Word Rewarding for Adequate Neural Machine Translation

Yuto Takebayashi[†], *Chu Chenhui*[‡], *Yuki Arase*[†], *Masaaki Nagata*[*]

[†]Graduate School of Information Science and Technology, Osaka University
[‡]Institute for Datability Science, Osaka University
[*]NTT Communication Science Laboratories, NTT Corporation

{takebayashi.yuto, arase}@ist.osaka-u.ac.jp, chu@ids.osaka-u.ac.jp, nagata.masaaki@lab.ntt.co.jp

Abstract

To improve the translation adequacy in neural machine translation (NMT), we propose a rewarding model with target word prediction using bilingual dictionaries inspired by the success of decoder constraints in statistical machine translation. In particular, the model first predicts a set of target words promising for translation; then boosts the probabilities of the predicted words to give them better chances to be output. Our rewarding model minimally interacts with the decoder so that it can be easily applied to the decoder of an existing NMT system. Extensive evaluation under both resource-rich and resource-poor settings shows that (1) BLEU score improves more than 10 points with oracle prediction, (2) BLEU score improves about 1.0 point with target word prediction using bilingual dictionaries created either manually or automatically, (3) hyper-parameters of our model are relatively easy to optimize, and (4) under-generation problem can be alleviated in exchange for increasing over-generated words.

1. Introduction

Neural machine translation (NMT) [1, 2, 3] has dramatically improved machine translation quality compared to statistical machine translation (SMT). However, current NMT systems still suffer from the *adequacy* problem due to inappropriate lexical choice, under-generation, and over-generation [4]. In SMT, bilingual dictionaries have been used to improve adequacy in translation as decoder constraints. Typical example is the XML markup function implemented on MOSES [5].

Inspired by the decoding constraints for SMT, we propose a rewarding model using bilingual dictionaries to address the adequacy problem in NMT. Our model *rewards* target words that are promising to be used in correct translations by boosting their probabilities to be output by a decoder. It predicts such target words using bilingual dictionaries that are created manually or automatically. By applying byte pair encoding (BPE) [6] to dictionaries, our model can benefit from both BPE and dictionaries.

While previous studies incorporate bilingual dictionaries into NMT for translation of rare words [7, 8] and domain-specific words [9], we do so to improve the adequacy of NMT. Hence, dictionaries are made use of translating not only specific types of words but also all words. In addition, these are methodologically different; our model simply biases the trained decoder while previous models change the inside NMT architectures and require training of the entire systems. Due to this design, our model is easy to add to trained NMT systems and compatible with BPE.

Extensive evaluation on Japanese-to-English and English-to-Japanese translation has been conducted using two datasets; IWSLT (TED Talk) [10], spoken language domain with a small set of bilingual sentences (223k), and ASPEC [11], a scientific domain with a large set of bilingual sentences (3M). We refer to the former as a *resource-poor domain* and the latter as a *resource-rich domain,* hereafter. The results show that the rewarding model with oracle prediction of target words, where all and only target words in references are predicted, BLEU score improves more than 10 points on average in both of the resource-poor and resource-rich domains. When using bilingual dictionaries created manually or automatically in the rewarding model to predict target words, BLEU scores improve about 1.0 point on average in both domains.

Detailed analysis of our model reveals that it is relatively insensitive to settings of its hyper-parameters and easy to optimize. In addition, it is shown that our model decreases the number of under-generated words while tends to increase the number of over-generated words.

2. Neural Machine Translation

The encoder-decoder model with attention [3, 12] is one of the most popular architectures in NMT. It takes an input sentence $X = \{x_1, ..., x_n\}$ and generates its translation $Y = \{y_1, ..., y_m\}$ as:

$$p(Y|X;\theta) = \prod_{j=1}^{m} p(y_j|y_{<j}, X;\theta),$$

where θ is a set of parameters and $y_{<j} = \{y_1, \cdots, y_{j-1}\}$. Given a parallel corpus $C = \{(X, Y)\}$, the training objective minimizes the cross-entropy loss with regard to θ:

