Improving Mid-Range Reordering using Templates of Factors

Hieu Hoang
School of Informatics
University of Edinburgh
h.hoang@sms.ed.ac.uk

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Abstract

We extend the factored translation model (Koehn and Hoang, 2007) to allow translations of longer phrases composed of factors such as POS and morphological tags to act as templates for the selection and reordering of surface phrase translation. We also reintroduce the use of alignment information within the decoder, which forms an integral part of decoding in the Alignment Template System (Och, 2002), into phrase-based decoding.

Results show an increase in translation performance of up to 1.0% BLEU for out-of-domain French–English translation. We also show how this method compares and relates to lexicalized reordering.

1 Introduction

One of the major issues in statistical machine translation is reordering due to systematic word-ordering differences between languages. Often reordering is best explained by linguistic categories, such as part-of-speech tags. In fact, prior work has examined the use of part-of-speech tags in pre-reordering schemes, Tomas and Casacuberta (2003).

Re-ordering can also be viewed as composing of a number of related problems which can be explained or solved by a variety of linguistic phenomena. Firstly, differences between phrase ordering account for much of the long-range reordering. Syntax-based and hierarchical models such as (Chiang, 2005) attempts to address this problem. Shorter range re-ordering, such as intra-phrasal word re-ordering, can often be predicted from the underlying property of the words and its context, the most obvious property being POS tags.

In this paper, we tackle the issue of shorter-range re-ordering in phrase-based decoding by presenting an extension of the factored translation which directly models the translation of non-surface factors such as POS tags. We shall call this extension the *factored template model*. We use the fact that factors such as POS-tags are less sparse than surface words to obtain longer phrase translations. These translations are used to inform the re-ordering of surface phrases.

Despite the ability of phrase-based systems to use multi-word phrases, the majority of phrases used during decoding are one word phrases, which we will show in later sections. Using word translations negates the implicit capability of phrases to re-order words. We show that the proposed extension increases the number of multi-word phrases used during decoding, capturing the implicit ordering with the phrase translation, leading to overall better sentence translation. In our tests, we obtained 1.0% increase in absolute for French-English translation, and 0.8% increase for German-English translation, trained on News Commentary corpora [1].

We will begin by recounting the phrase-based and factored model in Section 2 and describe the language model and lexicalized re-ordering model and the advantages and disadvantages of using these models to influence re-ordering. The proposed model is described in Section 4.

2 Background

Let us first provide some background on phrase-based and factored translation, as well as the use of part-of-speech tags in reordering.

2.1 Phrase-Based Models

Phrase-based statistical machine translation has emerged as the dominant paradigm in machine translation research. We model the translation of a given source language sentence $\mathbf{s}$ into a target language sentence $\mathbf{t}$ with a probability distribution $p(\mathbf{t}|\mathbf{s})$. The goal of translation is to find the best translation according to the model

$$\mathbf{t}_{\text{BEST}} = \text{argmax}_{\mathbf{t}}\ p(\mathbf{t}|\mathbf{s}) \quad (1)$$

The argmax function defines the search objective of the decoder. We estimate $p(\mathbf{t}|\mathbf{s})$ by decom-

[1] http://www.statmt.org/wmt07/shared-task.html

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 372–379,
Athens, Greece, 30 March – 3 April 2009.

posing it into component models

$$p(\mathbf{t}|\mathbf{s}) = \frac{1}{Z} \prod_m h'_m(\mathbf{t}, \mathbf{s})^{\lambda_m} \quad (2)$$

where $h'_m(\mathbf{t}, \mathbf{s})$ is the feature function for component m and λ_m is the weight given to component m. Z is a normalization factor which is ignored in practice. Components are translation model scoring functions, language model, reordering models and other features.

