Context Sensitive Word Deletion Model for Statistical Machine Translation

Qiang Li, Yaqian Han, Tong Xiao, and Jingbo Zhu

NiuTrans Laboratory, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
{liqiangneu,hanyaqianneu}@gmail.com
{xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract. Word deletion (WD) errors can lead to poor comprehension of the meaning of source translated sentences in phrase-based statistical machine translation (SMT), and have a critical impact on the adequacy of the translation results generated by SMT systems. In this paper, first we classify the word deletion into two categories, wanted and unwanted word deletions. For these two kinds of word deletions, we propose a maximum entropy based word deletion model to improve the translation quality in phrase-based SMT. Our proposed model are based on features automatically learned from a real-word bitext. In our experiments on Chinese-to-English news and web translation tasks, the results show that our approach is capable of generating more adequate translations compared with the baseline system, and our proposed word deletion model yields a +0.99 BLEU improvement and a -2.20 TER reduction on the NIST machine translation evaluation corpora.

Keywords: natural language processing, statistical machine translation, word deletion

1 Introduction

Recently, although researchers have shown an increasing interest in neural machine translation (NMT) [1–4], statistical machine translation (SMT) also draw a lot of attention. SMT has been applied to many applications, and the phrase-based translation model has been widely used in modern SMT systems due to its simplicity and strong performance [5, 6]. To evaluate the quality of machine translation systems, we often consider both adequacy and fluency [7], and measure the number of edits required to change a system output into one of the references [8]. For poor translation results with low adequacy, the problem is mainly caused by word deletion problems, word insertion problems, and incorrect word choices. Word insertion problems are not common during translation [9]. Incorrect word choices are eliminated by improving translation model [10, 11] or using domain adaptation [12]. As word deletion (WD) problems have not gotten enough attention in research community, in this paper we propose a context sensitive word deletion model to address these problems.

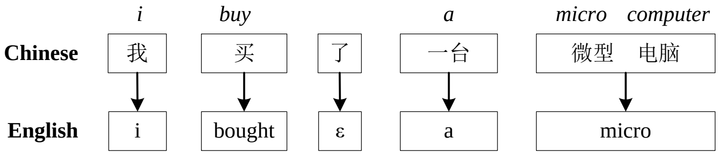


Fig. 1. Example of wanted and unwanted word deletion.

Table 1. Statistics of unwanted word deletion determined by human evaluators from 200 randomly selected machine-translated sentences on the news translation task.

Corpus		Unwanted WD	
# Sentences	# Words	Frequency	Ratio
200	3, 293	450	13.67%

There are two kinds of word deletions. First, every language has some *spurious* words that do not need to be translated, referred to as *wanted word deletion*. It is correct that the source spurious words are not translated during translation. See Fig. 1 for an example. The source Chinese spurious word '了' has no counterparts in the other language and will be translated to empty word ϵ by decoder. It is possible to learn phrase pairs '了, ϵ' for wanted word deletion from a word aligned bilingual training corpus. Consequently, SMT systems realize the function of wanted word deletion. However, *unwanted word deletion* appears along with wanted word deletion during phrase extraction. For example, the phrase pair '微型 电脑, micro' in Fig. 1 is an unwanted word deletion as the source *meaningful* word '电脑' has no counterparts in the target language and results in a translation error. An unwanted word deletion seriously influences the adequacy of translation results generated by SMT systems. Unwanted word deletion is very common in SMT systems, from Table 1 we can see that the ratio of unwanted word deletion is as high as 13.67% of all 3, 293 words in our 200 randomly selected sentences determined by human evaluators, and there are about 2.5 meaningful words that are not translated at all in each sentence on average. In this paper, we try to explore the research into wanted and unwanted word deletions for the phrase-based SMT system.