$$L_\theta = \sum_{(X,Y)\in C} -\log p(Y|X).$$

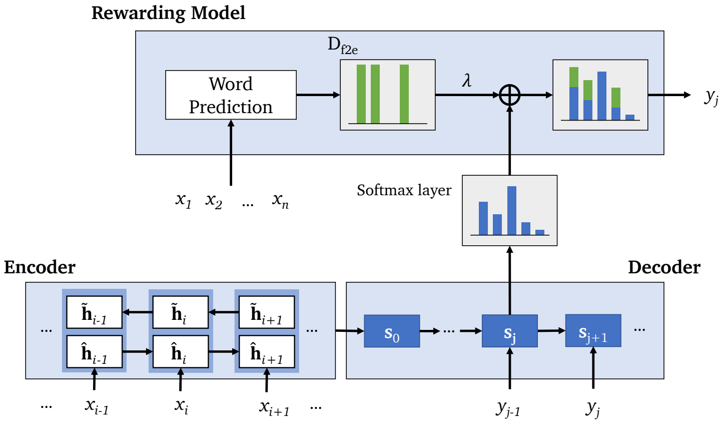


Figure 1: Rewarding model at decoding step j: predicted target words D_{f2e} are rewarded to have better chances to be output at each decoding time step. Note that the attention model is omitted for clarity.

The model consists of three parts, namely, an encoder, a decoder, and an attention model. The encoder has an embedding layer and an recurrent neural network (RNN) layer. The former converts words into their continuous space representations. Taking these embeddings, the RNN layer then computes a state that represents the input sequence till the current time step. Specifically, we use the bi-directional long short-term memory (LSTM) [13] that encodes the source sentence by forward and backward directions. At time step i, the state is represented by concatenating the forward hidden state $\hat{\mathbf{h}}_i$ and the backward one $\tilde{\mathbf{h}}_i$ as $\mathbf{h}_i = [\hat{\mathbf{h}}_i; \tilde{\mathbf{h}}_i]$. In this manner, X can be represented as $\mathbf{h} = \{\mathbf{h}_1, ..., \mathbf{h}_n\}$.

The decoder remembers all the history of translation and its softmax layer computes the posterior probability $p(y_j|y_{<j}, X)$ of a word y_j to output as translation. In order to focus on specific parts of the input sentence necessary for translation, the attention model is incorporated. We use the global attention mechanism proposed in [12].

3. Rewarding Model

On top of a decoder, our model rewards predicted words so that they have better chances to be output as translations as shown in Figure 1. Specifically, it first predicts a set of target words D_{f2e} that are promising to be used in translations using bilingual dictionaries. Then, our model *rewards* a target word if it is contained in D_{f2e} by adding weight to the posterior probability:

$$Q(y_j|y_{<j}, X) = \quad \log p(y_j|y_{<j}, X) + \lambda r_{y_j}, \tag{1}$$

where λ is the weight of reward that will be tuned using a development set. This means that our model boosts the probabilities of predicted words that might have been slipped away during beam search in the conventional decoder. In [14], a similar rewarding model is proposed, but rewards are based on remaining sequence lengths.

We use a simple binary rewarding in this paper:

$$r_{y_j} = \begin{cases} 1 & (y_j \in D_{f2e}), \\ 0 & (\text{otherwise}). \end{cases} \tag{2}$$

We also tried to model the rewarding function using lexical translation probabilities that can be estimated for automatically created dictionaries. However, preliminary experiments empirically showed that this simple form of rewarding worked best. This may be because these probabilities are modeled in completely different ways, *i.e.*, $p(y_j|y_{<j}, X)$ in Equation (1) is conditioned on the entire source sentence while lexical translation probabilities are conditioned on source words. Further investigation is our future work.

Finally, a target word is output as:

$$y_j = \arg\max_{y_j} Q(y_j|y_{<j}, X).$$

Accurate prediction of D_{f2e} is crucial for our rewarding model. In the next section, we discuss practical implementations to obtain D_{f2e} from dictionaries.

4. Target Word Prediction with Dictionaries

In this study, we look up bilingual dictionaries created manually or automatically as word prediction, which allows to make our model minimally interact with the original NMT system. We will consider a sophisticated prediction model using an information in the encoder in future [15].

4.1. Prediction with Manually Created Dictionary

Thanks to the accumulated efforts by the academia and industry, bilingual dictionaries have been manually created for language pairs of English and Japanese. Such manual bilingual dictionaries provide reliable translation knowledge, although their coverage is limited. One disadvantage of manual dictionaries is that conjugation and derivative forms are generally not provided in such dictionaries. As a simple way to predict the target word set, we look up source words in a manual bilingual dictionary.

4.2. Prediction with Automatically Created Dictionary

Previous studies have proposed methods to automatically construct bilingual dictionaries. Especially, word alignment techniques for SMT [16, 5] allow us to construct a dictionary directly from a parallel corpus. Similar alignment may be possible using the attention model in NMT, however, reliability is not assured because the attention model is rather soft as a constraint [17, 18].

