The Report of NLP^{2}CT for CWMT 2015 Evaluation Task

Yiming Wang, Derek F. Wong, Lidia S. Chao, Ben Ao, Connie Y. Tong, Yi Lu
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory,
Department of Computer and Information Science,
University of Macau, Macau S.A.R., China
wang2008499@gmail.com, {derekfw,lidiasc}@umac.mo
{nlp2ct.ben,yutongconnie,takamachi660}@gmail.com

Abstract

This paper presents our work in the evaluation task of the 11th China Workshop of Machine Translation (CWMT). The MT system is a phrase-based system, and built based on Moses. In this task, all the translation systems were constructed based on our data selection method using recursive neural network (RNN). We evaluated different selection methods and submitted three language pairs systems: Chinese-to-English (CE), English-to-Chinese (EC) and double blind (DB).

1 Introduction

CWMT aims to provide a platform for the communications and academic exchanges in the field of machine translation (MT), with a special emphasis on the Chinese language. The evaluation task is held every two years. This year, the evaluations focus on six language pairs including Chinese-to-English, English-to-Chinese, Mongol-to-Chinese, Tibet-to-Chinese, Uygur-to-Chinese and *Double Blind*. In this report, we present our submitted MT systems on English-to-Chinese, Chinese-to-English, and the Double Blind translation. The Double Blind translation task is a new trial in the CWMT, where both the source and target languages are encoded into unknown languages pair. The parallel training and development data are provided in form of sentences and parse structures, to accommodate both phrase-based and syntax-based translation paradigms. An encoded monolingual corpus for target language is also provided for training the language model (LM). The objective of this evaluation aims to assess the performance of the core model in a more focus way. The participants can put more attention on the design of the translation model itself rather than any language dependent pre- and post-processing.

We carefully inspected all of the concerned corpora, in particular the Chinese-English and English-Chinese data. A series of preprocessing steps are conducted. After processed, the number of Chinese and English sentences are nearly 8 million. Constructing translation models using such a large corpus is not only time consuming, but also costs a lot of computational resources, in particular, when we experimented with different model settings and tunings. To cope with this, our strategy is to use a small but relevant data for constructing the translation systems. Data selection (Wang et al., 2014a; Axelrod et al., 2011) is an ideal approach. We not only can save time and resource in building multiple models, but also can obtain a comparable or even better models. Most importantly, we gain more time in spotting the problems and determining the model with the best parameters. Different from previous works (Lu et al., 2014a; Wang et al., 2013) that relied on some amount of in-domain parallel data (i.e. tens of thousands parallel sentences), in this evaluation task we used the official test set as the *in-domain* data which is monolingual and consists of 1,100 sentences only. While the training data is treated as *general-domain* data. We participate the constrained track of the evaluation. All the resources used in building the MT systems are constrained to training data provided by the organizer.

The rest of this report is as follows: Section 2 presents the background and the data selection methods investigated in this work. Section 3 describes the construction details of MT models for three language pairs. Section

4 gives the official results, followed by a conclusion to end this paper.

2 Domain Adaptation and Data Selection

The state-of-the-art phrase-based machine translation uses data-driven method to build the translation system. According to this scenario, the more the training data it uses, the better the model it is. But the problem is that the performance of statistical machine translation (SMT) system is not only relevant to the quantity of training data, but also depends on the quality of data. In reality, it is very difficult to collect a large amount of training data which is highly relevant to the domain of text to be translated. On the other hand, small amount of domain specific data is easier to collect and construct, but which is far from enough for training a good translation model due to insufficient of the data. This establishes a research direction of data selection for domain adaptation in the context of SMT.

In domain adaptation, training data is treated as general-domain data. The general-domain data is usually considered as a large corpus that embraces data which is relevant to the domain of target text, i.e. in-domain data. Given a small in-domain data, data selection is to extract domain specific sentences from the general-domain corpus by taking the given in-domain data as reference. The extracted content is used together with the in-domain data for building a domain specific translation system. In data selection, there are two application scenarios. The first application is to enlarge the content of in-domain data by adding relevant parallel sentences selected from a rich data set, e.g. crawled from online sources. While in the second application, instead of using a large and (target domain) irrelevant data, data selection model is used to filter out such irrelevant sentences and preserve only the consistent and relevant data. Both of those application objectives aim to produce a better translation system. Our work in this evaluation belongs to the latter case.