To address wanted and unwanted word deletion problems, first of all, we should judge whether a source word is a spurious or meaningful word. For a source spurious word, we do not want to translate it during translation, which is referred to as 'deleted'. On the contrary, a source meaningful word that we want it to be correctly translated is referred to as 'reserved'. Obviously, this problem can be cast as a binary classification problem. Consequently, we propose a novel maximum entropy based context sensitive model to improve the translation quality. The proposed word deletion model based on maximum entropy automatically learns features from a real-world bitext. During decoding, our proposed model is embedded inside a log-linear phrase-based model of translation. Finally, experimental results demonstrate that our proposed methods achieve significant improvements in BLEU [7] and TER [8] score on the Chinese-to-English news and web translation tasks. For example, it yields a +0.99 BLEU improvement and a -2.20 TER reduction on the NIST MT evaluation corpora.

2 Statistical Machine Translation

The goal of machine translation is to automatically translate from a source string f_1^J to a target string e_1^I. In SMT, this problem can be stated as: we find a target string $\hat{e}_1^I$ from all possible translations by the following equation:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{Pr\left(e_1^I | f_1^J\right)\} \tag{1}$$

where $Pr\left(e_1^I | f_1^J\right)$ is the probability that e_1^I is the translation of the given source string f_1^J. To model the posterior probability $Pr\left(e_1^I | f_1^J\right)$, our decoder utilizes the log-linear model [13]. $Pr\left(e_1^I | f_1^J\right)$ is calculated as follows:

$$Pr\left(e_1^I | f_1^J\right) = \frac{\exp(\sum_a \lambda_a h_a(f_1^J, e_1^I))}{\sum_{e_1^{I^*}} \exp(\sum_a \lambda_a h_a(f_1^J, e_1^{I^*}))} \tag{2}$$

where $\{h_a(f_1^J, e_1^I) | a = 1, ...\}$ is a set of features, and λ_a is the feature weight corresponding to the a-th feature. $h_a(f_1^J, e_1^I)$ can be regarded as a function that maps each pair of source string f_1^J and target string e_1^I into a non-negative value, and λ_a can be regarded as the contribution of $h_a(f_1^J, e_1^I)$ to $Pr\left(e_1^I | f_1^J\right)$. Ideally, λ_a indicates the pairwise correspondence between the feature $h_a(f_1^J, e_1^I)$ and the overall score $Pr\left(e_1^I | f_1^J\right)$. A positive value of λ_a indicates a positive correlation between $h_a(f_1^J, e_1^I)$ to $Pr\left(e_1^I | f_1^J\right)$, while a negative value indicates a negative correlation.

In a general pipeline of SMT, λ is learned on a tuning data set to obtain an optimized weight vector λ^*. To learn the optimized weight vector λ^*, λ is usually optimized according to a certain objective function. The objective function should take the translation quality into account and can be automatically learned from MT outputs and reference translations. Therefore, we use BLEU to define the error function and learn optimized feature weights using the minimum error rate training method [14].

3 Context Sensitive Word Deletion Model

3.1 Word Deletion Model

As mentioned in section 2, all the features used in our system are combined in a log-linear fashion. So, Given a derivation v, the corresponding model score is calculated as follows:

$$Pr\left(e_1^I, v | f_1^J\right) = \prod_{(f_{j_1}^{j_2}, e_{i_1}^{i_2}) \in v} Pr^l(f_{j_1}^{j_2}, e_{i_1}^{i_2}) \times Pr_r(v)^{\lambda_r} \times Pr_{lm}(e_1^I)^{\lambda_{lm}} \times \mathrm{WD}^{\lambda_{\mathrm{WD}}} \tag{3}$$

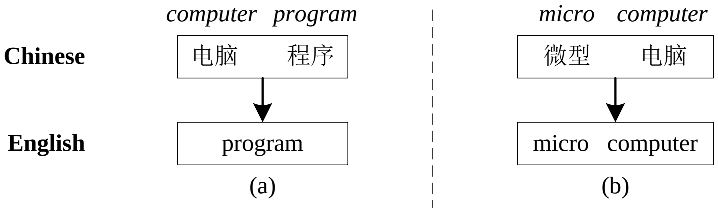


Fig. 2. Example of phrase pairs for illustrating context condition.

where $1 \leq i_1, i_2 \leq I$, $1 \leq j_1, j_2 \leq J$, $Pr_r(v)$ is the maximum entropy based reordering model proposed in [15], and λ_r is its weight. $Pr_{lm}(e_1^I)$ is the n-gram language model score, and λ_{lm} is its weight. We use $Pr^l(f_{j_1}^{j_2}, e_{i_1}^{i_2})$ to calculate the probability of the lexical rule.

