Factored Translation Models

Philipp Koehn and Hieu Hoang
pkoehn@inf.ed.ac.uk, H.Hoang@sms.ed.ac.uk
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW
Scotland, United Kingdom

Abstract

We present an extension of phrase-based statistical machine translation models that enables the straight-forward integration of additional annotation at the word-level — may it be linguistic markup or automatically generated word classes. In a number of experiments we show that factored translation models lead to better translation performance, both in terms of automatic scores, as well as more grammatical coherence.

1 Introduction

The current state-of-the-art approach to statistical machine translation, so-called phrase-based models, is limited to the mapping of small text chunks without any explicit use of linguistic information, may it be morphological, syntactic, or semantic. Such additional information has been demonstrated to be valuable by integrating it in pre-processing or post-processing steps.

However, a tighter integration of linguistic information into the translation model is desirable for two reasons:

- Translation models that operate on more general representations, such as lemmas instead of surface forms of words, can draw on richer statistics and overcome the data sparseness problems caused by limited training data.

- Many aspects of translation can be best explained on a morphological, syntactic, or semantic level. Having such information available to the translation model allows the direct modeling of these aspects. For instance: reordering at the sentence level is mostly driven by general syntactic principles, local agreement constraints show up in morphology, etc.

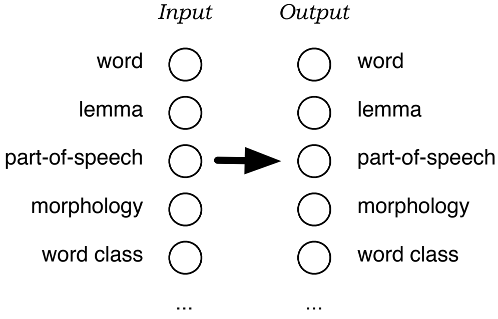

Figure 1: Factored representations of input and output words incorporate additional annotation into the statistical translation model.

Therefore, we extended the phrase-based approach to statistical translation to tightly integrate additional information. The new approach allows additional annotation at the word level. A word in our framework is not only a token, but a vector of factors that represent different levels of annotation (see Figure 1).

We report on experiments with factors such as surface form, lemma, part-of-speech, morphological features such as gender, count and case, automatic word classes, true case forms of words, shallow syntactic tags, as well as dedicated factors to ensure agreement between syntactically related items.

This paper describes the motivation, the modeling aspects and the computationally efficient decoding methods of factored translation models. We present briefly results for a number of language pairs. However, the focus of this paper is the description of the approach. Detailed experimental results will be described in forthcoming papers.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, June 2007. ©2007 Association for Computational Linguistics

2 Related Work

Many attempts have been made to add richer information to statistical machine translation models. Most of these focus on the pre-processing of the input to the statistical system, or the post-processing of its output. Our framework is more general and goes beyond recent work on models that back off to representations with richer statistics (Nießen and Ney, 2001; Yang and Kirchhoff, 2006; Talbot and Osborne, 2006) by keeping a more complex representation throughout the translation process.

Rich morphology often poses a challenge to statistical machine translation, since a multitude of word forms derived from the same lemma fragment the data and lead to sparse data problems. If the input language is morphologically richer than the output language, it helps to stem or segment the input in a pre-processing step, before passing it on to the translation system (Lee, 2004; Sadat and Habash, 2006).

Structural problems have also been addressed by pre-processing: Collins et al. (2005) reorder the input to a statistical system to closer match the word order of the output language.

On the other end of the translation pipeline, additional information has been used in post-processing. Och et al. (2004) report minor improvements with linguistic features on a Chinese-English task, Koehn and Knight (2003) show some success in re-ranking noun phrases for German-English. In their approaches, first, an n-best list with the best translations is generated for each input sentence. Then, the n-best list is enriched with additional features, for instance by syntactically parsing each candidate translation and adding a parse score. The additional features are used to rescore the n-best list, resulting possibly in a better best translation for the sentence.

The goal of integrating syntactic information into the translation model has prompted many researchers to pursue tree-based transfer models (Wu, 1997; Alshawi et al., 1998; Yamada and Knight, 2001; Melamed, 2004; Menezes and Quirk, 2005; Galley et al., 2006), with increasingly encouraging results. Our goal is complementary to these efforts: we are less interested in recursive syntactic structure, but in richer annotation at the word level. In future work, these approaches may be combined.