Cross-entropy is widely used in the study of data selection methods. Moore and Lewis (2010) consider the cross-entropy of source sentences with respective to an in-domain language model, LM_I, and the non-specific-domain language model, LM_N. The sentences are scored based on the difference of cross-entropies. The sentences whose score are less than a threshold T are considered as domain specific data and selected.

$$T = H_I(s) - H_N(s) \tag{1}$$

where $H_I(s)$ and $H_N(s)$ are the cross-entropies of sentence, s, according to the in-domain language model, LM_I, and non-specific-domain language model, LM_N, respectively.

Axelrod et al. (2011) extend the work to bilingual application. They score a parallel sentence pair by taking the sum of cross-entropy differences from both the source and target sentences, according to the following equation:

$$\begin{aligned} T = &[H_{I-src}(s) - H_{N-src}(s)] \\ &+[H_{I-tgt}(s) - H_{N-tgt}(s)] \end{aligned} \tag{2}$$

Duh et al. (2013) use a neural language model (Bengio et al., 2003) instead of conventional n-gram model. He observed that a large amount of words are not seen from the general-domain sentences according to the in-domain data, and more than 60% sentences contain at least one unknown word. Although this problem can be solved by using back-off and other smooth techniques, the neural language model provides a better way in smoothing the probabilities of unseen words but similar context. However, these methods based on language models suffer from a problem that they do not take the syntactic and semantic information of sentences into consideration. Especially in the context of data selection, such linguistic features of sentences are the vital clues. In this work, we design a data selection model based on recursive neural network (Lu et al., 2014b). The model is first trained by using both *in-domain* and *out-domain* data for inducing the representation of sentences. Then a logistic regression classifier is constructed from these sentence representations and domain label. Finally, the recursive neural network model and the classifier are used for generating the representations of sentences and scoring. The sentences whose score are higher than a threshold T are selected as relevant data.

3 Experiments

3.1 Chinese and English Translations

In this section, we describe the details in constructing the translation systems for Chinese-to-English and English-to-Chinese.

3.1.1 Data Preparation

As mentioned, phrase-based machine translation is data-driven. So careful preparation for training data may lead to very significant improvements in translation quality. Considering the difference between Chinese and English, a series of preparation steps are carried out for these two languages separately.

Tokenization: For English, we use the `Moses` (Koehn et al., 2007) accompanied scripts to tokenize the text. Tokenization separates the words from punctuation and change some special symbols to escape codes which can be recognized by the decoder, for example, converting apostrophe "′" to "*&apos*". Another adjustment is the truecasing. The intention of this is to prevent the word alignment model from generating spurious candidates caused by initial capital of words. That may affect the alignment distributions learned from the model. Take the two English sentences for example, *I want to buy some books*, and *Books are our friends*, the word *books* in these sentences refers to the same meaning 书. Because the initial capital of the second sentence, two possible alignments, *books* ↔ 书 and *Books* ↔ 书, are produced. The toolkit we used is accompanied with the `Moses`.

Word Segmentation: For Chinese, the tokenization falls into the segmentation of text. Empirical experiments show that different Chinese segmentation models lead to different translation quality (Zeng et al., 2013; Zeng et al., 2014). In this task, we try different off-the-self segmentation models including ANSJ[1], Stanford Tokenizer (Chang et al., 2008) and our SMT-based Chinese Word Segmentation model (Zeng et al., 2014), which takes the word alignment information of the training data into consideration. Please refer to Zeng et la. (2014) for more details.

Deduplication: According to observation, duplicate sentences are found in both Chinese and English data. This may lead to unpredictable result of the model, so we remove all the duplicates from the training data, including the sentences pair caused by case (case-insensitive criteria).

