Phrase Table as Recommendation Memory for Neural Machine Translation

Abstract

Neural Machine Translation (NMT) has drawn much attention due to its promising translation performance recently. However, several studies indicate that NMT often generates fluent but unfaithful translations. In this paper, we propose a method to alleviate this problem by using a phrase table as recommendation memory. The main idea is to add bonus to words worthy of recommendation, so that NMT can make correct predictions. Specifically, we first derive a prefix tree to accommodate all the candidate target phrases by searching the phrase translation table according to the source sentence. Then, we construct a recommendation word set by matching between candidate target phrases and previously translated target words by NMT. After that, we determine the specific bonus value for each recommendable word by using the attention vector and phrase translation probability. Finally, we integrate this bonus value into NMT to improve the translation results. The extensive experiments demonstrate that the proposed methods obtain remarkable improvements over the strong attentionbased NMT.

Introduction

The past several years have witnessed a significant progress in Neural Machine Translation (NMT). Most NMT methods are based on the encoder-decoder architecture BIBREF0 , BIBREF1 , BIBREF2 and can achieve promising translation performance in a variety of language pairs BIBREF3 , BIBREF4 , BIBREF5 .

However, recent studies BIBREF6 , BIBREF7 show that NMT often generates words that make target sentences fluent, but unfaithful to the source sentences. In contrast, traditional Statistical Machine

Translation (SMT) methods tend to rarely make this kind of mistakes. Fig. 1 shows an example that NMT makes mistakes when translating the phrase “jinkou dafu xiahua (the sharp decline in imports)” and the phrase “maoyi shuncha (the trade surplus)”, but SMT can produce correct results when translating these two phrases. BIBREF6 argues that the reason behind this is the use of distributed representations of words in NMT makes systems often generate words that seem natural in the context, but do not reflect the content of the source sentence. Traditional SMT can avoid this problem as it produces the translations based on phrase mappings.

Therefore, it will be beneficial to combine SMT and NMT to alleviate the previously mentioned problem. Actually, researchers have made some effective attempts to achieve this goal. Earlier studies were based on the SMT framework, and have been deeply discussed in BIBREF8 . Later, the researchers transfers to NMT framework. Specifically, coverage mechanism BIBREF9 , BIBREF10 , SMT features BIBREF11 , BIBREF12 , BIBREF13 , BIBREF14 , BIBREF15 and translation lexicons BIBREF6 , BIBREF16 , BIBREF17 have been fully explored. In contrast, phrase translation table, as the core of SMT, has not been fully studied. Recently, BIBREF18 and BIBREF19 explore the possibility of translating phrases in NMT. However, the “phrase” in their approaches are different from that used in phrase-based SMT. In BIBREF18 's models, the phrase pair must be a one-to-one mapping with a source phrase having a unique target phrase (named entity translation pairs). In BIBREF19 's models, the source side of a phrase pair must be a chunk. Therefore, it is still a big challenge to incorporate any phrase pair in the phrase table into NMT system to alleviate the unfaithfulness problem.

In this paper, we propose an effective method to incorporate a phrase table as recommendation memory into the NMT system. To achieve this, we add bonuses to the words in recommendation set to help NMT make better predictions. Generally, our method contains three steps. 1) In order to find out which words are worthy to recommend, we first derive a candidate target phrase set by searching the phrase table according to the input sentence. After that, we construct a recommendation word set at each decoding

step by matching between candidate target phrases and previously translated target words by NMT. 2) We then determine the specific bonus value for each recommendable word by using the attention vector produced by NMT and phrase translation probability extracted from phrase table. 3) Finally we integrate the word bonus value into the NMT system to improve the final results.

In this paper, we make the following contributions:

1) We propose a method to incorporate the phrase table as recommendation memory into NMT system. We design a novel approach to find from the phrase table the target words worthy of recommendation, calculate their recommendation scores and use them to promote NMT to make better predictions.

2) Our empirical experiments on Chinese-English translation and English-Japanese translation tasks show the efficacy of our methods. For Chinese-English translation, we can obtain an average improvement of 2.23 BLEU points. For English-Japanese translation, the improvement can reach 1.96 BLEU points. We further find that the phrase table is much more beneficial than bilingual lexicons to NMT.

Neural Machine Translation

NMT contains two parts, encoder and decoder, where encoder transforms the source sentence INLINEFORM0 into context vectors INLINEFORM1 . This context set is constructed by INLINEFORM2 stacked Long Short Term Memory (LSTM) BIBREF20 layers. INLINEFORM3 can be calculated as follows: DISPLAYFORM0

The decoder generates one target word at a time by computing the probability of INLINEFORM0 as follows: DISPLAYFORM0

where INLINEFORM0 is the score produced by NMT: DISPLAYFORM0

and INLINEFORM0 is the attention output: DISPLAYFORM0

the attention model calculates INLINEFORM0 as the weighted sum of the source-side context vectors: DISPLAYFORM0 DISPLAYFORM1

INLINEFORM0 is computed using the following formula: DISPLAYFORM0

Phrase Table as Recommendation Memory for NMT

In section 2 we described how the standard NMT models calculate the probability of the next target word (Eq. (2)). Our goal in this paper is to improve the accuracy of this probability estimation by incorporating information from phrase tables. Our main idea is to find the recommendable words and increase their probabilities at each decoding time step. Thus, three questions arise:

1) Which words are worthy to recommend at each decoding step?

