English to Chinese Translation: How Chinese Character Matters?

Rui Wang[1,2], **Hai Zhao**[1,2] *† and **Bao-Liang Lu**[1,2]
[1]Center for Brain-Like Computing and Machine Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
wangrui.nlp@gmail.com and {zhaohai, blu}@cs.sjtu.edu.cn

Abstract

Word segmentation is helpful in Chinese natural language processing in many aspects. However it is showed that different word segmentation strategies do not affect the performance of Statistical Machine Translation (SMT) from English to Chinese significantly. In addition, it will cause some confusions in the evaluation of English to Chinese SMT. So we make an empirical attempt to translation English to Chinese in the character level, in both the alignment model and language model. A series of empirical comparison experiments have been conducted to show how different factors affect the performance of character-level English to Chinese SMT. We also apply the recent popular continuous space language model into English to Chinese SMT. The best performance is obtained with the BLEU score 41.56, which improve baseline system (40.31) by around 1.2 BLEU score.

*Correspondence author.

†Thank all the reviewers for valuable comments and suggestions on our paper. This work was partially supported by the National Natural Science Foundation of China (No. 61170114, and No. 61272248), the National Basic Research Program of China (No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (No. 13511500200), the European Union Seventh Framework Program (No. 247619), the Cai Yuanpei Program (CSC fund 201304490199 and 201304490171), and the art and science interdiscipline funds of Shanghai Jiao Tong University, No. 14X190040031, and the Key Project of National Society Science Foundation of China, No. 15-ZDA041.

1 Introduction

Word segmentation is necessary in most Chinese language processing doubtlessly, because there are no natural spaces between characters in Chinese text (Xi et al., 2012). It is defined in this paper as character-based segmentation if Chinese sentence is segmented into characters, otherwise as word segmentation.

In Statistical Machine Translation (SMT) in which Chinese is target language, few work have shown that better word segmentation will lead to better result in SMT (Zhao et al., 2013; Chang et al., 2008; Zhang et al., 2008). Recently Xi et al. (2012) demonstrate that Chinese character alignment can improve both of alignment quality and translation performance, which also motivates us the hypothesis whether word segmentation is not even necessary for SMT where Chinese as target language.

From the view of evaluation, the difference between the word-based segmentation methods will also makes the evaluation of SMT where Chinese as target language confusing. The automatic evaluation methods (such as BLEU and NIST BLEU score) in SMT are mostly based on n-gram precision. If the segmentation of test sets are different, the elements of the n-gram of test sets will also be different, which means that the evaluation is made on different test sets. To evaluate the quality of Chinese translation output, the International Workshop on Spoken Language Translation in 2005 (IWSLT'2005) used the word-level BLEU metric (Papineni et al.,2002). However, IWSLT'08 and NIST'08 adopted character-level evaluation metrics

29th Pacific Asia Conference on Language, Information and Computation pages 270 - 280
Shanghai, China, October 30 - November 1, 2015

to rank the submitted systems. Although there are also a lot of other works on automatic evaluation of SMT, such as METEOR (Lavie and Agarwal, 2007), GTM (Melamed et al., 2003) and TER (Snover et al., 2006), whether word or character is more suitable for automatic evaluation of Chinese translation output has not been systematically investigated (Li et al., 2011). Recently, different kinds of character-level SMT evaluation metrics are proposed, which also support that character-level SMT may have its own advantage accordingly (Li et al., 2011; Liu and Ng, 2012).

Traditionally, Back-off *N*-gram Language Models (BNLM) (Chen and Goodman, 1996; Chen and Goodman, 1998; Stolcke, 2002) are being widely used for probability estimation. For a better probability estimation method, recently, Continuous-Space Language Models (CSLM), especially Neural Network Language Models (NNLM) (Bengio et al., 2003; Schwenk, 2007; Le et al., 2011) are being used in SMT (Schwenk et al., 2006; Son et al., 2010; Schwenk et al., 2012; Son et al., 2012; Wang et al., 2013). These works have shown that CSLMs can improve the BLEU scores of SMT when compared with BNLMs, on the condition that the training data for language modeling are in the same size. However, in practice, CSLMs have not been widely used in SMT mainly due to high computational costs of training and using CSLMs. Since the using costs of CSLMs are very high, it is difficult to use CSLMs in decoding directly. A common approach in SMT using CSLMs is the two pass approach, or n-best reranking. In this approach, the first pass uses a BNLM in decoding to produce an n-best list. Then, a CSLM is used to rerank those n-best translations in the second pass (Schwenk et al., 2006; Son et al., 2010; Schwenk et al., 2012; Son et al., 2012). Nearly all of the previous works only conduct CSLMs on English, we conduct CSLM on Chinese in this paper. Vaswani et al. propose a method for reducing the training cost of CSLM and apply it into SMT decoder (Vaswani et al., 2013). Some other studies try to implement neural network LM or translation model for SMT (Gao et al., 2014; Devlin et al., 2014; Zhang et al., 2014; Auli et al., 2013; Liu et al., 2013; Sundermeyer et al., 2014; Cho et al., 2014; Zou et al., 2013; Lauly et al., 2014; Kalchbrenner and Blunsom, 2013).

