2.2: Refined Factored Translation Model

Hieu Hoang, Philipp Koehn, Abhishek Arun, Barry Haddow

Distribution: Final

EuroMatrix
Statistical and Hybrid Machine Translation
Between All European Languages
IST 034291 Deliverable 2.2

February 27, 2009

Project funded by the European Community
under the Sixth Framework Programme for
Research and Technological Development.

Information Society
Technologies

Project ref no.	IST-034291
Project acronym	EuroMatrix
Project full title	Statistical and Hybrid Machine Translation Between All European Languages
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 September 2006 / 30 Months

Distribution	Final
Contractual date of delivery	March 1, 2009
Actual date of delivery	March 1, 2009
Deliverable number	2.2
Deliverable title	Refined Factored Translation Model
Type	Report
Status & version	
Number of pages	20
Contributing WP(s)	WP2
WP / Task responsible	Task 2.3, 2.6
Other contributors	
Author(s)	Hieu Hoang, Philipp Koehn, Abhishek Arun, Barry Haddow
EC project officer	Xavier Gros
Keywords	

The partners in EuroMatrix are:
Saarland University (USAAR)
University of Edinburgh (UEDIN)
Charles University (CUNI-MFF)
CELCT
GROUP Technologies
MorphoLogic

For copies of reports, updates on project activities and other EuroMatrix-related information, contact:
The EuroMatrix Project Co-ordinator
Prof. Hans Uszkoreit
Universität des Saarlandes, Computerlinguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
uszkoreit@coli.uni-sb.de
Phone +49 (681) 302-4115- Fax +49 (681) 302-4700

Copies of reports and other material can also be accessed via the project's homepage:
http://www.euromatrix.net/

© 2009, The Individual Authors
No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.

This deliverable addresses the tasks:

- **Task 2.3**: Exploit factored translation models for reordering
- **Task 2.6**: Maintain and develop a highly competitive machine translation system

The deliverable consists of three papers,
published at academic conferences and workshops:

- **Improving Mid-Range Re-Ordering using Templates of Factors**, *Hieu Hoang and Philipp Koehn*, EACL 2009
- **Edinburghs Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses**, *Philipp Koehn and Barry Haddow*, EACL Workshop on Statistical Machine Translation 2009
- **Towards better Machine Translation Quality for the German-English Language Pairs**, *Philipp Koehn, Abhishek Arun and Hieu Hoang*, ACL Workshop on Statistical Machine Translation 2008

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

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/

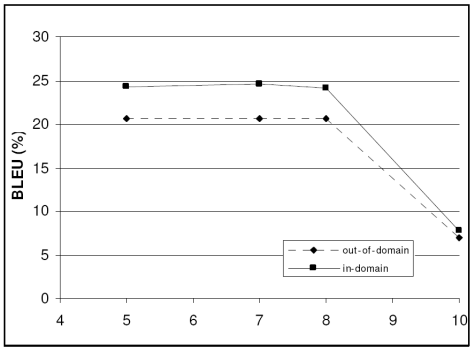


Figure 1: Varying max phrase length

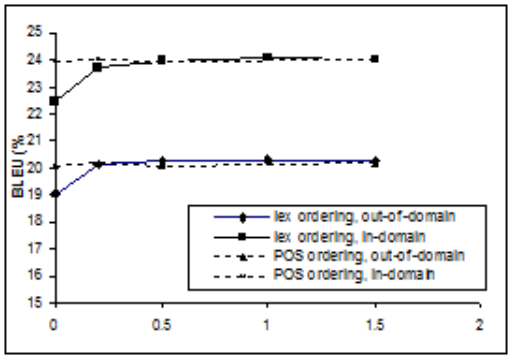


Figure 2: Effect of smoothing on lexicalized reordering

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).