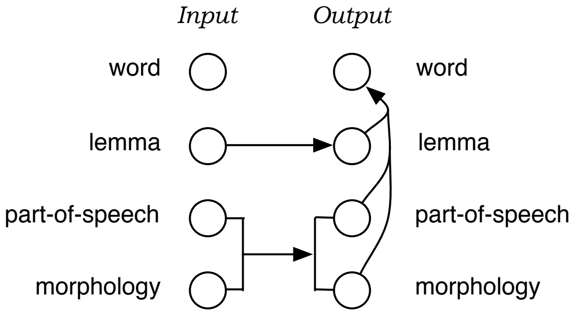

Figure 2: Example factored model: morphological analysis and generation, decomposed into three mapping steps (translation of lemmas, translation of part-of-speech and morphological information, generation of surface forms).

3 Motivating Example: Morphology

One example to illustrate the short-comings of the traditional surface word approach in statistical machine translation is the poor handling of morphology. Each word form is treated as a token in itself. This means that the translation model treats, say, the word *house* completely independent of the word *houses*. Any instance of *house* in the training data does not add any knowledge to the translation of *houses*.

In the extreme case, while the translation of *house* may be known to the model, the word *houses* may be unknown and the system will not be able to translate it. While this problem does not show up as strongly in English — due to the very limited morphological inflection in English — it does constitute a significant problem for morphologically rich languages such as Arabic, German, Czech, etc.

Thus, it may be preferably to model translation between morphologically rich languages on the level of lemmas, and thus pooling the evidence for different word forms that derive from a common lemma. In such a model, we would want to translate lemma and morphological information separately, and combine this information on the output side to ultimately generate the output surface words.

Such a model can be defined straight-forward as a factored translation model. See Figure 2 for an illustration of this model in our framework.

Note that while we illustrate the use of factored translation models on such a linguistically motivated example, our framework also applies to models that incorporate statistically defined word classes, or any other annotation.

4 Decomposition of Factored Translation

The translation of factored representations of input words into the factored representations of output words is broken up into a sequence of **mapping steps** that either **translate** input factors into output factors, or **generate** additional output factors from existing output factors.

Recall the example of a factored model motivated by morphological analysis and generation. In this model the translation process is broken up into the following three mapping steps:

1. **Translate** input lemmas into output lemmas
2. **Translate** morphological and POS factors
3. **Generate** surface forms given the lemma and linguistic factors

Factored translation models build on the phrase-based approach (Koehn et al., 2003) that breaks up the translation of a sentence into the translation of small text chunks (so-called phrases). This approach implicitly defines a segmentation of the input and output sentences into phrases. See an example in Figure 3.

Our current implementation of factored translation models follows strictly the phrase-based approach, with the additional decomposition of phrase translation into a sequence of mapping steps. Translation steps map factors in input phrases to factors in output phrases. Generation steps map output factors within individual output words. To reiterate: all translation steps operate on the phrase level, while all generation steps operate on the word level. Since all mapping steps operate on the same phrase segmentation of the input and output sentence into phrase pairs, we call these **synchronous factored models**.

Let us now take a closer look at one example, the translation of the one-word phrase *häuser* into English. The representation of *häuser* in German is: surface-form *häuser* | lemma *haus* | part-of-speech *NN* | count *plural* | case *nominative* | gender *neutral*.

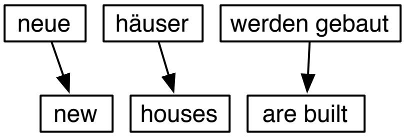

Figure 3: Example sentence translation by a standard phrase model. Factored models extend this approach.

The three mapping steps in our morphological analysis and generation model may provide the following applicable mappings:

1. **Translation:** Mapping lemmas
   - *haus → house, home, building, shell*
2. **Translation:** Mapping morphology
   - *NN|plural-nominative-neutral → NN|plural, NN|singular*
3. **Generation:** Generating surface forms
   - *house|NN|plural → houses*
   - *house|NN|singular → house*
   - *home|NN|plural → homes*
   - ...

