Domain Dependent Statistical Machine Translation

Jia Xu[†], **Yonggang Deng**[*], **Yuqing Gao**[*] and **Hermann Ney**[†]

[†]Computer Science VI
RWTH Aachen University
D-52056 Aachen, Germany
{xujia,ney}@cs.rwth-aachen.de

[*]IBM TJ.Watson Research Center
1101 Kitchawan Road,
Yorktown Heights, N.Y. 10598
{ydeng,yuqing}@us.ibm.com

Abstract

While statistical machine translation (SMT) has advanced significantly with better modeling techniques and much more training data, domain specific SMT has received much less attention and leaves much room for further improvements. In this work, we address domain issues and propose to use the combination of feature weights and language model adaptation, to distinguish multiple domains, which share a general translation engine with phrase-based log-linear models. The proposed method requires much less parallel data than what is typically used to build a domain independent system, which makes it easy, cheap and efficient to capture as many domains as required. Domain adaptation during decoding is approached with source text classification methods. Our results on the GALE tasks show significant improvements with the proposed domain dependent translation than domain independent translation.

1. Introduction

Statistical machine translation (SMT) addresses the problem of automatically translating a text in one language into a text in another language using machine learning techniques and statistical modeling approaches. In SMT, models are trained from parallel and monolingual corpora. The quality and quantity of the data and the underline modeling approach together mostly determines the quality of the translation output. With the increasing availability of parallel corpora and a better modeling approach, a significant improvement of the translation quality has been achieved in the recent years.

While translation performance has been advanced substantially in general, translation style and domain issue leave much room for further improvements. For instance, translating an utterance can be quite different than translating a written sentence in selecting words and phrases and their orders. Short phrases such as "what's up" are more likely to be observed in an informal situation than in written form. This offers a challenge to genre adaptation of SMT systems but causes at the same time a rise to potential improvement if the issue can be handled properly.

In this work, we approach domain adaptation in machine translation with classification methods. Two main problems need to be solved: the first one is how to build domain specific SMT systems in training; the second is how to perform domain adaptation during decoding. For the first problem, we use the domain dependent language modeling or feature weights combination. When translating a test document, we are going to automatically identify its domain and then apply a corresponding decoding setup. Different text classification methods are going to be investigated and compared. Furthermore, we are going to show their impact on translation performance.

We are going to review our baseline translation system in section 2., then we are going to discuss how to build domain specific SMT systems in section 3. and how to do domain adaptation during testing in section 4. In section 5., we are going to present the experimental setup and are going to show the classification and the translation results, followed by the conclusions and future work.

2. Review of the Baseline Translation System

In statistical machine translation, we are given a source ('Foreign') language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language ('English') sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\begin{aligned} \hat{e}_1^I &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \end{aligned} \quad (1)$$

The decomposition into two knowledge sources in Equation 1 is known as the source-channel approach to statistical machine translation (Brown et al., 1990). It allows an independent modeling of the target language model $Pr(e_1^I)$ and the translation model $Pr(f_1^J|e_1^I)$. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language.

The notational convention will be as follows: we use the symbol $Pr(\cdot)$ to denote general probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(\cdot)$.

2.1. Log-linear Model for SMT

As an alternative to the classical source-channel model, the log-linear model (Och and Ney, 2003) has become popular and proved to be effective in directly modeling the distribution of a target sentence when given a source sentence:

$$Pr(e_1^I|f_1^J) = \frac{1}{Z_{f_1^J}} \exp\big(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\big)$$

The denominator is a normalization factor, which guarantees a proper probability distribution over all possible target sentences conditioned on the source sentence. Since f_1^J is given during the decoding process, the denominator can be ignored, and the searching criterion is simplified as

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}} \Big\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \Big\} \quad (2)$$

The log-linear model is a generalization of the source-channel model. One advantage of the log-linear model is easy integration of multiple feature functions h_m, which guarantees a direct statistical decision during decoding for an accurate output and fast search. We will discuss active features used in the translation engine in section 2.2.

