A Quantitative Analysis of Reordering Phenomena

Alexandra Birch a.c.birch-mayne@sms.ed.ac.uk
Phil Blunsom pblunsom@inf.ed.ac.uk
Miles Osborne miles@inf.ed.ac.uk
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB, UK

Abstract

Reordering is a serious challenge in statistical machine translation. We propose a method for analysing syntactic reordering in parallel corpora and apply it to understanding the differences in the performance of SMT systems. Results at recent large-scale evaluation campaigns show that synchronous grammar-based statistical machine translation models produce superior results for language pairs such as Chinese to English. However, for language pairs such as Arabic to English, phrase-based approaches continue to be competitive. Until now, our understanding of these results has been limited to differences in BLEU scores. Our analysis shows that current state-of-the-art systems fail to capture the majority of reorderings found in real data.

1 Introduction

Reordering is a major challenge in statistical machine translation. Reordering involves permuting the relative word order from source sentence to translation in order to account for systematic differences between languages. Correct word order is important not only for the fluency of output, it also affects word choice and the overall quality of the translations.

In this paper we present an automatic method for characterising syntactic reordering found in a parallel corpus. This approach allows us to analyse reorderings quantitatively, based on their number and span, and qualitatively, based on their relationship to the parse tree of one sentence. The methods we introduce are generally applicable, only requiring an aligned parallel corpus with a parse over the source or the target side, and can be extended to allow for more than one reference sentence and derivations on both source and target sentences. Using this method, we are able to compare the reordering capabilities of two important translation systems: a phrase-based model and a hierarchical model.

Phrase-based models (Och and Ney, 2004; Koehn et al., 2003) have been a major paradigm in statistical machine translation in the last few years, showing state-of-the-art performance for many language pairs. They search all possible reorderings within a restricted window, and their output is guided by the language model and a lexicalised reordering model (Och et al., 2004), both of which are local in scope. However, the lack of structure in phrase-based models makes it very difficult to model long distance movement of words between languages.

Synchronous grammar models can encode structural mappings between languages which allow complex, long distance reordering. Some grammar-based models such as the hierarchical model (Chiang, 2005) and the syntactified target language phrases model (Marcu et al., 2006) have shown better performance than phrase-based models on certain language pairs.

To date our understanding of the variation in reordering performance between phrase-based and synchronous grammar models has been limited to relative BLEU scores. However, Callison-Burch et al. (2006) showed that BLEU score alone is insufficient for comparing reordering as it only measures a partial ordering on n-grams. There has been little direct research on empirically evaluating reordering.

We evaluate the reordering characteristics of these two paradigms on Chinese-English and Arabic-English translation. Our main findings are as follows: (1) Chinese-English parallel sentences exhibit many medium and long-range reorderings, but less short range ones than Arabic-English, (2) phrase-based models account for short-range reorderings better than hierarchical models do, (3)

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 197–205,
Athens, Greece, 30 March – 31 March 2009.

by contrast, hierarchical models clearly outperform phrase-based models when there is significant medium-range reordering, and (4) none of these systems adequately deal with longer range reordering.

Our analysis provides a deeper understanding of why hierarchical models demonstrate better performance for Chinese-English translation, and also why phrase-based approaches do well at Arabic-English.

We begin by reviewing related work in Section 2. Section 3 describes our method for extracting and measuring reorderings in aligned and parsed parallel corpora. We apply our techniques to human aligned parallel treebank sentences in Section 4, and to machine translation outputs in Section 5. We summarise our findings in Section 6.

2 Related Work

There are few empirical studies of reordering behaviour in the statistical machine translation literature. Fox (2002) showed that many common reorderings fall outside the scope of synchronous grammars that only allow the reordering of child nodes. This study was performed manually and did not compare different language pairs or translation paradigms. There are some comparative studies of the reordering restrictions that can be imposed on the phrase-based or grammar-based models (Zens and Ney, 2003; Wellington et al., 2006), however these do not look at the reordering performance of the systems. Chiang et al. (2005) proposed a more fine-grained method of comparing the output of two translation systems by using the frequency of POS sequences in the output. This method is a first step towards a better understanding of comparative reordering performance, but neglects the question of what kind of reordering is occurring in corpora and in translation output.

