# Deploying Part-of-Speech Patterns to Enhance Statistical Phrase-Based Machine Translation Resources

**Christina Lioma**
Department of Computing Science
University of Glasgow
G12 8QQ
xristina@dcs.gla.ac.uk

**Iadh Ounis**
Department of Computing Science
University of Glasgow
G12 8QQ
ounis@dcs.gla.ac.uk

## Abstract

Part-of-Speech patterns extracted from parallel corpora have been used to enhance a translation resource for statistical phrase-based machine translation.

## 1 Introduction

The use of structural and syntactic information in language processing implementations in recent years has been producing contradictory results. Whereas language generation has benefited from syntax [Wu, 1997; Alshawi et al., 2000], the performance of statistical phrase-based machine translation when relying solely on syntactic phrases has been reported to be poor [Koehn et al., 2003].

We carry out a set of experiments to explore whether heuristic learning of part-of-speech patterns from a parallel corpus can be used to enhance phrase-based translation resources.

## 2 System

The resources used for our experiments are as follows. The statistical machine translation GIZA++ toolkit was used to generate a bilingual translation table from the French-English parallel and sentence-aligned Europarl corpus. Additionally, a phrase table generated from the Europarl French-English corpus, and a training test set of 2000 French and English sentences that were made available on the webpage of the ACL 2005 workshop[1] were also used. Syntactic tagging was realized by the TreeTagger, which is a probabilistic part-of-speech tagger and lemmatizer. The decoder used to produce machine translations was Pharaoh, version 1.2.3.

We used GIZA++ to generate a translation table from the parallel corpus. The table produced consisted of individual words and phrases, followed by their corresponding translation and a unique probability value. Specifically, every line of the said table consisted of a French entry (in the form of one or more tokens), followed by an English entry (in the form of one or more tokens), followed by $P(f|e)$, which is the probability $P$ of translation to the French entry $f$ given the English entry $e$. We added the GIZA++-generated table to the phrase-based translation table downloaded from the workshop webpage. During this merging of translation tables, no word or phrase was omitted, replaced or altered. We chose to combine the two aforementioned translation tables in order to achieve better coverage. We called the resulting merged translation table *lexical phrase table*.

In order to utilize the syntactic information stemming from our resources, we used the TreeTagger to tag both the parallel corpus and the *lexical phrase table*. The probability values included in the *lexical phrase table* were not tagged. The TreeTagger uses a slightly modified version of the Penn Treebank tagset, different for each language. In order to achieve tag-uniformity, we performed the following dual tag-smoothing operation.

[1] The Europarl French-English corpus and phrase table, and the training test set are available at: http://www.statmt.org/wpt05/mt-shared-task/



*Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 163–166,
Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

Firstly, we changed the French tags into their English equivalents, i.e. NOM (noun – French) became NN (noun – English). Secondly, we simplified the tags, so that they reflected nothing more than general part-of-speech information. For example, tags denoting predicate-argument structures, wh-movement, passive voice, inflectional variation, and so on, were simplified. For example, NNS (noun – plural) became NN (noun).

Once our resources were uniformly tagged, we used them to extract part-of-speech correspondences between the two languages. Specifically, we extracted a sentence-aligned parallel corpus of French and English part-of-speech patterns from the tagged Europarl parallel corpus. We called this corpus of parallel and corresponding part-of-speech patterns *pos-corpus*. The format of the *pos-corpus* remained identical to the format of the original parallel corpus, with the sole difference that individual words were replaced by their corresponding part-of-speech tag. Similarly, we extracted a translation table of part-of-speech patterns from the tagged *lexical phrase table*. We called this part-of-speech translation table *pos-table*. The *pos-table* had exactly the same format as *the lexical phrase table*, with the unique difference that individual words were replaced by their corresponding part-of-speech tag. The translation probability values included in the *lexical phrase table* were copied onto the *pos-table* intact.

Each of the part-of-speech patterns contained in the *pos-corpus* was matched against the part-of-speech patterns contained in the *pos-table*. Matching was realized similarly to conventional left-to-right string matching operations. Matching was considered to be successful not simply when a part-of-speech pattern was found to be contained in, or part of a longer pattern, but when patterns were found to be absolutely identical. When a perfect match was found, the translation probability value of the specific pattern in the *pos-table* was increased to the maximum value of 1. If the score were already 1, it remained unchanged. When there were no matches, values remained unchanged. We chose to match identical part-of-speech patterns, and not to accept partial pattern matches, because the latter would require a revision of our probability recomputation method. This point is discussed in section 3 of this paper.

Once all matching was complete, the newly enhanced *pos-table*, which now contained translation probability scores reflecting the syntactic features of the relevant languages, was used to update the original *lexical phrase table*. This update consisted in matching each and every part-of-speech pattern with its original lexical phrase, and replacing the initial translation probability score with the values contained in the *pos-table*. The identification of the original lexical phrases that generated each and every part-of-speech pattern was facilitated by the use of pattern-identifiers (*pos-ids*) and phrase-identifiers (*phrase-ids*), which were introduced at a very early stage in the process for that purpose. The resulting translation phrase table contained exactly the same entries as the *lexical phrase table*, but had different probability scores assigned to some of these entries, in line with the parallel part-of-speech co-occurrences and correspondences found in the Europarl corpus. We called this table *enhanced phrase table*. Table 1 illustrates the process described above with the example of a phrase, the part-of-speech analysis of which has been used to increase its original translation probability value from 0.333333 to 1.

| *Lexical phrase table* |
|---|
| actions extérieures \| external action \| 0.333333 |
| Tagged *lexical phrase table* |
| actions_NN extérieures_JJ \| external_JJ action_NN \| 0.333333 |
| *pos-corpus* |
| NN JJ \| JJ NN |
| *Enhanced phrase table* |
| actions extérieures \| external action \| 1 |

Table 1: Extracting and matching a part-of-speech pattern to increase translation probability.

