Part-of-Speech Tagging Based on Machine Translation Techniques

Guillem Gascó i Mora and Joan Andreu Sánchez Peiró

Departament de Sistemes Informàtics i Computació
Universitat Politècnia de València
Camí de Vera s/n, 46022 València (Spain)
{ggasco,jandreu}@dsic.upv.es

Abstract. In this paper, a new approach to the Part-of-Speech (PoS) tagging problem is proposed. The PoS tagging problem can be viewed as a special translation process where the source language is the set of strings being considered and the target language is the sequence of POS tags. In this work, we have used phrase-based machine translation technology to tackle the PoS tagging problem. Experiments on the Penn Treebank WSJ task were carried out and very good results were obtained.

1 Introduction

Most Natural Language Processing (NLP) systems have to deal with the problem of ambiguity. Ambiguity increases the number of possible interpretations of natural language, and hence, the number of possible translations of a sentence. Part-of-Speech tagging can help identify the correct meaning (or translation) of a word in a sentence.

Part-of-Speech tagging is a well studied task. Many different approaches have been proposed to solve this problem [1,2,3,4]. A POS tagger can be considered as a translator between two "special" languages: the language that has to be tagged and a language formed by the PoS tags. Therefore, when we tag a sentence, we are translating that sentence into its corresponding tags. This idea has led us to use machine translation techniques to tackle the tagging process. Machine translation techniques are also used in [5] to tackle a different problem: Natural Language Understanding.

There are several different machine translation approaches. Currently, the best reported translation results are those acquired using phrase-based systems [6]. These systems use phrases to carry out the translation process. The phrases convey contextual information that can be very useful in the PoS tagging problem. Thus, we consider this sort of information to be useful in the PoS tagging process.

Section 2 introduces the Part-of-Speech tagging problem. In section 3, we explain the phrase-based machine translation procedure. Section 4 presents the application of these techniques to the PoS tagging problem. Finally, in section 5, experimental results are reported for the PennTreebank WSJ task.

J. Martí et al. (Eds.): IbPRIA 2007, Part I, LNCS 4477, pp. 257–264, 2007.

2 Part-of-Speech Tagging

Part-of-Speech tagging is the process of assigning a syntactic class marker to each word in a text. In a certain tagged sentence, every word has one, and only one, tag assigned to it. Nevertheless, more than one tag can be assigned to a single word depending on its context. The input of a tagging algorithm is a string of words, and the output is the most appropriate syntactic tag for each word.

The first approaches that were proposed for this task used handwritten linguistic rules to assign a tag to a word [1,7,8]. The main problem with these approaches is the high cost associated with obtaining the rules. Inductive approaches were later proposed to solve this problem. For example, a rule-based system that automatically learned rules from a tagged corpus was presented in [2]. Nevertheless, the best results reported are those from stochastic inductive approaches. The most relevant of these use Hidden Markov Models [9,10,3] and Maximum Entropy Models [4]. Although these approaches have obtained very good results, a further improvement is still possible.

One important factor in POS tagger performance is its behaviour with out-of-vocabulary (OOV) words, especially when training and test sets belong to different fields. An OOV word is a word that does not appear in the training set.

The kind of information that is useful when tagging an OOV word depends on the language being tagged. OOV words in English tend to be proper nouns. In other languages, like Mandarin Chinese or German, most of the OOV words are nouns or verbs. Previous approaches to the problem of tagging OOV words use different morphological features like prefixes, suffixes[1], capitalization, or hyphenation.

3 Phrase-Based Machine Translation

The main advantage of phrase-based machine translation over single-word-based approaches is that they take contextual information into account. The translation of one word usually depends heavily on its context. The basic idea of phrase-based translation is to segment a sentence into phrases, translate them individually, and compose the target sentence from these phrase translations.

The phrase translation model is based on the noisy channel model. Using Bayes rule, we reformulate the probability for translating a foreign sentence f into a target language sentence t as

$$\underset{t}{\operatorname{argmax}}\, p(t|f) = \underset{t}{\operatorname{argmax}}\, p(f|t)p(t)\ . \tag{1}$$

Then, we have a translation model $p(f|t)$ and a separate language model $p(t)$.