The biggest advantage of using word alignment for dictionary construction is that the domain of the dictionary matches that of translation targets. In addition, conjugations are available in the dictionary. A disadvantage is that alignment errors may decrease the quality of the dictionary.

We apply the GIZA++ toolkit[1] that is an implementation of the IBM alignment models [16] on a parallel corpus to automatically create a bilingual dictionary. To control the precision and recall of target word prediction, we introduce a threshold δ, which is tuned on development data. Target words with lower translation probability than δ are discarded.

4.3. Exact and Partial Matching with BPE

Conducting translation on sub-words is effective to address the unknown word problem [19]. We apply BPE [6] to dictionaries for word prediction to make our rewarding model compatible to BPE-based NMT. For both the dictionary entries and source sentences, we first apply a BPE model trained on a parallel corpus and then match the entries in dictionaries and source sentences.

We use two types of matching methods between an input sentence and dictionary entries: *exact match* and *partial match*. The former is precision-oriented and the latter is recall-oriented. After applying BPE, a dictionary headword (lemma) consists of multiple sub-words; a lemma w is denoted as $w = w_1, \ldots, w_k$. *Exact match* regards w as matched to a source sentence X if and only if: $w_1, \ldots, w_k \in X$, *s.t.*, for $\forall i \in \{1, \ldots, k-1\}, w_i = x_j \Leftrightarrow w_{i+1} = x_{j+1}$. On the other hand, *partial match* regards w as matched to X if $w_i \in X$ for $\exists w_i \in w$. In both matching methods, translations of w are added to the target word set as predictions. Obviously, target word predictions by *partial match* subsumes those by *exact match*.

5. Experiment Settings

To investigate the effects of our model, we conducted Japanese-to-English and English-to-Japanese translation experiments on resource-poor and resource-rich domains.

5.1. Translation Tasks

The resource-poor task used the IWSLT 2017 Japanese-English task from the WIT project [10]. The IWSLT task provides 223k parallel sentences for training. We used the dev 2010 and test 2010 sets for development and testing, containing 871 and $1,549$ sentences, respectively.

The resource-rich task used the Japanese-English paper excerpt corpus (ASPEC)[2] [11], which is one subtask of the workshop on Asian translation (WAT)[3] [20]. For training, we used the first 2M parallel sentence pairs among the entire 3M pairs sentences following [21], because the remaining 1M sentences were noisy. The ASPEC task provides $1,790$, and $1,812$ sentences for development and testing, respectively. We conducted both Japanese-to-English and English-to-Japanese translation experiments on these two tasks, referred to as *IWSLT-JE*, *IWSLT-EJ*, *ASPEC-JE*, and *ASPEC-EJ* for short, hereafter.

5.2. NMT and Rewarding Model

We used the mlpnlp-nmt system[4] that is an LSTM based encoder-decoder NMT model with attention, which achieved the best translation performance in human evaluations for both the ASPEC-JE and ASPEC-EJ tasks at WAT 2017 [20].[5] We implemented our rewarding model on top of the mlpnlp-nmt system (our implementation will be public upon acceptance of the paper). We followed the hyper-parameter settings of [21]. The sizes of the source and target side embeddings, the LSTM hidden states, the attention hidden states were all set to 512. We used 2-layer LSTMs for both the encoder and decoder with beam size of 5. Stochastic gradient descent was used as the learning algorithm, with an initial learning rate of 1.0, gradient clipping of 5.0, and a dropout rate of 30% for the inter-layer dropout. The mini batch size was 128. The training epochs for IWSLT-JE, IWSLT-EJ, ASPEC-JE, and ASPEC-EJ were all set to 20, and we chose the model with the best development BLEU score among all the epochs as the baseline systems.[6]

For the rewarding models, λ in Equation (1) was tuned on the development sets from 0.1 to 1.0 by 0.1 interval. The threshold δ that prunes the automatically constructed dictionaries in Section 4.2 was tuned on 0, 0.0001, 0.001, 0.01 and 0.1. We selected the best combination among all combinations of δ and λ on the development set for each model.

We investigate the upper-bound performance of our rewarding model using oracle target word prediction. On this oracle model, predicted target words are all and only words in a reference translation, *i.e.*, precision and recall of prediction are both 100%. The best weight of λ was searched from 0.1 increasing the value by 0.1 until we observed a decrease in BLEU scores.

