Modeling Indicative Context for Statistical Machine Translation

Shuangzhi Wu[1], Dongdong Zhang[2] Shujie Liu[2], and Ming Zhou[2]

[1] Harbin Institute of Technology, Harbin, China,
v-shuawu@microsoft.com
[2] Microsoft Research Asia, Bejing, China
dozhang, shujliu, mingzhou@microsoft.com

Abstract. Contextual information is very important to select the appropriate phrases in statistical machine translation (SMT). The selection of different target phrases is sensitive to different parts of source contexts. Previous approaches based on either local contexts or global contexts neglect impacts of different contexts and are not always effective to disambiguate translation candidates. As a matter of fact, the indicative contexts are expected to play more important roles for disambiguation. In this paper, we propose to leverage the indicative contexts for translation disambiguation. Our model assigns phrase pairs confidence scores based on different source contexts which are then intergraded into the SMT log-linear model to help select translation candidates. Experimental results show that our proposed method significantly improves translation performance on the NIST Chinese-to-English translation tasks compared with the state-of-the-art SMT baseline.

1 Introduction

In statistical machine translation (SMT) [12,3,5,14] the probabilities of target translation candidates are estimated based on the co-occurrence frequencies with source phrases. Meanwhile the selection of target candidates is affected by limited target contexts computed based on a variant of language models. In practice, source contexts of source phrases are proven helpful in making disambiguation of target translation candidates. Much work has been done to incorporate source context information to improve the translation performance [2,6,13,16,20]. However, these methods just leverage limited local contexts for translation disambiguation, which are often insufficient to model the disambiguation of translation candidates that may have long distance dependencies with the source phrases. Taking the sentence pairs of Chinese and English in Figure 1 as an example, the source phrase "**东西**" in bold occurs in the same sub-sentences in both Figure 1(a) and Figure 1(b), but it is translated into different target phrases denoted by the solid lines. It is impossible to distinguish these two target translations during SMT decoding merely using the local context with the distance less than 3 words. So it needs longer distance dependencies beyond local contexts to help make disambiguation.

Recently, with the success of distributed representation modeling by neural networks, source- and target-side local contexts are fully used to better generate the

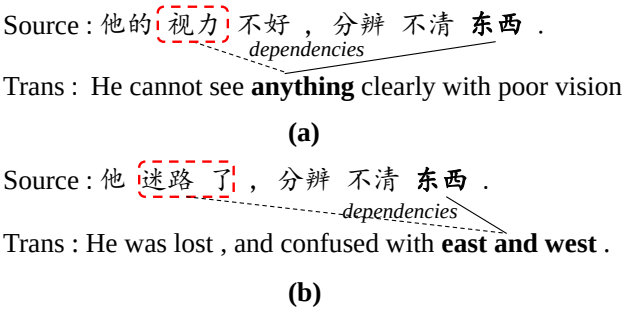

Fig. 1. Two examples of Chinese-English sentence pairs. The phrases in bold are translation pairs connected with solid lines. The indicative context for translation disambiguation are marked in dashed boxes .

target translation [4], where global sentence-level information is not adequately concerned. Additionally, global contexts over source sentences have been modeled to help with local translation prediction [22,7]. In these models, the contexts are generated over the entire source sentence for each local translation disambiguation, where they do not pay much attention to the effects of the critical context information named as indicative context which is expected to play more important roles for disambiguation than other contexts. For example in Figure 1, the translation disambiguation of the same source phrase "东西" should be mainly dominated by its indicative contexts marked in dashed boxes, which are more useful than other context words in these sentences.

In this paper, we propose an indicative context based translation disambiguation model (ICDM) to identify the indicative source context for disambiguating translation candidates for SMT. Our method models both the intertranslation quality of translation pairs and whether the target candidate is suitable for the specific source context. Then a confidence score is calculated for each phrase pair and integrated into SMT decoder as an extra feature. Experimental results demonstrate that our model significantly improves translation accuracy over a state-of-the-art SMT baseline on Chinese-English task.

2 Indicative Context Based Translation Disambiguation

We present an indicative context based translation disambiguation model (ICDM) to help with translation disambiguation for phrase pairs. Our model consists of three parts, a Context-RNN used to map source context into vector space, a Phrase-RNN used to map source phrase into vector space, and a RNNLM which is used to calculate the confidence score for a target phrase given the source context and phrase. Figure 2 gives a graphical overview of our model. Due to space limitation, we only detail the connections at time step t. In this section we will explain ICDM and how to incorporate ICDM into the SMT decoding in detail.