$$\begin{aligned} Pr^l(f_{j_1}^{j_2}, e_{i_1}^{i_2}) =& p(e_{i_1}^{i_2}|f_{j_1}^{j_2})^{\lambda_1} \times p(f_{j_1}^{j_2}|e_{i_1}^{i_2})^{\lambda_2} \times \\ & p_{lex}(e_{i_1}^{i_2}|f_{j_1}^{j_2})^{\lambda_3} \times p_{lex}(f_{j_1}^{j_2}|e_{i_1}^{i_2})^{\lambda_4} \times \\ & exp(1)^{\lambda_5} \times exp(|e_{i_1}^{i_2}|)^{\lambda_6} \end{aligned} \tag{4}$$

where $p(\cdot)$ are the phrase translation probabilities in both directions, $p_{lex}(\cdot)$ are the lexical translation probabilities in both directions, and $exp(1)$ and $exp(|e_{i_1}^{i_2}|)$ are the phrase penalty and word penalty, respectively.

WD in Equation 3 is our proposed context sensitive word deletion model, and λ_{WD} is its weight. For the word deletion model WD, we define it on source word f_j, context context(f_j) and deletion $d \in \{\textit{deleted}, \textit{reserved}\}$. WD is defined as follows:

$$\text{WD} = f(d, f_j, \text{context}(f_j)) \tag{5}$$

where f_j is a source word that needs to be translated. context(f_j) is the context words around the source word f_j. d is deletion or not that based on the word alignments between the source sentence s_1^I and the target sentence t_1^J, which covers the value over *deleted* and *reserved*. If f_j is a source spurious word and we do not want to translate it during translation, $d = \textit{deleted}$. On the other hand, if f_j is a source meaningful word and we want it to be correctly translated, $d = \textit{reserved}$. We will describe the proposed model in the next section.

3.2 Maximum Entropy Based Word Deletion Model

In this section, first, we will explain why we use context to address word deletion problems. Through data analysis determined by human evaluators, we can see that incorrect context condition is the main cause for the incorrect word deletion. See Fig. 2 for an example. The phrase pair '电脑 程序, program', whose source meaningful word '电脑' is translated to ϵ based on this context, is correct. '电脑' that can be regarded as a spurious source word based on this context is referred to as wanted word deletion. If our bilingual training corpus contains this phrase pair, then the phrase pair '电脑, ϵ' will be produced during phrase extraction.

But for phrase pairs ‘微型 电脑, micro computer’, ‘电脑’ can not be translated to ϵ based on current context. During translation, ‘微型 电脑’ will be translated into ‘micro’ if the phrase pair ‘电脑, ϵ’ is selected by the decoder. Then an unwanted word deletion occurs. Consequently, based on the context of a given source word, we want to automatically learn correct wanted and unwanted word deletion models. Therefore, we propose a context sensitive word deletion model to address word deletion problems.

As described above, we defined the word deletion model WD on three factors: source word f_j, context(f_j), and the word deletion d. The central problem is that given f_j and context(f_j), how to predict $d \in \{\textit{deleted}, \textit{reserved}\}$. This is a typical problem of two-class classification. To be consistent with the whole model, the conditional probability $p(d|f_j, \text{context}(f_j))$ is calculated. A good way to this problem is to use features of source lexical words and word alignments between source and target sentence as word deletion evidences. It is very straight to use maximum entropy model to integrate features to predicate word deletions of the source word f_j. Under the maximum entropy based model, we have:

$$\text{WD} = p_\theta(d|f_j, \text{context}(f_j)) = \frac{\exp(\sum_b \theta_b h_b(d, f_j, \text{context}(f_j)))}{\sum_{d^*} \exp(\sum_b \theta_b h_b(d^*, f_j, \text{context}(f_j)))} \tag{6}$$

where the functions $h_b \in \{0, 1\}$ are model features and the θ_b are weights of the model features which can be trained by different algorithms [16].