The remainder is organized as follows: In Section 2, we will review the background of English to Chinese SMT. The character based SMT will be proposed in Section 3. In Section 4, the experiments will be conducted and the results will be analyzed . We will conclude our work in the Section 5.

2 Background

The ancient Chinese (or Classical Chinese, 文言文) can be conveniently split into characters, for most characters in ancient Chinese still keep understood by one who only knows modern Chinese (or Written Vernacular Chinese, 白话文) words. For example, "三人行，则必有我师焉。" is one of the popular sentences in the Analects (论语), and its corresponding modern Chinese words and English meaning are shown in TABLE 1. From the table, we can see that the characters in ancient Chinese have independent meaning, but most of the characters in modern Chinese do not, and they must combine together into words to make sense. If we split modern Chinese sentences into characters, the semantic meaning in the words will partially lose. Whether or not this semantic function of Chinese word can be partly replaced by the alignment model and Language Model (LM) of character-based SMT will be shown in this paper.

Ancient Chinese	**Modern Chinese**	**English Meaning**
三	三个	three
人	人	people
行	走路	walk
，	，	，
则	那么	so
必	一定	must
有	存在	be
我	我的	my
师	老师	teacher/tutor
焉	在其中	there
。	。	.

Table 1: Ancient Chinese and Modern Chinese

SMT as a research domain started in the late 1980s at IBM (Brown et al., 1993), which maps individual words to words and allows for deletion and insertion of words. Lately, various research-

es have shown better translation quality with phrase translation. Phrase-based SMT can be traced back to Och's alignment template model (Och and Ney, 2004), which can be re-framed as a phrase translation system. Other researchers augmented their systems with phrase translation, such as Yamada and Knight (Yamada and Knight, 2001), who used phrase translation in a syntax-based model.

The phrase translation model is based on the noisy channel model. Bayes rule is mostly used to reformulate the translation probability for translating a foreign sentence f into target e as:

$$argmax_e p(e|f) = argmax_e p(f|e)p(e) \quad (1)$$

This allows for the probabilities of an LM $p(e)$ and a separated translation model $p(f|e)$. During decoding, the foreign input sentence f is segmented into a sequence of phrases f_1^i. It is assumed a uniform probability distribution over all possible segmentations. Each foreign phrase f_i in f_1^i is translated into an target phrase e_i. The target phrases may be reordered. Phrase translation is modeled by a probability distribution $\Omega(f_i|e_i)$. Recall that due to the Bayes rule, the translation direction is inverted.

Reordering of the output phrases is modeled by a relative distortion probability distribution $d(start_i, end_{i-1})$, where $start_i$ denotes the start position of the foreign phrase that is translated into the ith target phrase, and end_{i-1} denotes the end position of the foreign phrase that was translated into the $(i-1)-th$ target phrase. A simple distortion model $d(start_i, end_{i-1}) = \alpha^{|start_i - end_{i-1} - 1|}$ with an appropriate value for the parameter α is set.

In order to calibrate the output length, a factor ω (called word cost) for each generated English word in addition to the tri-gram LM p_{LM} is proposed. This is a simple means to optimize performance. Usually, this factor is larger than 1, biasing toward longer output. In summary, the best output sentence given a foreign input sentence f according to the model is:

$$argmax_e p(e|f) = argmax_e p(f|e)p_{LM}(e)\omega^{length(e)}, \quad (2)$$

where $p(f|e)$ is decomposed into:

$$p(f_1^i|e_1^i) = \phi_1^i \Omega(f_i|e_i) d(start_i, end_{i-1}). \quad (3)$$

In this paper, the f stands for English and the e stands for Chinese. In short, there are three main parts both in the English to Chinese and Chinese to English SMT: the alignment $p(f|e)$, the LM $p(e)$ and the parameters training (tuning). When Chinese is the foreign language, there is only the alignment model $p(f|e)$ containing Chinese language processing. Contrarily, when Chinese is the target language, both the the alignment part $p(f|e)$ and the LM $p(e)$ will help retrieve the sematic meaning in the characters which is originally represented by words. So it is possible that we can process the English to Chinese in character level without word segmentation, which may also avoid the confusion in the evaluation part as proposed above.