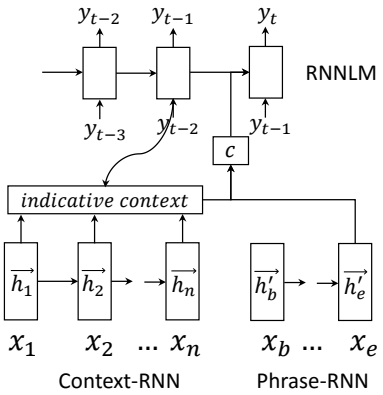

Fig. 2. Overview of ICDM. Only the details at time step t are listed.

2.1 Model Structure

Given the source sentence $X_1^n := x_1, x_2, x_3, ..., x_n$, source phrase $P_b^e := x_b, x_{b+1}, ..., x_e$, and the target phrase $T_1^m := y_1, y_2, ..., y_m$, where n, m is the length of source sentence and target phrase, x_i, y_i represent the source word and target word, $1 \leq b \leq e \leq n$. The translation disambiguation task is to estimate confidence scores for phrase pairs conditioned on the source sentence X_1^n and phrase P_b^e, which is denoted by Conf(P_b^e, T_1^m). Thus the task can be modeled as,

$$\text{Conf}(X_1^n, P_b^e, T_1^m) = p(T_1^m | X_1^n, P_b^e) \tag{1}$$

We use two recurrent neural network (RNN) named context-RNN and phrase-RNN to map X_1^n and P_b^e into context vectors denoted by $H_1^n := h_1, h_2, h_3 ... h_n$ and ${H'}_b^e := h'_b, h'_{b+1} ... h'_e$ respectively. Both RNNs are bidirectional. Empirically, the final hidden vectors h_n and h'_e can be used to represent the source sentence and source phrase. For the target phrase, we use a recurrent neural network language model (RNNLM) to score the target phrase. When integrating h_n and h'_e into the RNNLM, Equation 1 can be rewritten as Equation 2

$$\text{Conf}(X_1^n, P_b^e, T_1^m) = \prod_{t=1}^{m} p(y_t | y_{<t-1}, h_n, h'_e) \tag{2}$$

In Equation 2, the whole source sentence is utilized to provide global source context and the source phrase is used to measure the intertranslation of the phrase pair. For different phrase pairs of one source sentence, this model uses the same source context h_n for translation selection, thus we call this model Single Context-based Disambiguation Model (SCDM). However, different phrase pairs of one source sentence usually depends on different parts of source contexts, the single context is insufficient to distinguish different contexts.

Identifying Indicative Context To better leverage source contexts to facilitate translation disambiguation, we propose to highlight the indicative context which is dynamically constructed for different phrase pairs of one source sentence. We use the hidden vectors H_1^n of context-RNN to represent the local context of each source word in the source sentence. Our goal is to identify which local context plays a more important role in selecting targets for a certain source phrase. We assign different weights for each local context aiming at highlighting the indicative context. When calculating the RNNLM score of a target phrase, at each time step t, we use a single bilinear [15] transformation to compute a score between the RNNLM hidden state and each source context vector h_i, which is similar with the attention mechanism in [1].

$$e_{it} = h_i W_a s_t \tag{3}$$

where W_a is the weight matrix, s_t is the state of RNNLM at time step t. Then the weight of each context is modeled by,

$$a_{it} = \frac{\exp(e_{it})}{\sum_k \exp(e_{kt})} \tag{4}$$

We expect that the indicative context could have a higher weight than others and dominate the translation disambiguation. However, sometimes there is no explicit indicative context in selecting target phrase, and sometimes the model would make wrong decisions. Thus we define the final source context as the weighted sum of each local context to fully use the source contexts.

$$c_t = \sum_{i=1}^{n} a_{it} h_i \tag{5}$$

We use c_t to replace h_n as

$$\text{Conf}(X_1^n, P_b^e, T_1^m) = \prod_{t=1}^{m} p(y_t | y_{<t-1}, c_t, h'_e) \tag{6}$$

We named Equation 6 as the indicative context based translation disambiguation model (ICDM). In Equation 6, the source context c_t and the source phrase representation h'_e are first combined by

$$c = \tanh(W c_t + U h'_e) \tag{7}$$

where U and W are weight matrices. Then c is integrated into RNNLM as context as shown in Figure 2. Though there are several kinds of combinations for the two context vectors, such as concatenating, accumulating, etc, this kind of combination achieves better results in our work.