We call the application of these mapping steps to an input phrase **expansion**. Given the multiple choices for each step (reflecting the ambiguity in translation), each input phrase may be expanded into a list of translation options. The German *häuser|haus|NN|plural-nominative-neutral* may be expanded as follows:

1. **Translation:** Mapping lemmas
   { *?|house|?|?, ?|home|?|?, ?|building|?|?, ?|shell|?|?* }
2. **Translation:** Mapping morphology
   { *?|house|NN|plural, ?|home|NN|plural, ?|building|NN|plural, ?|shell|NN|plural, ?|house|NN|singular,* ... }
3. **Generation:** Generating surface forms
   { *houses|house|NN|plural, homes|home|NN|plural, buildings|building|NN|plural, shells|shell|NN|plural, house|house|NN|singular,* ... }

5 Statistical Model

Factored translation models follow closely the statistical modeling approach of phrase-based models (in fact, phrase-based models are a special case of factored models). The main difference lies in the preparation of the training data and the type of models learned from the data.

5.1 Training

The training data (a parallel corpus) has to be annotated with the additional factors. For instance, if we want to add part-of-speech information on the input and output side, we need to obtain part-of-speech tagged training data. Typically this involves running automatic tools on the corpus, since manually annotated corpora are rare and expensive to produce.

Next, we need to establish a word-alignment for all the sentences in the parallel training corpus. Here, we use the same methodology as in phrase-based models (typically symmetrized GIZA++ alignments). The word alignment methods may operate on the surface forms of words, or on any of the other factors. In fact, some preliminary experiments have shown that word alignment based on lemmas or stems yields improved alignment quality.

Each mapping step forms a component of the overall model. From a training point of view this means that we need to learn translation and generation tables from the word-aligned parallel corpus and define scoring methods that help us to choose between ambiguous mappings.

Phrase-based translation models are acquired from a word-aligned parallel corpus by extracting all phrase-pairs that are consistent with the word alignment. Given the set of extracted phrase pairs with counts, various **scoring functions** are estimated, such as conditional phrase translation probabilities based on relative frequency estimation or lexical translation probabilities based on the words in the phrases.

In our approach, the models for the translation steps are acquired in the same manner from a word-aligned parallel corpus. For the specified factors in the input and output, phrase mappings are extracted. The set of phrase mappings (now over factored representations) is scored based on relative counts and word-based translation probabilities.

The generation distributions are estimated on the output side only. The word alignment plays no role here. In fact, additional monolingual data may be used. The generation model is learned on a word-for-word basis. For instance, for a generation step that maps surface forms to part-of-speech, a table with entries such as *(fish,NN)* is constructed. One or more scoring functions may be defined over this table, in our experiments we used both conditional probability distributions, e.g., $p(\textit{fish}|\textit{NN})$ and $p(\textit{NN}|\textit{fish})$, obtained by maximum likelihood estimation.

An important component of statistical machine translation is the language model, typically an n-gram model over surface forms of words. In the framework of factored translation models, such sequence models may be defined over any factor, or any set of factors. For factors such as part-of-speech tags, building and using higher order n-gram models (7-gram, 9-gram) is straight-forward.

5.2 Combination of Components

As in phrase-based models, factored translation models can be seen as the combination of several components (language model, reordering model, translation steps, generation steps). These components define one or more feature functions that are combined in a log-linear model:

$$p(\mathbf{e}|\mathbf{f}) = \frac{1}{Z} \exp \sum_{i=1}^{n} \lambda_i h_i(\mathbf{e},\mathbf{f}) \tag{1}$$

Z is a normalization constant that is ignored in practice. To compute the probability of a translation **e** given an input sentence **f**, we have to evaluate each feature function h_i. For instance, the feature function for a bigram language model component is (m is the number of words e_i in the sentence **e**):

$$\begin{aligned} h_{\text{LM}}(\mathbf{e},\mathbf{f}) &= p_{\text{LM}}(\mathbf{e}) \\ &= p(e_1)\; p(e_2|e_1)..p(e_m|e_{m-1}) \end{aligned} \tag{2}$$

Let us now consider the feature functions introduced by the translation and generation steps of factored translation models. The translation of the input sentence **f** into the output sentence **e** breaks down to a set of phrase translations $\{(\bar{f}_j, \bar{e}_j)\}$.

