C4: HMM POS TAGGER CON POS GUESSER E LESSICO ESTERNO

C4: HMM POS TAGGER WITH POS GUESSER AND EXTERNAL LEXICON

SIMONE ROMAGNOLI

SOMMARIO/ ***ABSTRACT***

In questo articolo si presenta un Part of Speech tagger che utilizza sia una risorsa lessicale esterna che l'implementazione di una algoritmo specifico per incrementare la precisione nella elaborazione delle parole mai incontrate nella fase di training.

In this article we present a statistical part of speech tagger combined both with an external lexical resource and a specific algorithm to improve the processing of words never encountered in the training phase.

Keywords: pos tagging, successive abstraction, lexical resource.

1. Tagger description

C4 is a portable statistical part of speech tagger (STL) based on a second order Markov model technique, implemented in C++ using standard template libraries. To improve tagging quality and efficiency we implemented the following solutions, as suggested in [2]:

1) Added beginning/ending sequence markers to bound each sentence to analyze.

2) To avoid setting to zero a complete word sequence analysis, we estimated the probability of unknown trigrams through context-independent linear interpolation, estimating lambda values with deleted interpolation.

$$\hat{P}(t_3|t_1,t_2) = \lambda_1\hat{P}(t_3) + \lambda_2\hat{P}(t_3|t_2) + \lambda_3\hat{P}(t_3|t_1,t_2)$$

3) To deal with unknown words we implemented suffix analysis smoothed by successive abstraction [3].

$$\hat{P}(x \mid C_k) = \frac{\sigma(C_k)^{-1} f(x \mid C_k) + \hat{P}(x \mid C_{k-1})}{\sigma(C_k)^{-1} + 1}$$

$$\sigma(C_k)^{-1} = \sqrt{12}\sqrt{|C_k|}e^{-\mathrm{H}(C_{k-1})}$$

4) To speed up the tagging process we chose the Viterbi algorithm with beam search.

C4 can take advantage of external linguistic resources to enrich the set of "known" words. For the EVALITA task we used "MorphIt!", a free corpus-based morphological resource for Italian [1] automatically mapped onto task tag sets. The lexicon in the current version (0.47) contains 504,906 entries and 34,968 lemmas.

During the fine tuning step of linguistic model construction, we improved performance on the recognition of proper names, adding a simple but effective rule of thumb: every word that is upper case and not at the beginning of a sentence is marked as a proper name.

2. Task results

TAGGER		GLOBAL DATA	UNKNOWN TOKENS
C4	*DISTRIB*	901 differences on 17313 tokens Accuracy = 94.80 Error Rate = 5.20	123 differences on 1326 tokens UTAccuracy = 90.72 UTError Rate = 9.28
	EAGLES	556 differences on 17313 tokens **Accuracy = 96.79** Error Rate = 3.21	113 differences on 1326 tokens UTAccuracy = **91.48** UTError Rate = 8.52
TNT	*DISTRIB*	700 differences on 17313 tokens Accuracy = 95.96 Error Rate = 4.04	175 differences on 1326 tokens UTAccuracy = 86.80 UTError Rate = 13.20
	EAGLES	551 differences on 17313 tokens **Accuracy = 96.82** Error Rate = 3.18	176 differences on 1326 tokens UTAccuracy = **86.73** UTError Rate = 13.27

Table 1: C4 accuracy

According to *Table 1* the tagger performed quite well compared to baseline taggers. In particular for the EAGLES task we reached state of the art accuracy. C4 did not perform at the same level for the DISTRIB task, probably because of the linguistic complexities in mapping the "MorphIt!" resource to the tag set.

3. Discussion

To study and understand C4 behaviour we constructed the following tables starting from raw linguistic data:

1. *Tagging error classes*: we collected all errors in the tagging procedure and clustered them around the correct tag.
2. M*ost frequent errors in context*: we extracted the trigram lists from both gold standard and C4 results, selected and counted the differences in the analysis.

Table 2 - EAGLES: TAGGING ERROR CLASSES

N	GOLD	TEST	TOKEN
61	NN	ADJ	immobile| minimi| bianco| malato| politici
43	CONJ_S	PRON_REL	che
41	ADJ	NN	legislative| acido| politico| artefici
29	ADJ	V_PP	mulistato| accentuato| accorto

Table 3 - EAGLES: TAGGING ERROR CLASSES DETAILS

N	GOLD	TEST
113	NN	ADJ|V_PP|NN_P|V_GVRB|ADV|ADJ_NUM|PREP
103	ADJ	NN|V_PP|V_GVRB|ADV|ADJ_IND|ADJ_DIM|C_NUM
55	CONJ_S	PRON_REL|PREP|CONJ_C|PRON_PER|ADV|NN
35	V_PP	ADJ|V_GVRB|NN|V_CLIT

Table 4 - EAGLES: MOST FREQUENT ERRORS IN CONTEXT

N	GOLD	TEST	TRIGRAMS
10	___ADJ_DIM___	PREP PRON_DIM NN	in questo modo| in quella luce
5	_____ADJ	PREP_A NN NN	dalle elezioni legislative| al piano terreno
5	_____CONJ_S	PREP_A NN PRON_REL	nel fatto che| alla possibilità che

For the EAGLES task the most frequent errors are quite usual in part of speech tagging:

1. relative pronoun-subordinate conjunction (che);
2. adjective-noun inversion;
3. past participle-adjective inversion.