As preprocessing for the parallel corpora and bilingual dictionaries, we segmented Japanese sentences/entries using MeCab,[7] and tokenized and truecased the English sen-

[1] http://code.google.com/p/giza-pp

[2] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

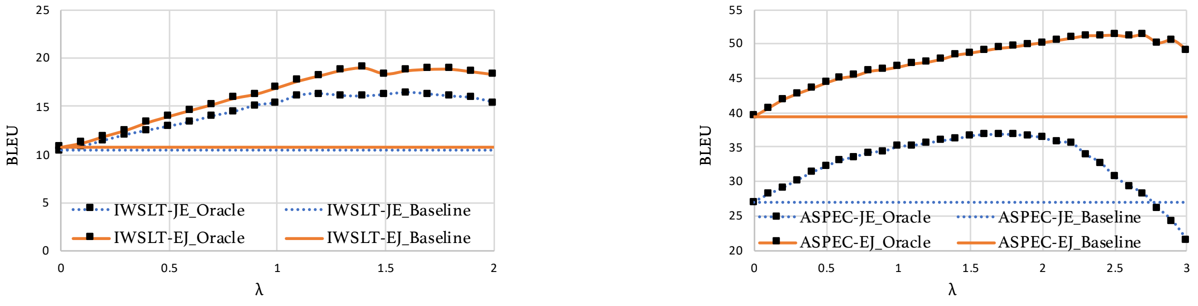
[3] http://orchid.kuee.kyoto-u.ac.jp/WAT/

[4] https://github.com/mlpnlp/mlpnlp-nmt/

[5] Experiments on other NMT models as future work.

[6] Epoch #11, #20, #13 and #13 for IWSLT-JE, IWSLT-EJ, ASPEC-JE, and ASPEC-EJ, respectively.

[7] https://github.com/taku910/mecab



(a) IWSLT

(b) ASPEC

Figure 2: BLEU scores by the oracle rewarding model when changing the λ on the development set. BLEU scores dramatically improved on ASPEC task; 9.8 and 11.8 point improvements on ASPEC-JE and EJ, respectively.

tences/entries with the *truecase.perl* script in Moses[8] for both translation tasks. We further split the words into subwords using joint BPE [6] with $32,000$ merge operations. The vocabulary sizes of the IWSLT-JE task were $21,534$ and $18,022$, respectively. The vocabulary sizes of ASPEC-JE task were $28,852$ and $22,340$, respectively.

5.3. Bilingual Dictionaries

As the manual dictionary, we used EDR,[9] which is the publicly available English and Japanese bilingual dictionary.[10] The numbers of English-to-Japanese and Japanese-to-English entry pairs are 676k and $1,052$k, respectively. In EDR, only lemmas are provided and thus inflected forms of English verbs are unavailable. To address this issue, inflected forms of the EDR lemmas are extracted from the English dictionary of XTAG project,[11] which is used as the English morphological analysis dictionary for TreeTagger.[12] All the possible inflected forms are added into our dictionary.

For dictionary look-up, a source sentence is first lemmatized and matched with the dictionary. We used MeCab for Japanese and TreeTagger for English to lemmatize words.

To automatically construct bilingual dictionaries,[13] we used the GIZA++ toolkit on the training corpus in both English-to-Japanese and Japanese-to-English directions.[14] We applied the “grow-diag-final-and” heuristic and obtained lexical translation probabilities using Moses. We then prune translation pairs with low probabilities by δ.

6. Results

We first investigate the effect of λ using the development sets on both the oracle target word sets and our word prediction methods. Next, we evaluate the translation quality on the test sets using the optimized λ. Finally, we conduct detailed analysis of translation results by our rewarding model.

Throughout the section, the BLEU-4 score was used as the evaluation metric, which was computed using the *multi-bleu.perl* script in Moses on tokenized and truecased English and word-segmented Japanese sentences, respectively. The significance tests were performed using the bootstrap resampling [22] at $p < 0.01$.

6.1. Effects of λ

Figure 2 shows the BLEU scores by the oracle word rewarding on the development sets of the IWSLT-JE, IWSLT-EJ, ASPEC-JE, and ASPEC-EJ tasks. The BLEU scores significantly improved according to the λ. The best settings of λ improves 6.00, 8.25, 9.80, and 11.77 BLEU scores on the IWSLT-JE, IWSLT-EJ, ASPEC-JE, and ASPEC-EJ tasks from each baseline system, respectively.

