The RWTH Machine Translation System

Evgeny Matusov, Richard Zens, David Vilar, Arne Mauser, Maja Popović, Saša Hasan and Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{matusov,zens,vilar,mauser,popovic,hasan,ney}@cs.rwth-aachen.de

Abstract

We present the statistical machine translation system used by RWTH in the second TC-STAR evaluation. We give a short overview of the system as used in the first evaluation and then enumerate the improvements of the system over the last months. We then discuss the results obtained by our group in the evaluation.

1. Introduction

In this paper we will describe the system used by RWTH in the second TC-STAR evaluation that took place February 2006. We participated in the Spanish to English, English to Spanish and Chinese to English tracks, in all the conditions using a statistical machine translation (SMT) system.

We use a two pass approach. First we generate lists of the n best translation candidates using a phrase-based translation model combined log-linearly with additional models (e.g. language model and single word based models) and then apply additional rescoring models on these generated hypotheses in order to extract the final translation.

The paper is organized as follows. In Section 2 we will briefly describe our baseline system, which is the one we used for the first TC-STAR evaluation that took place in March 2005. A more thorough description of the system can be found in (Vilar et al., 2005). In Section 3 we will describe the main improvements of our current system when compared to the baseline. Section 4 presents the results obtained in the evaluation and some conclusions will be drawn in Section 5.

2. Baseline System

In this section we will briefly present the system used in the first TC-STAR evaluation, which we used as baseline system for the current evaluation. As usual, we will denote the (given) source sentence with $f_1^J = f_1 \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_I$. Our baseline system models the translation probability directly using a log-linear model (Och and Ney, 2002):

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}, \quad (1)$$

with a set of different models h_m, scaling factors λ_m and the denominator a normalization factor that can be ignored in the maximization process. We choose the λ_m by optimizing a performance measure on a development corpus using the downhill simplex algorithm.

The most important models in equation (1) are phrase-based models in both source to target and target to source directions. In order to extract these models, an alignment between source and target sentence is found by using the IBM-1, HMM and IBM-4 models in both directions (source-to-target and target-to-source) and combining the two obtained alignments (Och and Ney, 2003). Given this alignment, an extraction of contiguous phrases is carried out and their probabilities are computed by means of relative frequencies (Zens and Ney, 2004).

Another important model in the log-linear model is the language model, a 4-gram language model with Kneser-Ney smoothing in our case. Additionally we use single word based lexica (IBM-1 like) at the level of extracted sentences, also in source to target and target to source direction. This has the effect of smoothing the relative frequencies used as estimates of the phrase probabilities. A length and a phrase penalty are the last models in the set.

2.1. Rescoring of n-best lists

Instead of generating only the translation that obtains the highest probability according to Equation (1), we generate a list of the n highest scoring translations (Ueffing et al., 2002; Zens and Ney, 2005). We then proceed to rescore these generated sentences with additional models, which, due to their structure or their high computational costs cannot be directly integrated into the beam search algorithm used for the optimization process. The most important models used for rescoring are the IBM1 model and additional language models.

Although the IBM1 model is the easiest one of the single-word based translation models and the phrase-based models clearly outperform this approach, the inclusion of the scores of this model, i.e.

$$h_{\text{IBM1}}(f_1^J|e_1^I) = \log \frac{1}{(I+1)^J} \prod_{j=1}^J \sum_{i=0}^I p(f_j|e_i) \quad (2)$$

has been shown experimentally to improve the performance of a machine translation system.

During the generation process, a single language model is used. However, additional language models specific to each sentence can help to improve the machine translation quality (Hasan and Ney, 2005). The motivation behind this lies in the following observation: the syntactic structure of a

sentence is influenced by its type. We apply a method based on regular expressions to cluster the sentences into specific classes. This information is then used to train class-specific language models (5-grams) which are linearly interpolated with the main language model to avoid data sparseness.
Additionally we also include some variations of the length penalty score as additional rescoring models.

3. Recent Improvements

In this section we will discuss the improvements which have lead to the biggest performance gains for the second evaluation campaign.

3.1. Reordering

Subjective error analysis carried out on the results of the first TC-STAR evaluation campaign showed that the word order of the generated sentences is not correct (Vilar et al., 2006). For this year's evaluation, depending on the language pair, we used two different reordering strategies. For the Spanish-English language pair, morphosyntactic based local reorderings were applied. For the Chinese-English language pair, long-range reorderings are needed, and we applied phrase-based reorderings with maximum entropy estimation of the distortion parameters.