Zollmann et al. (2008) performed an empirical comparison of the BLEU score performance of hierarchical models with phrase-based models. They tried to ascertain which is the stronger model under different reordering scenarios by varying distortion limits the strength of language models. They show that the hierarchical models do slightly better for Chinese-English systems, but worse for Arabic-English. However, there was no analysis of the reorderings existing in their parallel corpora, or on what kinds of reorderings were produced in their output. We perform a focused evaluation of these issues.

Birch et al. (2008) proposed a method for extracting reorderings from aligned parallel sentences. We extend this method in order to constrain the reorderings to a derivation over the source sentence where possible.

3 Measuring Reordering

Reordering is largely driven by syntactic differences between languages and can involve complex rearrangements between nodes in synchronous trees. Modeling reordering exactly would be sparse and heterogeneous and thus we make an important simplifying assumption in order for the detection and extraction of reordering data to be tractable and useful. We assume that reordering is a binary process occurring between two blocks that are adjacent in the source. We extend the methods proposed by Birch et al. (2008) to identify and measure reordering. Modeling reordering as the inversion in order of two adjacent blocks is similar to the approach taken by the Inverse Transduction Model (ITG) (Wu, 1997), except that here we are not limited to a binary tree. We also detect and include non-syntactic reorderings as they constitute a significant proportion of the reorderings.

Birch et al. (2008) defined the extraction process for a sentence pair that has been word aligned. This method is simple, efficient and applicable to all aligned sentence pairs. However, if we have access to the syntax tree, we can more accurately determine the groupings of embedded reorderings, and we can also access interesting information about the reordering such as the type of constituents that get reordered. Figure 1 shows the advantage of using syntax to guide the extraction process. Embedded reorderings that are extracted without syntax assume a right branching structure. Reorderings that are extracted using the syntactic extraction algorithm reflect the correct sentence structure. We thus extend the algorithm to extracting syntactic reorderings. We require that syntactic reorderings consist of blocks of whole sibling nodes in a syntactic tree over the source sentence.

In Figure 2 we can see a sentence pair with an alignment and a parse tree over the source. We perform a depth first recursion through the tree, extracting the reorderings that occur between whole sibling nodes. Initially a reordering is detected between the leaf nodes P and NN. The block growing algorithm described in Birch et al. (2008) is then used to grow block A to include NT and NN, and block B to include P and NR. The source and target spans of these nodes do not overlap the spans

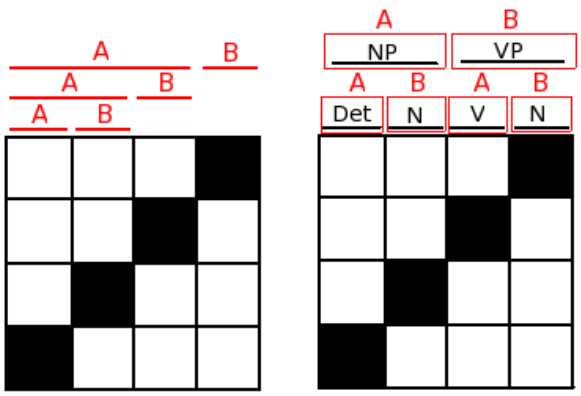

Figure 1. An aligned sentence pair which shows two different sets of reorderings for the case without and with a syntax tree.

of any other nodes, and so the reordering is accepted. The same happens for the higher level reordering where block A covers NP-TMP and PP-DIR, and block B covers the VP. In cases where the spans do overlap spans of nodes that are not siblings, these reorderings are then extracted using the algorithm described in Birch et al. (2008) without constraining them to the parse tree. These non-syntactic reorderings constitute about 10% of the total reorderings and they are a particular challenge to models which can only handle isomorphic structures.

