Parallel Treebanks in Phrase-Based Statistical Machine Translation

John Tinsley, Mary Hearne, and Andy Way

National Centre for Language Technology
Dublin City University, Ireland
{*jtinsley, mhearne, away*}*@computing.dcu.ie*

Abstract. Given much recent discussion and the shift in focus of the field, it is becoming apparent that the incorporation of syntax is the way forward for the current state-of-the-art in machine translation (MT). Parallel treebanks are a relatively recent innovation and appear to be ideal candidates for MT training material. However, until recently there has been no other means to build them than by hand. In this paper, we describe how we make use of new tools to automatically build a large parallel treebank and extract a set of linguistically motivated phrase pairs from it. We show that adding these phrase pairs to the translation model of a baseline phrase-based statistical MT (PBSMT) system leads to significant improvements in translation quality. We describe further experiments on incorporating parallel treebank information into PBSMT, such as word alignments. We investigate the conditions under which the incorporation of parallel treebank data performs optimally. Finally, we discuss the potential of parallel treebanks in other paradigms of MT.

1 Introduction

The majority of research in recent years in machine translation (MT) has centred around the phrase-based statistical approach. This paradigm involves translating by training models which make use of sequences of words, so-called phrase pairs, as the core translation model of the system [1]. These phrase pairs are extracted from aligned sentence pairs using heuristics over a statistical word alignment. While phrase-based models have achieved state-of-the-art translation quality, evidence suggests there is a limit as to what can be accomplished using only simple phrases, for example, satisfactory capturing of context-sensitive reordering phenomena between language pairs [2]. This assertion has been acknowledged within the field as illustrated by the recent shift in focus towards more linguistically motivated models.

Aside from the development of fully syntax-based models of MT, [3–6] to list a few, there have been many extensions and improvements to the phrase-based model which have endeavoured to incorporate linguistic information into the translation process. Examples of these can be seen in the work of [7] and [8] who make use of syntactic supertags and morphological information respectively. [9, 10] describes a phrase-based model which makes use of generalised templates while [11] exploit semantic information in the form of phrase-sense disambiguation. All of these approaches have a common starting point: the set of phrase pairs initially extracted in the phrase-based model.

Given this, we raise two questions: 1) would translation quality improve in a baseline phrase-based system if the translation model included linguistically motivated, constituent-based phrase pairs? and 2) would subsequent extensions to the phrase-based model, such as those outlined above, improve even further if they were implemented on a base of linguistically motivated phrase pairs? In this paper we will address the first question, with the second question being discussed in terms of future work.

We have shown previously that, on a small scale, incorporating linguistically motivated phrase pairs extracted from parallel treebanks can improve phrase-based statistical MT (PBSMT) systems [12]. We further examine this hypothesis by scaling up the experiments of [12] by approximately 2 orders of magnitude. We then carry out a detailed series of experiments to determine how to optimally use parallel treebank phrase pairs within the phrase-based model. In addition to this, we investigate some alternative ways of incorporating the information encoded in parallel treebanks, such as word alignments, into the translation process of a PBSMT system.

The remainder of this paper is outlined as follows: Section 2 gives some background on SMT phrase extraction and parallel treebanks. Section 3 describes the data used in all experiments in this paper. Section 4 details the experiments carried out along with results, analysis and discussion. Finally, we conclude and present some avenues for future work in Section 5.

2 Background

At the core of any phrase-based SMT system lies a table of translationally equivalent phrase pairs. These phrase pairs are extracted from parallel corpora, on a sentence pair by sentence pair basis, using heuristics which operate on a set of high-recall word alignments between the sentence pairs. The phrase pairs are then scored in a log linear model combining a number of different features. It was shown by [1] that restricting the set of extracted phrase pairs to those which correspond to syntactic constituents in a context-free phrase-structure tree harms translation accuracy. We carried out experiments previously [12], whereby rather than *restrict* the set of phrase pairs to those corresponding to constituents, we *supplement* the phrase-based translation model with all linked constituent pairs in a syntactically annotated version of the same parallel data used to train the PBSMT system. This led to improved accuracy across four translation tasks. The results of these experiments are summarised in Table 1.