3.3 Word Deletion Examples and Features

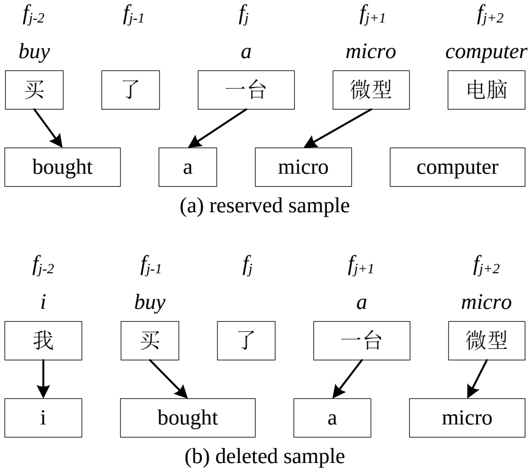
If we want to extract high-precision word deletion examples, first we should have a bilingual corpus with high-precision word alignments. We obtain the word alignments using the way in [15]. After running GIZA++ in both directions [17], we apply the *grow-diag-final-and* refinement rule on the intersection alignments for each sentence pair.

A *word deletion example* is a triple $(d, f_j, \text{context}(f_j))$. In the bilingual training corpus with high-precision word alignments, for the source word f_j, if there exists a target word e_i that is aligned to f_j, $d = \textit{reserved}$. Otherwise, $d = \textit{deleted}$.

With the extracted word deletion examples, we can obtain features for our maximum entropy based context sensitive word deletion model. See Fig. 3 for an example, we want to justify if the current source word ‘一台’ or ‘了’ should be deleted or reserved. The feature templates for our method is shown here:

- the lexical form of the source word f_j itself, here is ‘一台’ or ‘了’.
- the lexical forms f_{j-2}, f_{j-1}, f_{j+1}, and f_{j+2}. Where f_{j-2} and f_{j-1} are the two words to the left of f_j, and f_{j+1}, and f_{j+2} are the two words to the right of f_j.

Then we can obtain two word deletion examples for the source spurious and meaningful words, respectively:



(a) reserved sample

(b) deleted sample

Fig. 3. The lexical form of words in rectangle are the features that used in maximum entropy word deletion model.

- d=reserved, f_j =一台, f_{j-2} =买, f_{j-1} =了, f_{j+1} =微型, f_{j+2} =电脑
- d=deleted, f_j =了, f_{j-2} =我, f_{j-1} =买, f_{j+1} =一台, f_{j+2} =微型

We first extract word deletion evidences from our 2.43M bilingual training data with word alignments. Then, we use the MaxEnt toolkit[1] to tune the feature weights. Finally, the proposed maximum entropy based WD model is integrated into the standard log-linear model used by our decoder in Equation 3.

3.4 Decoder

After training the context sensitive word deletion model WD with the maximum entropy approach, we integrate it into our log-linear phrase-based model during decoding. Here, the source sentence f_1^J that is needed to be translated is '我买了一台微型电脑', after tokenization we can get a string '我 买 了 一台 微型 电脑'. For every word $f_{j_1}^{j_2}$ in f_1^J, $1 \leq j_1, j_2 \leq J$, we will use WD model to justify if it is needed to be *deleted* or *reserved*.

When we translate the Chinese source word '了', the WD model shows that it is more likely to be spurious word and needed to be 'deleted' based on current context during translation. When the decoder selects the phrase '买 了, bought' in Fig. 4(a) that contains source word '了' and does not have any correspondences for '了' in the target side, then the WD feature will be active during translation. But when the decoder selects the phrase in Fig. 4(b) that contains source word '了' and the target word 'a' is aligned to '了', then the WD feature will be inactive for the phrase (买 了, bought a) during translation.