RQuantity

The reordering extraction technique allows us to analyse reorderings in corpora according to the distribution of reordering widths and syntactic types. In order to facilitate the comparison of different corpora, we combine statistics about individual reorderings into a sentence level metric which is then averaged over a corpus. This metric is defined using reordering widths over the target side to allow experiments with multiple language pairs to be comparable when the common language is the target.

We use the average RQuantity (Birch et al., 2008) as our measure of the amount of reordering in a parallel corpus. It is defined as follows:

$$RQuantity = \frac{\sum_{r \in R} |r_{A_{\bar{t}}}| + |r_{B_{\bar{t}}}|}{I}$$

where R is the set of reorderings for a sentence, I is the target sentence length, A and B are the two blocks involved in the reordering, and $|r_{A_{\bar{s}}}|$ is the size or span of block A on the target side. RQuantity is thus the sum of the spans of all the reordering blocks on the target side, normalised

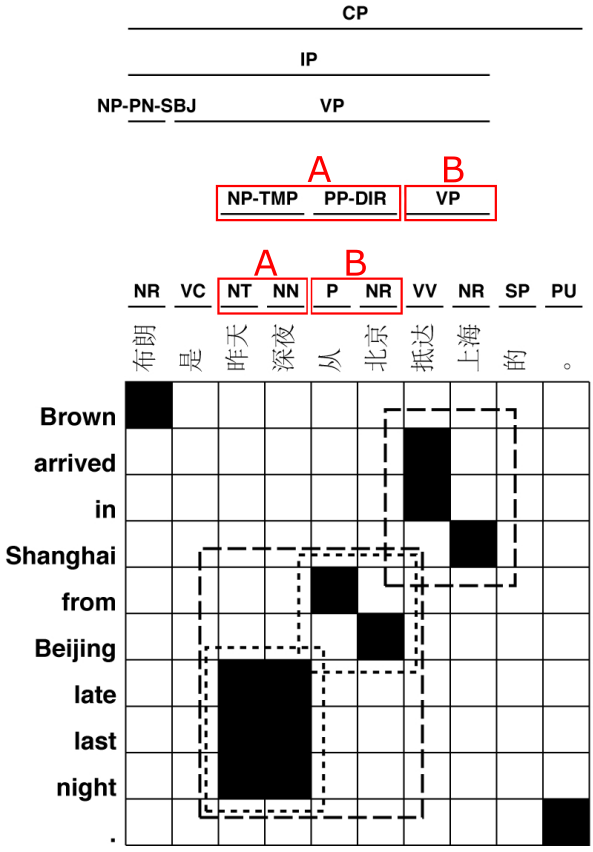

Figure 2. A sentence pair from the test corpus, with its alignment and parse tree. Two reorderings are shown with two different dash styles.

by the length of the target sentence. The minimum RQuantity for a sentence would be 0. The maximum RQuantity occurs where the order of the sentence is completely inverted and the RQuantity is $\sum_{i=2}^{I} i$. See, for example, Figure 1 where the RQuantity is $\frac{9}{4}$.

4 Analysis of Reordering in Parallel Corpora

Characterising the reordering present in different human generated parallel corpora is crucial to understanding the kinds of reordering we must model in our translations. We first need to extract reorderings for which we need alignments and derivations. We could use automatically generated annotations, however these contain errors and could be biased towards the models which created them. The GALE project has provided gold standard word alignments for Arabic-English (AR-EN) and Chinese-English (CH-EN) sentences.[1] A subset of these sentences come from the Arabic and Chinese treebanks, which provide gold standard parse trees. The subsets of parallel data for which we have both alignments and parse trees consist of