Config.	**en-es**	**es-en**	**en-de**	**de-en**
Baseline	0.1765	0.1754	0.1186	0.1622
+Tree	**0.1867**	**0.1880**	**0.1259**	**0.1687**

Table 1. Summary of translation results reported in [12] in terms of Bleu score.

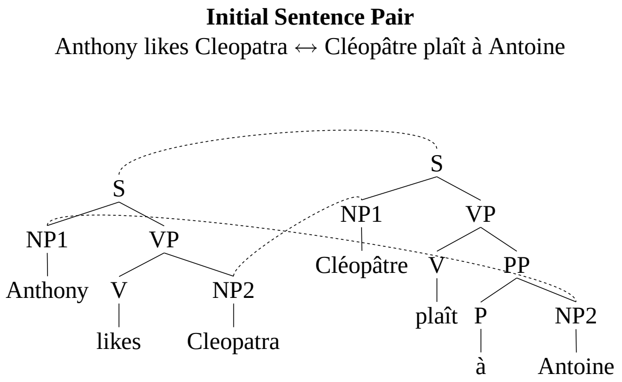


Fig. 1. An example English–French parallel treebank entry for the given sentence pair.

The acquisition of such syntactically annotated parallel resources, so-called parallel treebanks, has been the topic of much recent research [13–15]. A parallel treebank comprises syntactically parsed aligned sentences in two or more languages. In addition to this, sentences are aligned below the level of the clause [16], i.e. there are alignments between nodes in the tree pairs, which indicate translational equivalence between the surface strings dominated by the linked node pairs. An example parallel treebank entry is shown in Figure 1.

Until relatively recently parallel treebank acquisition was a manual task. It is a time-consuming, error-prone process which requires linguistic expertise in both the source and target languages. This makes it an impractical task on a large scale, such as the scale on which we may need to work in MT. For these reasons, parallel treebanks are thin on the ground and those that are available are relatively small [17, 18]. However, recent advances in technology, such as improvements in monolingual parsing and the development of subtree alignment tools, such as those described in the work referred to earlier in this section, have paved the way for the automatic creation of large high-quality parallel treebanks. In the following section we detail the construction of the parallel treebank used in our experiments.

3 Parallel Data

The principal resource used for the experiments described in this paper is the English–Spanish section of the Europarl corpus. After cleaning, which involved the removal of blank lines, erroneous alignments and sentences over 100 tokens in length, there were 729,891 aligned sentence pairs remaining. The process of building a parallel treebank from this parallel corpus was completely automated. Firstly, each monolingual corpus was parsed using freely available phrase-structure parsers. For the English corpus we used the Berkeley parser [19]. The Spanish corpus was parsed using Bikel's parser [20] trained on the Cast3LB Spanish treebank [21].

The final step in the annotation process was to automatically align the newly parsed parallel corpus at sub-sentential level. This is done by inserting links between constituent node pairs in the tree which imply translational equivalence between the surface strings dominated by the linked node pairs. Tree alignment is a precision-based task – the goal is not to aggressively align as many nodes as possible in the tree. To leave a node unaligned is not to say it has no translational equivalent. Instead, translational equivalences for unaligned nodes are encapsulated in wider contexts by links higher up in the tree pair. For example, looking back to the tree pair in Figure 1, although there is no direct link from the source tree V node, dominating *likes*, to the target tree, does not mean it has no translation in this sentence pair. Instead, its translational equivalence to the non-constituent *plaît à* is captured implicitly by the links between the S nodes and the NP nodes. To insert these links between the parallel tree pairs we used our own subtree alignment algorithm [22]. This algorithm automatically induces links between nodes (at both word- and phrase-level) in a tree pair by exploiting statistical word alignment probabilities estimated over the sentence pairs of the tree pairs to be aligned.

Given the parallel treebank is built automatically, the issue of its quality arises. Of course, there are parse errors and misalignments to be found, but we are satisfied that the quality is high enough to demonstrate our hypothesis. The papers describing the two parsers we use both report high accuracy: 90.05% labelled f-score for English, and 83.96% labelled f-score for Spanish. The reported accuracy of the sub-tree alignment algorithm is also high. We refer the interested reader to the original alignment paper for a more detailed evaluation.