The problem is typically presented in log-space, which simplifies computations, but otherwise does not change the problem due to the monotonicity of the log function ($h_m = \log h'_m$)

$$\log p(\mathbf{t}|\mathbf{s}) = \sum_m \lambda_m \, h_m(\mathbf{t}, \mathbf{s}) \quad (3)$$

Phrase-based models (Koehn et al., 2003) are limited to the mapping of small contiguous chunks of text. In these models, the source sentence $\mathbf{s}$ is segmented into a number of phrases $\bar{s}_k$, which are translated one-to-one into target phrases $\bar{t}_k$. The translation feature functions $h_{\text{TM}}(\mathbf{t}, \mathbf{s})$ are computed as sum of phrase translation feature functions $\bar{h}_{\text{TM}}(\bar{t}_k, \bar{s}_k)$:

$$h_{\text{TM}}(\mathbf{t}, \mathbf{s}) = \sum_k \bar{h}_{\text{TM}}(\bar{t}_k, \bar{s}_k) \quad (4)$$

where $\bar{t}_k$ and $\bar{s}_k$ are the phrases that make up the target and source sentence. Note that typically multiple feature functions for one translation table are used (such as forward and backward probabilities and lexical backoff).

2.2 Reordering in Phrase Models

Phrase-based systems implicitly perform short-range reordering by translating multi-word phrases where the component words may be reordered relative to each other. However, multi-word phrases have to have been seen and learnt from the training corpus. This works better when the parallel corpus is large and the training corpus and input are from the same domain. Otherwise, the ability to apply multi-word phrases is lessened due to data sparsity, and therefore most used phrases are only 1 or 2 words long.

A popular model for phrasal reordering is lexicalized reordering (Tillmann, 2004) which introduces a probability distribution for each phrase pair that indicates the likelihood of being translated monotone, swapped, or placed discontinuous to its previous phrase. However, whether a phrase is reordered may depend on its neighboring phrases, which this model does not take into account. For example, the French phrase *noir* would be reordered if preceded by a noun when translating into English, as in as in *chat noir*, but would remain in the same relative position when preceded by a conjunction such as *rouge et noir*.

The use of language models on the decoding output also has a significant effect on reordering by preferring hypotheses which are more fluent. However, there are a number of disadvantages with this low-order Markov model over consecutive surface words. Firstly, the model has no information about the source and may prefer orderings of target words that are unlikely given the source. Secondly, data sparsity may be a problem, even if language models are trained on a large amount of monolingual data which is easier to obtain than parallel data. When the test set is out-of-domain or rare words are involved, it is likely that the language model backs off to lower order n-grams, thus further reducing the context window.

2.3 POS-Based Reordering

This paper will look at the use of POS tags to condition reordering of phrases which are closely positioned in the source and target, such as intra-clausal reordering, however, we do not explicit segment along clausal boundaries. By mid-range reordering we mean a maximum distortion of about 5 or 6 words.

The phrase-based translation model is generally believed to perform short-range reordering adequately. It outperforms more complex models such as hierarchical translation when the most of the reordering in a particular language pair is reasonably short (Anonymous, 2008), as is the case with Arabic–English. However, phrase-based models can fail to reorder words or phrases which would seem obvious if it had access to the POS tags of the individual words. For example, a translation from French to English will usually correctly reorder the French phrase with POS tags NOUN ADJECTIVE if the surface forms exists in the phrase table or language model, e.g.,

Union Européenne → *European Union*

However, phrase-based models may not reorder even these small two-word phrases if the phrase is not in the training data or involves rare words. This situation worsens for longer phrases where the likelihood of the phrase being previously un-

seen is higher. The following example has a source POS pattern NOUN ADJECTIVE CONJUNCTION ADJECTIVE but is incorrectly ordered as the surface phrase does not occur in training,

difficultés économiques et socials
→ *economic and social difficulties*

However, even if the training data does not contain this particular phrase, it contains many similar phrases with the same underlying POS tags. For example, the correct translation of the corresponding POS tags of the above translation

NOUN ADJ CONJ ADJ
→ ADJ CONJ ADJ NOUN

is typically observed many times in the training corpus.