3.1.1. Morphosyntactic Reordering

For the English-Spanish language pair we used additional morphosyntactic knowledge in the form of part-of-speech (POS) tags. We tagged the English part using the Lingsoft tagger[1] and the Spanish part using the FreeLing tagger (Carreras et al., 2004). We then reorder the source language in order to come up with a sentence structure more similar to the one of the target language.
The motivation was that adjectives in Spanish are usually placed after the corresponding noun whereas in English adjectives preceed their nouns. Therefore local reorderings of nouns and adjective groups (adverb + adjective) are helpful for translation between these two languages. If Spanish is the source language, each Spanish noun is moved behind the correspondent adjective group. If English is the source language, each adjective group is moved behind the corresponding noun. Two examples of these reorderings can be found in Table 1. Such types of rule-based reorderings may sometimes be ambiguous and heavily depend on the quality of POS tagging. Nevertheless, significant improvements in translation quality were obtained by using these rules, see (Popović and Ney, 2006)
Note that these reorderings are applied as a preprocessing step, both in the training phase before the alignment computation, and before the translation of the test corpus. Additional local reorderings are applied in the style of (Kanthak et al., 2005) during the search process.

3.1.2. Lexicalized Reordering

The common phrase-based SMT systems use a very simple reordering model. Usually, the costs for phrase movements are linear in the distance, e.g. see (Och et al., 1999; Koehn, 2004; Zens et al., 2005). Recently, in (Tillmann and Zhang, 2005) and in (Koehn et al., 2005), a reordering model has been described that tries to predict the orientation of a phrase, i.e. it answers the question "should the next phrase be to the left or to the right of the current phrase?" This phrase orientation probability is conditioned on the current source and target phrase and relative frequencies are used to estimate the probabilities. We adopt the idea of predicting the orientation, but using a maximum-entropy based model.
The relative-frequency based approach of (Koehn et al., 2005) may suffer from the data sparseness problem, because most of the phrases occur only once in the training corpus. Our approach circumvents this problem by using a combination of phrase-level and word-level features and by using word classes or POS information. Maximum entropy is a suitable framework for combining these different features with a well-defined training criterion. A detailed description of this reordering model can be found in (Zens and Ney, 2006a).
This model has been used for the Chinese-English task.

3.2. Tuple LM

We also included a language model trained on the training corpus represented as bilingual tuples (as described in (Kanthak et al., 2005)) as an additional feature in the log-linear model. Thus, we combined two different translation model paradigms - conditional phrase translation probabilities and joint tuple language model probabilities. Given a segmentation $\tilde{e}_1^J$ of the target sentence e_1^I in a number of phrases given by the length of the source sentence f_1^J (the segmentation may include the empty target word ε), the joint probability is given by

$$h_{tuple}(f_1^J, e_1^I) = \log \prod_{f_j} p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}) \,. \qquad (3)$$

In our machine translation system, the translation is built during the search by concatenating target phrases corresponding to a segmentation of the source sentence into the matching source phrases. Each of the bilingual phrases can be represented as a sequence of bilingual tuples based on the within-phrase word alignment information and the single-word based lexicon costs. Thus, for the whole sentence a sequence of bilingual tuples can be built and scored with the tuple language model.

3.3. Sentence Segmentation for ASR output

Automatic speech recognition (ASR) systems normally do not include punctuation or sentence boundaries. Therefore the issue of sentence segmentation arises when translating such kind of output, because it is important to produce translations of sentences or sentence-like units to make the SMT output human-readable. At the same time, sophisticated speech translation algorithms (e. g. ASR word lattice translation, rescoring and system combination algorithms for (N-best) output of one or several SMT systems) may require that the number of words in the input source language segments/sentences is limited to about 30 or 40 words.
Our approach to the segmentation of ASR output originates from the work of (Stolcke et al., 1998). A decision for placing a segment boundary is made based on a log-linear com-

[1] http://www.lingsoft.fi/

original Spanish sentence	... este sistema no sea susceptible de ser usado como **arma política.**
reordered Spanish sentence	... este sistema no sea susceptible de ser usado como **política arma.**
generated English sentence:	
without reordering	... this system is not likely to be used as a **weapon policy.**
with reordering	... this system is not likely to be used as a **political weapon.**
reference English sentence	... the system cannot be used as a **political weapon.**

Table 1: Example of reorderings for the Spanish-English language pair.

bination of language model and prosodic features. However, in contrast to existing approaches, we explicitly optimize over the length of each segment (in words) and add a length model feature. This approach makes it possible to introduce restrictions on the minimum and the maximum length of a segment and nevertheless produce syntactically and semantically meaningful sentence-like units, which pass all the relevant context information on to the phrase-based SMT system.