4 Experiments

This section reports on the various experiments we carried out in which we incorporate phrase pairs extracted from the parallel treebank into a phrase-based SMT system. We first describe how we use the parallel treebank phrase pairs directly in translation, in Section 4.1. We follow this up in Sections 4.2–4.5 by examining a number of different approaches to incorporating the information encoded in the parallel treebank into the translation process.

For all translation experiments the setup included a development set of 1,000 sentence pairs, a test set of 2,000 sentence pairs,[1] all chosen at random, with the remaining 726,891 sentence pairs (and tree pairs where relevant) used for training. The baseline MT system was built using Moses [23]. For the phrase-extraction step of the training process, phrases pairs up to a maximum of 7 tokens in length were extracted using the *grow-diag-final* heuristic. 5-gram language modelling was carried out using the SRI language modelling toolkit [24]. System tuning was performed on the development set using minimum error-rate training as implemented in Moses. All translations were performed from English into Spanish and were automatically evaluated using the metrics BLEU [25], NIST [26] and METEOR [27]. Statistical significance was calculated using bootstrap resampling [28] (with p=0.05 unless otherwise stated).

[1] Test sentences were restricted in length to between 5 and 30 tokens.

4.1 Combining Phrase Resources

The first question we want to answer is: can linguistically motivated phrase pairs extracted from our parallel treebank improve translation when incorporated into a baseline phrase-based SMT system? To find out we must first extract the set of phrase pairs from the parallel treebank. These phrases correspond to the yields of all linked constituent pairs in the treebank. We then add these phrase pairs to the translation model of the baseline MT system and reestimate the phrase translation probabilities over the combined set of phrase pairs. We will illustrate this process with an example. In Figure 2 we see an example sentence pair from an English–French parallel corpus. Figure 2(a) illustrates the parallel treebank entry for this pair, while Figure 2(b) shows its statistical word alignment according to the PBSMT system. The combined set of extracted phrase pairs, to be added to the translation model, is given in Figure 2(c). We can see that while there is overlap between the two sets of phrase pairs, there are also a certain number of phrase pairs unique to the parallel treebank. Our hypothesis is that these unique constituent-based phrase pairs, along with the increase in probability mass given to those overlapping phrase pairs, will improve translation quality.

Table 2 shows the results of translation experiments using different combinations of data in the translation model.

Config.	**BLEU**	NIST	**%**METEOR
Baseline	0.3341	7.0765	57.39
+Tree	**0.3397**	**7.0891**	**57.82**
Tree only	0.3153	6.8187	55.98

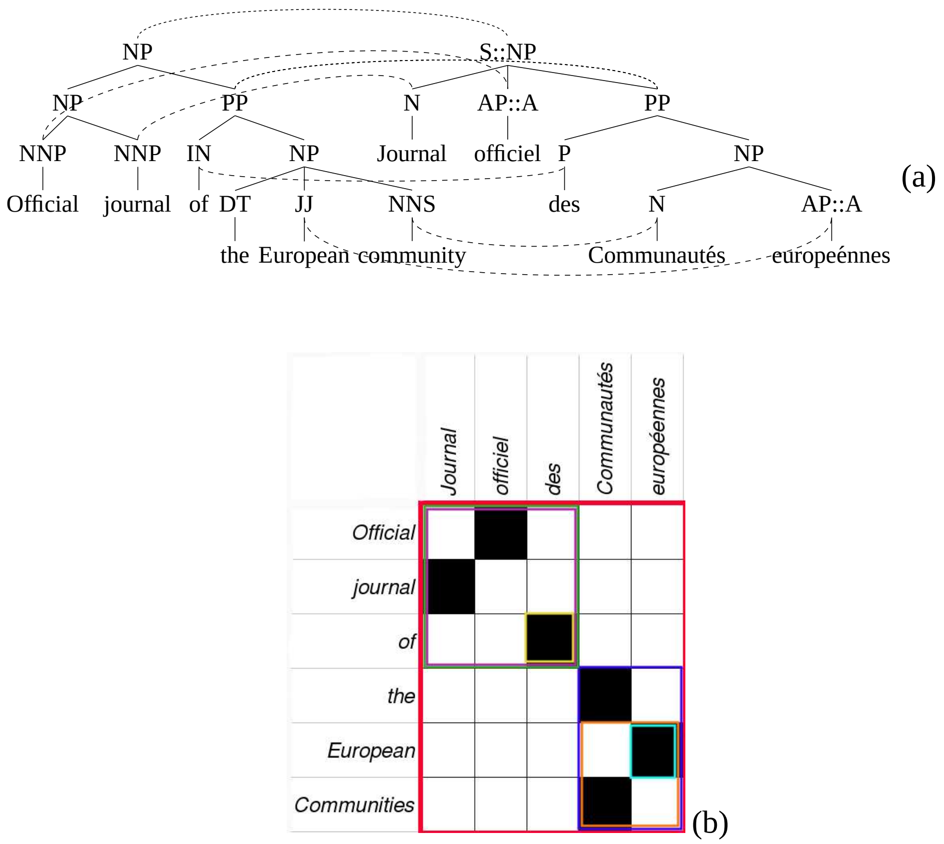
Table 2. Evaluation of combinations of data in translation models. Baseline = PBSMT phrase pairs. Tree = phrase pairs from the parallel treebank.