The alignment information in the training corpus shows exactly how the individual words in this phrase should be distorted, along with the POS tag of the target words. The challenge addressed by this paper is to integrate POS tag phrase translations and alignment information into a phrase-based decoder in order to improve reordering.

2.4 Factor Model Decomposition

Factored translation models (Koehn and Hoang, 2007) extend the phrase-based model by integrating word level factors into the decoding process. Words are represented by vectors of factors, not simple tokens. Factors are user-definable and do not have any specific meaning within the model. Typically, factors are obtained from linguistic tools such as taggers and parsers.

The factored decoding process can be decomposed into multiple steps to fully translate the input. Formally, this decomposes Equation 4 further into sub-component models (also called translation steps)

$$\bar{h}_{\text{TM}}(\bar{t}, \bar{s}) = \sum_i \bar{h}^i_{\text{TM}}(\bar{t}, \bar{s}) \quad (5)$$

with an translation feature function $\bar{h}^i_{\text{TM}}$ for each translation step for each factor (or sets of factors). There may be also generation models which create target factors from other target factors but we exclude this in our presentation for the sake of clarity.

Decomposition is a convenient and flexible method for integrating word level factors into phrase-based decoding, allowing source and target sentences to be augmented with factors, while at the same time controlling data sparsity. However, decomposition also implies certain independence assumptions which may not be justified. Various internal experiments show that decomposition may decrease performance and that better results can often be achieved by simply translating all factors jointly. While we can gain benefit from adding factor information into phrase-based decoding, our experience also shows the shortcomings of decomposing phrase translation.

3 Related Work

Efforts have been made to integrate syntactic information into the decoding process to improve reordering.

Collins et al. (2005) reorder the source sentence using a sequence of six manually-crafted rules, given the syntactic parse tree of the source sentence. While the transformation rules are specific to the German parser that was used, they could be adapted to other languages and parsers. Xia and McCord (2004) automatically create rewrite rules which reorder the source sentence. Zhang and Zens (2007) take a slightly different approach by using chunk level tags to reorder the source sentence, creating a confusion network to represent the possible reorderings of the source sentence. All these approaches seek to improve reordering by making the ordering of the source sentence similar to the target sentence.

Costa-jussà and Fonollosa (2006) use a two stage process to reorder translation in an n-gram based decoder. The first stage uses word classes of source words to reorder the source sentence into a string of word classes which can be translated monotonically to the target sentences in the second stage.

The Alignment Template System (Och, 2002) performs reordering by translating word classes with their corresponding alignment information, then translates each surface word to be consistent with the alignment. Tomas and Casacuberta (2003) extend ATS by using POS tags instead of automatically induced word classes.

Note the limitation of the existing work of POS-driven reordering in phrase-based models: the reordering model is separated from the translation model and the two steps are pipelined, with passing the 1-best reordering or at most a lattice to the translation stage. The ATS models do provide an integrated approach, but their lexical translation is

limited to the word level.

In contrast to prior work, we present a integrated approach that allows POS-based reordering and phrase translation. It is also open to the use of any other factors, such as driving reordering with automatic word classes.

Our proposed solution is similar to structural templates described in Phillips (2007) which was applied to an example-based MT system.

4 Translation Using Templates of Factors

A major motivation for the introduction of factors into machine translation is to generalize phrase translation over longer segments using less sparse factors than is possible with surface forms. (Koehn and Hoang, 2007) describes various strategies for the decomposition of the decoding into multiple translation models using the Moses decoder. We shall focus on POS-tags as an example of a less-sparsed factor.