[1] see LDC corpus LDC2006E93 version GALE-Y1Q4

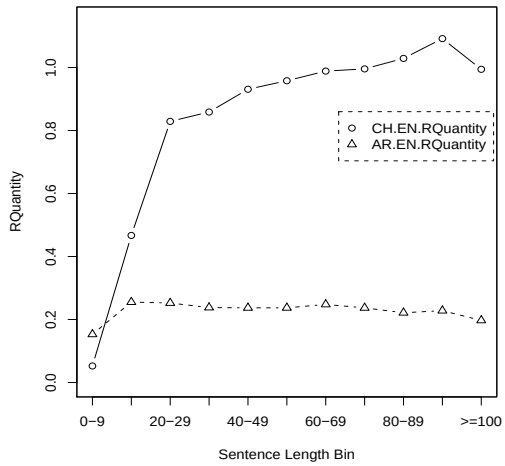

Figure 3. Sentence level measures of RQuantity for the CH-EN and AR-EN corpora for different English sentence lengths.

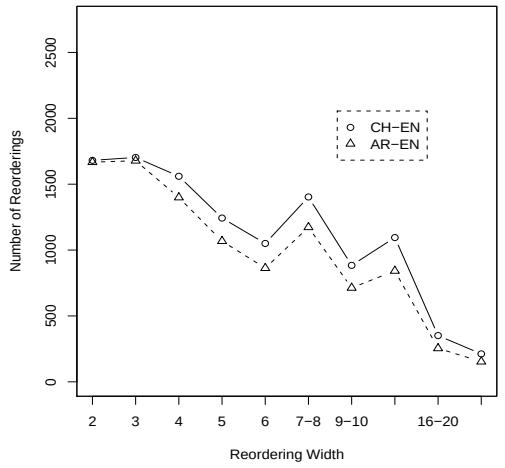

Figure 4. Comparison of reorderings of different widths for the CH-EN and AR-EN corpora.

3,380 CH-EN sentences and 4,337 AR-EN sentences.

Figure 3 shows that the different corpora have very different reordering characteristics. The CH-EN corpus displays about three times the amount of reordering (RQuantity) than the AR-EN corpus. For CH-EN, the RQuantity increases with sentence length and for AR-EN, it remains constant. This seems to indicate that for longer CH-EN sentences there are larger reorderings, but this is not the case for AR-EN. RQuantity is low for very short sentences, which indicates that these sentences are not representative of the reordering characteristics of a corpus. The measures seem to stabilise for sentences with lengths of over 20 words.

The average amount of reordering is interesting, but it is also important to look at the distribution of reorderings involved. Figure 4 shows the reorderings in the CH-EN and AR-EN corpora bro-

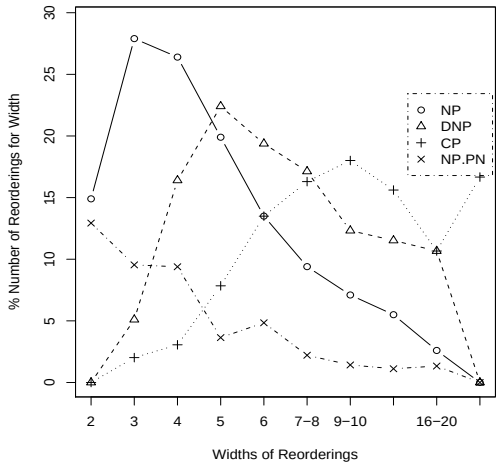

Figure 5. The four most common syntactic types being reordered forward in target plotted as % of total syntactic reorderings against reordering width (CH-EN).

ken down by the total width of the source span of the reorderings. The figure clearly shows how different the two language pairs are in terms of reordering widths. Compared to the CH-EN language pair, the distribution of reorderings in AR-EN has many more reorderings over short distances, but many fewer medium or long distance reorderings. We define *short*, *medium* or *long distance* reorderings to mean that they have a reordering of width of between 2 to 4 words, 5 to 8 and more than 8 words respectively.

Syntactic reorderings can reveal very rich language-specific reordering behaviour. Figure 5 is an example of the kinds of data that can be used to improve reordering models. In this graph we selected the four syntactic types that were involved in the largest number of reorderings. They covered the block that was moved forward in the target (block A). We can see that different syntactic types display quite different behaviour at different reordering widths and this could be important to model.