We see that adding parallel treebank phrase pairs to the baseline model (+Tree) significantly improves translation accuracy (1.68% relative increase in BLEU score) across all metrics. We attribute this to the increase in coverage of the translation model given the new phrase pairs combined with the increased probability mass of the phrase pairs in common between the two sets. This effect is desireable as we would assume those phrase pairs extracted by both methods would be more reliable. Of the treebank phrase pair types added to the translation model, 77.5% of these were not extracted in the baseline system. These ultimately constituted 16.79% of the total phrases in the translation model. The remaining treebank phrase pairs which were also extracted in the baseline system comprised 4.87% of the total phrase pairs. The full figures are provided in Table 3.

Using the data from the parallel treebank alone (Tree only) leads to a significant drop in translation accuracy (5.96% relative BLEU) compared the baseline. We attribute the drop to the insufficient translation coverage of this model. This was to be expected as we can maximally extract the number of phrase pairs from a tree pair as there are linked node pairs. This number will never approach the number necessary to achieve translation coverage competitive with that of the baseline system.

Training Sentence Pair

“Official journal of the ↔ ”Journal officiel des
European Community”   Communautés européennes“



† Official journal ↔ Journal officiel
† Official journal of ↔ Journal officiel des
∗ Official journal of the/ ↔ Journal officiel des/
European Communities   Communautés européennes
∗ of ↔ des
∗ of the European Communities ↔ des Communautés européennes
∗ the European Communities ↔ Communautés européennes
∗ European ↔ européennes
⋄ Communities ↔ Communautés
⋄ Official ↔ officiel
⋄ journal ↔ Journal

(c)

Fig. 2. Example of phrase extraction for the given sentence pair depicting: (a) the aligned parallel tree pair; (b) the word alignment matrix (the rectangled areas represent extracted phrase pairs); (c) the combined set of extracted phrase pairs where: ⋄ = only extracted from (a); † = only extracted from (b); ∗ = extracted from both (a) and (b).

Resource	**#Tokens**	**#Types**	$\bigcap$
Baseline	72,940,465	24,708,527	1,447,505
Treebank	21,123,732	6,432,771	

Table 3. Frequency information regarding the numbers of phrase pairs extracted from the baseline system and from the parallel treebank. $\bigcap$ is the number of phrase pair types extracted by both methods.

We carried out one further experiment where we added only strict phrase pairs[2] from the parallel treebank into the baseline phrase-based system. The motivation for this was the discovery that of all the data extracted from the parallel treebank, 20.3% were word alignments and 7.35% of these were alignments between function words and punctuation that occurred more than 1,000 times. By removing these high-risk alignments we reduce the potential for search errors while keeping the vast majority of useful translation units. The outcome of this experiment, presented in Table 4, was even further significant improvement (2.18% relative increase in BLEU score) across all metrics over the baseline phrase-based system than using all the parallel treebank data.