Decomposing the translation by separately decoding the POS tags and surface forms is be the obvious option, which also has a probabilistic interpretation. However, this combined factors into target words which don't exist naturally and bring down translation quality. Therefore, the decoding is constrained by decomposing into two translation models; a model with POS-tag phrase pairs only and one which jointly translates POS-tags and surface forms. This can be expressed using feature-functions

$$\bar{h}_{\text{TM}}(\bar{t},\bar{s}) = \bar{h}^{pos}_{\text{TM}}(\bar{t},\bar{s})\bar{h}^{surface}_{\text{TM}}(\bar{t},\bar{s}) \quad (6)$$

Source segment must be decoded by both translation models but only phrase pairs where the overlapping factors are the same are used. As an additional constraint, the alignment information is retained in the translation model from the training data for every phrase pair, and both translation models must produce consistent alignments. This is expressed formally in Equation 7 to 9.

An alignment is a relationship which maps a source word at position i to a target word at position j:

$$a : i \rightarrow j \quad (7)$$

Each word at each position can be aligned to multiple words, therefore, we alter the alignment relation to express this explicitly:

$$a : i \rightarrow j \quad (8)$$

where J is the set of positions, $j \epsilon J$, that I is aligned to in the other language. Phrase pairs for each translation model are used only if they can satisfy condition 9 for each position of every source word covered.

$$\forall a, b \, \epsilon \, T \quad \forall p : J^p_a J^p_b \neq \emptyset \quad (9)$$

where J^p_a is the alignment information for translation model, a, at word position, p and T is the set of translation models.

4.1 Training

The training procedure is identical to the factored phrase-based training described in (Koehn and Hoang, 2007). The phrase model retains the word alignment information found during training. Where multiple alignment exists in the training data for a particular phrase pair, the most frequent is used, in a similar manner to the calculation of the lexicalized probabilities.

Words positions which remain unaligned are artificially aligned to every word in the other language in the phrase translation during decoding to allow the decoder to cover the position.

4.2 Decoding

The beam search decoding algorithm is unchanged from traditional phrase-based and factored decoding. However, the creation of translation options is extended to include the use of factored templates. Translation options are the intermediate representation between the phrase pairs from the translation models and the hypotheses in the stack decoder which cover specific source spans of a sentence and are applied to hypotheses to create new hypotheses.

In phrase-based decoding, a translation option strictly contains one phrase pair. In factored decoding, strictly one phrase pair from each translation model is used to create a translation options. This is possible only when the segmentation is identical for both source and target span of each phrase pair in each translation model. However, this constraint limits the ability to use long POS-tag phrase pairs in conjunction with shorter surface phrase pairs.

The factored template approach extend factored decoding by constructing translation options from a single phrase pair from the POS-tag translation model, but allowing multiple phrase pairs from

other translation models. A simplified stack decoder is used to compose phrases from the other translation models. This so called intra-phrase decoder is constrained to creating phrases which adheres to the constraint described in Section 4. The intra-phrase decoder uses the same feature functions as the main beam decoder but uses a larger stack size due to the difficulty of creating completed phrases which satisfy the constraint. Every source position must be covered by every translation model.

The intra-phrase decoder is used for each contiguous span in the input sentence to produce translation options which are then applied as usual by the main decoder.

5 Experiments

We performed our experiments on the news commentary corpus[2] which contains 60,000 parallel sentences for German–English and 43,000 sentences for French–English. Tuning was done on a 2000 sentence subset of the Europarl corpus (Koehn, 2005) and tested on a 2000 sentence Europarl subset for out-of-domain, and a 1064 news commentary sentences for in-domain.

The training corpus is aligned using Giza++ (Och and Ney, 2003). To create POS tag translation models, the surface forms on both source and target language training data are replaced with POS tags before phrases are extracted. The taggers used were the Brill Tagger (Brill, 1995) for English, the Treetagger for French (Schmid, 1994), and the LoPar Tagger (Schmidt and Schulte im Walde, 2000) for German. The training script supplied with the Moses toolkit (Koehn et al., 2007) was used, extended to enable alignment information of each phrase pair. The vanilla Moses MERT tuning script was used throughout.

