAGIC fails to parse. Finally, whenever MAGIC outputs one morphological analysis only, the PoS of this parse is returned as the tagging result. It should be appreciated that Tagger-A assigns each morphologically ambiguous word form one of the PoS tags output by MAGIC. Tagger-B, on the other hand, assigns each unknown word form the most likely PoS out of the entire set of possible PoS tags.

1.1 Features for POS tagging

Tagger-A and Tagger-B are trained on the same training data, but with two different sets of features. Some features are common to both sets. Features can be distinguished into *local features* and *contextual ones*.

The local features used by both taggers include: the word form, its orthographic properties (i.e. "word-initial capital", "all upper case", "mixed upper/lower case"), its length, its maximal suffix (i.e. the last four letters of the word form, if the length of the word is more than four letters, or otherwise the word form itself), the presence of non-alphabetic characters (digits, math operators, etc.).

The contextual features common to both taggers include (w_t is the target word form to be tagged): i.) the lemma and the PoS of w_{t-1}, ii.) the bigram formed by w_t and the lemma of w_{t-1}, iii.) the PoS assigned by the tagger to w_{t-2} (this feature is used only for the EAGLES tagset), iv.) the form of w_{t+1}, v.) the possible parts of speech (as assigned by the morphological analyzer) of w_{t+1}, vi.) the bigram $<w_t , w_{t+1}>$.

For the DISTRIB tagset only, Tagger-A also uses a special feature to discriminate among the three different prepositional types (PREP_POLI, PREP_NA, PREP_VA). This feature is the bigram formed by the lemma of the preposition to be tagged and the lemma of the verb that precedes the preposition in the sentence, if any.

Tagger-B also uses the following extra set of local and contextual features: the prefix (max. length of three letters) of current unknown word, the PoS assigned to w_{t-2} (differently from Tagger-A, Tagger-B uses this feature for both tagsets), the bigram formed by of the PoS assigned to w_{t-1} and the PoS assigned to w_{t-2}.

2. Results

The performance achieved by the ILC-UniPi Tagger on the EVALITA test set (17.313 tokens) for both tagsets is reported in Table 1. The "unknown tokens" in Table 1 are the tokens present in the training set but not in the test set.

Table 1: Tagger official scores

	GLOBAL DATA		UNKNOWN TOKENS	
	accuracy	error rate	accuracy	error rate
DISTRIB	96.70%	3.30%	93.14%	6.86%
EAGLES	97.65%	2.35%	94.12%	5.88%

Table 2 and Table 3 show the main error types with the relative error rate for the EAGLES and DISTRIB tagsets respectively. The two tables report the 47% of the global errors, the remnants being represented by other different types of errors with a much lower frequency.

Table 2: Main error types in EAGLES tagset

Our result -> Correct	**%Error rate**
ADJ -> NN	9.8 %
NN -> ADJ	9.3 %
V_PP -> ADJ	8.1 %
PRON_REL -> CONJ_S	5.8 %
NN -> V_GVRB	3.9 %
NN_P -> NN	3.6 %
ADJ -> V_PP	3.4 %
V_PP -> NN	2.7 %

An important point worth remarking is that the target PoS in the EVALITA gold standard in some cases is not included among the morphological analyses returned by MAGIC. This mismatch results in tagging errors, since Tagger-A uses only the readings assigned by MAGIC to select the proper one (see section 2 above). In fact, the errors due to this problem cover about the 12% and 8% - respectively for EAGLES and DISTRIB tagsets - of the total error rate.

3. Conclusion

PoS Taggers relying on legacy lexical resources are clearly penalised by outstanding mismatches between background lexical information and unknown test data. This factor alone explains out the existing gap in performance between our system and the best-performing system on Evalita 2007 test data. On a more positive note, however, our strategy allows for a considerably finer-grained tagging, with information including the word form lemma and morpho-syntactic features.

Table 3: Main error types in DISTRIB tagset

Our result -> Correct	**%Error rate**
ADJ -> N	11.7 %
N -> ADJ	7.7 %
V -> ADJ	5.9 %
ADJ -> V	4.5 %
N -> V	4.4 %
ARG_DET -> ENTITIES	4.4 %
V -> N	4.4 %
REL -> SUB_ARG	4.2 %

REFERENCES

[1] M. Battista, V. Pirrelli. Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane. *ILC-CNR technical report*, 1999.

[2] A.L. Berger, S.A. Della Pietra, V.J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol.22, No.1, 1996.

[3] A. Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

[4] F. Dell'Orletta, A. Lenci, S. Montemagni, V. Pirrelli. Probing the Space of Grammatical Variation: Induction of Cross-Lingual Grammatical Constraints from Treebanks. *Proceedings of the ACL Workshop on Frontiers in Linguistically Annotated Corpora,* Sidney 2006.

CONTACTS

FELICE DELL'ORLETTA, MARIA FEDERICO, SIMONETTA MONTEMAGNI, VITO PIRRELLI
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
via Moruzzi 1 - 56125 Pisa
Email: {felice.dellorletta | maria.federico | simonetta.montemagni | vito.pirrelli }@ilc.cnr.it

ALESSANDRO LENCI
Dipartimento di Linguistica
Università di Pisa
via S. Maria 36 - 56126 Pisa
Email: alessandro.lenci@ilc.cnr.it

FELICE DELL'ORLETTA is a PhD student at the Department of Computer Science of the Pisa University and he works on machine-learning algorithms applied to Natural Language Processing at the Institute of Computational Linguistics of the Pisa CNR.