Having now characterised the space of reordering actually found in parallel data, we now turn to the question of how well our translation models account for them. As both the translation models investigated in this work do not use syntax, in the following sections we focus on non-syntactic analysis.

5 Evaluating Reordering in Translation

We are interested in knowing how current translation models perform specifically with regard to reordering. To evaluate this, we compare the reorderings in the parallel corpora with the reorderings that exist in the translated sentences. We com-

	None	Low	Medium	High
Average RQuantity				
CH-EN	0	0.39	0.82	1.51
AR-EN	0	0.10	0.25	0.57
Number of Sentences				
CH-EN	105	367	367	367
AR-EN	293	379	379	379

Table 1. The RQuantity and the number of sentences for each reordering test set.

pare two state-of-the-art models: the phrase-based system Moses (Koehn et al., 2007) (with lexicalised reordering), and the hierarchical model Hiero (Chiang, 2007). We use default settings for both models: a distortion limit of seven for Moses, and a maximum source span limit of 10 words for Hiero. We trained both models on subsets of the NIST 2008 data sets, consisting mainly of news data, totalling 547,420 CH-EN and 1,069,658 AR-EN sentence pairs. We used a trigram language model on the entire English side (211M words) of the NIST 2008 Chinese-English training corpus. Minimum error rate training was performed on the 2002 NIST test for CH-EN, and the 2004 NIST test set for AR-EN.

5.1 Reordering Test Corpus

In order to determine what effect reordering has on translation, we extract a test corpus with specific reordering characteristics from the manually aligned and parsed sentences described in Section 4. To minimise the impact of sentence length, we select sentences with target lengths from 20 to 39 words inclusive. In this range RQuantity is stable. From these sentences we first remove those with no detected reorderings, and we then divide up the remaining sentences into three sets of equal sizes based on the RQuantity of each sentence. We label these test sets: "none", "low", "medium" and "high".

All test sentences have only one reference English sentence. MT evaluations using one reference cannot make strong claims about any particular test sentence, but are still valid when used to compare large numbers of hypotheses.

Table 1 and Figure 6 show the reordering characteristics of the test sets. As expected, we see more reordering for Chinese-English than for Arabic to English.

It is important to note that although we might name a set "low" or "high", this is only relative to the other groups for the same language pair. The "high" AR-EN set, has a lower RQuantity than the "medium" CH-EN set. Figure 6 shows

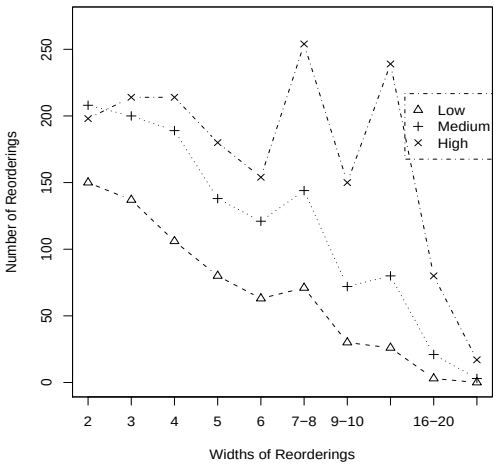

Figure 6. Number of reorderings in the CH-EN test set plotted against the total width of the reorderings.

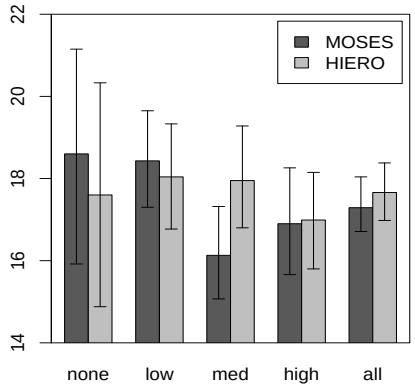

Figure 7. BLEU scores for the different CH-EN reordering test sets and the combination of all the groups for the two translation models.The 95% confidence levels as measured by bootstrap resampling are shown for each bar.

that the CH-EN reorderings in the higher RQuantity groups have more and longer reorderings. The AR-EN sets show similar differences in reordering behaviour.