The sentence f is segmented into a sequence of N phrases. Each of these phrases, f_n, is translated into a target language phrase t_n. Phrase translation is modeled by a probability distribution $\phi(f_n|t_n)$. The output phrases can be

[1] For inflectional languages like English or German derivational and inflectional affixes tend to be a strong indicator of word classes.

reordered. This reordering is modeled by a relative distortion probability distribution $d(a_i - b_{i-1})$, where a_i denotes the start position of the foreign phrase that was translated into the ith target language phrase and b_{i-1} denotes the end position of the foreign phrase translated into the (i-1)th target language phrase. A ω factor is introduced to calibrate the output length. In summary, the best target language output sentence t_{best}, given a source language input sentence, is

$$t_{best} = \underset{t}{\operatorname{argmax}}\left(\prod_{n=1}^{N} \phi(f_n|t_n)d(a_i - b_{i-1})p(t)\omega^{|t|}\right) . \tag{2}$$

Obtaining a phrase probability translation table from a bilingual corpus is an important stage of phrase-based machine translation systems. This table maps source phrases to target language phrases. One of the greatest difficulties in translation is the possible reordering of words from different languages. Hence, a phrase alignment is needed to obtain good translation tables. There are several methods available for extracting the phrase probability translation table [11]. The most widely used language model is a smoothed trigram model.

4 The PoS-Tagging Problem as a MT Problem

As stated above, the tagging process can be viewed as a translation process. Therefore, the probability of obtaining the set of POS tags t for a source language sentence s is

$$t_{best} = \underset{t}{\operatorname{argmax}}\, p(t|s) = \underset{t}{\operatorname{argmax}}\, p(s|t)p(t) . \tag{3}$$

where $p(t)$ is a tag language model and $p(t|s)$ is a translation/tagging model.

As in the translation process, the source language input sentence is segmented into a sequence of phrases. Each phrase is "translated" into tag phrases. However, tag phrases cannot be reordered because each tag must be aligned with its corresponding word in a sentence.

There are some special features that make phrase-based POS tagging simpler than translation:

1. Reordering is not necessary.
2. Given a source language phrase composed of n words, its corresponding tag phrase has exactly n tags. There are no insertions or deletions, so every *n-long* source language sentence is *translated* into a *n-long* tag sentence.

Thus, the reordering probability distribution, $d(a_i - b_{i-1})$, and the word penalty factor, ω, can be removed from the model.

4.1 Phrase Extraction

For the reasons mentioned above, phrase extraction is easier for PoS tagging than for translation. Thus, from the training sentences, we obtained all the sequential

parallel phrases that were shorter than a given length l. Then, we obtained the probability of each segment by counting and normalizing.

It is important to note that the size of the translation table increases as l increases. Therefore, only small values for l can be considered. Moreover, the inclusion of long phrases does not provide a significant improvement in quality [11] since the probability of the reapearence of long phrases is very low.

4.2 Treatment of OOV Words

The translation of a word that does not appear in the training set is very difficult in classical machine translation. In the POS tagging problem, the target language vocabulary is a closed, well-defined set of POS tags. Therefore, it is possible to assign a POS tag to OOV words. Of course, it is more difficult to tag a word that has not appeared in the training set because there is no history of assigned tags. For this reason, other information such as context or morphological features should be taken into account. Tagging of OOV words is similar to a classification problem where an OOV word is assigned to a class (POS tag) depending on its context and certain other features.

Thus, we need to obtain phrases to *translate* OOV words. If the training set is large enough, it can be assumed that OOV words are infrequent words. Therefore, OOV words will probably have a behaviour that is similar to the behaviour of infrequent words of the training set. These infrequent words will be used to obtain the OOV word translation phrases.

Hence, we proceeded in the following way: We createe a set V composed by the N most frequent words of the training set. Every word in the training and test sentences that was not in V was replaced by the symbol *unk*. Then, we obtained the probability translation table phrases from the training sentences as we have explained in section 4.1.