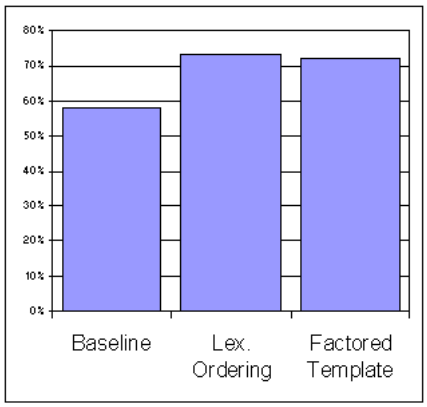


Figure 3: Percentage of correctly ordered NOUN ADJ phrases (100 samples)

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.

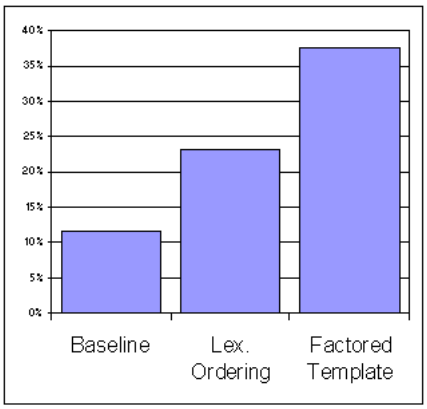


Figure 4: Percentage of correctly ordered NOUN ADJ CONJ ADJ phrases (69 samples)

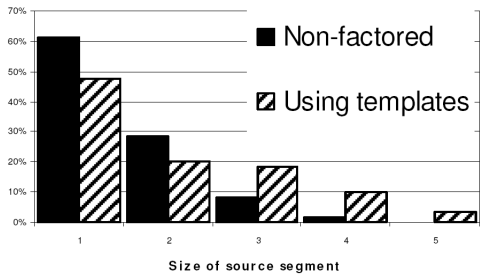


Figure 5: Length of source segmentation when decoding out-of-domain test set

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*.

Towards better Machine Translation Quality for the German–English Language Pairs

Philipp Koehn Abhishek Arun Hieu Hoang
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk a.arun@sms.ed.ac.uk h.hoang@sms.ed.ac.uk

Abstract

The Edinburgh submissions to the shared task of the Third Workshop on Statistical Machine Translation (WMT-2008) incorporate recent advances to the open source Moses system. We made a special effort on the German–English and English–German language pairs, leading to substantial improvements.

1 Introduction

Edinburgh University participated in the shared task of the Third Workshop on Statistical Machine Translation (WMT-2008), which is partly funded by the EUROMATRIX project, which also funds our work. In this project, we set out to build machine translation systems for all language pairs of official EU languages. Hence, we also participated in the shared task in all language pairs.

For all language pairs, we used the Moses decoder (Koehn et al., 2007), which follows the phrase-based statistical machine translation approach (Koehn et al., 2003), with default settings as a starting point. We recently added minimum Bayes risk decoding and reordering constraints to the decoder. We achieved consistent increase in BLEU scores with these improvements, showing gains of up to 0.9% BLEU on the 2008 news test set.

Most of our efforts were focused on the language pairs German–English and English–German. For both language pairs, we explored language-specific and more general improvements, resulting in gains of up to 1.5% BLEU for German–English and 1.4% BLEU for English–German.

2 Recent Improvements

Over the last months, we added minimum Bayes risk decoding and additional reordering constraints to the Moses decoder. The WMT-2008 shared task offered the opportunity to assess these components over a large range of language pairs and tasks.

For all our experiments, we trained solely on the Europarl corpus, which allowed us to treat the 2007 news commentary test set (nc-test2007) as a stand-in for the 2008 news test set (news-2008), for which we have no in-domain training data. This may have resulted in lower performance due to less (and very relevant) training data, but it also allowed us to optimize for a true out-of-domain test set.

The baseline training uses Moses default parameters. We use a maximum sentence length of 80, a phrase translation table with the five traditional features, lexicalized reordering, and lowercase training and test data. All reported BLEU scores are not case-sensitive, computed using the NIST tool.

2.1 Minimum Bayes Risk Decoding

Minimum Bayes risk decoding was proposed by Kumar and Byrne (2004). Instead of selecting the translation with the highest probability, minimum Bayes risk decoding selects the translation that is most similar to the highest scoring translations. Intuitively, this avoid the selection of an outlier as the best translation, since the decision rule prefers translations that are similar to other high-scoring translations.

