Incremental re-training of a hybrid English-French MT system with Customer Translation Memory data

Evgeny Matusov
Science Applications International Corporation (SAIC)
7990 Science Applications Ct.
Vienna, VA, USA
evgeny.matusov@saic.com

Abstract

In this paper, we present SAIC's hybrid machine translation (MT) system and show how it was adapted to the needs of our customer – a major global fashion company. The adaptation was performed in two ways: off-line selection of domain-relevant parallel and monolingual data from a background database, as well as on-line incremental adaptation with customer parallel and translation memory data. The translation memory was integrated into the statistical search using two novel features. We show that these features can be used to produce nearly perfect translations of data that fully or to a large extent partially matches the TM entries, without sacrificing on the translation quality of the data without TM matches. We also describe how the human post-editing effort was reduced due to significantly better MT quality after adaptation, but also due to improved formatting and readability of the MT output.

1 Introduction

We describe how a state-of-the-art hybrid MT system developed at SAIC is successfully being used to translate documents from English into Canadian French for our customer – a leading fashion designer company. This system has been designed especially for the needs of the customer which include:

- accurate translation of company-internal documents (e.g., cash register manuals, employee tutorials, etc.);
- error-free translation of templates for e-mails and other standardized documents;
- reduction of human translation costs.

Our customer employs the services of a leading human translation service provider (TSP). The goal was to make the work of the TSP's human translators more efficient by letting them post-edit the MT system output. To this end, the MT system had to fulfill the following requirements:

- The system had to be adapted to specific customer data, but the document topics were rather heterogeneous and included fashion, retail, and computer terminology; thus, the MT system had to be large enough to produce accurate translations for each of these domains.
- The documents already post-edited by human translators must be used to incrementally retrain the MT system to ensure high-quality translation of similar material in the future.
- Moreover, in case of full-sentence matches with already post-edited material, the MT system is expected to produce exactly the same translation.

To meet these customer requirements, we adapted the SAIC MT system to the customer data and introduced features into the system that allowed for integration of translation memory. Overtime, this has substantially improved the quality of the MT output and thus reduced the human post-editing effort. In addition, we addressed the formatting and MT output representation issues, which lead to further reduction of the post-editing costs of the TSP.

This paper is organized as follows. In Section 2, we describe the SAIC hybrid MT system and its

main features. Section 3 presents how the translation memory (TM) is integrated into the statistical search by using special flags which ensure that full TM matches are always selected, and partial TM matches are preferred during translation. In Section 4, we describe a method for off-line domain adaptation that allowed us to create a customer-tailored baseline system, as well as an on-line adaptation method which enables us to perform incremental updates of the system with new TM entries and other parallel data. In Section 5 we show how a number of simple, but useful techniques lead to improved readability of the MT output by humans. They involved punctuation spacing, capitalization, and adjustment to differences between European and Canadian French. Section 6 presents the experimental results. We show the impact of the adaptation on automatic MT metrics on different in-domain and out-of-domain test sets. We also present evidence that the improved MT quality significantly improved the productivity of human translators who post-edit the MT output. The paper concludes with a summary and a discussion of the future work in Section 7.

2 Baseline MT System

The SAIC MT system is a state-of-the-art hybrid system. The core of the system is a statistical search that employs a combination of multiple probabilistic translation models, including phrase-based and word-based lexicons, as well as reordering models and target n-gram language models. The actual search consists of two steps. First, those contiguous phrases in the source sentence are identified which have translation candidates in the phrase table. This phrase matching is done efficiently using an algorithm based on the work of (Zens, 2008). The second phase is the source cardinality-synchronous search (SCSS) implemented with dynamic programming. The goal of the search is to find the most probable segmentation of the source sentence into non-empty non-overlapping contiguous blocks, select the most probable permutation of those blocks, and choose the best phrasal translations for each of the blocks at the same time. The concatenation of the translations of the permuted blocks yields a translation of the whole sentence. The system is described in more detail in (Matusov and Köprü, 2010).