Figure 3 shows the BLEU scores with respect to the λ and precision/recall of word prediction on our model with word prediction using manually or automatically created dictionaries. EDR indicates the models predicting target words using EDR. XTAG indicates the models using EDR extended with XTAG, which are only for the Japanese-to-English direction. GIZA indicates the models that predict target words using automatically constructed dictionary by GIZA++. The suffixes *e* and *p* in the legends indicate *exact match* and *partial match*, respectively.

The results show that BLEU scores depend on precision and recall of target word prediction by different dictionaries. The weights of λ that achieved the best BLEU scores varied from 0.1 to 1.0. Notice that these weights are much smaller than the oracle prediction, which are 0.5, 0.4, 0.4, and 0.5 for IWSLT-JE, IWSLT-EJ, ASPEC-JE, and ASPEC-EJ on GIZA *partial-match*, respectively. This is because predicted words are less reliable and too much rewarding degrades the trans-

[8] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl

[9] http://www2.nict.go.jp/ipp/EDR/ENG/indexTop.html?

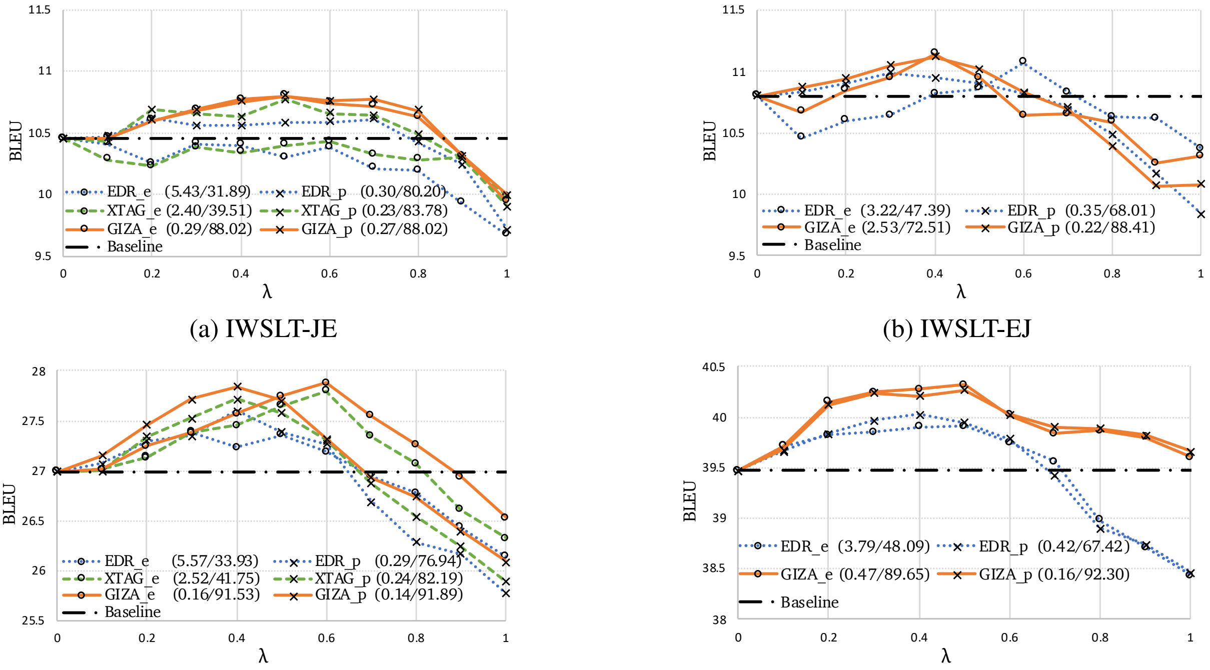
[10] https://www.nict.go.jp/en/about/

[11] https://www.cis.upenn.edu/∼xtag/

[12] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[13] Note that the corpora used for building the dictionaries are the same as the one used for training each NMT systems. Other resources have not been used to create automatic dictionaries.

[14] Note that GIZA++ was applied on the parallel corpora without BPE, which was only used for look up a source word in a dictionary.



(a) IWSLT-JE

(b) IWSLT-EJ

(c) ASPEC-JE

(d) ASPEC-EJ

Figure 3: BLEU scores by our rewarding models with word prediction using bilingual dictionaries when changing the λ on the development sets. The gentle convex curves of BLEU scores show that the weight of λ is tunable by a simple grid search.

lation quality. The gentle convex curves of BLEU scores also show that λ is easily tunable using a simple grid search.