We used the Pharaoh decoder firstly with our *lexical phrase table*, and secondly with our *enhanced phrase table* in order to generate statistical machine translations of source and target language variations of the French and English training test set. We measured performance using the BLEU score [Papineri et al., 2001], which estimates the accuracy of translation output with respect to a reference translation. For both source-target language combinations, the use of the *lexical phrase table* received a slightly lower score than the score achieved when using the *enhanced phrase table*. The difference between these two approaches is not significant (p-value > 0.05). The results of our

experiments are displayed in Table 2 and discussed in Section 3.

| Language Pair | Lexical | Enhanced |
|---|---|---|
| English-French | 25.50 | 25.63 |
| French-English | 26.59 | 26.89 |

Table 2: Our translation performance (measured with BLEU)

## 3 Discussion

The motivation behind this investigation has been to test whether syntactic or structural language aspects can be reflected or represented in the resources used in statistical phrase-based machine translation.

We adopted a line of investigation that concentrates on the correspondence of part-of-speech patterns between French and English. We measured the usability of syntactic structures for statistical phrase-based machine translation by comparing translation performance when a standard phrase table was used, and when a syntactically enhanced phrase table was used. Both approaches scored very similarly. This similarity in the performance is justified by the following three factors.

Firstly, the difference between the two translation resources, namely the *lexical phrase table* and the *enhanced phrase table*, does not relate to their entries, and thus their coverage, but to a simple alteration of the translation probability values of some of their entries. The coverage of these resources is exactly identical.

Secondly, a closer examination of the translation probability value alterations that took place in order to reflect part-of-speech correspondences reveals that the proportion of the entries of the phrase table that were matched syntactically to phrases from the parallel corpus, and thus underwent a modification in their translation probability score, was very low (less than 1%). The reason behind this is the fact that the part-of-speech patterns produced by the parallel corpus were long strings in their vast majority, while the part-of-speech patterns found in the phrase table were significantly shorter strings. The inclusion of phrases longer than three words in translation resources has been avoided, as it has been shown not to have a strong impact on translation performance [Koehn et al., 2003].

Thirdly, the above described translation probability value modifications were not parameterized, but consisted in a straightforward increase of the translation probability to its maximum value. It remains to be seen how these probability value alterations can be expanded to a type of probability value 'reweighing', in line with specific parameters, such as the size of the resources involved, the frequency of part-of-speech patterns in the resources, the length of part-of-speech patterns, as well as the syntactic classification of the members of part-of-speech patterns. If one is to compare the impact that such parameters have had upon the performance of automatic information summarisation [Mani, 2001] and retrieval technology [Belew, 2000], it may be worth experimenting with such parameter tuning when refining machine translation resources.

A note should be made to the choice of tagger for our experiments. A possible risk when attempting any syntactic examination of a large set of data may stem from the overriding role that syntax often assumes over semantics. Statistical phrase-based machine translation has been faced with instances of this phenomenon, often disguised as linguistic idiosyncrasies. This phenomenon accounts for such instances as when nouns appear in pronominal positions, or as adverbial modifiers. On these occasions, and in order for the syntactic examination to be precise, words would have to be defined on the basis of their syntactic distribution rather than their semantic function. The TreeTagger abides by this convention, which is one of the main reasons why we chose it over a plethora of other freely available taggers, the remaining reasons being its high speed and low error rate. In addition, it should be clarified that there is no statistical, linguistic, or other reason why we chose to adopt the English version of the Penn TreeBank tagset over the French, as they are both equally conclusive and transparent.

The overall driving force behind our investigation has been to test whether part-of-speech structures can be of assistance to the enhancement of translation resources for statistical phrase-based machine translation. We view our use of part-of-speech patterns as a natural extension to the introduction of structural elements to statistical machine translation by Wang [1998] and Och et al. [1999].

Our empirical results suggest that the use of part-of-speech pattern correspondences to enhance existing translation resources does not damage machine translation performance. What remains to be investigated is how this approach can be optimized, and how it would respond to known statistical machine translation issues, such as mapping nested structures, or the handling of 'unorthodox' language pairs, i.e. agglutinative-fusion languages.

## 4 Conclusion

Syntactic and structural language information contained in a bilingual parallel corpus has been extracted and used to refine the translation probability values of a translation phrase table, using simple heuristics. The usability of the said translation table in statistical phrase-based machine translation has been tested in the shared task of the second track of the ACL 2005 Workshop on Building and Using Parallel Corpora. Findings suggest that using part-of-speech information to alter translation probabilities has had no significant effect upon translation performance. Further investigation is required to reveal how our approach can be optimized in order to produce significant performance improvement.

## References


Alshawi, H., Bangalore, S., and Douglas, S. (2000). Learning Dependency Translation Models as Collections of Finite State Head Transducers. *Computational Linguistics*, 26(1).

Belew, R. K. (2000). *Finding Out About: Search Engine Technology from a Cognitive Perspective*. Cambridge University Press, USA.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference 2003 (HLT/NAACL 2003)*, pages 127-133.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company, Amsterdam.

Och, F. J., Tilmann, C., and Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora 1999 (EMNLP 1999)*, pages 20-28.

Papineri, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Report.

Wang, Y. (1998). Grammar Inference and Statistical Machine Translation. Ph.D. thesis, Carnegie Melon University.

Wu, D. (1997). Stochastic Inversion transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3).

Yamada, K. and Knight, K. (2001). A Syntax-based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 39)*, pages 6-11.