In theory, the permutations during the search are unlimited. In practice, the reordering is limited to allow for a maximum of M "gaps" (contiguous regions of uncovered word positions) at any time during the translation process. We set M to 2 for the English-to-French translation to model the most frequent type of reordering which is the reordering of an adjective-noun group.

The SAIC MT system also allows for partial or full rule-based translations. Specific source language entities can be identified prior to the search, and rule-based translations of these entities can be either forced to be chosen by the MT system, or can compete with phrase translation candidates from the phrase translation model. In both cases, the language model context at the boundaries of the rule-based translations is taken into account. Examples of entities which can be translated (or transfered to the target sentence) by rules are URLs, e-mail addresses, company names with trademark signs, document numbers, etc. SAIC offers its customers the possibility to specify rule-based translation using special tags in the input documents. Besides the actual rule-based translation of content, many customers use this possibility to pass formatting information such as XML tags through the translation system.

3 Translation Memory Integration

In addition to rule-based translation integration, the SAIC MT system is able to use translation memory (TM) when generating translations. In contrast to previous work (Koehn and Senellart, 2010), we decided not to apply approximate string matching to find the best and longest TM match for each input sentence. Instead, multiple *partial* TM matches are integrated in a "soft" way. When a new TM is uploaded to the system, each pair of source/target segments of the TM are word-aligned with a fast statistical alignment algorithm that employs probabilistic word lexica as its main knowledge source. Then, phrase pairs are extracted from the word-aligned segment pairs based on alignment constraints in the usual fashion (Och and Ney, 2003). The partial TM matches obtained in this way, as well as the

full matches are assigned special partial TM/full TM match flags, respectively. These flags are then used in the search to define binary TM features.

In the actual search, each phrase pair for which the full TM match feature has fired is used with a very high cost bonus that is quadratically proportional to the source length of the phrase. This means that this phrase pair is always selected for translation if there are no other competing phrase pairs which also originate from full TM matches. When such phrase pairs exist, then the longest TM match is selected because of the way the bonus is defined. In case of multiple full TM matches for the same phrase, the best translation is selected among them with the statistical search. In practice, we allow for a maximum full TM match length of 30 words.

The partial TM match feature fires only for phrase pairs from the TM of length ≥ 3. This constraint is introduced to deal with noise in the word alignment, as well as based on our experimental observations that very short partial TM matches introduce translations which might be incorrect in a context different from the one where they were extracted from, so that giving such phrase pairs a bonus may harm the translation quality. Similarly to the full TM match feature, the bonus assigned to partial TM matches is proportional to the length of the match, so that longer matches are favored. Usually, however, the weight of this feature is smaller than the weight of the full TM match feature. This means that partial TM matches still compete with the phrase pairs from the baseline phrase table. Thus, a partial match may not be necessarily selected if its context favors another translation from the background model.

4 Adaptation Techniques

4.1 Off-line Adaptation

At SAIC, we offer two possibilities to adapt the SAIC MT system to the domain/genre of the customer data. The first possibility is off-line adaptation. Given a small amount of customer in-domain training data, we sub-sample additional training data from large general-domain or out-of-domain parallel data collections. The sub-sampling is performed based on source or target perplexity values as computed with an n-gram language model trained on the customer data. For the particular case of English-to-French translation, we had very large parallel corpora available. They included the English-French Gigaword corpus crawled from the web, the proceedings of European and Canadian parliaments, news and news commentary articles, English-to-French movie titles, computer software manuals, etc. All of the background data we used is publicly available. We downloaded the Gigaword corpus from the Workshop on Machine Translation (WMT) web page[1]. The other data was downloaded from the OPUS collection of parallel corpora (Tiedemann, 2012).

From all of these sources, we selected a total of 1.7M sentence pairs (44M tokens), for which the corresponding French sentence had the lowest perplexity w.r.t. the 3-gram French language model trained on the 80K tokens of the French side of the parallel customer data. Then, we used the sampled data to train the baseline phrase table of the customer English-to-French MT system. The same approach was used to obtain an even larger amount of French monolingual data (186M tokens) for training of the 5-gram language model.