3 Character-based versus Word-based SMT

The standards of segmentation between word-based and character-based English to Chinese translation are different, as well as the standard of the evaluation of them. That is, the test data contains words as the smallest unit for word-based SMT, and characters for character-based SMT. So the translated sentences of word-based translation will be converted into character-based sentence, and evaluated together with character-based translation BLEU score for fair comparison. We select two popular segmentation segmenters, one of which is based on Forward Maximum Matching (FMM) algorithm with the lexicon of (Low et al., 2005), and the other is based on Conditional Random Fields (CRF) with the same implementation of (Zhao et al., 2006). Because most Chinese words contains 1 to 4 characters, so we set the word-based LM as default trigram in SRILM, and character-based LM for 5-gram. All the different methods share the same other default parameters in the toolkits which will be further introduced in Section 4.

There seems to be no ambiguity in different character segmentations, however English characters, numbers and other symbols are also contained in the corpus. If they are split into "characters" like "年 增 长 百 分 之 2 0 0" (200 % increment per year) or "J o r d a n 是 伟 大 的 篮 球 运 动 员" (Jordan is a great basketball player), they will cause a lot of misun-

derstanding. So the segmentation is only used for Chinese characters, and the foreign letters, numbers and other symbols in Chinese text are still kept consequent.

Shown in Table 2, the BLEU score of SMT system with character-based segmenter is much higher than both FMM and CRF segmenters. The word-based English to Chinese SMT system is trained and tuned in word level and evaluated in character level, so we use the character-based LM to re-score the nbest-list of the results of the FMM and CRF segmenters. Firstly we convert the translated 1000-best candidates for each sentence into characters. Then calculate their LM scores by the character-based LM, and replace the word-based LM score with character-based LM score. At last we re-calculate the global score to get the new 1-best candidate with the same tuning weight as before. The BLEU score of re-ranked method is slightly higher than before, but still much less than the result of character segmenter. Although we can not conclude the character-based segmenter is better simply according to this experiment, this result gives us the confidence that our approach is reasonable and feasible at least.

4 Comparison Experiment

We use the patent data for the Chinese to English patent translation subtask from the NTCIR-9 patent translation task (Goto et al., 2011). The parallel training, development, and test data consists of 1 million (M), 2,000, and 2,000 sentences, respectively[1].

The basic settings of the NTCIR-9 English to Chinese translation baseline system (Goto et al., 2011) was followed[2]. The Moses phrase-based SMT system was applied (Koehn et al., 2007), together with GIZA++ (Och and Ney, 2003) for alignment and MERT (Och, 2003) for tuning on the development data. 14 standard SMT features were used: five translation model scores, one word penalty score, seven distortion scores and one LM score. The translation performance was measured by the case-insensitive BLEU on the tokenized test data[3].

4.1 The Alignment

In this subsection we investigate two factors in the phrase alignment. Four different kinds of methods for heuristics and three kinds of maximum length of phrases in phrase table are used for word alignment, with other default parameters in the toolkits. The results are shown in Table 3. The $grow-diag-final-and$, which will be set as default without special statement in the following sections, is shown better than other settings, and the BLEU score do not increase as the maximum length of phrases increases.

Alignment Parameters	**BLEU (dev)**	**BLEU (test)**
union	42.24	39.33
intersect	40.64	38.08
grow-diag-final	42.70	39.78
grow-diag-final-and	42.80	**40.31**
Maximum Length	**BLEU (dev)**	**BLEU (test)**
7	42.80	**40.31**
10	42.78	40.04
13	42.85	40.30

Table 3: Different Heuristics Used for Word Alignment

4.2 The N-gram Language Model

In this part, we will investigate how the factors in the n-gram LM influence the whole system.

