EFFICIENT DATA SELECTION FOR MACHINE TRANSLATION

A. Mandal[1]*, D. Vergyri*[1]*, W. Wang*[1]*, J. Zheng*[1]*, A. Stolcke*[1,2]*, G. Tur*[1]*, D. Hakkani-Tür*[2]*, N. F. Ayan*[1]

[1] Speech Technology and Research Laboratory, SRI International, Menlo Park, CA

[2] International Computer Science Institute, Berkeley, CA

ABSTRACT

Performance of statistical machine translation (SMT) systems relies on the availability of a large parallel corpus which is used to estimate translation probabilities. However, the generation of such corpus is a long and expensive process. In this paper, we introduce two methods for efficient selection of training data to be translated by humans. Our methods are motivated by active learning and aim to choose new data that adds maximal information to the currently available data pool. The first method uses a measure of disagreement between multiple SMT systems, whereas the second uses a perplexity criterion. We performed experiments on Chinese-English data in multiple domains and test sets. Our results show that we can select only one-fifth of the additional training data and achieve similar or better translation performance, compared to that of using all available data.

Index Terms— machine translation, data selection

1. INTRODUCTION

Statistical machine translation (SMT) models [1] rely on the availability of parallel data to estimate probabilities of target language sentences given source language sentences. The "goodness" of the estimated probabilities depends on the quality, size and coverage of the parallel corpus. The conventional approach for improving performance of SMT systems is to increase the size of the parallel corpus by adding new training data. Therefore, a significant amount of work has been put into collecting source data and manually translating them into the target language. However, obtaining large amounts of high-quality parallel training data is an expensive process[1] since human involvement is needed to generate reliable translations. Therefore it is an important issue for the SMT community to be able to select the most beneficial data to be translated, and thus use the human resources in the most efficient way.

In this work, we propose two new methods to select new source data for manual translation to improve the performance of SMT systems that already achieve reasonable translation performance levels. Our first approach is motivated by active learning techniques [3], and uses a measure of disagreement between SMT systems to select the most informative training sentences to translate and add to the training corpus. The second method is similar to the n-gram coverage-based selection approach [2], and uses a perplexity measure to select sentences that seem more different from examples already observed, but robustly avoids selecting rare examples in the domain of interest. Our approaches are intended to choose the *best* source data from a pool of available source sentences, and they not rely on information from the target side. The selected data is not specific to any test set and is intended to generalize to all test sets by choosing data that is different from what is already available. Our experiments show that using either the disagreement-based approach or the perplexity-based approach, translation models trained using only one-fifth of the additional training corpus are able to achieve the same or better performance as those trained using all available additional parallel data, across multiple domains and evaluation sets.

2. RELATED WORK

Prior research on parallel data collection or data selection can be categorized into three groups: automatic data collection, domain/testset adaptation, and data selection for translation.

Under automatic data collection methods, Resnick et al. [4] identified web pages with similar content that appear in two or more languages and extracted sentences that are translations of each other from these web pages. Munteanu et al. [5] mined non-comparable corpora in two languages to identify sentences that are translations of each other. However, these automatic methods are not perfect and generate errorful parallel training data.

The most widely explored area in data selection for SMT is the adaptation of translation models or the language models to a given test set or domain. Hildebrand et al. [6] use information retrieval techniques to select parallel training data that is most similar to a given test set. Lu et al. [7] use an information retrieval system to assign weights to each training sentence pair according to their similarity to the sentences in a given test set, prior to estimating the translation model. Ittycheriah and Roukos [8] employ a data "subsampling" method by choosing the most similar sentences from the training data that have the highest n-gram overlap with a given test set. Adaptation can be performed at the level of the target language model [9, 10] by selecting data using information retrieval techniques and using it to adapt the language model.

A final group of prior work focuses on "smart" selection of new source sentences to be translated by humans to improve performance of SMT systems. To our best knowledge, the only work in this area is n-gram coverage-based data selection method proposed by Eck et al. [2], where sentences that have the highest number of previously unobserved n-grams are selected since these would have the greatest impact on coverage. In this paper, we follow the third approach and propose two new methods to select data for translation by humans. Our goal is to improve SMT systems by selecting data that is different from what is already available rather than choosing sentences that are similar to a given test set. Thus, our methods are general enough to improve performance on all test sets.