4.2 On-line adaptation

The second possibility for adaptation is on-line (and possibly incremental) system adaptation that can be performed with SAIC support or by customer alone. In this case, a second phrase table is generated from the customer TM and other parallel data. The phrase pairs from this customer phrase table compete with the phrase pairs from the (larger) baseline phrase table. The phrase translation probabilities are computed in the SAIC system on the fly as relative frequencies from the bilingual and marginal counts. This allows us to linearly interpolate the actual counts when multiple phrase table are used, so that more reliable probability estimates can be obtained. Thus, our phrase table combination technique is similar to the adaptation described by (Foster and Kuhn, 2007) and (Yasuda et al., 2008). The difference is that we linearly interpolate raw counts, not probabilities. Thus, our method can also be considered as an on-line corpus weighting technique in the style of (Koehn and Senellart, 2009).

The customer phrase table is generated based on

[1] http://www.statmt.org/wmt11/translation-task.html

a fast word alignment algorithm mentioned in Section 3. The customer can choose whether the TM match features should be used for the entries in the phrase table. Thus, either the (possibly more noisy) in-domain parallel corpus or a real TM can be used in on-line adaptation. Typically, it makes sense to use the on-line adaptation for customer data of less than 1M running words. In such cases, the adaptation time is 1-2 minutes on a single CPU core.

To perform incremental updates (i. e. to add additional parallel data or TM entries to the data previously uploaded to the system), the customer simply adds the new data to the old one and initiates the re-training of the customer phrase table under a new name. For each translation request, the customer can choose which customer-specific phrase table is to be used as the second phrase table, if any. The scaling factor for the counts from the second phrase table can also be adjusted by the customer.

In addition to the second phrase table, a second language model with its own scaling factor is used. This model is also trained on the customer-uploaded data. In our experiments, we used a 4-gram LM trained on the TM data plus the French side of the in-domain parallel data that was used for the off-line adaptation.

For the particular customer for which we created the English-to-French system, we performed incremental updates bi-weekly over a period of 6 months, combining up to 5K of new TM entries with the older TM entries at the time of each on-line adaptation. In addition to the TM entries, we also incrementally updated the customer's bilingual glossary, which was added to the system as phrase table entries with TM match features turned on. In addition, we inserted the single-word glossary entries and their single-word translations into the statistical word lexica with a fixed high probability before we word-aligned the TM data. This improved the word alignment quality, since many of the words found in the glossary also appeared in the TM data.

5 Localization and Formatting Issues

Besides focusing on the core technology for TM integration and domain adaptation, we had to deal with issues which are of importance when the final consumers of the MT output are humans (in this case, employees and clients of our customer). Previously, SAIC was mainly dealing with customers who needed MT to translate large amounts of data with satisfactory quality, for the purposes of information retrieval from the translated data. In the current case, we had to keep in mind that all of the MT output has to be post-edited with as few corrections as possible and has to be presented to humans in a well-readable form. This means that not only the content had to be transfered from English to French correctly, but also the formatting should be correct.

One important issue that we tried to solve is capitalization. The customer required that certain words (like company and product names) should start with a capital letter. Moreover, sometimes a whole sentence had to be written in all-capital letters. We implemented two solutions for this request. First, we trained a truecased translation model in the sense that we retained the original case of the French words. Only the first letter of the first word in each sentence was adjusted with heuristics relying on the frequencies with which the word appeared cased and uncased. These frequencies were obtained not only from the French side of the parallel corpus, but also from a much larger monolingual French data collection. Consequently, the French n-gram language model was also trained on the truecased data. This means that the statistical search had in some cases to choose from the same translation of a phrase, but with different casing.

The second solution was to transfer the case of the source English word to its target French translation. We implemented this feature by relying on within-phrase word alignment. The feature was quite useful when a whole (longer) phrase had to be translated with first-uppercase or all-uppercase letters. However, in several other cases the approach failed because of two reasons:

- the casing conventions/rules between English and French were different for the words involved;
- the word alignment was incorrect or only partially correct. For example, an English proper noun was aligned not only with the corresponding French proper noun, but also with its preceding article so that the article was wrongly capitalized together with the noun.