Config.	BLEU	NIST	**%**METEOR
Baseline	0.3341	7.0765	57.39
+Tree	0.3397	7.0891	57.82
Strict phrases	**0.3414**	**7.1283**	**57.98**

Table 4. Effect of using strictly phrase pairs from the parallel treebank.

Given these findings, which corroborate our findings in [12], we now describe further experiments we carried out to investigate additional ways to exploit the information encoded in the parallel treebank to use with the PBSMT framework.

4.2 Weighting Treebank Data

In the previous section we showed that we can improve over the baseline PBSMT system by simply adding parallel treebank phrases to the translation model. Our next set of experiments investigate whether giving more weight to the syntactic phrase pairs in the translation model will further improve performance. The motivation here is that the syntactic phrase pairs may be more reliable, as we suggested in [12], and thus preferable for use in translation. To do this we built 3 translation models – of the form Baseline+Tree – in which we count the parallel treebank phrase pairs twice, three times and five times when estimating phrase translation probabilities. The results of these experiments are shown in Table 5 [3].

[2] A strict phrase pair is an *m-to-n* alignment where both *m* and *n* are greater than 1.

[3] A * next to a particular configuration in the table indicates the results reported are statistically insignificant *compared to the baseline*. We assume this to be the case in all proceeding tables.

Config.	BLEU	NIST	%METEOR
Baseline+Tree	**0.3397**	**7.0891**	**57.82**
+Tree x2*	0.3386	7.0813	57.76
+Tree x3	0.3361	7.0584	57.56
+Tree x5*	0.3377	7.0829	57.71

Table 5. Effect of increasing relative frequency of parallel treebank phrase pairs in the translation model.

The findings here are slightly erratic. Doubling the presence of the parallel treebank phrase pairs (+Tree x2) leads to insignificant differences compared to the baseline across all metrics, while counting them three times (+Tree x3) leads to a significant drop (p=0.02) in translation accuracy. Counting them five times (+Tree x5) again leads to insignificant differences.

Given the ineffectiveness of this crude method of weighting, we built a system using two distinct phrase tables, one containing the baseline phrase-based SMT phrases and the other containing the phrase pairs from the parallel treebank. This allows the tuning process to choose the optimal weights for the two phrase tables and the decoder can chose phrase pairs from either table as the model dictates. Table 6 shows the performance of this system relative to the Baseline+Tree configuration. Again, no improvement was found. We see a significant decrease in translation accuracy but it is not uniform across the metrics.

Config.	BLEU	NIST	%METEOR
Baseline+Tree	**0.3397**	**7.0891**	**57.82**
Two Tables	0.3365	7.0812*	57.50

Table 6. Effect of using two separate phrase tables in the translation model.

We know from the experiments of Section 4.1 that adding parallel treebank data to the baseline phrase-based system can improve translation quality. However, simply increasing their frequency in the translation model has a detrimental effect on translation. This may be due to the fact that we are also increasing the influence of those treebank phrase pairs which are not as useful – such as those word alignments also mentioned in the previous section – and this is having a negative effect.

A potential way to proceed along these lines may be to find a more balanced compromise between the two sets of phrase pairs in the translation model, but for now we can conclude that when adding parallel treebank phrase pairs to the model, it is optimal to add them a single time into the baseline model.

4.3 Filtering Treebank Data

Phrase pairs extracted in the baseline system were restricted in length to 7 tokens as previous experiments have shown that phrases longer than this yield little improvement and are occasionally detrimental to translation quality [1]. In our previous experiments no such restriction was placed on the parallel treebank phrase pairs. To investigate whether longer treebank phrase pairs were harming translation quality, we built a translation model – Baseline+Tree – including parallel treebank phrase pairs up to a maximum of 7 tokens in length only. The filtered phrase table was 11.7% smaller than that which contained unrestricted phrase pairs. The effect of this filtering on translation performace is shown in Table 7 where we see statistically insignificant fluctuation across the metrics. This indicates that the longer phrases were inconsequential during decoding. Further analysis confirms this, with longer phrases rarely being used in the Baseline+Tree configuration, and only a small percentage (8%) of the sentences being translated differently when filtering them out. From this we can conclude that when adding treebank phrase pairs, we need only add in those phrase pairs of similar length to the ones in the baseline model.