Long Sentences Removal: The last step is deleting sentences limited with maximum length, as long sentences will degrade the alignment quality. This year, the organizer does not offer the baseline system and configurations for reference, we empirically set the maximum sentence length to 40 and 50, respectively, and included it in our experiments. Table 1 shows the statistics of the training data. SegSMT stands for our SMT-based Chinese Word Segmentation model. We prepared 8 different sets of training corpora, according to the different Chinese word segmentation models and the maximum sentence length constraint. We found that more (i.e. around ten thousand) sentences were discarded in using ANSJ segmentation model, compared with other segmentation methods, showing that more single Chinese characters are being recognized as words by ANSJ tool.

3.1.2 Language Model

Language model (LM) is an essential part to a translation system, whether a LM is good or not plays a vital role in hypothesis decoding (translation). So preparation of training data is important. The data we used for training the LM includes the target sentences of the parallel corpus, and the monolingual data provided by the CWMT organizer. We adopted the same preprocessing steps as described in § 3.1.1, except the step of constraining the maximum sentence length. For English, we conducted the text tokenization and truecasing. While for Chinese, we used $\text{Stanford}_{\text{PKU}}$ word segmentation, as we found that with the given training data, the Stanford Tokenizer trained on Peking University Corpus (Yu et al., 2003) exhibits highest BLEU score among the four Chinese-to-English translation systems. Table 2 shows the statistic of LM model data.

In the table, **Len.** stands for the average sentence length. It is noted that the average sentence length of Chinese text is nearly double the English ones. But, there is no any correlation with the size of Chinese and English data, where the number of Chinese sentences

[1] `https://github.com/NLPchina/ansj_seg`

Tokenizer	Language	Length	Sentences	Tokens	Average Length
ANSJ	CH	1-40	6,765,089	104,840,261	15.49
		1-50	7,003,963	114,504,506	16.34
	EN	1-40	6,765,089	106,671,600	15.76
		1-50	7,003,963	116,670,472	16.65
SegSMT	CH	1-40	6,790,063	96,641,697	14.23
		1-50	7,015,937	105,046,355	14.97
	EN	1-40	6,790,063	107,581,354	15.84
		1-50	7,015,937	117,250,116	16.71
Stanford$_{\mathbf{CTB}}$	CH	1-40	6,782,361	102,168,309	15.06
		1-50	7,012,291	111,279,369	15.87
	EN	1-40	6,782,361	107,256,599	15.81
		1-50	7,012,291	117,019,697	16.69
Stanford$_{\mathbf{PKU}}$	CH	1-40	6,778,457	102,860,336	15.17
		1-50	7,010,847	112,062,894	15.98
	EN	1-40	6,778,457	107,125,133	15.80
		1-50	7,010,847	116,960,760	16.68

Table 1: Statistic Summary of the Prepared Training Data

is twice as many as English.

3.1.3 Experiment Setup

In the training pipeline, Fast Align (Dyer et al., 2013) with grow-diag-final-and was used to generate the word-to-word alignments. The 5-gram LM models were built using the SRILM toolkit (Stolcke and others, 2002), smoothed with improved Kneser-Ney discounting. The lexicalized reordering model was trained by using msd-bidirectional-fe. The weighting parameters of the model were tuned via minimum-error-rate (Och, 2003) on the provided development set. We used BLEU (Papineni et al., 2002) as the metric to evaluate the performance of all constructed systems.

3.1.4 Evaluation and Analysis

To evaluate various methods, models and configurations, we implemented the experiemnt on Chinese-to-English language pair and randomly selected 1,000 Chinese sentences from the training data as the test set.

Baseline System: The baseline model was built on the original and unprocessed training data. That is we did not do any preprocessing to the data, i.e. the steps described in § 3.1.1. For Chinese segmentation, we employed the Stanford$_{\text{ctb}}$ tokenizer that trained on Penn Chinese Treebank (CTB) (Xue et al., 2005). The total number of sentences of the training data is 7,780,333. We want to see the impact of data preparation work for the final result.