3. INTER-SYSTEM DISAGREEMENT

As mentioned in Section 1, our approach to selecting parallel training data is motivated by active learning techniques. Active learning has been used in a wide range of applications including spoken language processing [11], parsing [12], and automatic speech recognition [13]. One class of active learning approaches introduced by Seung et al. [3] uses a set of distinct learners, each trained with a

[1] Eck et al. [2] report that it costs up to 250,000 USD to translate 1 million words into English from a new language.

small number of annotated examples, to label new training examples. The training examples for which the distinct learners agree the least (in labeling) are chosen for careful labeling and used as training data for the learner. This ensures that the learner is provided with the most informative training examples. Such approaches are referred to as *query-by-committee* methods and form the basis of our work.

First, we start with a set of N different SMT systems $\mathcal{S}$, which are trained using the same (or similar) initial source language training data corpus $\mathcal{D}_i$, and use them to translate a held-out corpus $\mathcal{D}_h$. Thus, the SMT systems in $\mathcal{S}$ differ only in the algorithms they use in generating hypothesized translations. Next, we estimate the inter-system disagreement described in Algorithm 1.

Algorithm 1 Computing SMT system disagreement

```
1: Given a set of $\mathcal{S} \equiv \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$ of $N$ distinct SMT systems.
   The translation model for each system in $\mathcal{S}$ is trained using the same
   (or similar) set of initial source language sentences $\mathcal{D}_i$. Given a set of
   held-out source language sentences $\mathcal{D}_h$: $\mathcal{D}_h \nsubseteq \mathcal{D}_i$.
2: Translate the source sentences in $\mathcal{D}_h$ using each system in $\mathcal{S}$ to obtain a
   set of $N$ translations $\mathcal{T} \equiv \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$.
3: Compute the translation error $\mathcal{X} : \mathcal{X}_j \forall \mathcal{T}_j \in \mathcal{T}$ using as translation
   references the translation in each $\mathcal{T}_k$: $\forall k \in \{1, \ldots, N\}, k \neq j$. Each
   translation error $\mathcal{X}_j$ reflects the degree of disagreement between system
   $\mathcal{S}_j$ and remaining systems in $\mathcal{S}$. These error metrics are referred to as
   inter-system translation errors.
4: Also, evaluate the translation error $\mathcal{Y} : \mathcal{Y}_j \forall \mathcal{T}_j \in \mathcal{T}$ using as translation
   references the target language side of $\mathcal{D}_h$ (this step is required for
   Algorithm 2).
```

Finally, we learn a function f (linear regression in this work) that can predict the actual translation performance on new (and untranslated) source language data, $\mathcal{D}_u$, using measures of inter-system translation disagreement, as described in Algorithm 2. A threshold on the predicted translation performance on sentences in $\mathcal{D}_u$ is used to select candidate sentences for careful translations that can be used as additional parallel training data. We will refer to the $\mathcal{D}_u$ as the corpora of interest for the rest of this article.

Algorithm 2 Data selection based on predicted SMT error

```
1: Given a set $\mathcal{S}$ of $N$ distinct SMT systems (Step 1 of Algorithm 1) and a
   set of source language sentences $\mathcal{D}_u$: $\mathcal{D}_u \nsubseteq \mathcal{D}_i$. Given a set of inter-
   system translation disagreement metrics $\mathcal{X}$ and a set of actual translation
   error metric $\mathcal{Y}$ (Steps 3 and 4 of Algorithm 1 respectively)
2: Learn a function $f$:$\mathcal{X} \rightarrow \mathcal{Y}$
3: Obtain a set of translations for $\mathcal{D}_u$ using each system in $\mathcal{S}$, and inter-
   system translation disagreement metrics $\mathcal{U}$ (same as Steps 3 and 4 of
   Algorithm 1)
4: Use the function $f$ and disagreement metrics $\mathcal{U}$ to predict translation
   error $\mathcal{V}$
5: Use a threshold on predicted translation error metrics $\mathcal{V}$ to choose the
   most informative sentences from $\mathcal{D}_u$
```