Usually the scaling factors λ_m are estimated iteratively to maximize the likelihood of the training data under the log-linear model using, i.e., GIS algorithm. Alternatively, they can be trained discriminatively on a development set to directly maximize the translation performance measured by an error criterion (Och, 2003).

2.2. Decoding Process and Active Features

In SMT, translation is implemented as a statistical decision making process to search for the best target sentence within all possibilities. We use phrase-based translation (Och et al., 1999; Och and Ney, 2004; Koehn et al., 2003), which takes phrases as basic translation units rather than words. A phrase is a consecutive sequence of words, which captures word context naturally. In a typical phrase-based SMT system, the translation process begins with the segmentation of the source sentence into phrases, and then translates each source phrase into a target phrase, and finally reorders the target phrases to generate the output hypothesis.

Our translation system employs the phrase-based log-linear model. The decoder generates target sentences from left to right by covering source phrases in a certain order under the heuristic function. The underline feature functions h_m in Equation 2 are crucial to the search procedure. We briefly describe some of them:

Phrase Translation Model:

The phrase translation model is the most crucial component, sometimes it is referred as phrase translation table, which specifies alternative translation candidates and their probabilities for each source phrase. To build a phrase translation model, we start from a collection of parallel sentences and word alignments between sentence pair. We use IBM Model-4 word alignments (Brown et al., 1990) trained with the GIZA++ toolkit (Och and Ney, 2003). All phrase pairs with respect to word alignment boundaries are identified (more details are in (Zens et al., 2002)) and pooled to estimate a phrase translation table by their relative frequency. For better performance, we use phrase translation models in both translation directions.

Word-based lexicon Model:

To alleviate the data sparseness problem, a word-based lexicon model (Zens et al., 2005) is usually introduced to smooth the phrase translation probabilities. We assume all words in the source phrase generate all words in the phrase equally as in IBM Model-1. The lexicon probability is estimated as relative frequency from the word-aligned training corpora. Like a phrase translation model, we also apply a word-based lexicon model in both directions.

Target Language Model:

A target language model helps to discriminate alternative target hypothesis by assigning, ideally, higher probability to a sentence which is more likely to be spoken/written. In our system, we use a statistical n-gram language model with modified Kneser-Ney smoothing trained with the SRI language model toolkit (Stolcke, 2002).

Word and Phrase Penalty Model:

The word and the phrase penalty model simply counts the number of target words and the number of target phrases. These two heuristics affect the average sentence and the phrase lengths.

3. Building Domain Specific SMT Systems

After reviewing the baseline translation system, we are now going to discuss how to build domain specific SMT systems. Ideally, we would have domain specific training data and could build separate SMT system for each domain. Practically, this is hardly the case. We assume that we have a collection of training corpora with the general domain including a variety of different domains. And at the same time, we have some domain specific parallel documents to be used for building the domain specific systems.

We avoid building separate systems for multiple domains due to a lack of training data. But we will build a general SMT system to be shared among all domain specific systems, which are constructed in two ways:

1. **Domain dependent model combination**

   We distinguish domain specific SMT systems by the combination of the feature weights, i.e., scaling factors in Equation 2:

   $$\{\lambda_m, m = 1, \cdots, M\}$$

   The feature functions are described in Section 2.2. Domain specific scaling factors will be trained discriminatively on the domain specific development set.

2. **Domain dependent language modeling**

   The documents to be translated can have different language styles, e.g. the language style of broadcast conversation is very different from that of the newswire text. Therefore, we can generate a specific language model for each domain.