When the phrase-based translation software finds an *unk* in a test sentence, it uses the *unk* phrases obtained from the training set. The words of the training set that have been replaced are considered to be OOV words. This is why N must be carefully chosen. If N is too large, there will not be enough *unk phrases* to get a reliable translation. On the other hand, if N is too small, the information from important words will be lost and the system performance will decrease.

5 Experiments

In this section, we present some preliminary experiments with the phrase-based tagger (PBT). The aim of these experiments was to test the system performance and to compare it with the performance of other taggers.

For these experiments, we chose the Penn Treebank corpus [12] to evaluate our phrase-based POS tagging system. This corpus has been automatically tagged and manually checked. Its tagset is composed of 45 different tags. We divided the corpus into three sets: training, tuning and test. Table 1 shows some features of these sets.

Table 1. Corpus and partition features

	Directories	Sentences	Words	Vocabulary	OOV words
Training	00-20	42,075	1,004,073	40,695	-
Tuning	21,22	3,371	80,156	9,942	2,172
Test	23,24	3,762	89,537	10,568	2,045

The phrase-based machine translation software Pharaoh [13] was used in these experiments. Pharaoh implements a beam search decoding for phrase-based statistical machine translation.

We have used a tag trigram model that was smoothed using the Kneser-Ney discount technique with Chen-Goodman modifications as the language model. Perplexity values of tuning and test sets are 8.36 and 8.28, respectively.

All the experiments except the ones in subsection 5.3 used the tuning set for tagging. To avoid an excessively large translation table, we used phrases whose maximum length was 3 in all the experiments. Note that, the larger the translation table, the slower the tagging system.

5.1 OOV Words

An important factor that affects system performance is the treatment of OOV words. As we stated in section 4.2, the size of the vocabulary of OOV words must be chosen carefully. Table 2 shows the results of the system using OOV word treatment with different vocabulary sizes. The accuracy of the basic system, which does not use OOV word treatment, is displayed in the first row.

Table 2. Performance of PBT with OOV word treatment for different vocabulary sizes

Vocabulary Size	OOV Words	Accuracy		
		Known Words	OOV Words	Global
Basic System	2172	96.83	0%	94.2%
10,000	6001	97.12%	60.52%	94.39%
20,000	3778	97.03%	60.43%	95.31%
30,000	3081	96.96%	62.60%	**95.72**%

The performance of the basic system for OOV words was 0%. As stated above, the use of OOV word treatment increases the number of OOV words. In spite of this, the system performance increased because OOV word accuracy improved to 60.52% with a vocabulary of only 10,000 words. With a larger vocabulary of 30,000 words, the OOV word accuracy did not improve but the number of OOV words decreased significantly. This increased the global system accuracy to 95.72%.

OOV word treatment can be improved if some morphological features are taken into account. Following other works, we chose capitalization, hyphenation, suffixes, and prefixes. To use capitalization information, a new OOV word

class (with its corresponding replacement word) was created. Infrequent training words and OOV test words were replaced using two different special *unk* symbols depending on their capitalization.

The same strategy was used with hyphenation and suffixes. Only the most significant English suffixes were taken into account. Table 3 shows some of the suffixes that were used. English prefixes usually only provide semantic information. Therefore, if pw is a OOV word and p is a semantic prefix and w is in the training set vocabulary, then pw will be replaced by w. Hence, we used w tag history information to tag pw because p does not modify the POS tag.

Table 3. Frequent significant English suffixes

Suffix	Usual tags
-ed	VBN VBD JJ
-ing	VBG
-er	JJR
-ion	NN
-an	JJ
-ness	NN
-al	JJ
-able	JJ
-est	JJS
-ly	RB
-s	NNS NNPS

Table 4 shows the improvement in tagging obtained by using these features. When all of them were used, the performance of the tagger on OOV words increased to 76.53%. The global system accuracy was 96.54%.

In the experiments described below, the tagger used OOV word treatment with a 30,000-word vocabulary and all the morphological features.

Table 4. Performance of PBT with OOV word treatment for different morphological information. B: Basic system; C: Capitalization; S: Suffixes; H: Hyphenation; P: Prefixes.