4. PERPLEXITY-BASED SELECTION

We also explored an n-gram based approach for selecting new parallel corpora. This approach required the prior availability of two language models: $\mathcal{L}_i$, which is trained on source sentences in corpora $\mathcal{D}_i$, and $\mathcal{L}_{i+u}$, which is trained on source sentences in both $\mathcal{D}_i$ and $\mathcal{D}_u$. The data selection approach is described in Algorithm 3. The aim of the perplexity ratio metric α of Algorithm 3 is to avoid selecting sentences that are rare (outliers) in the corpora of interest $\mathcal{D}_u$. Sentences that are rare in both $\mathcal{D}_i$ and $\mathcal{D}_u$, will have high perplexity values (for both $\mathcal{L}_i$ and $\mathcal{L}_{i+u}$) and a low ratio α, while sentences of interest, i.e., the ones with high perplexity (and hence rare) according to $\mathcal{L}_i$ and low perplexity according to $\mathcal{L}_{i+u}$ (not rare in the domain of interest), will have high perplexity ratios. Using a pre-determined threshold on perplexity ratio α of sentences, candidates for careful translation can be selected from $\mathcal{D}_u$ and subsequently used as additional parallel training data.

Algorithm 3 Data selection based on perplexity

```
1: Given a language model $\mathcal{L}_i$ trained on the source language sentences $\mathcal{D}_i$
   and a language model $\mathcal{L}_{i+u}$ trained on source sentences in $\mathcal{D}_u$ and $\mathcal{D}_i$
2: Compute the set of perplexities $\mathcal{P}_i$ of the sentences in $\mathcal{D}_u$ using $\mathcal{L}_i$ and
   the perplexities $\mathcal{P}_{i+u}$ of the sentences in $\mathcal{D}_u$ using $\mathcal{L}_{i+u}$
3: Compute the ratio $\alpha = \mathcal{P}_i / \mathcal{P}_{i+u}$ for each sentence in $\mathcal{D}_u$. Use a thresh-
   old on $\alpha$ to choose informative sentences from $\mathcal{D}_u$
```

4.1. Hybrid Selection

To leverage potential benefits of complementary data sets selected by the inter-system disagreement-based and perplexity methods, we considered an approach that combines the two. In this approach, we select equal amounts of additional training data using each of the previous approaches.

5. SYSTEMS AND DATA SETS

For this work, we focused on Mandarin-to-English translation and used $N = 3$ different SMT systems, each developed independently and achieving competitive translation performance in recent NIST benchmark evaluations. System $\mathcal{S}_1$ [14] is based on hierarchical phrase-based translation [15]. System $\mathcal{S}_2$ [16] and System $\mathcal{S}_3$ [17] are phrase-based translation systems. For all three systems both translation models and target language models were trained on Mandarin-to-English parallel newswire and web-text data sets released by LDC prior to 2005 comprising about 29 million English words.These sets correspond to the set $\mathcal{D}_i$. The corpora of interest for data selection, $\mathcal{D}_u$ consisted of the Mandarin-to-English parallel data released by LDC after 2005, also including both newswire and web text sources, and comprised approximately 16 million English words.Since the focus of this study was selecting training data for translation models, we kept the target-language model fixed in all experiments. It was trained using the English side of the corresponding parallel data of $\mathcal{D}_i$ corpora, the English Gigaword third edition corpus comprising 3.62 billion words with a vocabulary size 163,941. It used modified Kneser-Ney smoothing and was trained using the SRILM toolkit [18].

The proposed data selection methods were evaluated by retraining one of the three systems, $\mathcal{S}_1$ and translating four test sets in use by the GALE program (number of English words in parentheses): Dev 2007 Newswire (18,784), Dev 2007 Web Text (17,556), Eval 2006 Newswire (10,565), and Eval 2006 Web Text (9,953).The log-linear optimization [19] of the translation models was performed using held-out data drawn from Dev 2007 Newswire and Dev 2007 Web Text using SRILM tools. All systems are evaluated with the BLEU metric using a single reference.