Lang.	Sent.	Tokens	Len.
English	7,885,206	131,350,289	16.66
Chinese	15,663,416	487,604,950	31.13

Table 2: Statistics of LM Training Data

Segmentation Models: In the first experiment, we would like to explore how the segmentation of Chinese text affects the end-to-end translation performance of the trained models. In this evaluation, we carried out the data preprocessing procedures as described in § 3.1.1. Two sets of training data were prepared by considering the maximum sentence length, $MaxLen = 40$ and $MaxLen = 50$. For each of which, four different segmentation models were applied, and a translation system was trained on it. Table 3 presents the performance of the models, $SYS_{\langle * \rangle - \langle + \rangle}$, where $\langle * \rangle$ represents different segmentation models, and $\langle + \rangle$ indicates the cutoff threshold of the maximum sentence length.

From the results, we observed that most of the models gain improvements over the baseline on the development data, except the models of $SYS_{SegSMT-\langle + \rangle}$. The probable reason of this drop is due to the training data and the induced word alignments, which we used for training the segmentation model. In the fu-

ture, we will look into this problem. Regarding the test data, only three of the models exhibited statistically significant improvements over the baseline model. Among the evaluated systems, the $SYS_{StanPKU-50}$ achieved the best performance, giving a BLEU point gain of 1.48 and 1.89 on development and test data over the baseline system, we also extend the limited sentence length to 60, but result is not better than that of 50.

Model	BLEU (dev)	BLEU (test)
Baseline	23.86	31.59
$SYS_{ANSJ-40}$	24.97	30.20
$SYS_{SegSMT-40}$	22.33	29.85
$SYS_{StanPKU-40}$	**25.17**	31.38
$SYS_{StanCTB-40}$	24.92	**32.64**
$SYS_{ANSJ-50}$	24.68	31.96
$SYS_{SegSMT-50}$	22.36	32.30
$SYS_{StanPKU-50}$	**25.34**	**33.48**
$SYS_{StanCTB-50}$	25.31	33.43

Table 3: The Performance of Data Preparation

Data Selection: In this section, we compared the data selection methods on the task of translation based on the setting of $SYS_{StanPKU-50}$, that we described in § 2. This includes the cross-entropy (CrEn) model of Moore and Lewis (2010) with the same LM settings as § 3.1.3, neural LM (nLM) model of Duh et al. (2013) and the recursive neural network (RNN) model of Lu et al. (2014b). In determining the selection threshold, we experimented with different selection rates from 20% to 80% of the training data with a step of 20%. We wanted to investigate the translation performance of systems that trained on different size of the data. The results are presented in Table 4. From the evaluations, it displays our findings. First, we confirm the result of Duh et al. (2013), that using a recurrent neural language model to replace the conventional n-gram model improves the translation performance. Second, our proposed recursive neural network data selection model outperforms other models in all data sets. These findings can be witnessed from the translation results that evaluated on the development and test data, as shown in Table 4.

Model	BLEU(dev)	BLEU(test)
Baseline	23.86	31.59
$SYS_{StanPKU-50}$	25.34	33.48
$SYS_{CrEn-20}$	21.63	26.18
SYS_{nLM-20}	21.89	26.30
SYS_{RNN-20}	**22.66**	**26.46**
$SYS_{CrEn-40}$	22.95	29.48
SYS_{nLM-40}	23.30	29.42
SYS_{RNN-40}	**24.09**	**29.6**
$SYS_{CrEn-60}$	23.62	31.89
SYS_{nLM-60}	23.73	31.03
SYS_{RNN-60}	**24.54**	**32.15**
$SYS_{CrEn-80}$	24.28	32.96
SYS_{nLM-80}	24.72	33.29
SYS_{RNN-80}	**25.19**	**33.50**

Table 4: Translation Results of Systems Trained on Selected Data

Normalization of Punctuation: During the development, we discovered the English tokenizer produced the special words as Figure 1 shows, it seriously affects the quality of word alignment model, and as a consequence, this affection propagates to the subsequent processes and impacts the overall translation quality. We toke a closer look at the problem

The Original Sentence1:
I **wasn't** aware you are having a session .
Sentence1 after tokenization:
I **wasn鈥檛** aware you are having a session .
The Original Sentence2:
we **didn't** tell him he probably would never make the team , so he **didn't** know .
Sentence2 after tokenization:
we **didn鈥檛** tell him he probably would never make the team , so he **didn鈥檛** know .