Here is a short overview of the approach. We are given an (automatic) transcription of speech, denoted by the words $w_1^N := w_1, w_2, \ldots, w_N$. To score a hypothesized segment w_{i+1}^j starting with word position $i+1$ and ending on position j, we interpolate log-linearly the following probabilistic features.

The language model probability $p_{LM}(w_{i+1}^j)$ for a segment is computed as a product of the following three probabilities:

$$p_{LM}(w_{i+1}^j) = p_S(w_{i+1}^j) \cdot p_I(w_{i+1}^j) \cdot p_E(w_{i+1}^j) \quad (4)$$

These probabilities are modelled as follows (assuming a trigram language model): the probability for the first two words of a segment (segment **S**tart), conditioned on the last segment boundary <s>:

$$p_S(w_{i+1}^j) = p(w_{i+1}|\texttt{<s>}) \cdot p(w_{i+2}|w_{i+1}, \texttt{<s>}) \quad (5)$$

the probability for the other words within a segment (**I**nternal probability)

$$p_I(w_{i+1}^j) = \prod_{k=i+3}^{j} p(w_k|w_{k-1}, w_{k-2}) \quad (6)$$

and a LM probability for the segment boundary (**E**nd) in dependency on the last two words of a segment:

$$p_E(w_{i+1}^j) = p(\texttt{<s>}|w_j, w_{j-1})\,. \quad (7)$$

The extension to higher order language models is straightforward.

In addition to the language model probability, we use a prosodic feature, namely the normalized pause duration between any two consecutive words. Since the length of the segment is known, we also include an explicit parametric sentence length probability $p(j-i)$.

During the search, the word sequence w_1^N is processed from left to right. For all hypothesized segment end positions j, we optimize over the position of the last segment boundary i. The optimal sentence segmentation solution for words up to position i has already been computed in a previous recursion step. The globally optimal sentence segmentation for the document is determined when the last word of the document is reached. Note that the minimum and/or maximum sentence lengths l and L can be explicitly set by limiting the values of i to $l \leq j - i \leq L$.

The scaling factors in the log-linear combination of the models are tuned on a development set by computing precision/recall with respect to manually defined sentence-like units. At the moment, the algorithm achieves a performance level of up to 70% precision and 65% recall using the ASR output for the EPPS Spanish test corpus (2005 TC-STAR Evaluation). Further refinements of the algorithm are planned.

3.4. Rescoring Models

In addition to the already presented rescoring models, for the TC-STAR 2006 evaluation we used two new additional rescoring models.

3.4.1. n-gram Posterior Probabilities

The idea is similar to the word posterior probabilities: we sum the sentence posterior probabilities for each occurrence of an n-gram.

We define the fractional count $C(e_1^n, f_1^J)$ of an n-gram e_1^n for a source sentence f_1^J as:

$$C(e_1^n, f_1^J) := \sum_{I, e'^I_1} \sum_{i=1}^{I-n+1} p(e'^I_1|f_1^J) \cdot \delta(e'^{i+n-1}_i, e_1^n)\,, \quad (8)$$

with $\delta(\cdot,\cdot)$ the Kronecker function. The sums over the target language sentences are limited to an N-best list, i.e. the N best translation candidates according to the baseline model. In this equation, the term $\delta(e'^{i+n-1}_i, e_1^n)$ is one if and only if the n-gram e_1^n occurs in the target sentence e'^I_1 starting at position i.

Then, the posterior probability of an n-gram is obtained as:

$$p(e_1^n|f_1^J) = \frac{C(e_1^n, f_1^J)}{\sum_{e'^n_1} C(e'^n_1, f_1^J)} \quad (9)$$

The widely used word posterior probability is obtained as a special case, namely if n is set to one.

The n-gram posterior probabilities can be used similar to an n-gram language model:

$$h_n(f_1^J, e_1^I) = \frac{1}{I} \log \left(\prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}, f_1^J) \right) \quad (10)$$

with:

$$p(e_i|e_{i-n+1}^{i-1}, f_1^J) := \frac{C(e_{i-n+1}^i, f_1^J)}{C(e_{i-n+1}^{i-1}, f_1^J)} \quad (11)$$

Note that the models do not require smoothing as long as they are applied to the same N-best list they are trained on. A detailed description of the n-gram posterior probabilities can be found in (Zens and Ney, 2006b).