Config.	BLEU	NIST	**%**METEOR
Baseline+Tree	**0.3397**	7.0891	**57.82**
-Filtered*	0.3387	**7.0926**	57.67

Table 7. Effect of using filtering longer phrase pairs from the parallel treebank data.

4.4 Treebank-Driven Phrase Extraction

In this section we describe experiments in which we used the alignment information encoded in the parallel treebank to seed the phrase extraction heuristic in the PBSMT system.

One oft-cited reason for the inability of syntactic translation models to improve upon the state-of-the-art is that only using constituent-based phrase pairs is too restrictive [1, 9]. Translation units such as the English–German pair *there is* ↔ *es gibt* will never be extracted as a constituent phrase pair despite being a perfectly acceptable translation pair. To attempt to overcome this problem, we sought some ways in which to use the linguistic information encoded in the parallel treebank to extract a set of non-constituent-based phrase pairs. By doing this we would have "linguistically informed" phrase pairs as opposed to purely constituent phrase pairs.

In order carry this out, we built a translation model by seeding Moses' phrase extraction heuristic with the word alignments from the parallel treebank. The motivation for this is that we have syntax-based word alignments in the parallel treebank guided by the non-lexical links higher up in the tree [22] and thus subsequent phrases extracted based on these would possibly have more of a linguistic foundation than those based on statistical word alignments, and be potentially more reliable.

We also built a translation model using the union of the Moses word alignments and the parallel treebank word alignments. Finally, we built two more translation models in which both of the models above were supplemented with the phrase pairs extracted from the parallel treebank, as this was the original hypothesis we were examining. The results of translation experiments using all of these models are presented in Table 8.

Config.	BLEU	NIST	**%**METEOR
Baseline	0.3341	7.0765	57.39
+Tree	**0.3397**	**7.0891**	57.82
TBX	0.3102	6.6990	55.64
+Tree	0.3199	6.8517	5639
UnionX	0.3277	6.9587	56.79
+Tree	0.3384*	7.0508	**57.88**

Table 8. Evaluation of translations using different word alignments to seed phrase extraction. TBX = extraction seeded by parallel treebank word alignments. UnionX = extraction seeded by union of parallel treebank and Moses word alignments.

The first two rows in the table showing the results from Section 4.1 represent our baseline here. In the third row (TBX), we see that seeding the phrase extraction with the treebank alignments leads to a significant drop in translation performance compared to the baseline. Adding the treebank phrase pairs (+Tree) to this model significantly improves performance as we would expect given our previous findings, however, it still does not approach the performance of the baseline.

Seeding the phrase extraction using parallel treebank word alignments leads to an unwieldy amount of phrase pairs in the translation model – approximately 88.5 million (92.9 million when including treebank phrase pairs)– many of which are useless e.g. *framework for olaf, in order that* ↔ *marco*. This is due to the fact that the parallel treebank word alignments have quite low recall and thus the phrase extraction heuristic is free to extract a large number of phrases anchored by a single word alignment.[4] This tells us that the parallel treebank word alignments are too sparse to be used to seed the phrase extraction heuristics.

The intuition behind the next experiment – using the union of the parallel treebank and Moses word alignments to seed phrase extraction – was to simultaneously increase the recall of statistical word alignments and the precision of the parallel treebank word alignments and creating a more robust, reliable word alignment overall.

We see from the fifth row (UnionX) of Table 8 that using the union of alignments led to a small, but significant, drop in translation accuracy compared to the baseline. More interestingly we note that adding the parallel treebank phrase pairs to this model (UnionX+Tree) led to comparable performance to the baseline. [5] This is interesting

[4] In the example *framework for olaf, in order that* ↔ *marco* the only word alignment was between *framework* and *marco*.