6. BASELINE SYSTEMS

Our data selection pool $\mathcal{D}_u$ consisted only of parallel corpora released by LDC. Since $\mathcal{D}_u$ has been translated manually, we were able to evaluate the effect of including all of $\mathcal{D}_u$ as training data in addition to $\mathcal{D}_i$. The results, shown in Table 1, can be considered an approximate upper bound on potential performance improvements from data selection.[2]

[2] Assuming that translation performance does not improve by excluding all available data. In practice this might not be true if "bad" or unrepresentative data is included.

	$\mathcal{D}_i$	$\mathcal{D}_u + \mathcal{D}_i$
Dev 2007 Newswire	15.2	16.4
Dev 2007 Web Text	12.6	13.5
Eval 2006 Newswire	16.9	17.2
Eval 2006 Web Text	12.7	13.7

Table 1. Potential lower and upper bounds from including additional parallel data (numbers are BLEU scores)

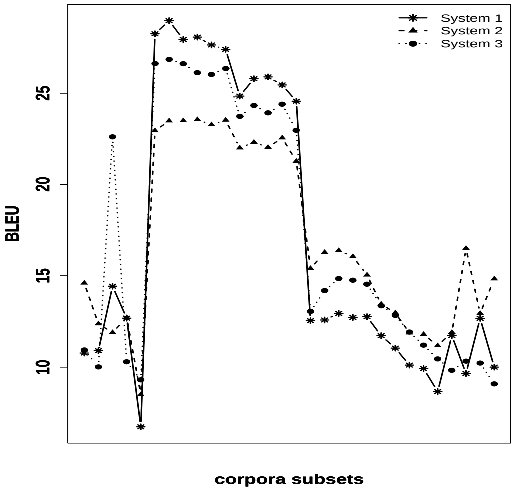

Fig. 1. Comparison of SMT systems when translating $\mathcal{D}_u$(Each point corresponds to 15000 sentences)

7. DATA SELECTION EXPERIMENTS

The next step was to apply the various data selection methods and evaluate their impact on SMT performance. We first describe our experimental configuration for each method.

7.1. Inter-System Disagreement

First, we used the three systems in $\mathcal{S}$ to translate the sentences in $\mathcal{D}_u$. The translation performance of each of the three systems is shown in Figure 1. As can be seen, the performance of the three systems is quite similar across the different subsets of $\mathcal{D}_u$, though they disagree, which allows us to learn the function f (Step 2 of Algorithm 2) that predicts translation performance based on inter-system disagreement. Since data selection is performed at the sentence level (and BLEU is not robust for individual sentences), we decided to use translation error rate (TER) [20] as the target variable to estimate for f. We applied Algorithm 1 by using a rotating subset of $\mathcal{D}_u$ as our held-out set $\mathcal{D}_h$ for estimating f, and then applying f on the complement subset, in a jack-knifing fashion. The function f learned was a linear regression over two input variables: the sentence level inter-system TER and the inter-system BLEU score. The linear regression function thus estimated performed better than predicting the mean sentence level TER of $\mathcal{D}_u$ by 10% relative in root mean square error,[3] which was statistically significant at the level of $p = 10^{-5}$. Since we wanted to select certain targeted amounts of data we did not perform selection by a given threshold (Step 5 of Algorithm 2). Instead, we ranked all sentences by their predicted TER and chose the sentences with the highest TER up until the desired amount of data was reached.

7.2. Perplexity-based Selection

For the perplexity-ratio-based selection approach, we needed two source-side (Mandarin) language models, $\mathcal{L}_i$ and $\mathcal{L}_{i+u}$, as described in Section 4. $\mathcal{L}_i$ was trained on 33 million words with a vocabulary size of 60,453, and $\mathcal{L}_{i+u}$ was trained on 55 million words with a vocabulary size of 93,134, both using modified Kneser-Ney smoothing. We followed the steps outlined in Algorithm 3 for selecting parallel training data. Similar to agreement-based selection, all sentences in $\mathcal{D}_u$ were ranked according to their perplexity ratios, and the sentences with the highest values were chosen up to the desired training corpus size.