Based on the discussions with the customer, we decided to make this feature optional in our system.

Another issue that is related to capitalization is specific to Canadian French as the target language. In European French, the accents on the French letters disappear if the letter is capitalized; in Canadian French, the accents remain. Since the bulk of our training data comes from European French sources, we had to extract a list of all possible words which start with an accented letter and make a postprocessing mapping that would correctly write those words if they are to be capitalized.

Next, we focused on the representation of punctuation marks. We switched from ASCII to the proper French UTF-8 symbols for apostrophes and quotation marks. We also implemented the Canadian French punctuation spacing rules as provided by the customer. The rules define if a whitespace before/after a punctuation mark is to be inserted or not. We implemented these rules in postprocessing.

One more complicated issue that involved also the translation model was differentiating between hyphen and dash. In most statistical MT systems, including ours, the dash is normalized to hyphen and all words are split at hyphen. Since hyphen and dash placement is ambiguous, such normalization step eliminates noise in the training data and reduces the number of different word forms, which leads to a better model estimation. However, this normalization also means that it is not straightforward to determine in postprocessing, which hyphens should remain as separate tokens and be replaced with dashes, and which ones should be spliced with the surrounding content words. Our solution was to pass the dash as the category content with the hyphen being its corresponding category label. This means that the hyphen is translated as usual by the MT system. If the translation of it is also the hyphen, then this hyphen on the target side is replaced with the category content, which is the dash. If the translation of the hyphen is not a hyphen, then no replacement is made. With this trick, we were able to produce MT output which had both hyphens (originating from hyphens not appearing as separate tokens in the raw English sentence) and dashes (originating from dashes and hyphens appearing as separate tokens in the raw English sentence). This made it possible to implement detokenization rules which left dashes as separate tokens, but joined the hyphens with the other words.

Finally, we also paid attention to the word order of specific expressions like number sequences in parentheses, monetary amounts, etc. To limit the reordering within parentheses, we employed the concept of zones similar to the one implemented in the Moses decoder (Koehn et al., 2007). Reordering across zone boundaries (which correspond to the parentheses in our particular case) is not allowed. For dollar amounts (e.g. "$ 500"), we implemented a reordering rule that forced the decoder to produce the French version with the dollar sign following the number: "500 $". We also implemented postprocessing rules for adjusting the number representations (e.g. "1234.15" to "1 234,15").

All of these adjustments do not have a significant effect on automatic measures of MT quality, but very much increase the readability and acceptance of the MT output by humans. In addition, they further reduce the human post-editing effort.

6 Experimental Results

6.1 Automatic Evaluation

The first part of the experimental evaluation involved computing the established automatic MT measure BLEU (Papineni et al., 2002) on several in-domain and out-of-domain test sets, using a single reference translation. The evaluation was case-insensitive. The test sets were:

- TM set (185 sentences): entries from the translation memory - 85 full-sentence matches and 100 partial matches created manually by introducing slight modifications to the original TM sentences and the corresponding reference translations.
- Main set (710 sentences): customer data representing marketing material about the customer products, as well as employee manuals. The domain of the sentences corresponds to the domain of the majority of the customer-provided parallel training data. However, none of the sentences appear in the TM data used for the on-line adaptation.
- Cash register set (300 segments): technical phrasal terminology and instructions on how to

use a cash register. These data are from a different domain than the bulk of customer training data.

- Hansards (200 sentences): out-of-domain set consisting of randomly selected sentences from the standard test set of Canadian Hansards.
- News articles (150 sentences): out-of-domain set consisting of randomly selected news commentary articles from the WMT 2010 news test set.