The scale of the training corpus is one of the most important factors to LM. And *"more data is better data"* (Brants and Xu, 2009) has been proved to be one of the most important rules for constructing a LMs. First we randomly divide the whole training sets into 4 parts equally. We build the LM with 1, 2 and 4 parts (i.e. for 1/4, 1/2 and the whole corpus respectively), with other setting as default. Then, we add the dictionary information to the LM. The pr stands for the size of the dictionary and the pf stands for the characters' frequency in the dictionary. The results in Table 4 show that using the whole corpus

[1]Since we are the participants of NTCIR-9, so we have the bilingual sides of the evaluation data.

[2]We are aware that the original NTCIR patentMT baseline is designed for Chinese-English translation. In this paper, we follow the same setting of the baseline system, only convert the source language and the target language.

[3]It is available at `http://www.itl.nist.gov/iad/mig/tests/mt/2009/`

Segmentation Methods	BLEU
FMM Segmenter	34.56
FMM Segmenter + Character-based LM Re-rank	35.08
CRF Segmenter	38.28
CRF Segmenter + Character-based LM Re-rank	38.78
Character Segmenter	**40.31**

Table 2: Comparison Between Word-based Translation and Character-based Translation

for language training is necessary and using the dictionary information does not improve the translation performance.

Size of The Corpus	BLEU (dev)	BLEU (test)
1/4 Corpus	42.30	39.76
1/2 Corpus	42.51	40.19
the whole Corpus	42.80	**40.31**
Dictionaries		
pr=10k pf=5	42.63	40.01
pr=10k pf=10	42.60	40.17
pr=20k pf=10	42.73	40.02
No Dictionary	42.80	**40.31**

Table 4: Scale of Corpus for LM

We select the three most popular smoothing algorithms, Witten-Bell, Kneser-Ney (KN), and improved Kneser-Ney (improved KN), and compare their performance in the character-level English to Chinese SMT task. As shown in Table 5, when n is too small , the result is less satisfactory, and the BLEU score continues increase as n increases. However, the BLEU score begins to decrease when the LM becomes too *long*. The best 9-gram LM with Witten-Bell smoothing, corresponding to 5-gram to 7-gram in word-based LM, which is the widestly used in word-bases English to Chinese SMT.

4.3 The Tuning

We have shown that the different lengths of n-gram LMs make a significant influence in the English to Chinese translation. The 4-gram BLEU score is broadly accepted as the evaluate standard when we tune the other parameters using the minimum error rate training, which means that the MERT stage will not stop until it reaches the highest 4-gram BLEU on the development set. However, the same sentence

Smoothing Method	n-gram LM	BLEU (dev)	BLEU (test)
Kneser-Ney	9	42.55	39.91
Improved KN	7	42.95	40.30
Improved KN	9	42.84	40.55
Improved KN	11	42.44	40.07
Witten-Bell	7	42.72	40.10
Witten-Bell	9	42.71	**40.62**
Witten-Bell	11	42.44	39.67

Table 5: Different Smoothing Methods for LM

becomes longer if the character based segmentation is applied. That is, four words may be segmented into around 10 characters. Will the system gain a better performance if the n-gram of BLEU score in the MERT convergence standard increases as the n-gram in the LM increases?

To evaluate this hypothesis, the alignment model is set the same as the best performance in Table 3, and 5-gram LM with improved KN smoothing is set for LM. The results in Table 6 show that singly increasing the n-gram of MERT can not improve the performance of SMT.

n-gram MERT	n-gram BLEU (dev)	4-gram BLEU (test)
4	42.80	**40.31**
7	25.45	40.30
10	15.02	40.17

Table 6: Different Setting on MERT

4.4 Parameter Combinations

We have investigated how different factors affect the performance of English to Chinese SMT. However, most of the other factors are fixed when we discuss

one single factor. So in this subsection, we analyze how the combined factors perform in the whole system.

Firstly, we combine the parameters of the smoothing methods and the maximum length of phrases together. The LM is set to 9-gram and $grow-diag-final-and$ is set for alignment, which has the best BLEU score in n-gram LM experiments. Other factors is set as default in the toolkits. The results are shown in Table 7.