Minimum Bayes risk decoding is defined as:

$$\mathbf{e}_{\text{MBR}} = \text{argmax}_{\mathbf{e}} \sum_{\mathbf{e}'} L(\mathbf{e}, \mathbf{e}')\ p(\mathbf{e}'|\mathbf{f})$$

As similarity function L, we use sentence-level BLEU with add-one smoothing. As highest scoring translations, we consider the top 100 distinct translations, for which we convert the translation scores into a probability distribution p (with a scaling factor of 1). We tried other n-best list sizes and scaling factors, with very similar outcomes.

Language Pair	Baseline	MBR	MP	MBR+MP
Spanish–German news	11.7	11.8 (+0.1)	11.9 (+0.2)	12.0 (+0.3)
Spanish–German ep	20.7	21.0 (+0.3)	20.8 (+0.1)	21.0 (+0.3)
German–Spanish news	16.2	16.3 (+0.1)	16.4 (+0.2)	16.6 (+0.4)
German–Spanish ep	28.5	28.6 (+0.1)	28.5 (±0.0)	28.6 (+0.1)
Spanish–English news	19.8	20.2 (+0.4)	20.2 (+0.4)	20.3 (+0.5)
Spanish–English ep	33.6	33.7 (+0.1)	33.6 (±0.0)	33.7 (+0.1)
English–Spanish news	20.1	20.5 (+0.4)	20.5 (+0.4)	20.7 (+0.6)
English–Spanish ep	33.1	33.1 (±0.0)	33.0 (–0.1)	33.1 (±0.0)
French–English news	18.5	19.1 (+0.6)	19.1 (+0.6)	19.2 (+0.7)
French–English ep	33.5	33.5 (±0.0)	33.4 (–0.1)	33.5 (±0.0)
English–French news	17.8	18.0 (+0.2)	18.2 (+0.4)	18.3 (+0.5)
English–French ep	31.1	31.1 (±0.0)	31.1 (±0.0)	31.1 (±0.0)
Czech–English news	14.2	14.4 (+0.2)	14.3 (+0.1)	14.5 (+0.3)
Czech–English nc	22.8	23.0 (+0.2)	22.9 (+0.2)	23.0 (+0.2)
English–Czech news	9.6	9.6 (±0.0)	9.7 (+0.1)	9.6 (±0.0)
English–Czech nc	12.9	13.0 (+0.1)	12.9 (±0.0)	13.0 (+0.1)
Hungarian–English news	7.9	8.3 (+0.4)	8.5 (+0.6)	8.8 (+0.9)
English–Hungarian news	6.1	6.3 (+0.2)	6.4 (+0.3)	6.5 (+0.4)
average news	-	+0.26	+0.33	+0.46
average ep	-	+0.08	–0.02	+0.08

Table 1: Improvements in BLEU on the test sets test2008 (ep), newstest2008 (news) and nc-test2008 (nc) for minimum Bayes risk decoding (MBR) and the monotone-at-punctuation reordering (MP) constraint.

2.2 Monotone at Punctuation

The reordering models in phrase-based translation systems are known to be weak, since they essentially relies on the interplay of language model, a general preference for monotone translation, and (in the case of lexicalized reordering) a local model based on a window of neighboring phrase translations. Allowing any kind of reordering typically reduces translation performance, so reordering is limited to a window of (in our case) six words.

One noticeable weakness is that the current model frequently reorders words beyond clause boundaries, which is almost never well-motivated, and leads to confusing translations. Since clause boundaries are often indicated by punctuation such as comma, colon, or semicolon, it is straight-forward to introduce a reordering constraint that addresses this problem.

Our implementation of a monotone-at-punctuation reordering constraint (Tillmann and Ney, 2003) requires that all input words before clause-separating punctuation have be translated, before words afterwards are covered. Note that this constraint does not limit in any way phrase translations that span punctuation.

2.3 Results

Table 1 summarizes the impact of minimum Bayes risk decoding (MBR) and the monotone-at-punctuation reordering constraint (MP). Scores show higher gains for out-of-domain news test sets (+0.46) than for in-domain Europarl sets (+0.08).