6.2. Word Prediction and Translation Results

Table 1 shows the comparison of BLEU scores on the test sets of the baseline and the rewarding models. We also report the results that use a merged dictionary. We chose the XTAG partial and GIZA partial for Japanese-to-English, EDR partial and GIZA partial for English-to-Japanese for merging because of their individual good performance. We tuned the λ for merged dictionary using the development set.

We can see that compared to the baselines, most of our methods significantly improve BLEU scores. Overall, a word prediction method with high recall shows a larger improvement in BLEU score as consistently shown by comparing exact matching *v.s.* partial matching, as well as comparing EDR *v.s.* XTAG, EDR or XTAG *v.s.* GIZA, and GIZA *v.s.* merged dictionary. However, there is still a gap between rewarding by our target word prediction and rewarding by oracle prediction. Our GIZA and merged dictionary models achieve a high recall of about 90% but a very low precision of 0.1%. Improving the precision for word prediction while keeping a recall high is our future work.

The baselines on ASPEC-JE and ASPEC-EJ are our reproduction of the state-of-the-art at WAT competition as single models, which are reported as achieved 27.62 and 39.71 BLEU scores in the paper. Compared to these scores, our rewarding model improved 0.67 and 0.36 points, respectively.

6.3. Under and Over Generation

We investigated the rate of under-generation and over-generation that are the major adequacy problems in NMT [23] using Translation Edit Rate (TER) [24]. TER aligns a reference and translation result. We counted the number of *Deletion* and *Insertion* regarding these are caused by under and over generation, respectively. This is an approximation to detect under and over generations, but we consider it is useful as an automatic and handy evaluation metric.

Table 2 shows the average numbers of under and over generations per sentence. The under-generation decreases on all the rewarding models in exchange of increasing over-generation. The rewarding model with oracle target word prediction reduces under generation about 1.2 word on average. This result shows that our rewarding model is also effective for alleviating the under-generation problem. The over-generation can be reduced by adding global constraint to the rewarding model, which prohibits rewarding the same predicted target. This is our future work.

Example translations of the baseline and our rewarding model (GIZA partial match) are shown in the following. The phrase of “congenital immunity” and “cancer of” were successfully translated by our model.

Source IL - 1 2 の 癌 に対する 抵抗 性 (先天 免疫) の 生物 反応 について も 考察 し た

Reference biological response of the resistance (*congenital immunity*) to *cancer of* IL - 12 was also examined .

		IWSLT				ASPEC			
		JE		EJ		JE		EJ	
		BLEU	(Pre. / Rec.)	BLEU	(Pre. / Rec.)	BLEU	(Pre. / Rec.)	BLEU	(Pre. / Rec.)
Baseline		9.97	(- / -)	10.26	(- / -)	27.21	(- / -)	39.50	(- / -)
EDR	exact	9.99	(5.19 / 30.58)	**10.75**	(3.14 / 46.08)	**27.82**	(5.57 / 34.10)	39.74	(3.70 / 48.33)
	partial	10.06	(0.27 / 79.50)	**10.73**	(0.33 / 67.09)	**27.94**	(0.28 / 77.15)	**40.05**	(0.42 / 67.37)
XTAG	exact	9.94	(2.25 / 38.10)	-	(- / -)	**27.73**	(2.50 / 41.97)	-	(- / -)
	partial	**10.30**	(0.20 / 82.88)	-	(- / -)	**28.00**	(0.23 / 82.42)	-	(- / -)
GIZA	exact	**10.36**	(0.27 / 85.98)	**10.88**	(2.56 / 72.92)	**28.29**	(0.15 / 91.48)	**39.96**	(0.46 / 89.73)
	partial	**10.32**	(0.25 / 87.61)	**10.83**	(0.21 / 87.38)	**28.28**	(0.13 / 91.80)	**40.07**	(0.15 / 91.65)
Merged dictionary		**10.33**	(0.16 / 89.25)	**10.81**	(0.17 / 88.45)	**28.29**	(0.14 / 91.77)	**40.05**	(0.17 / 92.12)
Oracle		**17.68**	(100 / 100)	**20.26**	(100 / 100)	**37.13**	(100 / 100)	**52.22**	(100 / 100)

Table 1: Comparison of BLEU scores on the test sets (The scores in bold indicate that the results are significantly better than the baseline at $p < 0.01$). The best improvement in BLEU score is 1.08 point when using GIZA *exact-match* in ASPEC-JE.