Results are also presented for models trained on the larger Europarl corpora[3].

5.1 German–English

We use as a baseline the traditional, non-factored phrase model which obtained a BLEU score of 14.6% on the out-of-domain test set and 18.2% on the in-domain test set (see Table 1, line 1).

POS tags for both source and target languages were augmented to the training corpus and used in the decoding and an additional trigram language model was used on the target POS tags. This increased translation performance (line 2). This model has the same input and output factors, and the same language models, as the factored model we will present shortly and it therefore offers a fairer comparison of the factored template model than the non-factored baseline.

#	Model	out-domain	in-domain
1	Unfactored	14.6	18.2
2	Joint factors	15.0	18.8
3	Factored template	15.3	18.8

Table 1: German–English results, in %BLEU

#	Model	out-domain	in-domain
1	Unfactored	19.6	23.1
2	Joint factors	19.8	23.0
3	Factored template	20.6	24.1

Table 2: French–English results

The factored template model (line 3) outperforms the baseline on both sets and the joint factor model on the out-of-domain set.

However, we believe the language pair German–English is not particularly suited for the factored template approach as many of the short-range ordering properties of German and English are similar. For example, ADJECTIVE NOUN phrases are ordered the same in both languages.

5.2 French–English

Repeating the same experiments for French–English produces bigger gains for the factored template model. See Table 4 for details. Using the factored template model produces the best result, with gains of 1.0 %BLEU over the unfactored baseline on both test sets. It also outperforms the joint factor model.

5.3 Maximum Size of Templates

Typical phrase-based model implementation use a maximum phrase length of 7 but such long phrases are rarely used. Long templates over POS may be more valuable. The factored template models were retrained with increased maximum phrase length but this made no difference or negatively impacted translation performance, Figure 1.

However, using larger phrase lengths over 5 words does not increase translation performance,

[2] http://www.statmt.org/wmt07/shared-task.html

[3] http://www.statmt.org/europarl/

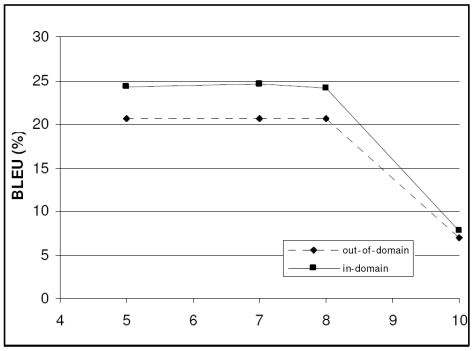

Figure 1: Varying max phrase length

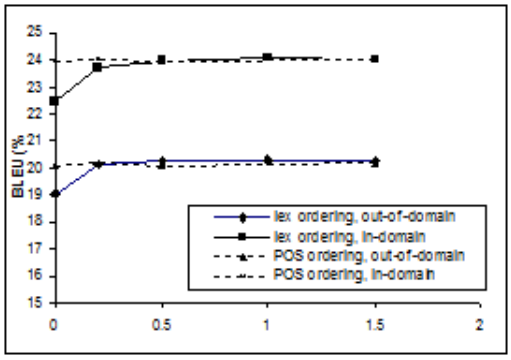

Figure 2: Effect of smoothing on lexicalized reordering

as had been expected. Translation is largely unaffected until the maximum phrase length reaches 10 when performance drops dramatically. This results suggested that the model is limited to mid-range reordering.

6 Lexicalized Reordering Models

There has been considerable effort to improve reordering in phrase-based systems. One of the most well known is the lexicalized reordering model (Tillmann, 2004).

The model uses the same word alignment that is used for phrase table construction to calculate the probability that a phrase is reordered, relative to the previous and next source phrase.