For a translation step component, each feature function h_{T} is defined over the phrase pairs $(\bar{f}_j, \bar{e}_j)$

given a scoring function τ:

$$h_{\mathrm{T}}(\mathbf{e},\mathbf{f}) = \sum_j \tau(\bar{f}_j, \bar{e}_j) \qquad (3)$$

For a generation step component, each feature function h_{G} given a scoring function γ is defined over the output words e_k only:

$$h_{\mathrm{G}}(\mathbf{e},\mathbf{f}) = \sum_k \gamma(e_k) \qquad (4)$$

The feature functions follow from the scoring functions (τ, γ) acquired during the training of translation and generation tables. For instance, recall our earlier example: a scoring function for a generation model component that is a conditional probability distribution between input and output factors, e.g., $\gamma(\textit{fish,NN,singular}) = p(\textit{NN}|\textit{fish})$.

The feature weights λ_i in the log-linear model are determined using a minimum error rate training method, typically Powell's method (Och, 2003).

5.3 Efficient Decoding

Compared to phrase-based models, the decomposition of phrase translation into several mapping steps creates additional computational complexity. Instead of a simple table look-up to obtain the possible translations for an input phrase, now multiple tables have to be consulted and their content combined.

In phrase-based models it is easy to identify the entries in the phrase table that may be used for a specific input sentence. These are called **translation options**. We usually limit ourselves to the top 20 translation options for each input phrase.

The beam search decoding algorithm starts with an empty hypothesis. Then new hypotheses are generated by using all applicable translation options. These hypotheses are used to generate further hypotheses in the same manner, and so on, until hypotheses are created that cover the full input sentence. The highest scoring complete hypothesis indicates the best translation according to the model.

How do we adapt this algorithm for factored translation models? Since all mapping steps operate on the same phrase segmentation, the **expansions** of these mapping steps can be efficiently pre-computed prior to the heuristic beam search, and stored as translation options. For a given input phrase, all possible translation options are thus computed before decoding (recall the example in Section 4, where we carried out the expansion for one input phrase). This means that the fundamental search algorithm does not change.

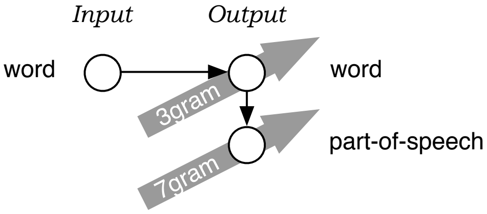

Figure 4: Syntactically enriched output: By generating additional linguistic factors on the output side, high-order sequence models over these factors support syntactical coherence of the output.

However, we need to be careful about combinatorial explosion of the number of translation options given a sequence of mapping steps. In other words, the expansion may create too many translation options to handle. If one or many mapping steps result in a vast increase of (intermediate) expansions, this may be become unmanageable. We currently address this problem by early pruning of expansions, and limiting the number of translation options per input phrase to a maximum number, by default 50. This is, however, not a perfect solution. We are currently working on a more efficient search for the top 50 translation options to replace the current brute-force approach.

6 Experiments

We carried out a number of experiments using the factored translation model framework, incorporating both linguistic information and automatically generated word classes.

This work is implemented as part of the open source Moses[1] system (Koehn et al., 2007). We used the default settings for this system.

6.1 Syntactically Enriched Output

In the first set of experiments, we translate surface forms of words and generate additional output factors from them (see Figure 4 for an illustration). By adding morphological and shallow syntactic infor-

[1] available at `http://www.statmt.org/moses/`

English–German

Model	BLEU
best published result	18.15%
baseline (surface)	18.04%
surface + POS	18.15%
surface + POS + morph	18.22%

English–Spanish

Model	BLEU
baseline (surface)	23.41%
surface + morph	24.66%
surface + POS + morph	24.25%

English–Czech

Model	BLEU
baseline (surface)	25.82%
surface + all morph	27.04%
surface + case/number/gender	27.45%
surface + CNG/verb/prepositions	27.62%

Table 1: Experimental results with syntactically enriched output (part of speech, morphology)

mation, we are able to use high-order sequence models (just like n-gram language models over words) in order to support syntactic coherence of the output. Table 1 summarizes the experimental results.

The English–German systems were trained on the full 751,088 sentence Europarl corpus and evaluated on the WMT 2006 test set (Koehn and Monz, 2006). Adding part-of-speech and morphological factors on the output side and exploiting them with 7-gram sequence models results in minor improvements in Bleu. The model that incorporates both POS and morphology (18.22% Bleu vs. baseline 18.04% Bleu) ensures better local grammatical coherence. The baseline system produces often phrases such as *zur*(to) *zwischenstaatlichen*(inter-governmental) *methoden*(methods), with a mismatch between the determiner (singular) and the noun (plural), while the adjective is ambiguous. In a manual evaluation of intra-NP agreement we found that the factored model reduced the disagreement error within noun phrases of length ≥ 3 from 15% to 4%.