5.2 Performance on Test Sets

In this section we compare the translation output for the phrase-based and the hierarchical system for different reordering scenarios. We use the test sets created in Section 5.1 to explicitly isolate the effect reordering has on the performance of two translation systems.

Figure 7 and Figure 8 show the BLEU score results of the phrase-based model and the hierarchical model on the different reordering test sets. The 95% confidence intervals as calculated by bootstrap resampling (Koehn, 2004) are shown for each of the results. We can see that the models show quite different behaviour for the different test sets and for the different language pairs. This demonstrates that reordering greatly influences the

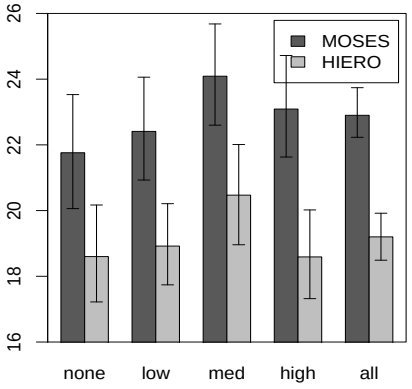

Figure 8. BLEU scores for the different AR-EN reordering test sets and the combination of all the groups for the two translation models. The 95% confidence levels as measured by bootstrap resampling are shown for each bar.

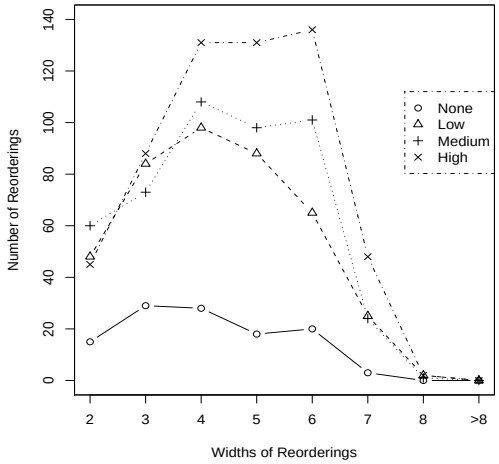

Figure 9. Reorderings in the CH-EN MOSES translation of the reordering test set, plotted against the total width of the reorderings.

BLEU score performance of the systems.

In Figure 7 we see that the hierarchical model performs considerably better than Moses on the "medium" CH-EN set, although the confidence interval for these results overlap somewhat. This supports the claim that Hiero is better able to capture longer distance reorderings than Moses.

Hiero performs significantly worse than Moses on the "none" and "low" sets for CH-EN, and for all the AR-EN sets, other than "none". All these sets have a relatively low amount of reordering, and in particular a low number of medium and long distance reorderings. The phrase-based model could be performing better because it searches all possible permutations within a certain window whereas the hierarchical model will only permit reorderings for which there is lexical evidence in the training corpus. Within a small window, this exhaustive search could discover the best reorderings, but within a bigger window, the more constrained search of the hierarchical model produces better results. It is interesting that Hiero is not always the best choice for translation performance, and depending on the amount of reordering and the distribution of reorderings, the simpler phrase-based approach is better.

The fact that both models show equally poor performance on the "high" RQuantity test set suggests that the hierarchical model has no advantage over the phrase-based model when the reorderings are long enough and frequent enough. Neither Moses nor Hiero can perform long distance reorderings, due to the local constraints placed on their search which allows performance to be linear with respect to sentence length. Increasing the window in which these models are able to perform reorderings does not necessarily improve performance, due to the number of hypotheses the models must discriminate amongst.

The performance of both systems on the "high" test set could be much worse than the BLEU score would suggest. A long distance reordering that has been missed, would only be penalised by BLEU once at the join of the two blocks, even though it might have a serious impact on the comprehension of the translation. This flaw seriously limits the conclusions that we can draw from BLEU score, and motivates analysing translations specifically for reordering as we do in this paper.