   We use a sentence-level mixture of K sublanguage models (Iyer and Ostendorf, 1996), each of which can be identified with the n-gram statistics for a specific topic of sentences. The probability of a word sequence $w_1, .., w_N$ is modelled as

   $$P(w_1, .., w_N) = \sum_{k=1}^{K} \gamma_k [\prod_{i=1}^{N} P_k(w_i|w_{i-n+1}^{i-1})], \quad (3)$$

where γ_k are the mixture weights and $P_k()$ is the n-gram model for the k-th topic.

The advantage of both methods is that we only need a small amount of data from each domain.

4. Domain Adaptation

Before decoding a test document, we decide which domain specific SMT system is to be applied by examining the source text of the test document. So the domain adaptation is transformed into a monolingual text classification problem: Which domain is the test document most similar to?

In theory, any text classification method can be applied here. We investigate and compare two text classification techniques: One technique is based on domain specific language models, the other is based on an information retrieval approach.

4.1. Language Model Based Domain Identification

We will consider two domains as an example, e.g. newsgroup text and newswire text. We build domain specific language models P_d ($d \in \{1, 2\}$) for the source side of the development corpora. Since the development sets are usually small, each of the models P_d is linearly intepolated with a general domain independent language model P_g:

$$P_d^*(w|h) = (1 - \alpha)P_d(w|h) + \alpha P_g(w|h) \quad (4)$$

For a test document to be translated, we compute the perplexity of each domain specific language model P_d^* and select the domain with the lowest perplexity.

4.2. Information Retrieval Approach

The second method for text classification is based on concepts from information retrieval. We use a simplified version of the method described in (Iyer and Ostendorf, 1996). We compute the similarity $S(d)$ between a test document and the development set of domain d:

$$S(d) = \sum_{w \in A \cap A_d} \frac{1}{(|A^w| + 1)|A|}, \quad (5)$$

where A_d is the set of words for the development set of domain d, A is the set of words in the test document, $|A|$ is the vocabulary size of the test document, and $|A^w|$ is the number of documents in the test corpus containing the word w.

For each test document, we select the domain with the highest score $S(d)$.

5. Experimental Results

The experiments have been carried out on the GALE Chinese-English tasks of 2006 and 2007. In both tasks, we use similar training corpora as well as the training and decoding processes. In task I, there are two domains, newswire text and newsgroup text, the domain of test data is not given, so we performed the domain adaptation methods described in Section 4. In task II, there are four domains in the test data, newswire, broadcast news, broadcast conversation and web text. The domain boundaries are provided already, so there is no need to perform the domain classification any more.

5.1. Task and Corpus

The corpus statistics of the bilingual training data and the test sets are shown in Table 1. The preprocessing step includes the tokenization and the categorization on the numbers and dates. Long sentences are segmented into short sentences using the binary segmentation method (Xu et al., 2006) to reduce the training time. After the preprocessing and segmentation, the parallel training data contains more than 20 million sentences and approximately 250 million words in each language.

The six-gram language model was trained on the English part of the bilingual training corpus and on the monolingual data from the LDC GigaWord corpus. The total amount of the language model training data is more than 1.5 billion running words.

In task II, using the method described in Section 3., we mix this general language model with a domain specific language model trained with speech transcription domain corpora containing over 100 million running words. The language model mixture weights are optimized using Powell algorithm with the respect of the PPL measured on the development corpus for each domain.

We use the BLEU (Papineni et al., 2002) score as the primary evaluation criterion. Model scaling factors (feature weights) are optimized with the respect of the BLEU score using the Downhill Simplex algorithm.

In task I, we use the NIST 2002 evaluation set as the newswire and the GALE 2006 dryrun development corpus as the newsgroup development set. The evaluation set is the GALE 2006 evaluation data. The sentences in some Chinese test sets are segmented but not in the English references. Our purpose is to show that the system optimized on the domain specific development corpus outperforms the one optimized on the general or out of domain development corpus.