3.4.2. Sentence-level Mixtures

As an additional rescoring model we used sentence level mixtures language models, as presented in (Iyer and Ostendorf, 1990). The goal is to represent topic dependencies combining M different language models with a global one, corresponding to the index $m = 0$ in the following equation (for the case of trigram language models)

$$p(e_1^I) = \sum_{m=0}^{M} \lambda_m \left[\prod_{i=1}^{I} p_m(e_i|e_{i-1}, e_{i-2}) \right] . \qquad (12)$$

The training sentences are automatically divided into a fixed number M of clusters (representing different topics) using a maximum likelihood approach and the weights λ_m are trained on the development data. We used 4-grams for this rescoring model.

4. Experimental Results

4.1. Experimental Setup

The EPPS training corpus used for this evaluation is the same we used for the previous evaluation, extended with the data corresponding to the period between December 2004 and May 2005. The statistics can be found in Table 2. The data has been further preprocessed to adapt it to the different conditions.

For the FTE, hardly any preprocessing of the data is needed. To aid the translation system, a categorization of the text has been carried out where numbers, dates, proper names, etc. have been detected and marked. The text is also lowercased to reduce the vocabulary size when computing the alignments, but the translation models are trained on the true case corpus.

For the Verbatim transcriptions, we did some additional preprocessing and normalization, like expanding contractions ("I am" instead of "I'm", "we will" instead of "we'll", etc.) and eliminating hesitations ("uhm-", "ah-", etc.). Additionally, all numbers are written out (e.g. "forty-two" instead of "42").

Note that for the test data, there is near twice as much running words for the Spanish to English translation direction as for the English to Spanish. This is due to the fact that data from the Spanish Parliament has also been included in this year's evaluation campaign.

For Chinese–English task, a large variety of bilingual corpora is provided by the Linguistic Data Consortium (LDC). The domain is news, the vocabulary is very large and the sentences have an average length of about 30 words. The Chinese part is word segmented using the LDC segmentation tool. After the preprocessing, our training corpus consists of about seven million sentences with somewhat more than 200 million running words. The corpus statistics of the preprocessed training and test corpora are shown in Table 3. For the ASR translation task, we removed punctuation marks from the training corpora. Despite of that the used corpora are identical for the text and the ASR condition.

4.2. Detailed Results

The results obtained by the RWTH in the 2006 TC-STAR evaluation are presented in Table 4. The results for Spanish to English and English to Spanish are in the same range. This is in clear contrast to the results of last year's evaluation, where the results for the translation direction Spanish to English were clearly superior compared to the opposite direction, mainly due to the simpler structure of the English language. However, note that the Spanish to English results of this year's evaluation also include an important part consisting of the texts of the Spanish Parliament. The system was not specially tuned for this kind of data, and there is a certain mismatch between these data and the training data originating from the European Parliament. The results for each of these conditions alone (EPPS and Cortes) can be found in Table 5. It can be seen that the translations for the EPPS test corpus, as expected, have a much higher BLEU score (up to 12.2% absolute BLEU difference for the FTE condition, around 9% absolute for Verbatim and ASR).

Another (unexpected) feature of the results presented in Table 4 is that the results for Spanish to English Verbatim condition are better than the results for the FTE condition. This contradicts the experience gained from last year's evaluation campaign where the scores for the output of the FTE condition were consistently better than the ones of the Verbatim condition. As of yet we have no clear explanation for this effect.

The effect of the new methods applied in Section 3 can be seen in Table 6 on the EPPS Verbatim task.

5. Conclusions

We have described the RWTH machine translation system used in the second TC-STAR evaluation campaign. We have put special emphasis on the improvements of the system with respect to the first evaluation campaign and have shown the results obtained with these methods.

The main differences with respect to the system used in the 2005 evaluation are more advanced reorderings models, an additional tuple language model, and new rescoring models. Additionally, for the ASR condition an automatic segmentation of the input has been carried out.

6. Acknowledgments

This work has been funded by the integrated project TC-STAR– Technology and Corpora for Speech-to-Speech Translation – (IST-2002-FP6-506738).

7. References

X. Carreras, I. Chao, L. Padr, and M. Padr. 2004. Freeling: An open-source suite of language analyzers. In *Proc. of the Fourth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 239–242, Lisbon, Portugal, May.