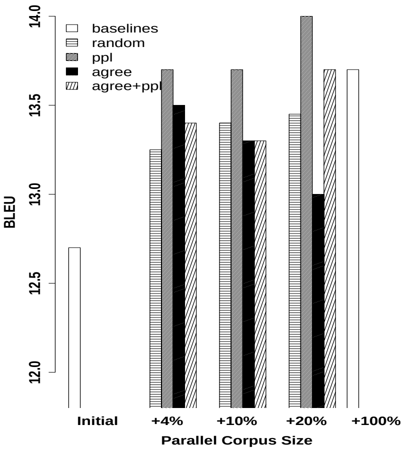

Fig. 2. BLEU scores for Eval2006 webtext

7.3. Hybrid Selection

For the combined selection method, equal amounts of data (in terms of number of English words) were selected using the agreement-based and the perplexity-based approaches. Sentences selected by both methods were included only once in training.

7.4. Random Selection

This approach involves selecting source language sentences from the corpora of interest $\mathcal{D}_u$ at random as candidates for careful translation and subsequent inclusion in the parallel training corpus. This is equivalent to the chance condition and allows us to determine the effectiveness of our proposed approaches (Section 3 and 4). For the purpose of estimating robust translation performance metrics, random sampling of parallel training data was repeated several times (four in our experiments) and a translation model was built using each sample. The translation performance metrics were then averaged over all random selection trials to obtain the final performance estimate.

7.5. Results

Using the four data selection approaches, we selected varying amounts of additional parallel training data[4]to add to our existing hierarchical phrase-based SMT system. For each training set, we recomputed phrase alignments, retrained the translation models, and reestimated the parameters for log-linear model combination, while keeping the target language model fixed. We compared three sizes of additional data: 4%, 10% and 20% of the available selection pool $\mathcal{D}_u$, corresponding to approximately 640,000, 1.6 million, and 3.2 million English words, respectively. The results for the four evaluation sets are shown in Figures 2 and 3. Each figure includes the result with only initial data set (“Initial”) and with all additional data (“+100%”). Results with increasing intermediate amounts of additional data are shown from left to right. Different shading patterns correspond to different selection methods. The hybrid approach is denoted “agree+ppl”.

At 20% additional data, one or more of the proposed selection methods perform better than the putative upper bound (“+100%”),

[3] This quantifies the margin by which the linear regression is better than predicting chance.

[4] Manual translations of selected sentences are used in training.

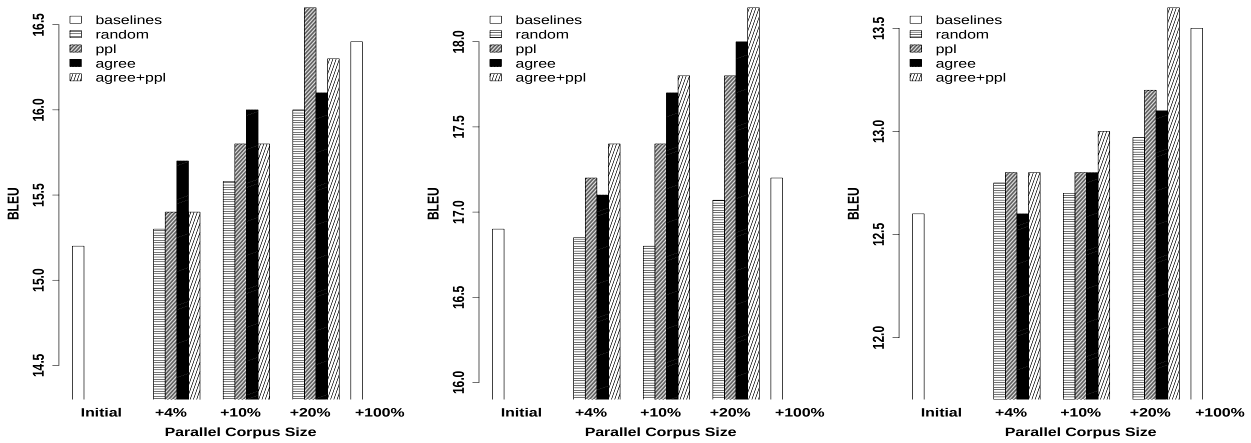

Fig. 3. BLEU scores for (from left to right): Dev2007 newswire, Eval2006 newswire, and Dev2007 webtext

for all four test sets. The perplexity-based approach performs best for Dev 2007 Newswire (by 0.2 BLEU over "+100%") and for Eval2006 Web Text (by 0.3 BLEU over "+100%"). Hybrid selection performs best for Eval 2006 Newswire (by 1.0 BLEU over "+100%") and Dev2007 Web Text (by 0.1 BLEU over "+100%"). In addition, it can be observed that with only 10% of additional data, one of the proposed methods still outperforms the all-data system for two of the four evaluation sets. The hybrid method performs best for Eval 2006 Newswire (by 0.6 BLEU over "+100%") and perplexity-based selection matches the performance of the "+100%" case for Eval 2006 Web Text.