[5] Differences were either statistically insignificant or inconsistent across the evaluation metrics.

as the baseline translation model including treebank phrases, Baseline+Tree, has approximately 29.7M entries. However, the UnionX+Tree translation model contains only 13.1M phrase pair entries. This constitutes a 56% decrease in translation model size without any significant decrease in translation accuracy. These figures, and those for the other models described in this section, are given in Table 9. This discovery is a very positive by-product of these experiments. We can conclude that using the union of statistical and treebank-based word alignments may be effective for producing smaller translation models without suffering a reduction in translation performance. We intend to investigate these findings in greater depth in the near future.

Word Alignment	#Phrases	#Phrases+Tree
Moses	24.7M	29.7M
Treebank	88.5M	92.89M
Union	7.5M	13.1M

Table 9. Comparison of the phrase table size for each model. #Phrase = number of phrases extracted using a given word alignment. #Phrase+Tree = size of model when treebank phrases are included.

4.5 Alternative Lexical Weighting

In this section we discuss experiments carried out in which we used the information encoded in the parallel treebank to calculate the values for the lexical weighting feature in the log-linear model.

The translation model in a phrase-based SMT system, in addition to calculating a phrase translation probability, calculates a lexical weighting score for each phrase pair. This feature checks how well the words in the source and target phrases translate to one another by scoring each phrase pair according to its word alignment using the word translation table extracted during training.

In order to potentially improve these lexical weighting scores, we recalculate them according to the word alignments found in the parallel treebank, as opposed to the statistical word alignment. Firstly we reassign each phrase pair in the translation model (Baseline+Tree) a word alignment according to the parallel treebank word alignments. We then estimate a word translation distribution over the word alignments in the parallel treebank and use this to calculate new lexical weights for the phrase pairs in the translation model.

We then replicate this setup by assigning the phrase pairs new alignments according to the union of the statistical and parallel treebank word alignments – as we did in Section 4.4 – and scoring them from a word translation probability distribution over all the word alignments from both resources. The results of these experiments are given in Table 10.

Config.	BLEU	NIST	%METEOR
Baseline+Tree	**0.3397**	**7.0891**	**57.82**
TB_words	0.3356	7.0355	57.32
Union_words	0.3355	7.0272	57.41

Table 10. Effect of using linguistically motivated word alignments to calculate lexical weighting for phrase pairs in the translation model. TB_words = lexical weights according to treebank word alignments. Union_words = lexical weights according to union of treebank and Moses word alignments.

We see from these results that performance degrades slightly, but significantly, when using the new lexical weights and that the results are almost identical between the two new methods of scoring.[6]

The ineffectiveness of this approach can be attributed to the fact the the majority of the phrase-pairs, i.e. those extracted in Moses, were extracted according to the statistical word alignments and thus would have a high-recall word alignment. To replace these word alignments with the parallel treebank alignments, however precise, will give a much lower recall word alignment between the extracted phrase pairs. This, coupled with the fact that the lower recall word alignments give a less reliable word translation table, leads to poorer lexical weights and, ultimately, a decrease in translation quality.

5 Conclusions and Future Work

Augmenting the standard phrase-based model with linguistically motivated phrase pairs from a parallel treebank can improve translation quality. Some ongoing work we are carrying out along these lines involves investigating the effect of the treebank phrase pairs on translation performance as the size of the training set increases. Early results seem to indicate that increasing the training set leads to a decrease in the influence of the treebank phrases.

As per the second question we raised in Section 1, it would be interesting to investigate whether some of the approaches mentioned in the introduction, which improved over the standard model, would yield further improvement by building on the treebank-induced model described here.

In Section 4.2 we saw that simply increasing the relative frequency of the treebank phrases in the model did not help, nor did using separate phrase tables. But we believe there is still a better compromise to be found between all the phrase resources in the model. Section 4.3 indicated that filtering the phrase table has negligible effect on translation accuracy.

We saw in Sections 4.4 and 4.5 that variations on incorporating the parallel treebank data, specifically the word alignments, did not lead to any improvements. The phrase-based model is tailored to high-recall statistical word alignments and reductions

[6] This is not to say there was no difference between them. 18.5% of the sentences in the test set were translated differently.

in recall as seen here, regardless of the precision, do not lend themselves to improved translations. However, we also saw that we can induce much smaller translation models from the parallel treebank without a significant drop in MT performance.