	Accuracy		
Information used	Known Words	OOV Words	Global
B	96.96%	61.6%	95.72%
B+C	96.94%	72.92%	96.21%
B+C+S	96.92%	74.47%	96.45%
B+C+S+H+P	96.96%	76.53%	**96.54**%

5.2 Maximum Phrase Length

As we stated in section 4.1, long phrases produce very large translation tables and slow taggers. In addition, long phrases are rarely used in translation systems.

Table 5. Accuracy and number of phrases produced by different maximum phrase length values

Maximum phrase length	Phrases produced	Accuracy
1	35,882	96.22%
2	358,791	**96.57%**
3	973,992	96.54%
4	1,694,364	96.54%
5	2,427,061	96.53%

Therefore, we carried out a serie of experiments to test how the length of the phrases affected the performance of the tagger. Table 5 displays these results.

As the table indicates, the maximum phrase length 2 had the best performance: 96.57%. Learning longer phrases did not yield much improvement. This may be due to the fact that long training phrases rarely appear in test sentences.

5.3 Final Results and Comparison

In this section, we compare the accuracy of PBT with two well-known taggers: the Brill rule-based tagger[2], and the TnT[3] which is a HMM-based tagger. In order to make an adequate comparison we trained three taggers with the same training set and tagged the same test set. The results are shown in Table 6.

As the table indicates, the phrase-based tagger (PBT) performed significantly better than the other taggers.

Table 6. Comparison between PBT and other taggers

Tagger:	TnT	Brill	**PBT**
Tagging accuracy:	96.61%	96.39%	**96.78%**

Experiments reported in other publications used some typical training, tuning and test sets[2]. To compare their results with the results obtained with PBT, we carried out a final experiment using these sets, and we obtained an accuracy of 96.97%.

6 Conclusions

We have presented a new approach to the POS tagging problem using machine translation techniques. The phrase-based translation model has been chosen since the use of phrase information improves tagging. Some special tagging features simplify the translation model by making the word penalty and the reordering model unnecessary. In addition, OOV words can be tagged using context and morphological information.

[2] Directories 2-20 for training, 23 for tuning and 24 for test.

The experiments reported here show that phrase-based Part-of-Speech tagging is a viable approach for achieving state-of-the-art accuracy. Moreover, with these techniques it is very easy to obtain taggers for other languages.

Acknowledgements. This work has been partially supported by the *Universitat Politècnica de València* with the ILETA project and by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01.

References

1. Harris, Z.: String analysis of sentence structure. Mouton, The Hague (1962)
2. Brill, E.: A Simple Rule-Based Part-of-speech Tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing. ANLP (1992)
3. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA (2000)
4. Ratnaparkhi, A.: A Maximum Entropy Part-of.Speech Tagger. In: Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing, EMNLP 1996 (1996)
5. Bender, O., Macherey, K., Och, F., Ney, H.: Comparison of Alignment Templates and Maximum Entropy Models for Natural Language Understanding. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2003 (2003)
6. Zens, R., Och, F., Ney, H.: Improvements in phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference, HLT-NAACL'2004 (2004)
7. Klein, S., Simons, F.: A computational approach to grammatical coding of English words. Journal of the Association for Computing Machinery. vol. 10(3) (1963)
8. Greene, B., Rubin, M.: Automatic tagging of English. Technical report, Department of Linguistics. Providence, Rhode Island. 1071 (1962)
9. Weischedel, R., Schwartz, R., Palmucci, J., Meteer, M., Ramsaw, L.: Coping with ambiguity and unknown words trhough probabilistic models. Computational Linguistics. vol. 19(2) (1993)
10. Merialdo, B.: Tagging English text with a probabilistic model. Computational Linguistics. vol. 20(2) (1994)
11. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference, HLT-NAACL'2003 (2003)
12. Marcus, M., Santorini, B., Marcinkiewicz, M.: Building a large annotated corpus of english: The Penn Treebank. Computational Linguistics, vol. 19(2) (1994)
13. Koehn, P.: Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In: Frederking, R.E., Taylor, K.B. (eds.) AMTA 2004. LNCS (LNAI), vol. 3265, Springer, Heidelberg (2004)