Part-of-Speech Tagging with Lexicalized HMM*

Ferran Pla and **Antonio Molina**
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
46020 València, Spain
{fpla,amolina}@dsic.upv.es

Abstract

We present a method to build Lexicalized Hidden Markov Models (*L-HMM*) for improving the precision of Part-of-Speech (POS) tagging. This lexicalization increased the tagging precision from 96.28% to 96.71% (for bigrams) and from 96.58% to 96.80% (for trigrams) on the *Penn Treebank* corpus. Furthermore, we have conducted an experimental comparison that shows that *L-HMM* yields results which are better than or similar to other state-of-the-art POS tagging approaches.

1 Lexicalized HMM

Lexicalization techniques have been applied to POS tagging on different paradigms. The Maximum Entropy (ME) model (Ratnaparkhi 96) includes features for the words in different contexts. In addition, this model is refined by means of specializing some features for "difficult" words (words with a high error rate). However, this additional specialization got an improvement which was lower than 0.1%. Memory-based learning (MBL) methods improve their precision for certain languages including the focus word in the feature set (Zavrel & Daelemans 99). Transformation-based learning (TBL) (Brill 95) also improves its performance when it takes into account lexical rules (from 97.0% to 97.2% for known words). Several techniques for lexicalizing a HMM have been proposed, such as (Kim *et al.* 99) in which lexicalized states are made for "uncommon" words within a certain category. It improves the precision from 95.79% to 95.99%. A fully lexicalized HMM model is used in (Lee *et al.* 00).

The aim of this work is to present a lexicalization technique to enrich HMM. This technique consists of incorporating a set of selected words into the Language Model (LM) in addition to the POS tags. These words can be selected empirically from the training set or following other criteria. Although this lexicalization increases the size of the LMs, the performance of the tagging process improves. The lexicalization process is carried out on the training set as follows.

Let $\mathcal{C} = \{c_1, c_2, ..., c_N\}$ be a set of POS tags and $\mathcal{V} = \{w_1, w_2, ..., w_m\}$ the vocabulary of the application. Let $\mathcal{W}_s \subset \mathcal{V}$ be the word set to be incorporated in the LM. Taking into account this set, a specialization function f_s is defined over the training set $\mathcal{T}$ as follows:

$$f_s : \mathcal{T} \subset (\mathcal{V} \times \mathcal{C})^* \rightarrow \widetilde{\mathcal{T}} \subset \left(\mathcal{V} \times \widetilde{\mathcal{C}}\right)^*$$

$$f_s(\langle w_i, c_i \rangle) = \begin{cases} \langle w_i, (w_i, c_i) \rangle \text{ if } w_i \in \mathcal{W}_s \\ \langle w_i, (\lambda, c_i) \rangle \text{ if } w_i \notin \mathcal{W}_s \end{cases}$$

This function produces a new training set $\widetilde{\mathcal{T}}$ in which a POS tag c_i, is replaced by the new tag (w_i, c_i), if w_i is tagged with c_i and belongs to the set $\mathcal{W}_s$. If this word does not belong to $\mathcal{W}_s$, the POS tag is not changed. In this case, the POS tag has been represented as (λ, c_i), where λ stands for the null string. When this function is applied, there is a new set of POS tags $\widetilde{\mathcal{C}} \subset ((\mathcal{W}_s \cup \lambda) \times \mathcal{C})$.

The parameters of the *L-HMM* are estimated by Maximum Likelihood from this new training set ($\widetilde{\mathcal{T}}$), and this process does not depend on the HMM order.

Note that, when *L-HMM* are used, no change is needed in the tagging process. You simply have to map the sequence of output POS tags (that belongs to $\widetilde{\mathcal{C}}$) to the original POS tag set $\mathcal{C}$.

2 Experimental Work

We present the evaluation of tagging performance using the models described above. We considered first-order HMM (bigrams) and second-order HMM (trigrams). We tested how the lexicalization improves the tagging precision with respect to non-lexicalized HMM. In this sense, we defined different lexicalization criteria that are independent of the language.