Reorderings in Translation

At best, BLEU can only partially reflect the reordering performance of the systems. We therefore perform an analysis of the distribution of reorderings that are present in the systems' outputs, in order to compare them with each other and with the source-reference distribution.

For each hypothesis translation, we record which source words and phrase pairs or rules were used to produce which target words. From this we create an alignment matrix from which reorderings are extracted in the same manner as previously done for the manually aligned corpora.

Figure 9 shows the distribution of reorderings that occur between the source sentence and the translations from the phrase-based model. This graph is interesting when compared with Figure 6, which shows the reorderings that exist in the original reference sentence pair. The two distributions are quite different. Firstly, as the models use phrases which are treated as blocks, reorderings which occur within a phrase are not recorded. This reduces the number of shorter distance reorderings in the distribution in Figure 6, as mainly short

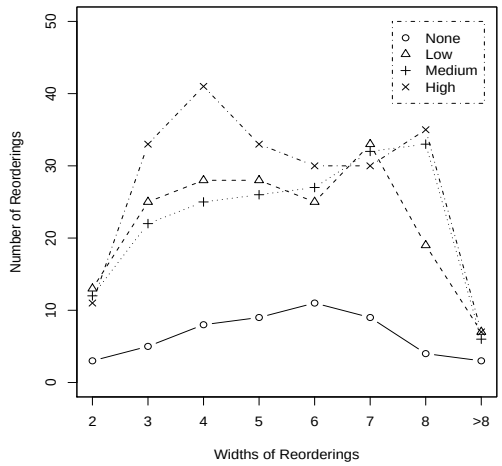

Figure 10. Reorderings in the CH-EN Hiero translation of the reordering test set, plotted against the total width of the reorderings.

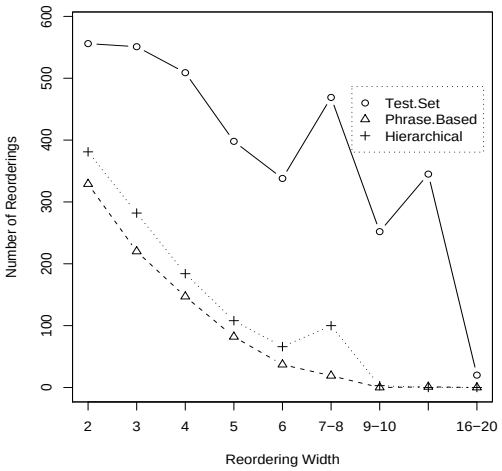

Figure 11. Number of reorderings in the original CH-EN test set, compared to the reorderings retained by the phrase-based and hierarchical models. The data is shown relative to the length of the total source width of the reordering.

phrases pairs are used in the hypothesis. However, even taking reorderings within phrase pairs into account, there are many fewer reorderings in the translations than in the references, and there are no long distance reorderings.

It is interesting that the phrase-based model is able to capture the fact that reordering increases with the RQuantity of the test set. Looking at the equivalent data for the AR-EN language pair, a similar pattern emerges: there are many fewer reorderings in the translations than in the references.

Figure 10 shows the reorderings from the output of the hierarchical model. The results are very different to both the phrase-based model output (Figure 9) and to the original reference reordering distribution (Figure 6). There are fewer reorderings here than even in the phrase-based output. However, the Hiero output has a slightly higher BLEU score than the Moses output. The number of reorderings is clearly not the whole story. Part of the reason why the output seems to have few reorderings and yet scores well, is that the output of hierarchical models does not lend itself to the analysis that we have performed successfully on the reference or phrase-based translation sentence pairs. This is because the output has a large number of non-contiguous phrases which prevent the extraction of reorderings from within their span. Only 4.6% of phrase-based words were blocked off due to non-contiguous phrases but 47.5% of the hierarchical words were. This problem can be ameliorated with the detection and unaligning of words which are obviously dependent on other words in the non-contiguous phrase.