Smoothing Method (LM)	**Maximum Length (align)**	**BLEU (dev)**	**BLEU (test)**
KN	7	42.55	39.91
KN	10	42.80	40.49
KN	13	42.89	39.93
Improved KN	7	42.84	40.55
Improved KN	10	43.00	40.24
Improved KN	13	40.07	40.56
Witten-Bell	7	42.71	**40.62**
Witten-Bell	10	42.85	40.06
Witten-Bell	13	42.85	40.09

Table 7: Parameter Combinations of Smoothing Methods and Maximum Length of Phrase Alignment

Then, the length of n-gram MERT and the different order n-gram LM are tuned together. We set the Improved KN as the smoothing method, and others as default in the toolkits. The results are shown in Table 8.

n-gram LM	**n-gram MERT**	**BLEU (dev)**	**4-gram BLEU (test)**
7	4	42.95	40.30
7	7	25.54	39.91
9	4	42.84	40.55
9	7	25.93	**40.75**
9	10	15.82	40.37
13	7	25.41	40.47

Table 8: Parameter Combinations of n-gram LM and n-gram MERT

At last, the length of n-gram MERT and the smoothing methods are tuned together. The LM is set as 9-gram, the best BLEU score in n-gram LM experiments, and other factors set as default in the toolkits. The results are shown in Table 9.

Among different parameters-combined setting,

Smoothing Method	**n-gram MERT**	**BLEU (dev)**	**BLEU (test)**
KN	4	42.55	39.91
KN	7	25.33	40.65
Improved KN	4	42.84	40.55
Improved KN	7	25.93	**40.75**
Improved KN	10	15.82	40.37
Witten-Bell	4	42.71	40.62
Witten-Bell	7	25.45	40.30

Table 9: Parameter Combinations of n-gram MERT and Smoothing Methods

BLEU score is from 38.08 to 40.75, and the best performance is not gained when all the factors which singly perform best are put together. The highest BLEU score occurs when the 9-gram LM, the 7-gram MERT method and the improved KN smoothing algorithm. This BLEU score is about one percent higher than our baseline. At last, we show three parameter combinations with their NIST scores that bring the best performance up to now in Table 10.

4.5 Continues Space Language Model

Traditional Backoff N-gram LMs (BNLMs) have been widely used in many NLP tasks (Jia and Zhao, 2014; Zhang et al., 2012; Xu and Zhao, 2012).

Recently, Continuous-Space Language Models (CSLMs), especially Neural Network Language Models (NNLMs) (Bengio et al., 2003; Schwenk, 2007; Mikolov et al., 2010; Le et al., 2011), are actively used in SMT (Schwenk et al., 2006; Schwenk et al., 2006; Schwenk et al., 2012; Son et al., 2012; Niehues and Waibel, 2012). These models have demonstrated that CSLMs can improve BLEU scores of SMT over *n*-gram LMs with the same sized corpus for LM training. An attractive feature of CSLMs is that they can predict the probabilities of *n*-grams outside the training corpus more accurately.

A CSLM implemented in a multi-layer neural network contains four layers: the input layer projects all words in the context h_i onto the projection layer (the first hidden layer); the second hidden layer and the output layer achieve the non-liner probability estimation and calculate the LM probability $P(w_i|h_i)$ for the given context (Schwenk, 2007).

The CSLM calculates the probabilities of al-

Factors vs BLEU	**(1) 40.75**	**(2) 40.65**	**(3) 40.62**
Maximum Length of Phrases	7	10	10
Heuristic for Alignment	grow-diag-final-and	grow-diag-final-and	grow-diag-final-and
Scales of LM	whole	whole	whole
Dictionary of LM	none	none	none
n-gram of LM	9	9	9
Smoothing of LM	Improved KN	Kneser-Ney	Witten-Bell
n-gram MERT	7	7	4
NIST Score	9.32	**9.40**	9.23

Table 10: Parameters for TOP Performance

Methods vs BLEU	**(1) 40.75**	**(2) 40.65**	**(3) 40.62**
CSLM Re-rank	41.15	**41.27**	41.18
CSLM Decoding	41.34	41.34	**41.57**

Table 11: CSLM Re-rank and decoding for TOP Performance

l words in the vocabulary of the corpus given the context at once. However, due to too high computational complexity, the CSLM is only used to calculate the probabilities of a subset of the whole vocabulary. This subset is called a *short-list*, which consists of the most frequent words in the vocabulary. The CSLM also calculates the sum of the probabilities of all words not in the short-list by assigning a neuron. The probabilities of other words not in the short-list are obtained from an Backoff *N*-gram LM (BNLM) (Schwenk, 2007; Schwenk, 2010; Wang et al., 2013; Wang et al., 2015).