2.2 Model Training

To train the model, we collect the training instances from the bilingual corpus according to the word alignment results. Each training instance consists of three parts,

source context (i.e. **source sentence excluding the source phrase**, otherwise the target phrase always aligns to its source phrase.) , source phrase, and the target phrase which is extracted according to the method in [11]. We use the the cross entropy loss function to train the target phrase RNNLM as equation (7):

$$J = \sum_{(X,P,T)\in D} -\log\ p(T|X,P) \tag{8}$$

where D is the training corpus, T, X, P are the target phrase, source sentence and source phrase. In the update procedure, we leverage the stochastic gradient descent (SGD) algorithm, and Adadelta [21] is used to automatically adapt the learning rate. To speedup decoding when adding ICDM as extra feature to SMT, we use the self-normalize technique [4] for the softmax layer and a shortlist [9] of 10K is used to reduce the output dimension. In addition, we precompute the source context H for all phrase pairs.

2.3 Integration into SMT Decoding

We incorporate the confidence scores into the standard log-linear framework for SMT. Given the context of source sentence, the higher the confidence score is, the better the translation quality of phrase pairs is expected to be. For SMT system, the best translation candidate $\hat{e}$ is calculated by:

$$\hat{e} = \text{argmax}_e P(e|f) \tag{9}$$

where the translation score is given by

$$\begin{aligned} P(e|f) &\propto \sum_i w_i \cdot \log \phi_i(f,e) \\ &= \underbrace{\sum_k w_k \cdot \log \phi_k(f,e)}_{\text{Standard feature scores}} + \underbrace{w_p \cdot \log \text{Conf}_p(f,e)}_{\text{Confidence scores}} \end{aligned} \tag{10}$$

where $\phi_k(f,e)$ denotes the standard feature function and $\text{Conf}_p(f,e)$, w_p are our confidence feature functions and its weights. The detailed feature description is as follows:

Features: Translation model, including translation probabilities and lexical weights for both directions (4 features), 5-gram language model (1 feature), word count (1 feature), phrase count (1 feature), NULL penalty (1 feature), number of hierarchical rules used (1 feature), phrase confidence score (1 feature).

3 Experiments

In this section, we evaluate the performance of our disambiguation model on NIST Chinese-English tasks. The evaluation metric is the case-insensitive IBM BLEU-4 [19].

Settings	NIST 2005	NIST 2006	NIST 2008	NIST 2012	Average
HIERO	37.44	34.81	26.80	27.88	31.73
+NNJM	38.17	35.95	28.04	28.88	32.76
+SCDM	38.28	36.19	28.29	29.02	32.95
+ICDM	38.57	36.65	28.61	29.34	33.29

Table 1. Evaluation results of different methods in BLEU% on four NIST test sets. The "Average" setting is the averaged result of the four test sets.

3.1 Setup

The bilingual data we use is a set of LDC [3] corpus, which consists of around 1M sentence pairs. A 4-gram language model is trained over the English Gigaword corpus (LDC2009T13) and the target monolingual data of the bilingual corpus. The development data is the NIST 2003 dataset, and the test data contains NIST 2005, 2006, 2008 and 2012 datasets. For the disambiguation model training, in order to limit the number of training instances, we limit the frequency of phrase pairs to 3, and only keep top 10 target phrases for each source phrase. Besides, the maximum length of phrase is limited up to 5 The total number of training instance is about 10.3 million. In addition, we use the 30K most frequent words for both Chinese and English. All the remaining words are replaced by a special token "UNK". The hidden dimensions of all RNNs are set to 500.

3.2 Baselines

We have two baselines for comparison. The first is an in-house re-implementation of the hierarchical phrase-based SMT system (HIERO) [3]. The CKY decoding algorithm is used and cube pruning is applied [3,8]. Translation models are trained over the parallel corpus that is automatically aligned using GIZA++ [18] in both directions, and the grow-diag-final heuristic is used to refine symmetric word alignment. For language model, we use an in-house toolkit with modified Kneser-Ney smoothing [10]. The feature weights are tuned by MERT [17].

The second baseline is the HIERO system which incorporates the feed-forward neural network joint model [4] named as +NNJM. The target window is set to 3 and the source-side context is set to 11.

3.3 Evaluation on NIST Task

The evaluation results are shown in Table 1. According to the Table 1, +NNJM can improve HIERO by 1.03 points in average, which shows that source context significantly contributes to improving translation performance. Compared with +NNJM,

[3] LDC2003E14, LDC2005T10, LDC2005E83, LDC2006E26, LDC2006E34, LDC2006E85, LDC2006E92, LDC2003E07,LDC2005T06, LDC2004T08, LDC2005T06

+SCDM performs better on all the test sets, gaining about 0.2 BLEU point improvements in average, which shows global source contexts are more helpful for translation candidate prediction in SMT decoding. Meanwhile, the ICDM outperforms all the others on the whole test sets, where the average improvement is 1.56 compared with HIERO and 0.53 compared with +NNJM. The biggest improvement can be up to 1.86 BLEU points on NIST 2012. The main reason is that ICDM can capture long distance source dependencies and enhance the effect of indicative contexts in predicting translation candidates.