2) How to determine an appropriate bonus value for each recommendable word?

3) How to integrate the bonus value into NMT?

In this section, we will describe the specific methods to answer above three questions. As the basis of our work, we first introduce two definitions used by our methods.

Definition 1 (prefix of phrase): the prefix of a phrase is a word sequence which begins with the first word

of the phrase and ends with any word of the phrase. Note that the prefix string can be empty. For a phrase INLINEFORM0 , this phrase contains four prefixes: INLINEFORM1 .

Definition 2 (suffix of partial translation): the suffix of the partial translation INLINEFORM0 is a word sequence, which begins with any word belonging to INLINEFORM1 , and ends with INLINEFORM2 . Similarly, the suffix string can also be empty. For partial translation INLINEFORM3 , there are four suffixes INLINEFORM4 .

Word Recommendation Set

The first step is to derive a candidate target phrase set for a source sentence. The recommendation words are selected from this set.

Given a source sentence INLINEFORM0 and a phrase translation table (as shown in upper right of Fig. 2), we can traverse the phrase translation table and get all the phrase pairs whose source side matches the input source sentence. Then, for each phrase pair, we add the target phrases with the top INLINEFORM1 highest phrase translation probabilities into the candidate target phrase set.

In order to improve efficiency of the next step, we represent this candidate target phrase set in a form of prefix tree. If the phrases contain the same prefix (Definition 1), the prefix tree can merge them and represent them using the same non-terminal nodes. The root of this prefix tree is an empty node. Fig. 2 shows an example to illustrate how we get the candidate target phrase set for a source sentence. In this example, In phrase table (upper right), we find four phrases whose source side matches the source sentence (upper left). We add the target phrases into candidate target phrase set (middle). Finally, we use a prefix tree (bottom) to represent the candidate target phrases.

With above preparations, we can start to construct the word recommendation set. In our method, we need to construct a word recommendation set INLINEFORM0 at each decoding step INLINEFORM1 . The basic idea is that if a prefix INLINEFORM2 (Definition 1) of a phrase in candidate target phrase set matches a suffix INLINEFORM3 (Definition 2) of the partial translation INLINEFORM4 , the next word of INLINEFORM5 in the phrase may be the next target word INLINEFORM6 to be predicted and thus is worthy to recommend.

Here, we take Fig. 2 as an example to illustrate our idea. We assume that the partial translation is “he settled in the US, and lived in the suburb of”. According to our definition, this partial translation contains a suffix “suburb of”. Meanwhile, in candidate target phrase set, there is a phrase (“suburb of Milwaukee”) whose two-word prefix is “suburb of” as well. We can notice that the next word of the prefix ("Milwaukee") is exactly the one that should be predicted by the decoder. Thus, we recommend “Milwaukee” by adding a bonus to it with the hope that when this low-frequency word is mistranslated by NMT, our recommendation can fix this mistake.

Under this assumption, the procedure of constructing the word recommendation set INLINEFORM0 is illustrated in Algorithm 1. We first get all suffixes of INLINEFORM1 (line 2) and all prefixes of target phrases belonging to candidate target phrase set (line 3). If a prefix of the candidate phrase matches a suffix of INLINEFORM2 , we add the next word of the prefix in the phrase into recommendation set INLINEFORM3 (line 4-7).

In the definition of the prefix and suffix, we also allow them to be an empty string. By doing so, we can add the first word of each phrase into the word recommendation set, since the suffix of INLINEFORM0 and the prefix of any target phrase always contain a match part INLINEFORM1 . The reason we add the first word of the phrase into recommendation set is that we hope our methods can still recommend some possible words when NMT has finished the translation of one phrase and begins to translate another new

one, or predicts the first target word of the whole sentence.

[t] Construct recommendation word set Input: candidate target phrase set; already generated partial translation INLINEFORM0

Output: word recommendation set INLINEFORM0 [1] INLINEFORM1 Get all suffixes of INLINEFORM2 (denote each suffix by INLINEFORM3) Get all prefixes of each target phrase in candidate target phrase set (denote every prefix by INLINEFORM4) each suffix INLINEFORM5 and each prefix INLINEFORM6 INLINEFORM7 Add the next word of INLINEFORM8 into INLINEFORM9 INLINEFORM10

Now we already know which word is worthy to recommend. In order to facilitate the calculation of the bonus value (section 3.2), we also need to maintain the origin of each recommendation word. Here, the origin of a recommendation word contains two parts: 1) the phrase pair this word belongs to and 2) the phrase translation probability between the source and target phrases. Formally, for a recommendation word INLINEFORM0 , we can denote it by: DISPLAYFORM0

where INLINEFORM0 denotes the INLINEFORM1 -th phrase pair the recommendation word INLINEFORM2 belongs to (some words may belong to different phrase pairs and INLINEFORM3 denotes the number of phrase pairs). INLINEFORM4 is the source phrase and INLINEFORM5 is the target phrase. INLINEFORM6 is the phrase translation probability between the source and target phrases. Take Fig. 2 as an example. When the partial translation is “he”, word “settled” can be recommended according to algorithm 1. Word “settled” is contained in two phrase pairs and the translation probabilities are respectively 0.6 and 0.4. Thus, we can denote the word "settled" as follows: DISPLAYFORM0

Bonus Value Calculation

The next task is to calculate the bonus value for each recommendation word. For a recommendation word INLINEFORM0 denoted by Eq. (8), its bonus value is calculated as follows:

Step1: Extracting each phrase translation probability INLINEFORM0 .