3 German–English

Translating between German and English is surprisingly difficult, given that the languages are closely related. The main sources for this difficulty is the different syntactic structure at the clause level and the rich German morphology, including the merging of noun compounds.

In prior work, we addressed **reordering** with a pre-order model that transforms German for training and testing according to a set of hand-crafted rules (Collins et al., 2005). Employing this method to our baseline system leads to an improvement of +0.8 BLEU on the nc-test2007 set and +0.5 BLEU on the test2007 set.

German–English	nc-test2007	test2007
baseline	20.3	27.6
tokenize hyphens	20.1 (–0.2)	27.6 (±0.0)
tok. hyph. + truecase	20.7 (+0.4)	27.8 (+0.2)

Table 2: Impact of truecasing on case-sensitive BLEU

In a more integrated approach, factored translation models (Koehn and Hoang, 2007) allow us to consider grammatical coherence in form of **part-of-speech language models**. When translating into output words, we also generate a part-of-speech tag along with each output word. Since there are only 46 POS tags in English, we are able to train high-order n-gram models of these sequences. In our experiments, we used a 7-gram model, yielding improvements of +0.2/–0.1. We obtained the POS tags using Brill's tagger (Brill, 1995).

Next, we considered the problem of unknown input words, which is partly due to hyphenated words, noun compounds, and morphological variants. Using the baseline model, 907 words (1.78%) in nc-test2007 and 262 (0.47%) in test2007 are unknown. First we separate our **hyphens** by tokenizing words such as *high-risk* into *high @-@ risk*. This reduces the number of unknown words to 791/224. Unfortunately, it hurts us in terms of BLEU (–0.1/–0.1). Second, we **split compounds** using the frequency-based method (Koehn and Knight, 2003), reducing the number of unknown words to than half, 424/94, improving BLEU on nc-test2007 (+0.5/–0.2).

A final modification to the data preparation is **truecasing**. Traditionally, we lowercase all training and test data, but especially in German, case marks important distinctions. German nouns are capitalized, and keeping case allows us to make the distinction between, say, the noun *Wissen* (*knowledge*) and the verb *wissen* (*to know*). By truecasing, we only change the case of the first word of a sentence to its most common form. This method still needs some refinements, such as the handling of headlines or all-caps text, but it did improve performance over the hyphen-tokenized baseline (+0.3/+0.2) and the original baseline (+0.2/+0.1).

Note that truecasing simplifies the recasing problem, so a better way to gauge its effect is to look at the case-sensitive BLEU score. Here the difference are slightly larger over both the hyphen-tokenized baseline (+0.6/+0.2) and the original baseline (+0.4/+0.2). See the Table 2 for details.

German–English	nc-test2007	test2007
baseline	21.3	28.4
pos lm	21.5 (+0.2)	28.3 (–0.1)
reorder	22.1 (+0.8)	28.9 (+0.5)
tokenize hyphens	21.2 (–0.1)	28.3 (–0.1)
tok. hyph. + split	21.8 (+0.5)	28.2 (–0.2)
tok. hyph. + truecase	21.5 (+0.2)	28.5 (+0.1)
mp	21.6 (+0.3)	28.2 (–0.2)
mbr	21.4 (+0.1)	28.3 (–0.1)
big beam	21.3 (±0.0)	28.3 (–0.1)

Table 3: Impact of individual modifications for German–English, measured in BLEU on the development sets

German–English	nc-test2007	test2007
baseline	21.3	28.4
+ reorder	22.1 (+0.8)	28.9 (+0.5)
+ tokenize hyphens	22.1 (+0.8)	28.9 (+0.5)
+ truecase	22.7 (+1.3)	28.9 (+0.5)
+ split	23.0 (+1.7)	29.1 (+0.7)
+ mbr	23.1 (+1.8)	29.3 (+0.9)
+ mp	23.3 (+2.0)	29.2 (+0.8)

Table 4: Impact of combined modifications for German–English, measured in BLEU on the development sets

As for the other language pairs, using the **monotone-at-punctuation** reordering constraint (+0.3/–0.2) and **minimum Bayes risk decoding** (+0.1/–0.1) mostly helps. We also tried **bigger beam** sizes (stack size 1000, phrase table limit 50), but without gains in BLEU (±0.0/–0.1).