S. Hasan and H. Ney. 2005. Clustered language models based on regular expressions for SMT. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, May.

R. M. Iyer and M. Ostendorf. 1990. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Speech and Audio Processing*, 7(1):30–39.

		Spanish	English
Train	Sentences	1 167 627	
	Words + Punct. Marks	35 320 646	33 945 468
	Words	32 074 034	30 821 291
	Vocabulary	159 080	110 636
	Singletons	63 045	46 121
Test	Sentences	1 782	1 117
	Words + Punct. Marks	56 468	28 492
	Words	50 634	25 869
	OOV Words	363	72

Table 2: Statistics of the EPPS Corpora.

		Chinese	English
Train	Sentences	7.1M	
	Words + Punct. Marks	199M	213M
	Words	173M	191M
	Vocabulary	223K	351K
	Singletons	100K	162K
	Conv. Dictionary Entry Pairs	82K	
Monolingual training data (Words)			537M
Test Verbatim	Sentences	1 232	
	Words	29 736	61 861
	OOV Words	48	705
Test ASR (WER=9.8%)	Sentences	1 286	
	Words	32 641	62 037
	OOV Words	1	697

Table 3: Training data for the Chinese-English task: large variety of bilingual corpora from LDC

Language Pair	Condition	BLEU[%]	NIST	WER[%]	PER[%]
English to Spanish	FTE	49.4	10.16	39.8	30.5
	Verbatim	45.4	9.71	43.1	32.1
	ASR	35.9	8.72	50.5	38.7
Spanish to English	FTE	47.1	10.36	42.9	30.9
	Verbatim	50.6	10.87	40.7	28.8
	ASR	35.0	9.08	51.3	38.8
Chinese to English	Verbatim	16.3	6.43	77.6	56.0
	ASR	12.2	5.05	81.8	64.5

Table 4: Official RWTH results in the 2006 evaluation.

Condition	BLEU[%]	NIST	WER[%]	PER[%]
FTE	47.1	10.36	42.9	30.9
EPPS	53.1	10.65	37.1	27.0
Cortes	40.9	9.13	48.8	35.0
Verbatim	50.6	10.87	40.7	28.8
EPPS	55.1	10.94	36.4	25.9
Cortes	46.2	9.85	44.9	31.7
ASR	35.0	9.08	51.3	38.8
EPPS	39.4	9.38	46.5	35.6
Cortes	30.3	8.00	56.4	42.4

Table 5: Spanish to English translation results, split in EPPS and Cortes Corpora.

Task	Condition	BLEU[%]	NIST	WER[%]	PER[%]
Spanish to English, Verbatim	Single Best Translation	48.6	10.64	41.9	29.3
	+ POS Reordering	49.6	10.87	40.9	28.9
	+ Rescoring	50.6	10.87	40.7	28.8
English to Spanish, Verbatim	Single Best Translation	44.6	9.66	43.2	32.7
	+ POS Reordering	45.2	9.71	43.3	32.2
	+ Rescoring	45.4	9.71	43.1	32.1

Table 6: Effect of the different methods on the EPPS Verbatim tasks. The results correspond to the three official submissions of RWTH.

S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, Ann Arbor, MI, June.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, October.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. AMTA-04*, pages 115–124, Washington DC, September/October.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

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

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

M. Popović and H. Ney. 2006. Pos-based word reorderings for statistical machine translation. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, May. To appear.

A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP)*, Sidney, Australia.

C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 557–564, Ann Arbor, MI, June.

N. Ueffing, F. J. Och, and H. Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 156–163, Philadelphia, PA, July.

D. Vilar, E. Matusov, S. Hasan, R. Zens, and H. Ney. 2005. Statistical Machine Translation of European Parliamentary Speeches. In *Proceedings of MT Summit X*, pages 259–266, Phuket, Thailand, September. Asia-Pacific Association for Machine Translation (AAMT).

D. Vilar, J. Xu, L. F. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, May. To appear.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 257–264, Boston, MA, May.

R. Zens and H. Ney. 2005. Word graphs for statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 191–198, Ann Arbor, Michigan, June.

R. Zens and H. Ney. 2006a. Discriminative reordering models for statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL): Proc. Workshop on Statistical Machine Translation*, New York City, NY, June. To appear.

R. Zens and H. Ney. 2006b. N-gram posterior probabilities for statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL): Proc. Workshop on Statistical Machine Translation*, New York City, NY, June. To appear.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.