6.1 Smoothing

Tillmann (2004) proposes a block orientation model, where phrase translation and reordering orientation is predicted by the same probability distribution $p(o, \bar{s}|\bar{t})$. The variant of this implemented in Moses uses a separate phrase translation model $p(\bar{s}|\bar{t})$ and lexicalized reordering model $p(o|\bar{s}, \bar{t})$

The parameters for the lexicalized reordering model are calculated using maximum likelihood with a smoothing value α

$$p(o|\bar{s}, \bar{t}) = \frac{\text{count}(o, \bar{s}, \bar{t}) + \alpha}{\sum_{o'}(\text{count}(o, \bar{s}, \bar{t}) + \alpha)} \quad (10)$$

where the predicted orientation o is either monotonic, swap or discontinuous.

The effect of smoothing lexical reordering tables on translation is negligible for both surface forms and POS tags, except when smoothing is disabled (α=0). Then, performance decreases markedly, see Figure 2 for details. Note that the un-smoothed setting is closer to the block orientation model by Tillmann (2004).

#	Model	out-domain	in-domain
1	Unfactored	19.6	23.1
1a	+ word LR	20.2	24.0
2	Joint factors	19.8	23.0
2a	+ POS LR	20.1	24.0
2b	+ POS LR + word LR	20.3	24.1
3	Factored template	20.6	24.1
3a	+ POS LR	20.6	24.3

Table 3: Extending the models with lexicalized reordering (LR)

6.2 Factors and Lexicalized Reordering

The model can easily be extended to take advantage of the factored approach available in Moses. In addition to the lexicalized reordering model trained on surface forms (see line 1a in Table 3), we also conducted various experiments with the lexicalized reordering model for comparison.

In the joint factored model, we have both surface forms and POS tags available to train the lexicalized reordering models on. The lexicalized reordering model can be trained on the surface form, the POS tags, jointly on both factors, or independent models can be trained on each factor. It can be seen from Table 3 that generalizing the reordering model on POS tags (line 2a) improves performance, compared to the non-lexicalized reordering model (line 2). However, this performance does not improve over the lexicalized reordering model on surface forms (line 1a). The surface and POS tag models complement each other to give an overall better BLEU score (line 2b).

In the factored template model, we add a POS-

based lexicalized reordering model on the level of the templates (line 3a). This gives overall the best performance. However, the use of lexicalized reordering models in the factored template model only shows improvements in the in-domain test set.

Lexicalized reordering model on POS tags in factored models underperforms factored template model as the latter includes a larger context of the source and target POS tag sequence, while the former is limited to the extent of the surface word phrase.

7 Analysis

A simple POS sequence that phrase-based systems often fail to reorder is the French–English

NOUN ADJ → ADJ NOUN

We analyzed a random sample of such phrases from the out-of-domain corpus. The baseline system correctly reorders 58% of translations. Adding a lexicalized reordering model or the factored template significantly improves the reordering to above 70% (Figure 3).

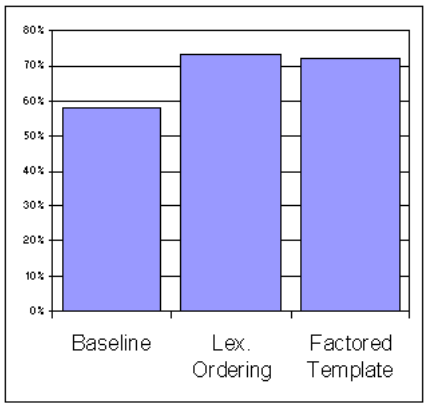

Figure 3: Percentage of correctly ordered NOUN ADJ phrases (100 samples)

A more challenging phrase to translate, such as

NOUN ADJ CONJ ADJ → ADJ CONJ ADJ NOUN

was judge in the same way and the results show the variance between the lexicalized reordering and factored template model (Figure 4).

The factored template model successfully uses POS tag templates to enable longer phrases to be used in decoding. It can be seen from Figure 5, that the majority of input sentence is decoded word-by-word even in a phrase-based system. However, the factored template configuration contains more longer phrases which enhances mid-range reordering.