Let w_i, h_i be the current word and history, respectively. The CSLM with a BNLM calculates the probability of w_i given h_i, $P(w_i|h_i)$, as follows:

$$P(w_i|h_i) = \begin{cases} \frac{P_c(w_i|h_i)}{\sum_{w \in V_0} P_c(w|h_i)} P_s(h_i) & \text{if } w_i \in V_0 \\ P_b(w_i|h_i) & \text{otherwise} \end{cases} \quad (4)$$

where V_0 is the short-list, $P_c(\cdot)$ is the probability calculated by the CSLM, $\sum_{w \in V_0} P_c(w|h_i)$ is the summary of probabilities of the neuron for all the words in the short-list, $P_b(\cdot)$ is the probability calculated by the BNLM, and

$$P_s(h_i) = \sum_{v \in V_0} P_b(v|h_i). \quad (5)$$

We may regard that the CSLM redistributes the probability mass of all words in the short-list, which is calculated by using the *n*-gram LM.

Due to too high computational cost, it is difficult to use CSLMs in decoding directly. As mentioned in the introduction, a common approach in SMT using CSLMs is a two-pass procedure, or *n*-best re-ranking. In this approach, the first pass uses a BNLM in decoding to produce an *n*-best list. Then, a CSLM is used to re-rank those *n*-best translations in the second pass (Schwenk et al., 2006; Son et al., 2010; Schwenk et al., 2012; Son et al., 2012).

Because CSLM outperforms BNLM in probability estimation accuracy and BNLM outperforms CSLM in computational time. To integrate CSLM more efficiently into decoding, some existing approaches calculate the probabilities of the *n*-grams before decoding and store them (Wang et al., 2013; Wang et al., 2014; Arsoy et al., 2013; Arsoy et al., 2014) in *n*-gram format. That is, n-grams from BNLM are used as the input of CSLM, and the output probabilities of CSLM together with the corresponding n-grams of BNLM constitute *converted CSLM*. The converted CSLM is directly used in SMT, and its decoding speed is as fast as the *n*-gram LM.

From the above tables, we find the most important parameter for character-based English to Chinese translation is the LM, and other parameters just have a minor influence. To verify this observation, we use 9-gram character based CSLM (Schwenk et al., 2006), with 4096 characters in the short list, the projection layer of dimension 256 and the hidden layer of dimension 192 are set in the CSLM exper-

iments. (1) We add the CSLM score as the additional feature to re-rank the 1000-best candidates in the top three performance In Table 10. The weight parameters were tuned by using Z-MERT (Zaidan, 2009). This method is called *CSLM Re-rank*. (2) We follow (Wang et al., 2013)'s method and convert CSLM into *n*-gram LM. This converted CSLM can be directly applied to SMT decoding and called *CSLM-decoding*.

It is shown in Table 11 that the BLEU score nearly improve by 0.4 point to 0.6 point (CSLM Re-rank) and 0.6 point to 0.9 point (CSLM-decoding). This indicates that the CSLMs affect the performance of character based SMT in a significant way. This may indicate that the LM can take part place of the segmentation for character based English to Chinese SMT. A better character-based English to Chinese translation can be obtained by building a better LM.

5 Conclusion

Because the role of word segmentation in English to Chinese translation is arguable, an attempt of character-based English to Chinese translation seems to be necessary. In this paper, we have shown why character-based English to Chinese translation is necessary and feasible, and investigated how different factors perform in the system from the alignment, LM and the tuning aspects. Several empirical studies, including recent popular CSLM, have been done to show how to determine a optimal parameters for better SMT performance, and the results show that the LM is the most important factor for character-based English to Chinese translation.

Acknowledgments

We appreciate the anonymous reviewers for valuable comments and suggestions on our paper. Rui Wang, Hai Zhao and Bao-Liang Lu were partially supported by the National Natural Science Foundation of China (No. 60903119, No. 61170114, and No. 61272248), the National Basic Research Program of China (No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (No. 13511500200), the European Union Seventh Framework Program (No. 247619), the Cai Yuanpei Program (CSC fund 201304490199 and 201304490171), and the art and science interdiscipline funds of Shanghai Jiao Tong University (A study on mobilization mechanism and alerting threshold setting for online community, and media image and psychology evaluation: a computational intelligence approach).

References

Ebru Arsoy, Stanley F. Chen, Bhuvana Ramabhadran, and Abhinav Sethy. 2013. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition. In *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May.