Table 3 summarizes the contributions of the individual modifications we described above. For our final system, we added the improvements one by one (see Table 4), except for the bigger beam size and the POS language model. This led to an overall increase of +2.0/+0.8 over the baseline. Due to a bug in splitting, the system we submitted to the shared task had a score of only +1.5/+0.6 over the baseline.

4 English–German

For English–German, we applied many of the same methods as for the inverse language pair. Tokenizing out **hyphens** has questionable impact (–0.1/+0.1), while **truecasing** shows minor gains (±0.0/+0.1), slightly higher for case-sensitive scoring (+0.2/+0.3). We have not yet developed a method that is the analog of the compound splitting

English–German	nc-test2007	test-2007
baseline	14.6	21.0
tokenize hyphens	14.5 (–0.1)	21.1 (+0.1)
tok. hyph. + truecase	14.6 (±0.0)	21.1 (+0.1)
morph lm	15.7 (+1.1)	21.2 (+0.2)
mbr	14.9 (+0.3)	21.0 (±0.0)
mp	14.8 (+0.2)	20.9 (–0.1)
big beam	14.7 (+0.1)	21.0 (±0.0)

Table 5: Impact of individual modifications for English–German, measured in BLEU on the development sets

method — compound merging. We consider this an interesting challenge for future work.

While the rich German morphology on the source side mostly poses sparse data problems, on the target side it creates the problem of which morphological variant to choose. The right selection hinges on grammatical agreement within noun phrases, the role that each noun phrase plays in the clause, and the grammatical nature of the subject of a verb. We use LoPar (Schmidt and Schulte im Walde, 2000), which gives us **morphological features** such as case, gender, count, although in limited form, it often opts for more general categories such as *not genitive*. We include these features in a sequence model, as we used a sequence model over part-of-speech tags previously. The gains of this method are especially strong for the out-of-domain set (+1.1/+0.2).

Minimum Bayes risk decoding (+0.3/±0.0), the **monotone-at-punctuation** reordering constraint (+0.2/–0.1), and **bigger beam sizes** (+0.1/±0.0) have similar impact as for the other language pairs. See Table 5 for a summary of all modifications. By combining everything except for the bigger beam size, we obtain overall gains of +1.4/+0.4 over the baseline. For details, refer to Table 6.

5 Conclusions

We built Moses systems trained on either only Europarl data or, for Czech and Hungarian, the available training data. We showed gains with minimum Bayes risk decoding and a reordering constraint involving punctuation. For German↔English, we employed further language-specific improvements.

Acknowledgements: This work was supported in part under the EuroMatrix project funded by the European Commission (6th Framework Programme).

English–German	nc-test2007	test2007
baseline	14.6	21.0
+ tokenize hyphens	14.5 (–0.1)	21.1 (+0.1)
+ truecase	14.6 (±0.0)	21.1 (+0.1)
+ morph lm	15.4 (+0.8)	21.3 (+0.3)
+ mbr	15.7 (+1.1)	21.4 (+0.4)
+ mp	16.0 (+1.4)	21.4 (+0.4)

Table 6: Impact of combined modifications for English–German, measured in BLEU on the development sets

References

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

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.

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. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

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)*.

Kumar, S. and Byrne, W. (2004). Minimum bayes-risk decoding 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)*.

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. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1).

Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses

Philipp Koehn and Barry Haddow
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk bhaddow@inf.ed.ac.uk

Abstract

Edinburgh University participated in the WMT 2009 shared task using the Moses phrase-based statistical machine translation decoder, building systems for all language pairs. The system configuration was identical for all language pairs (with a few additional components for the German-English language pairs). This paper describes the configuration of the systems, plus novel contributions to Moses including truecasing, more efficient decoding methods, and a framework to specify reordering constraints.

1 Introduction

The commitment of the University of Edinburgh to the WMT shared tasks is to provide a strong statistical machine translation baseline with our open source tools for all language pairs. We are again the only institution that participated in all tracks.