In task II, the GALE 2007 development corpus is taken as development set, and the evaluation set is the 2006 MT development corpus GALE part. Here each test set is separated into four domains, the domain dependent language modeling helps in translating the data in the broadcast conversation (B.C.) and web text (W.T.) domains but not in translating the newswire and broadcast news documents, since most of the training corpora are already newswire articles.

Because of the large amount of training data and the categorization, the out of vocabulary words (OOVs) on all Chinese test sets are low. The statistics of the English references are measured without preprocessing.

5.2. Classification Results

Our primary goal of addressing the domain issue in machine translation is to improve the translation quality. Since the domain adaptation is implemented as document classification, the classification accuracy can be indicative of translation performance. In this section, we are going to present the results of the classification methods described in Section 4. for task I.

From two aspects we separate the evaluation set into different domains, i.e. the newswire and newsgroup. There are 55 documents in the evaluation set, including 36 newswire and 19 newsgroup articles. We calculate the classification

Table 1: Corpus Statistics.
(RW: running words, OOVs: out of vocabulary words)

			Chinese	English
Train		Sentences	20.3 M	
		R.W.	249 M	269 M
		Vocabulary	251 K	430 K
		Singletons	109 K	160 K
Task I				
Dev	newswire	Sentences	878	
		R.W.	24 111	27 914
		Vocabulary	4 095	3 888
		OOVs	3	100
	newsgroup	Sentences	2 203	2 115
		R.W.	41 102	46 759
		Vocabulary	5 660	5 423
		OOVs	11	113
Eval	newswire	Sentences	460	364
		R.W.	9 979	10 344
		Vocabulary	2 636	3 155
		OOVs	11	1 279
	newsgroup	Sentences	441	415
		R.W.	9 606	10 526
		Vocabulary	2 594	3 042
		OOVs	11	1 378
Task II				
Dev	all	Sentences	3 166	
		R.W.	72 326	79 674
		Vocabulary	8 481	7 745
		OOVs	177	1 103
	B.C.	Sentences	1 431	
		R.W.	21 934	23 445
		Vocabulary	2 982	2 776
		OOVs	41	312
	W.T.	Sentences	657	
		R.W.	17 778	19 757
		Vocabulary	4 221	3 680
		OOVs	84	258
Eval	all	Sentences	2 276	
		R.W.	48 654	54 493
		Vocabulary	6 735	6 213
		OOVs	49	1 340
	B.C.	Sentences	979	
		R.W.	14 162	15 287
		Vocabulary	2 434	2 204
		OOVs	2	436
	W.T.	Sentences	415	
		R.W.	9 946	11 822
		Vocabulary	2 667	2 478
		OOVs	23	235

error rate by dividing the number of incorrectly classified documents by the number of all test documents.

1. Language model approach

   This method was presented in Section 4.1. We build a six-gram language model from the Dev newswire corpus and from the Dev group corpus respectively. Because of the limited resources in each domain, we also produce a trigram general language model LM_g in Equation 4 to cover some unknown words from the evaluation data. The vocabulary is constrained to the union of the vocabularies of the Dev newswire and Dev group. The general language model was trained on the Chinese side of the bilingual corpora in Table 1.

   Further more, as shown in Figure 1, the lowest error rate is 25.5%, if the value of α (see Equation 4) is set between 0.5 and 0.7 or between 0.9 and 0.95. In Figure 1, the blue (with plus symbol) and green line (with star symbol) plots the classification error rate on the newswire articles and on the newsgroup articles respectively. The red line (with circle symbol) indicates the error rate in both domains. We see that the error rate curve is flat until the value of α approaches one. As long as the general language model weight is not given a very high value, the results from the combination of the in-domain and the general language model are stable.

2. Information retrieval approach

   Using the information retrieval approach described in Section 4.2. we have a classification error rate of 34.5%, where none of the newswire article is classified wrongly, and 19 of the newsgroup articles are classified incorrectly as newswire.