Ebru Arsoy, Stanley F. Chen, Bhuvana Ramabhadran, and Abhinav Sethy. 2014. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):184–192.

Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. 2013. Joint language and translation modeling with recurrent neural networks. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1044–1054, Seattle, Washington, USA, October.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3:1137–1155, March.

Thorsten Brants and Peng Xu. 2009. Distributed language models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, NAACL-Tutorials '09, pages 3–4, Boulder, Colorado, USA. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, June.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232, Columbus, Ohio, USA. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on*

Association for Computational Linguistics, ACL '96, pages 310–318, Santa Cruz, California, USA. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard Univ.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1370–1380, Baltimore, Maryland, June.

Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning continuous phrase representations for translation modeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Baltimore, Maryland, June.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 559–578, Tokyo, Japan, December.

Zhongye Jia and Hai Zhao. 2014. A joint graph model for pinyin-to-chinese conversion with typo correction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1512–1523, Baltimore, Maryland, June.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Hai-Son Le, I. Oparin, A. Allauzen, J. Gauvain, and F. Yvon. 2011. Structured output layer neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5524–5527, Prague, Czech Republic, May. IEEE.

Maoxi Li, Chengqing Zong, and Hwee Tou Ng. 2011. Automatic evaluation of Chinese translation output: Word-level or character-level? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 159–164, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chang Liu and Hwee Tou Ng. 2012. Character-level machine translation evaluation for languages with ambiguous word boundaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 921–929, Jeju Island, Korea, USA. Association for Computational Linguistics.

Lemao Liu, Taro Watanabe, Eiichiro Sumita, and Tiejun Zhao. 2013. Additive neural networks for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 791–801, Sofia, Bulgaria, August.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea, October. Association for Computational Linguistics.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL-Short '03, pages 61–63, Edmonton, Canada. Association for Computational Linguistics.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurren-

t neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Jan Niehues and Alex Waibel. 2012. Continuous space language models using restricted boltzmann machines. In *Proceedings of the International Workshop for Spoken Language Translation*, IWSLT 2012, pages 311–318, Hong Kong.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449, December.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 723–730, Sydney, Australia. Association for Computational Linguistics.

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, WLM '12, pages 11–19, Montreal, Canada, June. Association for Computational Linguistics.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21(3):492–518.

Holger Schwenk. 2010. Continuous-space language models for statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, pages 137–146.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Le Hai Son, Alexandre Allauzen, Guillaume Wisniewski, and François Yvon. 2010. Training continuous space language models: some practical issues. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 778–788, Cambridge, Massachusetts, October. Association for Computational Linguistics.

Le Hai Son, Alexandre Allauzen, and François Yvon. 2012. Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 39–48, Montreal, Canada, June. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, Seattle, USA, November.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14–25, Doha, Qatar, October.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392, Seattle, Washington, USA, October.

Rui Wang, Masao Utiyama, Isao Goto, Eiichro Sumita, Hai Zhao, and Bao-Liang Lu. 2013. Converting continuous-space language models into n-gram language models for statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 845–850, Seattle, Washington, USA, October. Association for Computational Linguistics.

Rui Wang, Hai Zhao, Bao Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2014. Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 189–195, Doha, Qatar, October.

Rui Wang, Hai Zhao, Bao-Liang Lu, M. Utiyama, and E. Sumita. 2015. Bilingual continuous-space language model growing for statistical machine translation. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(7):1209–1220, July.

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2012. Enhancing statistical machine translation with character alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 285–290, Jeju Island, Korea, July. Association for Computational Linguistics.

Qiongkai Xu and Hai Zhao. 2012. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 1341–1350, Mumbai, India, December.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the*

39th Annual Meeting on Association for Computational Linguistics, ACL '01, pages 523–530, Toulouse, France. Association for Computational Linguistics.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Improved statistical machine translation by multiple Chinese word segmentation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 216–223, Columbus, Ohio, USA. Association for Computational Linguistics.

Xiaotian Zhang, Hai Zhao, and Cong Hui. 2012. A machine learning approach to convert CCGbank to Penn treebank. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 535–542, Mumbai, India, December.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, and Hai Zhao. 2014. Learning hierarchical translation spans. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 183–188, Doha, Qatar, October.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165, Sydney, Australia, July. Association for Computational Linguistics.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for Chinese machine translation. In *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 248–263, Berlin, Heidelberg. Springer-Verlag.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, October.