*This work has been supported by the Spanish research projects CICYT TIC2000-1599-C01-01 and TIC2000-0664-C02-01

The experimental work was conducted using the TnT[1] tagger (Brants 00). TnT is a very efficient statistical POS tagger based on Hidden Markov models. To deal with sparse problems, it uses linear interpolation as smoothing technique to estimate the LM. To handle unknown words, it uses a probabilistic method based on the analysis of the suffix of the words. All the following experiments were done with TnT's default options.

We used the part of the Wall Street Journal which had been processed in the *Penn Treebank*, release 2. This corpus was automatically labelled with POS tags and manually checked as described in (Marcus *et al.* 93). The POS tag set is composed of 45 different tags. In all the experiments carried out, the training set consisted of sections 00 to 19 (956,549 words) and the test set included sections 23 and 24 (89,529 words). From this training set, we learnt both bigram (BIG) and trigram (TRI) models. With these models, TnT achieved a precision of 96.24% (BIG) and 96.45% (TRI). These results are considered as the baseline system to contrast with the lexicalized models.

2.1 Lexicalization Criteria

We defined two criteria to determine the set of words to specialize the models. The first one is based on the frequency of the words in the training set (SWF). The second one only takes into account the words in the training set that belong to closed categories (SCC).

For the SWF criterion, we chose the words whose frequency in the training set was higher than a certain *threshold* (some words such as proper nouns, punctuation marks or numbers were not considered). With these words, we specialized the training set and learnt the corresponding lexicalized bigram (BIG-SWF) and trigram (TRI-SWF) models.

For this experiment, we divided the training set into two partitions: 90% for training and 10% for tuning (development set). In order to determine which threshold maximized the performance of the model (that is, the best set of words to specialize the model), we tested on the development partition with word sets of different sizes.

In Figure 1, we show the results obtained with these specialized models on the development set.

[1]TnT is available on the WWW at `http://www.coli.uni-sb.de/thorsten/tnt`.

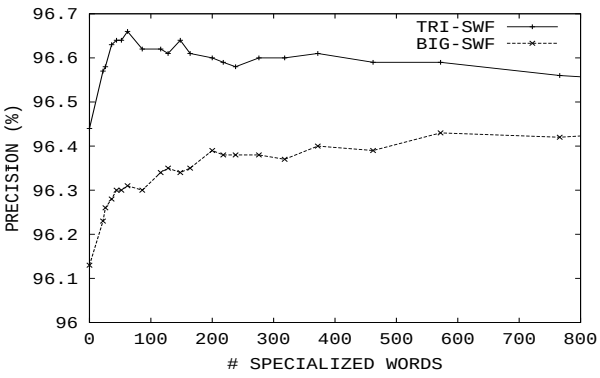

Figure 1: Performance of *L-HMM* using SWF criterion for different word set sizes on the development set.

The result for zero words corresponds to the baseline system. The precision for BIG-SWF and TRI-SWF was better than BIG and TRI, respectively. The best result for BIG-SWF was 96.43% using 572 words (those words whose frequency was higher than 250). The best precision for TRI-SWF was 96.66% using 62 words (with a frequency higher than 2,000). It can be observed that with a few words (around 60 words), lexicalized models obtain significant improvements. On the other hand, the use of more words in the models reduces the tagging precision. We think this is because the number of parameters of the models is larger and thus more poorly estimated for the same training set.

The main drawback of this criterion is that the threshold is very dependent on each training set and can only be known if a tagging experiment is carried out in advance.

Therefore, we proposed the SCC criterion which is based on more general properties. In particular, this criterion takes into account only the words from the training set that belong to closed categories[2]. The number of these words was 396 and the precision obtained was 96.34% for bigrams (BIG-SCC) and 96.60% for trigrams (TRI-SCC). These results were slightly lower than those obtained using the SWF criterion (96.43% for BIG-SWF and 96.66% for TRI-SWF), but the SCC criterion is more general and it can be established in advance.