We compared the following three MT systems or system configurations:

- *General MT*: a system trained on news data and Canadian Hansards data (4.1M sentences, 122M words) and optimized on a development set from the general news domain. This is a quite competitive baseline system that is probably more suited for translating into Canadian French than the other two systems evaluated below, which were trained mostly on European French data.
- *Off-line adaptation*: system trained on the data sub-sampled from background parallel corpora collections. Thus, the training data for the system reflected the different domains present in the customer training data, with more emphasis on the terminology from the world of fashion and sales. It is important to note that the customer training data itself used as the basis for sub-sampling was included only for language model training, but not for the translation model training.
- *On-line adaptation*: this system extends the off-line adapted system by a second phrase table trained on the 80K tokens of the customer parallel data and on all of the TM entries available after 11 cycles of incremental adaptation described in Section 4.2. Full and partial TM matches were extracted from a total of 61K sentence pairs (680K English tokens), and the TM matching features were used in the MT search.

The latter two systems were optimized on a separate development set representing a mix of the domains of all 5 test sets using the Downhill Simplex Algorithm (Cettolo and Federico, 2004). The scaling factors obtained in the optimization process showed that high weights were assigned to the customer phrase table, the customer language model, as well as, in case of the on-line adapted system, to the TM matching features[2].

Table 1: Effect of off-line and TM-based on-line adaptation on the BLEU score on different in-domain and out-of-domain test sets.

	BLEU [%]		
Test set	General	Adaptation	
	MT	off-line	+ on-line
TM	16.0	15.5	73.7
main	19.0	34.1	30.1
cash register	13.8	14.1	35.4
Hansards	32.0	20.5	20.6
news articles	21.2	16.3	16.8

The results of the translation experiments are presented in Table 1. We can see that the off-line adaptation does not improve the results on the TM set. The reason for this is that the TM entries are quite different from the general training data and were not used as the "ground truth" for sub-sampling. On the main test set, which was more similar to the customer training data, the improvement is more than significant (from 19.0 to 34.1% absolute). The results on the cash register set improve only slightly, since the domain of this set is underrepresented in the sub-sampled training data and in the language model training data. The results on the Hansards and news articles test sets deteriorate, which was expected, since the majority of the sentence pairs from the respective training sets were automatically considered irrelevant to the customer domain and were excluded from training.

When we add the on-line adaptation to the off-line adapted system, we first immediately notice the BLEU score of 73.7% on the TM set. This means that the translation is almost perfect both for full TM and partial TM matches. Thus, the TM features succeeded in forcing the MT systems to use the TM entries. Since the TM also contains a few entries from the technical domain, the cash register

[2]Note that the optimization of the on-line adapted system was performed after the first cycle of incremental adaptation with 21K TM sentence pairs. In subsequent adaptation cycles, the scaling factors were kept fixed.

set results also improve significantly. The results on the main test set which has no long matches with the TM go down. We attribute this to the fact that some translations from the TM are either incorrectly used out-of-context or do not match with the original reference translation from the customer training data, which might be different for several reasons. These reasons may include a poorer human translation quality as compared to the entries in the TM, as well as the fact that some TM translations include additional information that a statistical system can not produce given the source sentence even with an ideal model. Such information may include, for example, the translation of "it" into "elle/il", conversion of measurements in inches into centimeters, etc. When such translations are selected by the systems as partial TM matches, they do not correspond to the reference translation which is often more close to the original than a TM translation would be. Nevertheless, the MT quality of 30.1% BLEU on the main test set is still quite good, especially if compared with the 19.0% BLEU of the general model.

Finally, the results in Table 1 show that the MT quality does not deteriorate on the out-of-domain test sets when the TM features are used. This would allow the customer to obtain good translation results on heterogeneous texts with some sentences containing many TM matches and other sentences containing no TM matches at all.

6.2 Human Evaluation

We let our customer evaluate the system manually. The customer was very much satisfied with the MT results. In particular, SAIC was praised for the fact that the system was able to consistently perfectly translate full TM matches and almost full matches, even if such translations contradicted the logic of the statistical system core (for example, when an English computer-related term had to be translated into its French equivalent with the original English term following it in parentheses). At the same time, the system was still able to produce good translations with only a few errors when there was only little or no overlap with the TM entries known to the system. In addition, the customer TSP was very much pleased with the formatting/localization enhancements described in Section 5. These enhancements significantly reduced the post-editing time; the translator was not distracted anymore by formatting issues and could focus on the correction of the remaining few MT errors.