Figure 1: Problem of Punctuation

and found that the *single quotation* and *apostrophe* are incorrectly used (and interchangeably used) through the data. It causes the tokenizer unable to correctly distinguish the punctuation in the context, and resulted in turning it into a special word. We addressed this issue by normalizing the punctuation to the right practice. We then trained a model using the same data as SYS_{RNN-80} after normalization, and obtained a statistically significant improvement with a gain of 1.83 and 2.49 BLEU points on the development and test data compared with baseline system. Table 5

presents the results of the best models concluded from each of the experiments.

Model	BLEU (dev)	BLEU (test)
Baseline	23.86	31.59
$SYS_{StanPKU-50}$	25.34	33.48
SYS_{RNN-80}	25.19	33.50
$SYS_{PunNorm}$	**25.69**	**34.08**

Table 5: The Performance of Selected Models

$SYS_{PunNorm}$ has the best performance among all models, so we adopt the same setting for our Chinese-to-English submitted system, because all training and develop data for Chinese-to-English and English-to-Chinese are the same, we use the same setting for English-to-Chinese submitted system.

3.2 Double Blind Translation

This section presents the details in implementing the systems for the *Double Blind* translation task.

3.2.1 Data Preparation

In Double Blind translation task, all words including punctuation are encoded into *ciphertext* which are unreadable. As the affection of punctuation and initial capital is not taken into consideration any more, tokenization and truecasing are unnecessary. According to the experiments of Chinese to English translation, the length of sentences and duplication are the factors that we concerned about, the same preparation work as § 3.1.1 is adopted in order to limit sentence length and remove duplicated sentences. Table 6 shows the statistic of all training data after data preparation.

The total number of training corpus is 2,000,000, from this table we can see that 17.5% sentences which are longer than 40 tokens, and nearly 10% are longer than 50 That means long sentences still constitute a relatively large proportion of the training corpus.

3.2.2 Language Model

The same as Chinese to English translation task, the data we used for training the LM includes the target language sentences and the monolingual data provided by the CWMT organizer. Duplicated sentences are removed as they may lead to unpredictable result. After processing, the size of data for training the LM model is 10,070,854.

3.2.3 Evaluation and Analysis

For this double blind evaluation task, we adopt the same experiment settings as Chinese to English translation task. Due to the lack of reference in official test set, we randomly extract two test sets which contain 1000 sentences separately, and the remaining training corpus is our final training corpus.

Baseline System: The baseline system is built using all original training corpus without any pre-processing work. The same as Chinese to English translation, we want to see if the data preparation can impact the translation performance.

Limited Sentence Length: Long sentences may cause unpromising word alignment. In this section, we carried out more detailed experiments on limiting sentence length. Six SMT systems are trained on different training corpus which are filtered by various thresholds of maximum sentence length (from 30 to 80). Their BLEU scores are shown in Table 7.

In the column of model, the '+' of SYS_{+} stands for the longest sentence length in the training corpus. Limiting the maximum sentence length is helpful in improving the translation performance. However, when the threshold is 30, the result becomes nearly the same as baseline. The reason may be that it removed a large amount of sentence (about 30%).

Data Selection: The same as Chinese to

Length	Sentences	Lang.	Token
1-30	1,393,797	Source	25,174,482
		Target	26,685,038
1-40	1,670,239	Source	33,678,437
		Target	35,565,856
1-50	1,800,391	Source	38,597,509
		Target	41,094,934
1-60	1,879,521	Source	42,165,777
		Target	45,229,548
1-70	1,931,563	Source	44,905,359
		Target	48,436,555
1-80	1,965,383	Source	46,942,536
		Target	50,834,982

Table 6: The Statistic of Training Data

Model	BLEU(test1)	BLEU(test2)
Baseline	17.48	16.55
SYS_{30}	17.54	16.58
SYS_{40}	**18.26**	**17.23**
SYS_{50}	17.76	16.81
SYS_{60}	17.65	16.87
SYS_{70}	17.70	16.73
SYS_{80}	17.68	16.79

Table 7: The Performance of Data Preparation

English translation task, we also use data selection method for this task. According to Section 3.1 and 3.2.3, we built SMT systems by combining the best data selection model and max sentence length threshold found in previous experiments, which are RNN and 40. Table 8 shows the result of the data selection methods with different selection rates vary from 20% to 80% with a step of 20%.