Finally, what we have described here is only scratching the surface in terms of the exploitability of parallel treebanks in MT. We are currently working on the extraction of generalised translation templates and translation rules from parallel treebanks with a view to evaluating their performance in more syntax-aware models of MT, such as those of [4] and [6]. Such models are illustrative of the potential of this linguistically rich resource.

Acknowledgements

This work is supported by Science Foundation Ireland (grant no. 05/RF/CMS064) and the Irish Centre for High-End Computing [7].

References

1. Koehn, P., Och, F.J., Marcu, D.: Statistical Phrase-Based Translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada (2003) 48–54
2. Zollmann, A., Venugopal, A., Och, F., Ponte, J.: A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. In: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, England (2008) 1145–1152
3. Yamada, K., Knight, K.: A Syntax-Based Statistical Translation Model. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01), Toulouse, France (2001) 523–530
4. Hearne, M.: Data-Oriented Models of Parsing and Translation. PhD thesis, Dublin City University, Dublin, Ireland (2005)
5. Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., Thayer, I.: Scalable Inference and Training of Context-Rich Syntactic Translation Models. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006) 961–968
6. Lavie, A.: Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation. In: Proceedings of thr 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel (2008) 362–375
7. Hassan, H., Sima'an, K., Way, A.: Supertagged Phrase-based Statistical Machine Translation. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic (2007) 288–295
8. Koehn, P., Hoang, H.: Factored Translation Models. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic (2007) 868–876
9. Chiang, D.: A Hierarchical Phrase-Based Model for Statistical Machine Translation. In: 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, MI (2005) 263–270
10. Chiang, D.: Hierarchical Phrase-Based Translation. Computational Linguistics **33** (2007) 201–228

[7] http://www.ichec.ie/

11. Carpuat, M., Wu, D.: How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In: Proceedings of TMI-07, Skövde, Sweden (2007) 43–52
12. Tinsley, J., Hearne, M., Way, A.: Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07), Bergen, Norway (2007) 175–187
13. Samuelsson, Y., Volk, M.: Alignment Tools for Parallel Treebanks. In: Proceedings of the Biennial GLDV Conference, Tübingen, Germany (2007)
14. Lavie, A., Parlikar, A., Ambati, V.: Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In: Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation (SSST-2), Columbus, OH (2008)
15. Zhechev, V., Way, A.: Automatic Generation of Parallel Treebanks. In: Proceedings of the 22nd International Conference on Computational Linguistics (CoLing'08), Manchester, UK (2008) 1105–1112
16. Volk, M., Samuelsson, Y.: Bootstrapping Parallel Treebanks. In: Proceedings of the 7th Conference of the Workshop on Linguistically Interpreted Corpora (LINC), Geneva, Switzerland (2004) 71–77
17. Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kuboň, V.: Prague Czech-English Dependency Treebank. Syntactically Annotated Resources for Machine Translation. In: Proceedings of LREC-2004, Lisbon, Portugal (2004) 1597–1600
18. Gustafson-Čapková, S., Samuelsson, Y., Volk, M.: SMULTRON - The Stockholm MULtilingual parallel TReebank. www.ling.su.se/dali/research/smultron/index (2007)
19. Petrov, S., Klein, D.: Improved Inference for Unlexicalized Parsing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY (2007) 404–411
20. Bikel, D.: Design of a Multi-lingual, parallel-processing statistical parsing engine. In: Human Language Technology Conference (HLT), San Diego, CA (2002)
21. Civit, M., Martí, M.A.: Building Cast3LB: A Spanish Treebank. Research on Language and Computation **2(4)** (2004) 549–574
22. Tinsley, J., Zhechev, V., Hearne, M., Way, A.: Robust Language-Pair Independent Sub-Tree Alignment. In: Machine Translation Summit XI, Copenhagen, Denmark (2007) 467–474
23. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic (2007) 177–180
24. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of the International Conference Spoken Language Processing, Denver, CO. (2002)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA (2002) 311–318
26. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Human Language Technology: Notebook Proceedings, San Diego, CA (2002) 128–132
27. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL-05, Ann Arbor, MI (2005)
28. Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain (2004) 388–395