English–Spanish systems were trained on a 40,000 sentence subset of the Europarl corpus. Here, we also used morphological and part-of-speech factors on the output side with an 7-gram sequence model, resulting in absolute improvements of 1.25% (only morph) and 0.84% (morph+POS). Improvements on the full Europarl corpus are smaller.

English-Czech systems were trained on a 20,000 sentence Wall Street Journal corpus. Morphological features were exploited with a 7-gram language model. Experimentation suggests that it is beneficial to carefully consider which morphological features to be used. Adding all features results in lower performance (27.04% Bleu), than considering only case, number and gender (27.45% Bleu) or additionally verbial (person, tense, and aspect) and prepositional (lemma and case) morphology (27.62% Bleu). All these models score well above the baseline of 25.82% Bleu.

An extended description of these experiments is in the JHU workshop report (Koehn et al., 2006).

6.2 Morphological Analysis and Generation

The next model is the one described in our motivating example in Section 4 (see also Figure 2). Instead of translating surface forms of words, we translate word lemma and morphology separately, and generate the surface form of the word on the output side.

We carried out experiments for the language pair German–English, using the 52,185 sentence News Commentary corpus[2]. We report results on the development test set, which is also the out-of-domain test set of the WMT06 workshop shared task (Koehn and Monz, 2006). German morphological analysis and POS tagging was done using LoPar Schmidt and Schulte im Walde (2000), English POS tagging was done with Brill's tagger (Brill, 1995), followed by a simple lemmatizer based on tagging results.

Experimental results are summarized in Table 2. For this data set, we also see an improvement when using a part-of-speech language model — the Bleu score increases from 18.19% to 19.05% — consistent with the results reported in the previous section. However, moving from a surface word translation mapping to a lemma/morphology mapping leads to a deterioration of performance to a Bleu score of 14.46%.

Note that this model completely ignores the surface forms of input words and only relies on the

[2] Made available for the WMT07 workshop shared task `http://www.statmt.org/wmt07/`

German–English

Model	BLEU
baseline (surface)	18.19%
+ POS LM	19.05%
pure lemma/morph model	14.46%
backoff lemma/morph model	19.47%

Table 2: Experimental results with morphological analysis and generation model (Figure 2), using News Commentary corpus

more general lemma and morphology information. While this allows the translation of word forms with known lemma and unknown surface form, on balance it seems to be disadvantage to throw away surface form information.

To overcome this problem, we introduce an alternative path model: Translation options in this model may come either from the surface form model or from the lemma/morphology model we just described. For surface forms with rich evidence in the training data, we prefer surface form mappings, and for surface forms with poor or no evidence in the training data we decompose surface forms into lemma and morphology information and map these separately. The different translation tables form different components in the log-linear model, whose weights are set using standard minimum error rate training methods.

The alternative path model outperforms the surface form model with POS LM, with an BLEU score of 19.47% vs. 19.05%. The test set has 3276 unknown word forms vs 2589 unknown lemmas (out of 26,898 words). Hence, the lemma/morph model is able to translate 687 additional words.

6.3 Use of Automatic Word Classes

Finally, we went beyond linguistically motivated factors and carried out experiments with automatically trained word classes. By clustering words together by their contextual similarity, we are able to find statistically similarities that may lead to more generalized and robust models.

We trained models on the IWSLT 2006 task (39,953 sentences). Compared to a baseline English–Chinese system, adding word classes on the output side as additional factors (in a model as previously illustrated in Figure 4) to be exploited by a 7-gram sequence model, we observe a gain 1.5% BLEU absolute. For more on this experiment, see (Shen et al., 2006).