The first kind of error is typically caused by the presence of a long distance dependency. The second and third kinds of errors frequently originate from semantic ambiguities (ie. "*La vecchia porta la sbarra*"). Pos-taggers using Markov models can't solve this kind of ambiguity and C4 seems to suffer from the same blind spots.

Table 5 - DISTRIB: ERROR CLASSES

N	GOLD	TEST	TOKEN
76	ENTITIES	ARG_DET	le|la|una|quella|questo|gli|
73	SUB_ARG	ARG_PREP	a|da|come|passava|Da|A
70	N	ADJ	immobile|minimi|bianco|malato|politici| albanese|soggetto|giovani|tecnica
52	SUB_ARG	PREP_NA	di
41	ADJ	N	legislative|francese|acido|artefici| berlusconiano|continuo|fine|terreno

Table 6 - DISTRIB: TAGGING ERROR CLASSES

N	GOLD	TEST
168	SUB_ARG	ARG_PREP|PREP_NA|REL|SUB_ADJ|ENTITIES
148	ENTITIES	ARG_DET|ADJ|ADV|REL|SUB_ADJ|SUB_ARG
128	ADJ	N|V|ARG_DET|ENTITIES|ADV|SUB_ARG|NULL
63	V	N|ADJ|ADV

Table 7 - DISTRIB: MOST FREQUENT ERRORS

N	GOLD	TEST	TRIGRAMS
35	___SUB_ARG___	V ARG_PREP V	continua a funzionare| tende a occupare
29	___SUB_ARG___	N PREP_NA V	modo di guardare| Usa dimuoversi
23	___SUB_ADJ___	N PREP_VA V	Satelliti per studiare| flash per chiudere

For the DISTRIB task we have to point out that, in addition to the errors mentioned above, we had troubles dealing with the following ambiguities:

1. SUB_ARG-ARG_PREP (a,da)
2. ENTITIES-ARG_DET
3. SUB_ARG-PREP_NA (di)
4. SUB_ADJ-PREP_VA (per)

As we have already suggested the most likely mistake could be an unsound mapping between the Morph-It! lexical resource and the tagset. In *Table 8* we give an account of the rules used to perform the critical tag mapping.

Table 8 – rules for mapping MorphIt! on DISTRIB tagset

- **SUB_ARG/ENTITIES:** *<lemma>+PRO-PERS,CI,CE,NE,SI,WH-CHE,PRO-DEMO,PRO-INDEF,PRO-WH,PRO-POSS,PRO-NUM*
- **SUB_ADJ:** *<subordinating conjunction>+CON*
- **ARG_DET:** *<lemma>+DET-DEMO,DET-INDEF,DET-WH,INT,DET-POS,ART,DET-NUM-CARD*
- **ARG_PREP**: *<lemma>+PRE,ARTPRE*

4. Conclusion

C4 shows it can achieve high accuracy in analysing Italian. Performance seems to be heavily dependent on the associated lexical resource and in particular on the quality of the handcrafted mapping for the tag set in use. Future developments should aim to extend the size and quality of available lexical resources: ie. list of proper nouns, technical lexicons, list of abbreviations.

REFERENCES

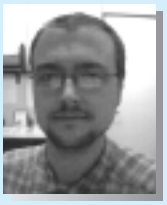
[1] E. Zanchetta and M. Baroni. 2006. Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings of Corpus Linguistics 2005*.

[2] T. Brants, 2000. TnT -- a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, April 29 -- May 3, 2000, Seattle, WA.

[3] C. Samuelsson, 1996. Handling Sparse Data by Successive Abstraction. In *Proceedings of COLING-96*, Kopenhagen, Denmark.

CONTACT

SIMONE ROMAGNOLI
Università di Bologna
Piazza S. Giovanni in Monte, 4
40124 Bologna
Email: simone.romagnoli3@unibo.it



SIMONE ROMAGNOLI collaborated with ExpertSystem (http://www.expertsystem.net/) as a professional consultant, with TCC division of ITC-IRST (http://tcc.itc.it/) as research consultant and with CILTA (http://www.cilta.unibo.it/). His main areas of interest are Natural Language Processing, Artificial intelligence ad its application to e-Learning.