3.4 Analyses of Translation Disambiguation

In this section, we give a case study to explain how our method works. Examples of translation disambiguation are shown in Table 3. We investigate the phrase "有" as an example. Two source sentences with different contexts are selected. Top five target translation candidates are selected from the translation table.

In Table 2, all the phrase pairs are ranked in terms of log confidence scores which vary with different source contexts. We can see that ⟨有, have⟩ gets the best confidence score for S1, and ⟨有, is there⟩ ranks first for S2, though the log translation probability for the same phrase pair is constant as shown in the $\log P$ column. This result significantly caters to the reference. This shows that our method can leverage source contexts to make a better translation selection.

	S1: 您好我想订个房间. 您**有** 空房吗?			S2: 请问这附近**有** 车站吗?		
Reference	hello , i 'd like to make a reservation . do you **have** any rooms ?			excuse me , **is there** a station near here ?		
Conf. Ranking	Phrase pairs	$\log P$	log Conf	Phrase pairs	$\log P$	log Conf
	⟨有, have⟩	-2.19	-1.83	⟨有, is there⟩	-2.89	-1.39
	⟨有, there is⟩	-3.37	-3.00	⟨有, there is⟩	-3.37	-2.82
	⟨有, is there⟩	-2.89	-3.22	⟨有, NULL⟩	-2.51	-4.35
	⟨有, do you have⟩	-2.73	-3.51	⟨有, do you have⟩	-2.73	-5.50
	⟨有, NULL⟩	-2.51	-3.82	⟨有, have⟩	-2.19	-5.53

Table 2. Translation example. log Conf denotes the log confidence score and $\log P$ denotes the log translation probability in the translation table. All the characters are lowercased, and phrases in bold is the phrase pairs.

4 Conclusion and Future Work

In this paper, we propose a translation disambiguation model for SMT. Our model can leverage indicative source contexts for target translation disambiguation. In SMT decoding, appropriate target phrases are selected to best match the source sentence according to the confidence scores. Experimental results show that our

method can significantly improvements the state-of-the-art hierarchical phrase-based system.

In the future, we will perform forced decoding for bilingual training sentences and collect the phrase pairs used in order to obtain high-quality pairs to train our model.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. ICLR 2015 (2015)
2. Carpuat, M., Wu, D.: Word sense disambiguation vs. statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 387–394. Association for Computational Linguistics (2005)
3. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 263–270. Association for Computational Linguistics (2005)
4. Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R.M., Makhoul, J.: Fast and robust neural network joint models for statistical machine translation. In: ACL (1). pp. 1370–1380. Citeseer (2014)
5. Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., Thayer, I.: Scalable inference and training of context-rich syntactic translation models. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. pp. 961–968. Association for Computational Linguistics (2006)
6. He, Z., Liu, Q., Lin, S.: Improving statistical machine translation using lexicalized rule selection. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 321–328. Association for Computational Linguistics (2008)
7. Hu, B., Tu, Z., Lu, Z., Li, H., Chen, Q.: Context-dependent translation selection using convolutional neural network. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 536–541. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-2088
8. Huang, L., Chiang, D.: Better k-best parsing. In: Proceedings of the Ninth International Workshop on Parsing Technology. pp. 53–64. Association for Computational Linguistics (2005)
9. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1–10. Association for Computational Linguistics, Beijing, China (July 2015), http://www.aclweb.org/anthology/P15-1001
10. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. vol. 1, pp. 181–184. IEEE (1995)
11. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP. pp. 388–395. Citeseer (2004)
12. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 48–54. Association for Computational Linguistics (2003)

13. Liu, Q., He, Z., Liu, Y., Lin, S.: Maximum entropy based rule selection model for syntax-based statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 89–97. Association for Computational Linguistics (2008)
14. Liu, Y., Liu, Q., Lin, S.: Tree-to-string alignment template for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. pp. 609–616. Association for Computational Linguistics (2006)
15. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
16. Marton, Y., Resnik, P.: Soft syntactic constraints for hierarchical phrased-based translation. In: ACL. pp. 1003–1011 (2008)
17. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. pp. 160–167. Association for Computational Linguistics (2003)
18. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. pp. 440–447. Association for Computational Linguistics (2000)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
20. Xiong, D., Zhang, M., Aw, A., Li, H.: A syntax-driven bracketing model for phrase-based translation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. pp. 315–323. Association for Computational Linguistics (2009)
21. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
22. Zhang, J.: Local translation prediction with global sentence representation. arXiv preprint arXiv:1502.07920 (2015)