Once the tagger was tuned, we applied it to a new unseen data set (sections 23 and 24 of WSJ). Table 1 shows the results on this test set. It can

[2]The closed categories considered are: *CC, DT, MD, POS, PP$, RP, TO, WDT, WP$, EX, IN, PDT, PRP, WP, WRB.*

Model	Precision
BIG	96.28%
BIG-SCC	96.52%
BIG-SWF	96.71%
TRI	96.58%
TRI-SCC	96.77%
TRI-SWF	96.80%

Table 1: Comparison among HMM and *L-HMM* models on WSJ corpus.

be observed that lexicalization improved tagging precision in all cases, and that SWF models performed better than SCC models. Moreover, we want to highlight that lexicalized bigram models outperformed trigram models in some cases (96.71% for BIG-SWF vs. 96.58% for TRI). This experimental result suggests that lower order *L-HMM* can perform as well as higher order HMM. This would be an important conclusion especially for small training sets, in which the problem of sparse data is more critical.

These criteria can be applied automatically to any training set, independently of the language and the tag set used. To confirm this, we have also applied *L-HMM* to the Spanish corpus *Lex-Esp* achieving similar improvements (Pla *et al.* 01).

2.2 Experimental Comparison

The results presented above are similar to the best tagging results reported in the literature on the WSJ corpus. However, these results cannot be reliably interpreted because the experimental conditions were different. Therefore, we performed some experiments in order to compare our system to other current tagging approaches (ME, TBL and MBL). The parameters of all taggers were set in order to optimize the tagging precision, but not the training and test time. The experiments for TRI-SFW, TBL and ME were run on a Pentium 266 Mhz with 256MB of RAM. The results for MBL were provided by Walter Daelemans on the same data sets.

Tagger	Precision	Training	Testing
TRI-SFW	96.80%	20 sec.	18,000 w/s
ME	96.92%	1 day	70 w/s
TBL	96.47%	9 days	750 w/s
MBL	96.45%	4.5 min.	11,200 w/s

Table 2: Comparison among different taggers on WSJ corpus.

Table 2 shows the results of this comparison among different taggers. We calculated tagging precision, training time and tagging speed (words per second) including file I/O. It can be observed that lexicalized models (TRI-SWF) perform as well or better than TBL and MBL. Only ME achieved a precision (96.92%) which was slightly better than TRI-SWF (96.80%), but on the other hand, the training time and testing speed for ME were much higher than TRI-SWF.

3 Conclusions

We have presented a method to build Lexicalized HMM incorporating a set of words into the LM. We used two different criteria which are independent of the language and the tag set used: the most frequent words in the training set (SWF) and the words that belong to closed categories (SCC).

In all the experiments conducted, the *L-HMM* outperformed the standard HMM tagger. This increment on the tagging precision is better than the results presented in other works (Kim *et al.* 99) that use more sophisticated lexicalization methods.

Finally, the experimental comparison conducted shows that our approach (*L-HMM*) outperforms other current approaches (HMM, MBL and TBL) and yields comparable results to the ME approach.

References

(Brants 00) Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

(Brill 95) E. Brill. Transformation–based Error–driven Learning and Natural Language Processing: A Case Study in Part–of–speech Tagging. *Computational Linguistics*, 21(4):543–565, 1995.

(Kim *et al.* 99) J.D. Kim, S.Z. Lee, and H.C. Rim. HMM Specialization with Selective Lexicalization. In *Proceedings of the join SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora(EMNLP-VLC-99)*, 1999.

(Lee *et al.* 00) Sang-Zoo Lee, Juni ichi Tsujii, and Hae-Chang Rim. Lexicalized Hidden Markov Models for Part-of-Speech Tagging. In *Proceedings of 18th International Conference on Computational Linguistics*, Saarbrucken, Germany, August 2000.

(Marcus *et al.* 93) M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 1993.

(Pla *et al.* 01) F. Pla, A. Molina, and N. Prieto. Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano. *Revista para el procesamiento del lenguaje natural (aceptado y pendiente de publicación)*, 2001.

(Ratnaparkhi 96) A. Ratnaparkhi. A Maximum Entropy Part–of–speech Tagger. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing, EMNLP*, 1996.

(Zavrel & Daelemans 99) J. Zavrel and W. Daelemans. Recent Advances in Memory–Based Part–of–Speech Tagging. In *Proceedings of the VI Simposio Internacional de Comunicacion Social*, Santiago de Cuba, Cuba, 1999.