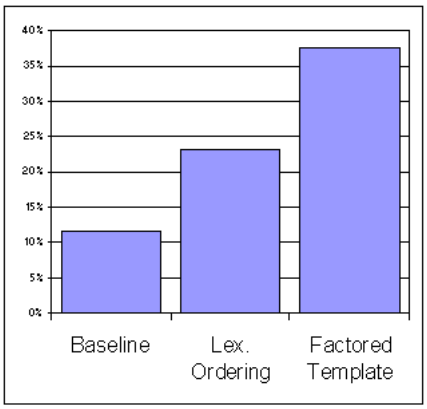

Figure 4: Percentage of correctly ordered NOUN ADJ CONJ ADJ phrases (69 samples)

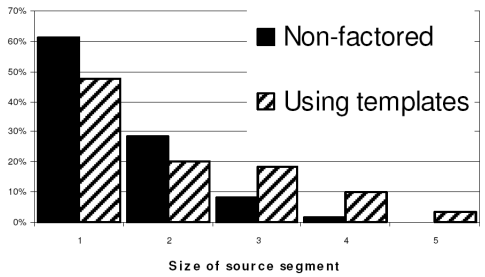

Figure 5: Length of source segmentation when decoding out-of-domain test set

8 Larger training corpora

It is informative to compare the relative performance of the factored template model when trained with more data. We therefore used the Europarl corpora to train and tuning the models for French to English translation. The BLEU scores are shown below, showing no significant advantage to adding POS tags or using the factored template model. This result is similar to many others which have shown that the large amounts of additional data negates the improvements from better models.

#	Model	out-domain	in-domain
1	Unfactored	31.8	32.2
2	Joint factors	31.6	32.0
3	Factored template	31.7	32.2

Table 4: French–English results, trained on Europarl corpus

9 Conclusion

We have shown the limitations of the current factored decoding model which restrict the use of long phrase translations of less-sparsed factors. This negates the effectiveness of decomposing the translation process, dragging down translation quality.

An extension to the factored model was implemented which showed that using POS tag translations to create templates for surface word translations can create longer phrase translation and lead to higher performance, dependent on language pair.

For French–English translation, we obtained a 1.0% BLEU increase on the out-of-domain and in-domain test sets, over the non-factored baseline. The increase was also 0.4%/0.3% when using a lexicalized reordering model in both cases.

In future work, we would like to apply the factored template model to reorder longer phrases. We believe that this approach has the potential for longer range reordering which has not yet been realized in this paper. It also has some similarity to example-based machine translation (Nagao, 1984) which we would like to draw experience from.

We would also be interested in applying this to other language pairs and using factor types other than POS tags, such as syntactic chunk labels or automatically clustered word classes.

Acknowledgments

This work was supported by the EuroMatrix project funded by the European Commission (6th Framework Programme) and made use of the resources provided by the Edinburgh Compute and Data Facility (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk/`).

References

Anonymous (2008). Understanding reordering in statistical machine translation. In *(submitted for publucation)*.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4).

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Costa-jussà, M. R. and Fonollosa, J. A. R. (2006). Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 70–76, Sydney, Australia. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

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., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In *Proceedings of Artificial and Human Intelligence*.

Och, F. J. (2002). *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Phillips, A. B. (2007). Sub-phrasal matching and structural templates in example-based mt. In *Theoretical and Methodological Issues in Machine Translation*, Prague, Czech Republic.

Schmid, H. (1994). Probabilistic part-of-speech tagger using decision trees. In *International Conference on New methods in Language Processing*.

Schmidt, H. and Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Tillmann, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Tomas, J. and Casacuberta, F. (2003). Combining phrase-based and template-based alignment models in statistical translation. In *IbPRIA*.

Xia, F. and McCord, M. (2004). Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland. COLING.

Zhang, Y. and Zens, R. (2007). Improved chunk-level reordering for statistical machine translation. In *International Workshop on Spoken Language Translation*.