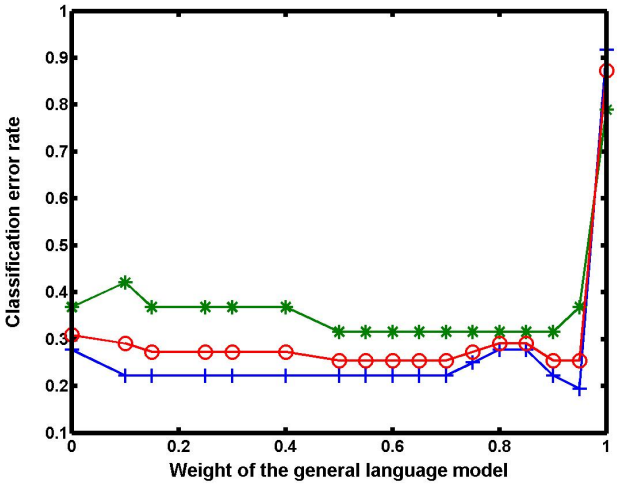

Figure 1: The documents classification error rate related to the weight of the general language model α on the newswire (+), newsgroup (*) and all (o) data sets.

5.3. Translation Performance

Since the language model approach outperforms the information retrieval approach in the test document domain classification results, we simply perform the translations with the classification results obtained using the language model approach.

5.3.1. Evaluation Metrics

So far, in machine translation research a single generally accepted criterion for the evaluation of the experimental results does not exist. Therefore, we used different criteria.

- WER (word error rate):
  The WER is computed as the minimum number of substitution, insertion and deletion operations which have to be performed to convert the generated sentence into the reference sentence.
- PER (position-independent word error rate):
  The PER is defined as the WER ignoring the word order.
- BLEU score:
  This score measures the precision of unigrams, bigrams, trigrams and fourgrams with respect to reference translations with a penalty for too short sentences (Papineni et al., 2002). The BLEU score measures the accuracy, i.e. larger BLEU scores are better.
- NIST score:
  NIST score (Doddington, 2002) is similar to BLEU, but it uses an arithmetic average of N-gram counts rather than a geometric average, and it weights more heavily those N-grams that are more informative.
- TER
  Translation Edit Rate (TER) (Snover et al., 2006) measures the amount of editing which a human being would have to perform to change a system output so it exactly matches a reference translation.

5.3.2. Translation Results

In task I, we distinguish the systems with different settings of the scaling factors of the log-linear model in the decoder.

In Table 2, in the baseline systems the scaling factors are optimized on the newswire development corpus. If all the newsgroup documents are translated with the feature weights optimized on the newsgroup development corpus, we receive oracle best (O.B.) translation results. Here we show the oracle best results optimized with the respect of the BLEU and the TER. Using our language model based document classification method with $\alpha = 0.5$, the BLEU score rises from 9% to 11%, which is an improvement of 18% relatively, while the TER score reduces too. The oracle best shows the BLEU score can still reach to 13.6%, if the documents are 100% correctly classified.

In task II, the domain dependent language models (genreLMs) instead of a general language model are applied in each domain specific system.

From Table 3, we see the perplexity of the domain specific language model measured on the development corpus in each domain reduces a lot. This results in the improvements of the translation performances in both domains, i.e. 0.6% in the BLEU score for the web text and 0.9% in the BLEU score for the broadcast conversation domain.

5.3.3. Classification Accuracy versus Translation Performance

For task I, we plot the BLEU scores of the newsgroup text translations and the classification error rates measured on all data sets (newsgroup and newswire text). The incorrectly classified documents are randomly selected. As shown Figure 2, a roughly proportional relationship exists between the document classification accuracy and the translation performance.