	under-generation				over-generation			
	IWSLT		ASPEC		IWSLT		ASPEC	
	JE	EJ	JE	EJ	JE	EJ	JE	EJ
Baseline	3.58	3.25	3.37	3.36	**1.64**	**2.04**	**2.27**	**1.69**
EDR_e	3.53	2.89	3.07	2.94	1.70	2.40	2.51	2.13
EDR_p	3.44	3.13	2.85	2.90	1.69	2.18	2.70	2.09
XTAG_e	3.48	-	2.92	-	1.90	-	2.64	-
XTAG_p	3.14	-	2.92	-	2.36	-	2.64	-
GIZA_e	3.18	2.90	2.75	**2.78**	2.33	2.58	2.67	2.15
GIZA_p	3.20	**2.86**	2.75	**2.78**	2.34	2.52	2.66	2.15
Oracle	**2.40**	**2.86**	**2.71**	3.01	4.26	2.80	3.04	3.06

Table 2: Numbers of under/over-generated words per sentence estimated by TER (The scores in bold indicate the best scores).

Baseline the biological response of the resistance to IL - 12 is also discussed .

Our Model the biological response of the resistance (*congenital immunity*) to the *cancer of* IL - 12 is also discussed .

7. Related Work

Our rewarding model can be viewed as a constraint on the decoder to output desired target words. There have been studies that aim to output predetermined words or phrases in neural language generation. For this purpose, the grid beam search in NMT is proposed [25] and the SMT lattice is combined into NMT [26]. In neural conversation generation, Wen et al. (2015) input a vector representing which information should be generated to an encoder [27], and a decoder is designed to explicitly control generation of emotional words [28].

Compared with these previous studies, one benefit of our rewarding model is that the predicted words are used as soft constraints on outputs with minimal interaction to the decoder. The most relevant study from the methodological point of view is [14] that also proposes a rewarding model in a decoder of NMT to improve the translation quality in general, such as remaining sequence lengths to output. We focus on the adequacy problem in NMT and combine word prediction with bilingual dictionaries. Some studies tackle the adequacy problem in NMT, but they require an independent SMT system [29, 30] or modification of the decoder [31]. Different from these, ours is simple and a cost-effective solution for the adequacy problem.

The under and over-generation problems have been recognized not only in NMT, but in other applications that use the encoder-decoder model for natural language generation. Different solutions have been proposed. First, a coverage vector is introduced in NMT [23, 32, 33] that tracks which source words have been translated by the attention mechanism. A sparse and constrained attention has been proposed [34], while word prediction, which are also used to reduce computational cost of softmax function at the decoder [35, 36], has been proposed to solve the under-generation problem. The decoder in [37] encourages to output predicted target words by initializing the decoder through word prediction, and the model in [38] predicts target words and their expected frequencies to resolve the under and over generation problems in NMT-based summarization.

8. Conclusion

We proposed a rewarding model with word prediction to boost the translation probabilities of the predicted target words that should be in correct translations. Our model allows incorporating bilingual dictionaries on a BPE-based NMT system. Extensive evaluation on both resource-poor and resource-rich domains showed its effectiveness.

As future work, first, we plan to improve the precision of word prediction preserving the recall at high. Second, we plan to improve our rewarding model to effectively incorporate translation probabilities and extend the model to reward not only words but also phrases. We will also consider a global constraint by predicting not only target words but their frequencies, and adjust rewards when a word has been used in translation. Finally, more experiments on datasets of various domains and language pairs will be conducted to investigate the generality of our approach.

9. References

[1] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "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)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: http://www.aclweb.org/anthology/D14-1179

[2] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2014, pp. 3104–3112. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969173

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. San Diego, USA: International Conference on Learning Representations, May 2015.

[4] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, August 2017, pp. 28–39. [Online]. Available: http://www.aclweb.org/anthology/W17-3204

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. . Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "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*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 177–180. [Online]. Available: http://www.aclweb.org/anthology/P/P07/P07-2045

[6] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 1715–1725. [Online]. Available: http://www.aclweb.org/anthology/P16-1162

[7] P. Arthur, G. Neubig, and S. Nakamura, "Incorporating discrete translation lexicons into neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 1557–1567. [Online]. Available: https://aclweb.org/anthology/D16-1162

[8] J. Zhang and C. Zong, "Bridging neural machine translation and bilingual dictionaries," *CoRR*, vol. abs/1610.07272, 2016. [Online]. Available: http://arxiv.org/abs/1610.07272

[9] M. Arcan and P. Buitelaar, "Translating domain-specific expressions in knowledge bases with neural machine translation," *CoRR*, vol. abs/1709.02184, 2017. [Online]. Available: http://arxiv.org/abs/1709.02184

[10] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[11] T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara, "Aspec: Asian scientific paper excerpt corpus," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA), May 2016.