Even taking blocked off phrases into account, however, the number of reorderings in the hierarchical output is still low, especially for the medium and long distance reorderings, as compared to the reference sentences. The hierarchical model's reordering behaviour is very different to human reordering. Even if human translations are freer and contain more reordering than is strictly necessary, many important reorderings are surely being lost.

Targeted Automatic Evaluation

Comparing distributions of reorderings is interesting, but it cannot approach the question of how many reorderings the system performed correctly. In this section we identify individual reorderings in the source and reference sentences and detect whether or not they have been reproduced in the translation.

Each reordering in the original test set is extracted. Then the source-translation alignment is inspected to determine whether the blocks involved in the original reorderings are in the reverse order in the translation. If so, we say that these reorderings have been retained from the reference to the translation.

If a reordering has been translated by one phrase pair, we assume that the reordering has been retained, because the reordering could exist inside the phrase. If the segmentation is slightly different, but a reordering of the correct size occurred at the right place, it is also considered to be retained.

Figure 11 shows that the hierarchical model retains more reorderings of all widths than the phrase-based system. Both systems retain few reorderings, with the phrase-based model missing almost all the medium distance reorderings, and both models failing on all the long distance re-

	Correct	Incorrect	NA
Retained	61	4	10
Not Retained	32	31	12

Table 2. Correlation between retaining reordering and it being correct - for humans and for system

orderings. This is possibly the most direct evidence of reordering performance so far, and again shows how Hiero has a slight advantage over the phrase-based system with regard to reordering performance.

Targeted Manual Analysis

The relationship between targeted evaluation and the correct reordering of the translation still needs to be established. The translation system can compensate for not retaining a reordering by using different lexical items. To judge the relevance of the targeted evaluation we need to perform a manual evaluation. We present evaluators with the reference and the translation sentences. We mark the target ranges of the blocks that are involved in the particular reordering we are analysing, and ask the evaluator if the reordering in the translation is correct, incorrect or not applicable. The not applicable case is chosen when the translated words are so different from the reference that their ordering is irrelevant. There were three evaluators who each judged 25 CH-EN reorderings which were retained and 25 CH-EN reorderings which were not retained by the Moses translation model.

The results in Table 2 show that the retained reorderings are generally judged to be correct. If the reordering is not retained, then the evaluators divided their judgements evenly between the reordering being correct or incorrect. It seems that the fact that a reordering is not retained does indicate that its ordering is more likely to be incorrect. We used Fleiss' Kappa to measure the correlation between annotators. It expresses the extent to which the amount of agreement between raters is greater than what would be expected if all raters made their judgements randomly. In this case Fleiss' kappa is 0.357 which is considered to be a fair correlation.

6 Conclusion

In this paper we have introduced a general and extensible automatic method for the quantitative analyse of syntactic reordering phenomena in parallel corpora.

We have applied our method to a systematic analysis of reordering both in the training corpus, and in the output, of two state-of-the-art translation models. We show that the hierarchical model performs better than the phrase-based model in situations where there are many medium distance reorderings. In addition, we find that the choice of translation model must be guided by the type of reorderings in the language pair, as the phrase-based model outperforms the hierarchical model when there is a predominance of short distance reorderings. However, neither model is able to capture the reordering behaviour of the reference corpora adequately. These result indicate that there is still much research to be done if statistical machine translation systems are to capture the full range of reordering phenomena present in translation.

References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The Hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 779–786, Vancouver, Canada.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics (to appear)*, 33(2).

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 304–311, Philadelphia, USA.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 127–133, Edmonton, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics Companion Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–450.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 161–168, Boston, USA. Association for Computational Linguistics.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the International Conference on Computational Linguistics and of the Association for Computational Linguistics*, pages 977–984, Sydney, Australia.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 144–151, Sapporo, Japan.

Andreas Zollmann, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of International Conference On Computational Linguistics*.