Now we switch to the source meaningful word. When we translate the Chinese source word '微型', the context sensitive model shows that it is more likely to

[1] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

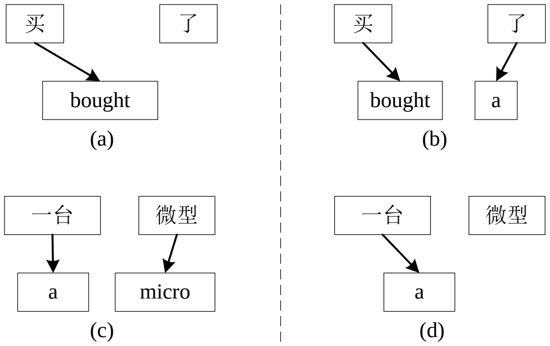


Fig. 4. Sample phrases pairs used during decoding.

be meaningful word and needed to be 'reserved' based on this context. When the decoder selects the phrase (一台 微型, a micro) in Fig. 4(c) that the target word 'micro' is aligned to the source word '微型', then the WD feature will be active during decoding. Otherwise, the WD feature will be inactive if the decoder selects the phrase (一台 微型, a) in Fig. 4(d) as there are no correspondences for the Chinese source meaningful word '微型'.

4 Evaluation

In this section, we describe our method of evaluating the context sensitive word deletion model to address word deletion issues in SMT. We applied the proposed methods to a state-of-the-art phrase-based SMT system [18] and carried out experiments on Chinese-to-English news and web translation tasks.

4.1 Experiment Setup

We developed a CKY style decoder that employed a beam search and cube pruning to build our phrase-based SMT system [19]. Our SMT system used all standard features adopted in the current state-of-the-art phrase-based system, including bidirectional phrase translation probabilities, bidirectional lexical weights, an n-gram language model, target word penalty and phrase penalty. In addition, the ME-based lexicalized reordering model was employed in our system [15]. The reordering limit was set to 8 and the beam size was set to 30. The maximum length of source and target phrases were limited to 5 words.

Our experiments were conducted on two Chinese-to-English translation tasks: news and web domains. In both domains, our bilingual data consisted of 2.43 million sentence pairs selected from the NIST portion of the bilingual data of NIST MT 2008 Evaluation. The 5-gram language model for both translation tasks was trained on the Xinhua portion of English Gigaword corpus (16.28M) in addition to the target side of the bilingual data. For the news domain, we used the NIST 2006 news MT evaluation set as our development set (616 sentences) and the NIST 2008 news, 2008 progress news, and 2012 news MT evaluation sets

Table 2. BLEU4 scores of the baseline and the proposed maximum entropy based model in Chinese-to-English news translation. Here, * indicates significantly better on test performance at the $p = 0.05$ level compared to the baseline method.

Method	BLEU on News Data [%]			
	2006	2008	2008 pro	2012
baseline	30.58	29.60	27.22	29.42
$+\ \epsilon$	31.28^*	29.68	27.46	30.08^*
$+\ \epsilon$ + maxent	31.57^*	30.25^*	28.13^*	30.28^*

Table 3. BLEU4 scores of the baseline and the proposed maximum entropy based model in Chinese-to-English web translation. Here, * indicates significantly better on test performance at the $p = 0.05$ level compared to the baseline method.

Method	BLEU on Web Data [%]			
	2006	2008	2008 pro	2012
baseline	28.03	21.28	22.41	19.68
$+\ \epsilon$	28.44	21.68	22.78	19.95
$+\ \epsilon$ + maxent	28.77^*	21.99^*	23.01^*	19.91

as our test sets (691, 688, and 400 sentences). For the web domain, we used the NIST 2006 webdata MT evaluation set as our development set (483 sentences) and the NIST 2008 web, 2008 progress web, and 2012 web MT evaluation sets as our test sets (666, 682, and 420 sentences).

The GIZA++ tool was used to perform the bidirectional word alignment between the source and target sentences [17]. After running GIZA++ in both directions, we applied the *grow-diag-final-and* refinement rule on the intersection alignments for each sentence pair.

4.2 Results

Table 2 and Table 3 depict the BLEU scores of the baseline approach (Row 'baseline') and the maximum entropy based WD model (Row '$+\epsilon$+maxent') on Chinese-to-English news and web translation tasks. In order to compare with the method proposed by Li et al. [20] to address spurious source word translation, a specific empty symbol ϵ on the target language side is posited and any source word is allowed to translate into ϵ (Row '$+\epsilon$'). This symbol is just visible in phrase table. That is, ϵ is not counted when calculating language model score, word penalty and any other feature values, and it is omitted in the final translation results. For our proposed context sensitive WD model, any source word is also allowed to translate into ϵ.