[12] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D15-1166

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[14] J. Li, W. Monroe, and D. Jurafsky, "Learning to decode for future success," *CoRR*, vol. abs/1701.06549, 2017. [Online]. Available: http://arxiv.org/abs/1701.06549

[15] S. Ma, X. SUN, Y. Wang, and J. Lin, "Bag-of-words as target for neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 332–338. [Online]. Available: http://www.aclweb.org/anthology/P18-2053

[16] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–312, 1993.

[17] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," in *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016)*, Osaka, Japan, December 2016, pp. 3093–3102.

[18] H. Mi, Z. Wang, and A. Ittycheriah, "Supervised attentions for neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 2283–2288. [Online]. Available: https://aclweb.org/anthology/D16-1249

[19] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in 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)*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 11–19. [Online]. Available: http://www.aclweb.org/anthology/P15-1002

[20] T. Nakazawa, S. Higashiyama, C. Ding, H. Mino, I. Goto, H. Kazawa, Y. Oda, G. Neubig, and S. Kurohashi, "Overview of the 4th workshop on asian translation," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, November 2017, pp. 1–54. [Online]. Available: http://www.aclweb.org/anthology/W17-5701

[21] M. Morishita, J. Suzuki, and M. Nagata, "Ntt neural machine translation systems at wat 2017," in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, November 2017, pp. 89–94. [Online]. Available: http://www.aclweb.org/anthology/W17-5706

[22] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 388–395.

[23] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 76–85. [Online]. Available: http://www.aclweb.org/anthology/P16-1008

[24] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of Association for Machine Translation in the Americas*, 2006.

[25] C. Hokamp and Q. Liu, "Lexically constrained decoding for sequence generation using grid beam search," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1535–1546. [Online]. Available: http://aclweb.org/anthology/P17-1141

[26] F. Stahlberg, A. de Gispert, E. Hasler, and B. Byrne, "Neural machine translation by minimising the bayes-risk with respect to syntactic translation lattices," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 362–368. [Online]. Available: http://www.aclweb.org/anthology/E17-2058

[27] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1711–1721. [Online]. Available: http://aclweb.org/anthology/D15-1199

[28] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," 2018.

[29] J. Zhang, M. Utiyama, E. Sumita, G. Neubig, and S. Nakamura, "Improving neural machine translation through phrase-based forced decoding," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, November 2017, pp. 152–162. [Online]. Available: http://www.aclweb.org/anthology/I17-1016

[30] L. Zhou, W. Hu, J. Zhang, and C. Zong, "Neural system combination for machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 378–384. [Online]. Available: http://aclweb.org/anthology/P17-2060

[31] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context gates for neural machine translation," *Transactions of the Association for Computational Linguistics*,

vol. 5, pp. 87–99, 2017. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/948

[32] F. Meng, Z. Lu, H. Li, and Q. Liu, "Interactive attention for neural machine translation," in *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016)*, Osaka, Japan, December 2016, pp. 2174–2185.

[33] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[34] C. Malaviya, P. Ferreira, and A. F. T. Martins, "Sparse and constrained attention for neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018, pp. 370–376. [Online]. Available: http://aclweb.org/anthology/P18-2059

[35] X. Shi and K. Knight, "Speeding up neural machine translation decoding by shrinking run-time vocabulary," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 574–579. [Online]. Available: http://aclweb.org/anthology/P17-2091

[36] B. Sankaran, M. Freitag, and Y. Al-Onaizan, "Attention-based vocabulary selection for NMT decoding," *CoRR*, vol. abs/1706.03824, 2017. [Online]. Available: http://arxiv.org/abs/1706.03824

[37] R. Weng, S. Huang, Z. Zheng, X.-Y. DAI, and J. CHEN, "Neural machine translation with word predictions," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 136–145. [Online]. Available: https://www.aclweb.org/anthology/D17-1013

[38] J. Suzuki and M. Nagata, "Cutting-off redundant repeating generations for neural abstractive summarization," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 291–297. [Online]. Available: http://www.aclweb.org/anthology/E17-2047