Unfortunately, we were not given any detailed post-editing speed and productivity increase information from our customer until the deadline for this paper. From the customer's TSP, who uses similarly adapted SAIC systems for other customers, we know that the human translation throughput per hour increased by a factor of 3 to 5 depending on the source language text quality, when the TSP switched from manual translation to post-editing.

Table 2 shows an example of improved translation quality when the on-line adapted system is used instead of the baseline general MT system. Except for the proper noun "Madison", there were no full TM/glossary matches for this sentence. There was also only one partial TM match: ("leather that has", "cuir qui a"). Yet the translation quality improved dramatically due to the off-line adaptation and the use of the customer parallel training data in the on-line adaptation.

When the system was first deployed, our customer only planned to translate company-internal documents such as manuals using our dedicated MT solution. After seeing the quality of the provided MT output and how this quality is improving over time due to incremental adaptation, the customer is now ready to use SAIC MT for more important documents. These include company-internal e-mails (where human-in-the-loop post-editing is not feasible because of time constraints), as well as company-external template-based e-mails and even publicly available marketing material which is very sensitive to any translation errors. This means that our customer is very pleased with the MT solution that SAIC has designed and implemented.

7 Summary and Future Work

In this paper, we described how a hybrid MT system with a statistical core can be adapted to the needs of a customer who requests high-quality translations in several narrow domains for fast subsequent post-editing. We described a perplexity-based method of training data selection for off-line adaptation, as well as a method for on-line adaptation that allows the customer to automatically train and use

Table 2: An example of improved translation quality when using the adapted vs. general MT system. Full and partial TM matches for the adapted system are shown in italic.

source	*Madison* money pieces are smooth and refined, sewn in *leather that has* a rich sheen and takes color beautifully.
general	Madison d'argent sont en douceur et raffiné, partie en cuir qui a un riche et reflets de couleur très bien.
adapted	Pièces d'argent *Madison* lisses et raffinées, cousues dans le *cuir qui a* un éclat riche et prend magnifiquement la couleur.
reference	Pièces d'argent Madison lisses et raffinées, cousues dans le cuir qui a un éclat riche et se colore magnifiquement.

a phrase table with entries from the customer parallel data, including glossaries and translation memories. The TM is integrated into the MT system in such a way that the longest full TM matches are always selected. Partial TM matches are given a preference when competing with other translation candidates in the statistical search. We showed that all of these enhancements to the baseline system significantly improve translation quality on the customer-relevant test data. Together with the additional improvements related to human readibility of the MT output, we were able to significantly reduce the post-editing costs of our customer's TSP. We received very positive comments about our MT system from our customer, who is now considering to use it for translation of more error-sensitive data.

In the future, as we obtain more customer data for the individual domains (marketing material, general and technical manuals, e-mails, etc.), we would like to adapt the system to each of these domains by using multiple domain-specific customer phrase tables or a mixture translation model.

References

Mauro Cettolo and Marcello Federico. 2004. Minimum Error Training of Log-Linear Translation Models. In *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan.

George Foster and Roland Kuhn. 2007. Mixture Model Adaptation for Statistical Machine Translation. In *ACL Workshop on Statistical Machine Translation. Prague, Czech Republic*, June.

Philipp Koehn and Jean Senellart. 2009. Discriminative Corpus Weight Estimation for Machine Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August.

Philipp Koehn and Jean Senellart. 2010. Fast Approximate String Matching with Suffix Arrays and A* Parsing. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA, November.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic. Association for Computational Linguistics.

Evgeny Matusov and Selçuk Köprü. 2010. Improving Reordering in Statistical Machine Translation from Farsi. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA, November.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model. In *Proceedings of the International Joint Conference on Natural Language Processing*.

Richard Zens. 2008. *Phrase-based Statistical Machine Translation: Models, Search, Training*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, February.