The shared task is also an opportunity to incorporate novel contributions and test them against the best machine translation systems for these language pairs. In this paper we describe the speed improvements to the Moses decoder (Koehn et al., 2007), as well as a novel framework to specify reordering constraints with XML markup, which we tested with punctuation-based constraints.

2 System Configuration

We trained a default Moses system with the following non-default settings:

- maximum sentence length 80
- grow-diag-final-and symmetrization of GIZA++ alignments
- interpolated Kneser-Ney discounted 5-gram language model
- msd-bidirectional-fe lexicalized reordering

Language	ep	nc	news	intpl.
English	449	486	216	192
French	264	311	147	131
German	785	821	449	402
Spanish	341	392	219	190
Czech	*:1475	1615	752	690
Hungarian	hung:2148		815	786

Table 1: Perplexity (ppl) of the domain-trained (ep = Europarl (CzEng for Czech), nc = News Commentary, news = News) and interpolated language models.

2.1 Domain Adaptation

In contrast to last year's task, where news translation was presented as a true out-of-domain problem, this year large monolingual news corpora and a tuning set (last year's test set) were provided. While still no in-domain news parallel corpora were made available, the monolingual corpora could be exploited for domain adaption.

For all language pairs, we built a 5-gram language model, by first training separate language models for the different training corpora (the parallel Europarl and News Commentary and new monolingual news), and then interpolated them by optimizing perplexity on the provided tuning set. Perplexity numbers are shown in Table 1.

2.2 Truecasing

Our traditional method to handle case is to lowercase all training data, and then have a separate recasing (or recapitalization) step. Last year, we used truecasing: all words are normalized to their natural case, e.g. *the, John, eBay*, meaning that only sentence-leading words may be changed to their most frequent form.

To refine last year's approach, we record the seen truecased instances and truecase words in test sentences (even in the middle of sentences) to seen forms, if possible.

Truecasing leads to small degradation in case-

language pair		baseline	w/ news	mbr/mp	truecased	big beam	ued'08	best'08
French-English	uncased	21.2	23.1	23.3	22.7	22.9	19.2	21.9
	cased			21.7	21.6	21.8		
English-French	uncased	17.8	19.4	19.6	19.6	19.7	18.2	21.4
	cased			18.1	18.7	18.8		
Spanish-English	uncased	22.5	24.4	24.7	24.5	24.7	20.1	22.9
	cased			23.0	23.3	23.4		
English-Spanish	uncased	22.4	23.9	24.2	23.8	24.4	20.7	22.7
	cased			22.1	22.8	23.1		
Czech-English	uncased	16.9	18.9	18.9	18.6	18.6	14.5	14.7
	cased			17.3	17.4	17.4		
English-Czech	uncased	11.4	13.5	13.6	13.6	13.8	9.6	11.9
	cased			12.2	13.0	13.2		
Hungarian-English	uncased	-	11.3	11.4	10.9	11.0	8.8	
	cased			8.3	10.1	10.2		
English-Hungarian	uncased	-	9.0	9.3	9.2	9.5	6.5	
	cased			8.1	8.4	8.7		

Table 2: Results overview for news-dev2009b sets: We see significant BLEU score increases with the addition of news data to the language model and using truecasing. As a comparison our results and the best systems from last year on the full news-dev2009 set are shown.

insensitive BLEU, but to a significant gain in case-sensitive BLEU. Note that we still do not properly address all-caps portions or headlines with our approach.

2.3 Results

Results on the development sets are summarized in Table 2. We see significant gains with the addition of news data to the language model (about 2 BLEU points) and using truecasing (about 0.5–1.0 BLEU points), and minor if any gains using minimum Bayes risk decoding (mbr), the monotone-at-punctuation reordering constraint (mp, see Section 3.2), and bigger beam sizes.