The pattern in these results, achieving better performance with fewer training samples, agrees with previously reported results using active learning techniques by Tür et al. [11]. When comparing the different selection methods, a broad generalization that emerges is that the perplexity-based method works better on web text, while the disagreement-based method works better on newswire. As might be expected, the hybrid approach avoids problems in cases where one of the methods alone has relatively poor results; in some cases it gives substantial gains over each of the simpler methods.

8. CONCLUSIONS AND FUTURE WORK

We have proposed three approaches for selecting parallel training data for translation models. The first approach is based on using a measure of inter-system translation disagreement to select data that is most difficult for existing SMT systems. The second approach uses a source-side language model perplexity measure to select sentences that have novel information while avoiding outliers with respect to the domain of interest. The third approach combines the first two by selecting equal amounts of data using both methods. In our experiments, these approaches are able to select 20% of available data to achieve Mandarin-to-English translation accuracy that is similar or better than training on the full data set.

The inter-system translation disagreement-based approach suggests some interesting research directions. In particular, we intend to retrain the distinct SMT systems (used for measuring disagreement) with the additional parallel training data (selected using our approach) and using the retrained systems for further data selection. This procedure would be closer to traditional active learning techniques.

Acknowledgments: This work was funded by DARPA under contract No. HR0011-06-C-0023 (approved for public release, distribution is unlimited). The views herein are those of the authors and do not reflect the views of the funding agencies.

9. REFERENCES

[1] P. F. Brown et al., "A statistical approach to machine translation," *Computational Linguistics*, vol. 16(2), pp. 79–85, 1990.

[2] M. Eck, S. Vogel, and A. Waibel, "Low cost portability for statistical machine translation based on n-gram coverage," in *MT Summit X*, 2005, pp. 227–234.

[3] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Computational Learning Theory*, 1992, pp. 287–294.

[4] P. Resnik and N. A. Smith, "The web as a parallel corpus," *Computational Linguistics*, vol. 29, no. 3, pp. 349–380, 2003.

[5] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting comparable corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[6] A. S. Hildebrand et al., "Adaptation of the translation model for statistical machine translation based on information retrieval," in *Proc. of EAMT*, 2005, pp. 133–142.

[7] Y. Lu, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization," in *Proc. of EMNLP-CoNLL*, 2007, pp. 343–350.

[8] A. Ittycheriah and S. Roukos, "Direct translation model 2," in *Proc. of HLT/NAACL*, 2007, pp. 57–64.

[9] B. Zhao, M. Eck, and S. Vogel, "Language model adaptation for statistical machine translation with structured query models," in *Proc. of COLING*, 2004.

[10] M. Eck, S. Vogel, and A. Waibel, "Language model adaptation for statistical machine translation based on information retrieval," *Proc. of LREC*, pp. 327–330, 2004.

[11] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Journal of Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[12] R. Hwa, "Sample selection for statistical parsing," *Computational Linguistics*, vol. 30, no. 3, 2004.

[13] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. of ICASSP*, 2002.

[14] J. Zheng, W. Wang, and N. F. Ayan, "Development of SRI's translation systems for broadcast news and broadcast conversations," in *Proceedings of Interspeech*, 2008.

[15] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[16] LanguageWeaver, "Language Weaver Inc's Chinese-to-English MT system, release v4.0.," `http://www.languageweaver.com`, 2005.

[17] R. Zens, *Phrase-based Statistical Machine Translation: Models, Search, Training*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2008.

[18] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002, pp. 901–904.

[19] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of ACL*, 2003, pp. 160–167.

[20] M. Snover et al., "A study of translation edit rate with targeted human annotation," in *Proc. of AMTA*, 2006.