On the Chinese-to-English news translation in Table 2, first we can see that, when empty symbol ϵ is posited on the target language side (Row '$+\epsilon$'), the baseline system achieved BLEU score increase of 0.70, 0.08, 0.24, and 0.66 on NIST 2006 news, 2008 news, 2008 progress news, and 2012 news, respectively. From this we can say that it is better to improve translation quality in BLEU score when any source words are allowed to translated to empty word ϵ. Second,

Table 4. TER scores of various methods in news translation. For TER, lower is better.

Method	TER on News Data [%]			
	2006	2008	2008 pro	2012
baseline	76.31	59.65	60.30	60.54
+ ϵ + maxent	74.15	58.20	59.10	58.34

Table 5. TER scores of various methods in web translation. For TER, lower is better.

Method	TER on Web Data [%]			
	2006	2008	2008*pro*	2012
baseline	62.31	63.92	63.06	66.95
+ ϵ + maxent	60.33	62.33	61.31	65.53

we can see that our proposed context sensitive WD model significantly improve the BLEU score on both development and test sets (Row '+ϵ+maxent'). Our proposed method achieved BLEU score increase of 0.99, 0.65, 0.91, and 0.86 on development and test datasets, respectively.

When we switch to Chinese-to-English web translation in Table 3, the experimental results are similar to those in Table 2. When introduced empty symbol ϵ (Row '+ϵ'), the baseline system achieved BLEU score increase of 0.41, 0.40, 0.37, and 0.27 on NIST 2006 web, 2008 web, 2008 progress web, and 2012 web, respectively. For our proposed context sensitive model (Row '+ϵ+maxent'), the achievements are 0.74, 0.71, 0.60, and 0.23 points on the BLEU score for the NIST 2006 web, 2008 web, 2008 progress web, and 2012 web, respectively.

Translation Edit Rate (TER) is an error metric for machine translation that measures the number of edits required to change a system output into one of the references [8]. Snover et al. [8] showed that the single-reference variant of TER correlates as well with human judgments of MT quality as the four-reference variant of BLEU. So, in addition to use BLEU score to evaluate the translation quality for statistical machine translation system, we also use TER metric to evaluate our proposed context sensitive WD model for the word deletion problems. Different from BLEU score, the lower is better for TER metric. Table 4 and Table 5 depict the TER scores of the baseline approach and the proposed methods on Chinese-to-English news and web translation tasks. On news translation in Table 4, we can see that our proposed method (Row '+ϵ+maxent') yields a gain of 2.16, 1.45, 1.20, and 2.20 TER score decrease on the development and test sets, respectively. When we switch to web translation in Table 5, the experimental results are similar to those in Table 4. For example, our proposed method (Row '+ϵ+maxent') achieved TER score decrease of 1.98, 1.59, 1.75 and 1.42 on the development and test sets, respectively. Therefore, we can conclude that our proposed maximum context sensitive word deletion model can significantly improve the translation quality in TER metric for Chinese-to-English news and web translation.

5 Related Work

Although researchers have shown an increasing interest in NMT [1–4], SMT also draw a lot of attention. There are many problems that SMT finds difficult to solve and the word deletion issue is among them. Several studies have addressed the word deletion problems, Li et al. [21] proposed four effective models to handle undesired word deletion. Parton et al. [22] presented a hybrid approach, APES, to target adequacy errors. Huck and Ney [23] investigated an insertion and deletion model that was implemented as phrase-level feature functions that counted the number of inserted or deleted word. Zhang et al. [24] focused on unaligned words only and applied hard deletion and optional deletion of the unaligned words on the source side before phrase extraction. Though easy to implement, this method introduced more noise into the phrase table. They showed that reducing the noise in phrase extraction is more effective than improving word alignment [25, 26]. Menezes and Quirk [27] presented an extension of the treelet translation method to include order templates with structural insertion and deletion. Li et al. [20] proposed three models to handle spurious source words. They utilized different methods to calculate the translation probability that the source words are translated into empty word ϵ.