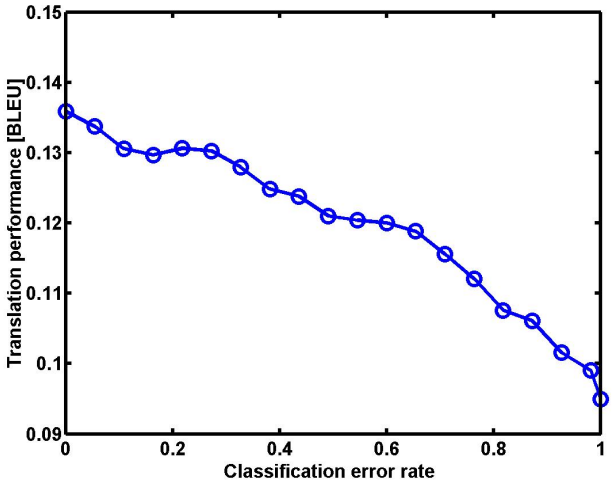

Figure 2: The translation performance in the BLEU score related to the documents classification error rate for newsgroup text in task I.

6. Conclusions and Future Work

We have discussed the domain issue in statistical machine translation and proposed an efficient method to build domain specific machine translation systems. We have used a combination of feature weights of the phrase-based log-linear translation model to discriminate multiple domains. The training of a domain specific system is a tuning process where the translation performance is to be maximized on a small amount of the domain specific development set. Moreover, the domain dependent language modeling also helps in enhancing the translation performance in each domain.

Domain adaptation during the translation of the test documents are implemented as solving monolingual text classification problems. We compared a language model based approach with an information retrieval based approach and found the former achieved lower document classification error rate.

Our results on the GALE Chinese-English translation tasks have showed that the domain adaption in the translation process achieved significant improvements over the domain independent translation, even with a pretty high document classification error rate in the domain adaptation.

We plan to exploit better document classification algorithms. Another future work is to perform dynamic domain selection and adaptation driven by the test data.

7. Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org).

8. References

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S.

Table 2: The translation results on the evaluation data with the domain specific optimizations in task I.

Domain	Method		WER[%]	PER[%]	BLEU[%]	NIST	Ins	Del	Sub	Shft	TER[%]
Newsgroup	Baseline		73.34	61.98	**9.50**	3.034	161	3705	3490	993	70.00
	O.B.	Opt-bleu grp	77.63	59.14	13.59	4.599	1076	1251	4963	1208	0.7132
		Opt-ter grp	75.58	61.36	9.39	3.588	264	2959	3818	1105	0.6837
	Opt-bleu $\alpha = 0.5$		73.70	60.79	**11.01**	3.771	350	3087	3816	1030	69.51

Table 3: The translation results with the genre language models in task II.

Test data	Domain	Method	WER[%]	PER[%]	BLEU[%]	TER[%]	LM-PPL
Dev	Web text	Baseline	77.54	55.68	14.87	71.91	116
		GenreLM	78.78	56.48	15.15	73.24	105
	Broadcast conversation	Baseline	71.66	51.79	17.57	66.42	127
		GenreLM	72.86	52.80	18.19	67.74	79
Eval	Web text	Baseline	72.44	53.16	**15.57**	68.07	
		GenreLM	73.43	53.46	**16.16**	68.67	
	Broadcast conversation	Baseline	74.06	55.55	**15.33**	69.74	
		GenreLM	75.81	56.75	**16.24**	71.46	

Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology*, pages 128–132, San Diego, California, March.

R. Iyer and M. Ostendorf. 1996. Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models. In *International Conference on Spoken Language Processing '96*, volume 1, pages 236–239, Philadelphia, PA.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

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 and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):135–244, December.

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

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

K. A. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, July.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference On Spoken Language Processing*, pages 901–904, Denver, Colorado, September.

J. Xu, R. Zens, and H. Ney. 2006. Partitioning parallel documents using binary segmentation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation*, pages 78–85, New York City, NY, June.

R. Zens, F. J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *25th German Conf. on Artificial Intelligence (KI2002)*, pages 18–32, Aachen, Germany, September. Springer Verlag.

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.