2.4 German–English

For German–English, we additionally incorporated

rule-based reordering — We parse the input using the Collins parser (Collins, 1997) and apply a set of reordering rules to re-arrange the German sentence so that it corresponds more closely English word order (Collins et al., 2005).

compound splitting — We split German compound words (mostly nouns), based on the frequency of the words in the potential decompositions (Koehn and Knight, 2003a).

part-of-speech language model — We use factored translation models (Koehn and Hoang, 2007) to also output part-of-speech tags with each word in a single phrase mapping and run a second n-gram model over them. The English part-of-speech tags are obtained using MXPOST (Ratnaparkhi, 1996).

German–English (ued'08: 17.1, best'08: 19.7)	**BLEU (uncased)**
baseline	16.6
+ interpolated news LM	20.6
+ minimum Bayes risk decoding	20.6
+ monotone at punctuation	20.9
+ truecasing	20.9
+ rule-based reordering	21.7
+ compound splitting	22.0
+ part-of-speech LM	22.1
+ big beam	22.3

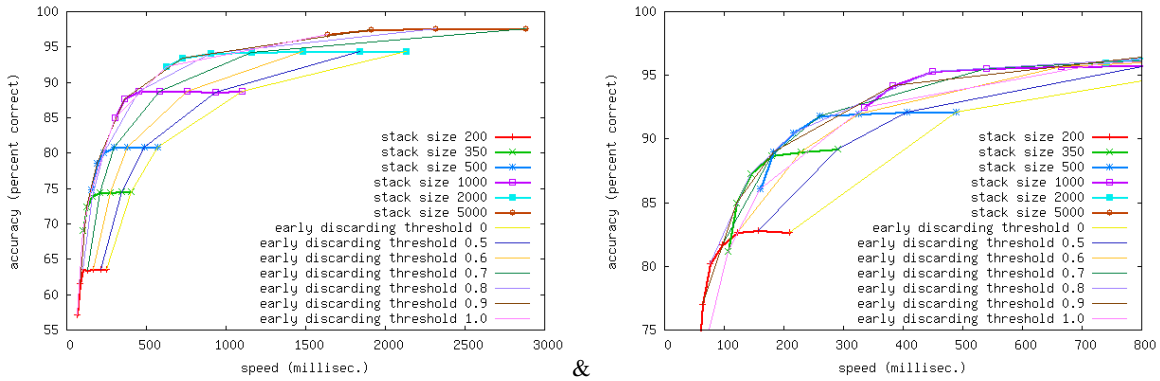
Table 3: Results for German–English with the incremental addition of methods beyond a baseline trained on the parallel corpus

English–German (ued'08: 12.1, best'08: 14.2)	**BLEU (uncased)**
baseline	13.5
+ interpolated news LM	15.2
+ minimum Bayes risk decoding	15.2
+ monotone at punctuation	15.2
+ truecasing	15.2
+ morphological LM	15.2
+ big beam	15.7

Table 4: Results for English–German with the incremental addition of methods beyiond a baseline trained on the parallel corpus

2.5 English-German

For English–German, we additionally incorporated a morphological language model the same way we incorporated a part-of-speech language model in the other translation direction. The morphological tags were obtained using LoPar (Schmidt and Schulte im Walde, 2000).



German-English & French-English

Figure 1: Early discarding results in speedier but still accurate search, compared to reducing stack size.

3 Recent Improvements

In this section, we describe recent improvements to the Moses decoder for the WMT 2009 shared task.

3.1 Early Discarding

We implemented in Moses a more efficient beam search, following suggestions by Moore and Quirk (2007). In short, the guiding principle of this work is not to build a hypothesis and not to compute its language model scores, if it is likely to be too bad anyway.

Before a hypothesis is generated, the following checks are employed:

1. the **minimum allowed score** for a hypothesis is the worst score on the stack (if full) or the threshold for the stack (if higher or stack not full) *plus* an early discarding threshold cushion
2. if (a) new hypothesis future score, (b) the current hypothesis actual score, and (c) the future cost of the translation option are worse than the allowed score, do not generate the hypothesis
3. if adding all real costs except for the language model costs (i.e., reordering costs) makes the score worse than the allowed score, do not generate the hypothesis.
4. complete generation of the hypothesis and add it to the stack

Note that check 1 and 2 mostly consists of adding and comparing already computed values. In our implementation, step 3 implies the somewhat costly construction of the hypothesis data structure, while step 4 performs the expensive language model calculation. Without these optimizations, the decoder spends about 60-70% of the search time computing language model scores. With these optimization, the vast majority of potential hypotheses are not built.