In contrast to previous studies, we first categorize word deletion problems into wanted and unwanted word deletion. Second, we proposed maximum entropy based context sensitive word deletion model to address both the wanted and unwanted word deletions.

6 Conclusion

In this paper, we tackled the word deletion issue for the phrase-based SMT. First of all, we classified word deletion problems into two categories, wanted and unwanted word deletion. For these two kinds of word deletion problems, we proposed a context sensitive WD model to address them. The proposed WD model are based on features automatically learned from a real-word bitext. We evaluated our proposed methods on Chinese-to-English news and web translation tasks, and the experimental results demonstrated that our proposed context sensitive model achieved significant improvements in BLEU and TER scores. On the NIST Chinese-to-English evaluation corpora, it achieved a +0.99 BLEU improvement and a -2.20 TER reduction on top of a baseline system.

Acknowledgements

This work was supported in part by the National Science Foundation of China (61672138 and 61432013), the Fundamental Research Funds for the Central Universities, and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research.

References

1. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. (2014) 3104–3112
2. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. (2015) 1412–1421
3. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. (2016)
4. Britz, D., Goldie, A., Luong, M., Le, Q.V.: Massive exploration of neural machine translation architectures. CoRR **abs/1703.03906** (2017)
5. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003. (2003)
6. Och, F.J., Ney, H.: The alignment template approach to statistical machine translation. Computational Linguistics **30**(4) (2004) 417–449
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. (2002) 311–318
8. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. Volume 200. (2006)
9. Vilar, D., Xu, J., d'Haro, L.F., Ney, H.: Error analysis of statistical machine translation output. In: Proceedings of LREC. (2006) 697–702
10. Chiang, D.: Hierarchical phrase-based translation. Computational Linguistics **33**(2) (2007) 201–228
11. Galley, M., Hopkins, M., Knight, K., Marcu, D.: What's in a translation rule? In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004. (2004) 273–280
12. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proceedings of the second workshop on statistical machine translation, Association for Computational Linguistics (2007) 224–227
13. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. (2002) 295–302
14. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan. (2003) 160–167
15. Xiong, D., Liu, Q., Lin, S.: Maximum entropy based phrase reordering model for statistical machine translation. In: ACL 2006, 21st International Conference

on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006. (2006)

16. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002. (2002)
17. Och, F.J., Ney, H.: Improved statistical alignment models. In: 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 1-8, 2000. (2000)
18. Xiao, T., Zhu, J., Zhang, H., Li, Q.: Niutrans: An open source toolkit for phrase-based and syntax-based machine translation. In: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea. (2012) 19–24
19. Huang, L., Chiang, D.: Forest rescoring: Faster decoding with integrated language models. In: Annual Meeting-Association For Computational Linguistics. Volume 45. (2007) 144
20. Li, C.H., Zhang, D., Li, M., Zhou, M., Zhang, H.: An empirical study in source word deletion for phrase-based statistical machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics (2008) 1–8
21. Li, Q., Zhang, D., Li, M., Xiao, T., Zhu, J.: Better addressing word deletion for statistical machine translation. In: Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings. (2016) 91–102
22. Parton, K., Habash, N., McKeown, K., Iglesias, G., De Gispert, A.: Can automatic post-editing make mt more meaningful? In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, EAMT 2012. (2012) 111–118
23. Huck, M., Ney, H.: Insertion and deletion models for statistical machine translation. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada. (2012) 347–351
24. Zhang, Y., Matusov, E., Ney, H.: Are unaligned words important for machine translation? In: Proceedings of The 13th Annual Conference of the EAM T. pages. Volume 226. (2009) 233
25. Liu, Y., Liu, Q., Lin, S.: Discriminative word alignment by linear modeling. Computational Linguistics **36**(3) (2010) 303–339
26. Zhu, J., Li, Q., Xiao, T.: Improving syntactic rule extraction through deleting spurious links with translation span alignment. Natural Language Engineering **21**(2) (2015) 227–249
27. Menezes, A., Quirk, C.: Syntactic models for structural word insertion and deletion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2008) 735–744