Model	BLEU(test1)	BLEU(test2)
Baseline	17.48	16.55
$SYS_{RNN\text{-}20}$	16.18	15.66
$SYS_{RNN\text{-}40}$	16.95	16.38
$SYS_{RNN\text{-}60}$	17.47	16.65
$SYS_{RNN\text{-}80}$	**17.72**	**16.76**

Table 8: The Performance of Data Selection

The '+' of $SYS_{RNN\text{-}+}$ stands for the selecting rate. From this table we can see that when selection rate is below 60%, the result is poor because the training corpus is too small to train a decent model. The system with selection rate of 80% does outperform the baseline. However, its score is still lower than that of SYS_{40}. It indicates that for the small training corpus, the data selection may not be a good choice.

4 Official Result and Conclusion

This paper discusses our work of the evaluation task of Chinese and English transaltion, double blind transaltion. Experiments show that a careful data preparation work can improve the performance of translation. We also found that data selection is an efficient method to select high quality data and achieve a comparable result to the systems using all training corpus. Finally, we submit the result of official test set by using the same setting as $SYS_{PunNorm}$ for Chinese-to-English and English-to-Chineses, and the same setting as $SYS_{RNN\text{-}80}$ for Double Blind.

Langauge	Metric	Score
Chinese-English	BLEU4	21.60
English-Chinese	BLEU5	36.00
Double Blind	BLEU4	18.88

Table 9: The Performance of Official Result

For Chinese to English and English to Chinese transaltion, both our systems rank 2_{rd} among six and four teams, while for double blind, the result is not good, the reason may be that data selection hardly impacts the result while the whole training corpus is relatively small. In this evaluation task, we also tried some other method, such as using neural LM model instead of traditional model, doing preordering and re-ranking(Wang et al., 2014b), but these methods didn't improve the result. In the future, we may try building an efficient LM model by applying data selection methods.

Acknowledgements

The authors are grateful to the Science and Technology Development Fund, Macao S.A.R. (FDCT) and the Research Committee of the University of Macau for the funding support for our research, under the reference No. 057/2014/A, MYRG2015-00175-FST and MYRG2015-00188-FST.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation via Pseudo In-domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chi-

nese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation. In *ACL (2)*, pages 678–683.

Chris Dyer, Victor Chahuneau, and Noah Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–649, June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi Lu, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Yiming Wang. 2014a. Domain adaptation for medical text translation using web resources. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 233–238, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Yi Lu, Derek F. Wong, Lidia S. Chao, and Longyue Wang. 2014b. Data Selection via Semi-supervised Recursive Autoencoders for SMT Domain Adaptation. In *Machine Translation*, pages 13–23. Springer.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Stolcke and others. 2002. SRILM-an extensible language modeling toolkit. In *INTERSPEECH*.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Junwen Xing, Yi Lu, and Isabel Trancoso. 2013. Edit Distance: A New Data Selection Criterion for Domain Adaptation in SMT. In *RANLP*, pages 727–732.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing. 2014a. A Systematic Comparison of Data Selection Criteria for SMT Domain Adaptation. *The Scientific World Journal*, 2014:e745485, February.

Yiming Wang, Longyue Wang, Derek F. Wong, and Lidia S. Chao. 2014b. Effective Hypotheses Re-ranking Model in Statistical Machine Translation. In Xiaodong Shi and Yidong Chen, editors, *Machine Translation*, number 493 in Communications in Computer and Information Science, pages 24–32. Springer Berlin Heidelberg, November.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Shiwen Yu, Huiming Duan, Xuefeng Zhu, Bin Swen, and Baobao Chang. 2003. Specification for corpus processing at peking university: Word segmentation, pos tagging and phonetic notation. *Journal of Chinese Language and Computing*, 13(2):121–158.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013. Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *ACL (1)*, pages 770–779.

Xiaodong Zeng, Lidia S. Chao, Derek F. Wong, Isabel Trancoso, and Liang Tian. 2014. Toward better chinese word segmentation for smt via bilingual constraints. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1360–1369, Baltimore, Maryland, June. Association for Computational Linguistics.