See Figure 1 for the time/search-accuracy tradeoffs using this early discarding strategy. Given a stack size, we can vary the threshold cushion mentioned in step 1 above. A tighter threshold (the factor 1.0 implies no cushion at all), results in speedier but worse search. Note, however, that the degradation in quality for a given time point is less severe than the alternative — reducing the stack size (and also tightening the beam threshold, not shown in the figure). To mention just two data points in the German-English setting: Stack size of 500 and early discarding threshold of 1.0 results in faster search (150ms/word) and better quality (73.5% search accuracy) than the default search setting of a stack size 200 and no early discarding (252ms/word for 62.5% seach accuracy). Accuracy is measured against the best translations found under any setting.

Note that this early discarding is related to ideas behind cube pruning (Huang and Chiang, 2007), which generates the top n most promising hypotheses, but in our method the decision not to generate hypotheses is guided by the quality of hypotheses on the result stack.

3.2 Framework to Specify Reordering Constraints

Commonly in statistical machine translation, punctuation tokens are treated just like words. For tokens such as commas, many possible translations are collected and they may be translated into any of these choices or reordered if the language model sees gains. In fact, since the comma is one

Requiring the translation of quoted material as a block:
`He said <zone> " yes " </zone> .`

Hard reordering constraint:
`Number 1 : <wall/> the beginning .`

Local hard reordering constraint within zone:
`A new idea <zone> ( <wall/> maybe not new <wall/> ) </zone> has come forward .`

Nesting:
`The <zone> " new <zone> ( old ) </zone> " </zone> proposal .`

Figure 2: Framework to specify reordering constraints with zones and walls. Words within zones have to be translated without reordering with outside material. Walls form hard reordering constraints, over which words may not be reordered (limited to zones, if defined within them).

the most frequent tokens in a corpus and not very consistently translated across languages, it has a very noisy translation table, often with 10,000s if not 100,000s of translations.

Punctuation has a meaningful role in structuring a sentence, and we see some gains exploiting this in the systems we built last year. By disallowing reordering over commas and sentence-ending punctuation, we avoid mixing words from different clauses, and typically see gains of 0.1–0.2 BLEU.

But also other punctuation tokens imply reordering constraints. Parentheses, brackets, and quotation marks typically define units that should be translated as blocks, meaning that words should not be moved in or out of sequences in quotes and alike.

To handle such reordering constraints, we introduced a framework that uses what we call **zones** and **walls**. A zone is a sequence of words that should be translated as block. This does not mean that the sequence cannot be reordered as a whole, but that once we start to translate words in a zone, we have to finish all its words before moving outside again. To put it another way: words may not reordered into or out of zones.

A wall is a hard reordering constraint that requires that all words preceeding it have to be translated before words after may be translated. If we specify walls within zones, then we consider them **local walls** where the before-mentioned constraint only applies within the zone.

Walls and zones may be specified with XML markup to the Moses decoder. See Figure 2 for a few examples. We use the extended XML framework to

1. limit reordering of clause-ending punctuation (walls)
2. define zones for quoted and parenthetical word sequences
3. limit reordering of quotes and parentheses (local walls within zones)
4. specify translations for punctuation (not comma).

Only (1) leads to any noticable change in BLEU in the WMT 2009 shared task, a slight gain 0.1–0.2.

Note that this framework may be used in other ways. For instance, we may want to revisit our work on noun phrase translation (Koehn and Knight, 2003b), and check if enforcing the translation of noun phrases as blocks is beneficial or harmful to overall machine translation performance.

Acknowledgements

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

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)*.

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.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic. Association for Computational Linguistics.

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. J., 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. and Knight, K. (2003a). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

Koehn, P. and Knight, K. (2003b). Feature-rich translation of noun phrases. In *41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Moore, R. C. and Quirk, C. (2007). Faster beam-search decoding for phrasal statistical machine translation. In *Proceedings of the MT Summit XI*.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

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)*.