English–Chinese

Model	BLEU
baseline (surface)	19.54%
surface + word class	21.10%

Table 3: Experimental result with automatic word classes obtained by word clustering

Chinese–English

Recase Method	BLEU
Standard two-pass: SMT + recase	20.65%
Integrated factored model (optimized)	21.08%

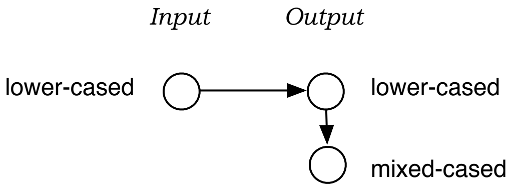

Table 4: Experimental result with integrated recasing (IWSLT 2006 task)

6.4 Integrated Recasing

To demonstrate the versatility of the factored translation model approach, consider the task of recasing (Lita et al., 2003; Wang et al., 2006). Typically in statistical machine translation, the training data is lowercased to generalize over differently cased surface forms — say, *the*, *The*, *THE* — which necessitates a post-processing step to restore case in the output.

With factored translation models, it is possible to integrate this step into the model, by adding a generation step. See Table 4 for an illustration of this model and experimental results on the IWSLT 2006 task (Chinese-English). The integrated recasing model outperform the standard approach with an BLEU score of 21.08% to 20.65%. For more on this experiment, see (Shen et al., 2006).

6.5 Additional Experiments

Factored translation models have also been used for the integration of CCG supertags (Birch et al., 2007), domain adaptation (Koehn and Schroeder, 2007) and for the improvement of English-Czech translation (Bojar, 2007).

7 Conclusion and Future Work

We presented an extension of the state-of-the-art phrase-based approach to statistical machine translation that allows the straight-forward integration of additional information, may it come from linguistic tools or automatically acquired word classes.

We reported on experiments that showed gains over standard phrase-based models, both in terms of automatic scores (gains of up to 2% BLEU), as well as a measure of grammatical coherence. These experiments demonstrate that within the framework of factored translation models additional information can be successfully exploited to overcome some short-comings of the currently dominant phrase-based statistical approach.

The framework of factored translation models is very general. Many more models that incorporate different factors can be quickly built using the existing implementation. We are currently exploring these possibilities, for instance use of syntactic information in reordering and models with augmented input information.

We have not addressed all computational problems of factored translation models. In fact, computational problems hold back experiments with more complex factored models that are theoretically possible but too computationally expensive to carry out. Our current focus is to develop a more efficient implementation that will enable these experiments.

Moreover, we expect to overcome the constraints of the currently implemented *synchronous* factored models by developing a more general *asynchronous* framework, where multiple translation steps may operate on different phrase segmentations (for instance a part-of-speech model for large scale reordering).

Acknowledgments

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No NR0011-06-C-0022 and in part under the EuroMatrix project funded by the European Commission (6th Framework Programme).

We also benefited greatly from a 2006 summer workshop hosted by the Johns Hopkins University and would like thank the other workshop participants for their support and insights, namely Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst Christine Moran, Wade Shen, and Richard Zens.

References

Alshawi, H., Bangalore, S., and Douglas, S. (1998). Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL).*

Birch, A., Osborne, M., and Koehn, P. (2007). CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Bojar, O. (2007). English-to-Czech factored machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic. Association for Computational Linguistics.

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.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia. Association for Computational Linguistics.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Hoang, H., Callison-Burch, C., Cowan, B., Zens, R., Dyer, C., Bojar, O., Moran, C., Constantin, A., and Herbst, E. (2006). Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. Technical report, John Hopkins University Summer Workshop.

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 Annual Meeting of the Association for Computational Linguistics, demonstation session.*

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

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. 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)*.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

Lee, Y.-S. (2004). Morphological analysis 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)*.

Lita, L. V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). tRuEcasIng. In Hinrichs, E. and Roth, D., editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

Melamed, I. D. (2004). Statistical machine translation by parsing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 653–660, Barcelona, Spain.

Menezes, A. and Quirk, C. (2005). Microsoft research treelet translation system: IWSLT evaluation. In *Proc. of the International Workshop on Spoken Language Translation*.

Nießen, S. and Ney, H. (2001). Toward hierarchical models for statistical machine translation of inflected languages. In *Workshop on Data-Driven Machine Translation at 39th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 47–54.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features 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)*.

Sadat, F. and Habash, N. (2006). Combination of arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

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)*.

Shen, W., Zens, R., Bertoldi, N., and Federico, M. (2006). The JHU Workshop 2006 IWSLT System. In *Proc. of the International Workshop on Spoken Language Translation*, pages 59–63, Kyoto, Japan.

Talbot, D. and Osborne, M. (2006). Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia. Association for Computational Linguistics.

Wang, W., Knight, K., and Marcu, D. (2006). Capitalizing 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)*.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3).

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Yang, M. and Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.