Step2: For each phrase pair INLINEFORM0 , we convert the attention weight INLINEFORM1 in NMT (Eq. (6)) between target word INLINEFORM2 and source word INLINEFORM3 to phrase alignment probability INLINEFORM4 between target word INLINEFORM5 and source phrase INLINEFORM6 as follows: DISPLAYFORM0

where INLINEFORM0 is the number of words in phrase INLINEFORM1 . As shown in Eq. (10), our conversion method is making an average of word alignment probability INLINEFORM2 whose source word INLINEFORM3 belongs to source phrase INLINEFORM4 .

Step3: Calculating the bonus value for each recommendation word as follows: DISPLAYFORM0

From Eq. (11), the bonus value is determined by two factors, i.e., 1) alignment information INLINEFORM0 and 2) translation probability INLINEFORM1 . The process of involving INLINEFORM2 is important because the bonus value will be influenced by different source phrases that systems focus on. And we take INLINEFORM3 into consideration with a hope that the larger INLINEFORM4 is, the larger its bonus value is.

Integrating Bonus Values into NMT

The last step is to combine the bonus value with the conditional probability of the baseline NMT model (Eq.(2)). Specifically, we add the bonuses to the words on the basis of original NMT score (Eq. (3)) as

follows: DISPLAYFORM0

where INLINEFORM0 is calculated by Eq. (11). INLINEFORM1 is the bonus weight, and specifically, it is the result of sigmoid function (INLINEFORM2), where INLINEFORM3 is a learnable parameter, and this sigmoid function ensures that the final weight falls between 0 and 1.

Experimental Settings

In this section, we describe the experiments to evaluate our proposed methods.

Dataset

We test the proposed methods on Chinese-to-English (CH-EN) translation and English-to-Japanese (EN-JA) translation. In CH-EN translation, we test the proposed methods with two data sets: 1) small data set, which includes 0.63M sentence pairs; 2) large-scale data set, which contains about 2.1M sentence pairs. NIST 2003 (MT03) dataset is used for validation. NIST2004-2006 (MT04-06) and NIST 2008 (MT08) datasets are used for testing. In EN-JA translation, we use KFTT dataset, which includes 0.44M sentence pairs for training, 1166 sentence pairs for validation and 1160 sentence pairs for testing.

Training and Evaluation Details

We use the Zoph_RNN toolkit to implement all our described methods. In all experiments, the encoder and decoder include two stacked LSTM layers. The word embedding dimension and the size of hidden layers are both set to 1,000. The minibatch size is set to 128. We limit the vocabulary to 30K most frequent words for both the source and target languages. Other words are replaced by a special symbol “UNK”. At test time, we employ beam search and beam size is set to 12. We use case-insensitive 4-gram

BLEU score as the automatic metric BIBREF21 for translation quality evaluation.

Phrase Translation Table

Our phrase translation table is learned directly from parallel data by Moses BIBREF22 . To ensure the quality of the phrase pair, in all experiments, the phrase translation table is filtered as follows: 1) out-of-vocabulary words in the phrase table are replaced by UNK; 2) we remove the phrase pairs whose words are all punctuations and UNK; 3) for a source phrase, we retain at most 10 target phrases having the highest phrase translation probabilities.

Translation Methods

We compare our method with other relevant methods as follows:

1) Moses: It is a widely used phrasal SMT system BIBREF22 .

2) Baseline: It is the baseline attention-based NMT system BIBREF23 , BIBREF24 .

3) Arthur: It is the state-of-the-art method which incorporates discrete translation lexicons into NMT model BIBREF6 . We choose automatically learned lexicons and bias method. We implement the method on the base of the baseline attention-based NMT system. Hyper parameter INLINEFORM0 is 0.001, the same as that reported in their work.

Translation Results

Table 1 reports the detailed translation results for different methods. Comparing the first two rows in

Table 1, it is very obvious that the attention-based NMT system Baseline substantially outperforms the phrase-based SMT system Moses on both CH-EN translation and EN-JA translation. The average improvement for CH-EN and EN-JA translation is up to 3.99 BLEU points (32.71 vs. 28.72) and 3.59 BLEU (25.99 vs. 22.40) points, respectively.

Effect of Integrating Phrase Translation Table

The first question we are interested in is whether or not phrase translation table can improve the translation quality of NMT. Compared to the baseline, our method markedly improves the translation quality on both CH-EN translation and EN-JA translation. In CH-EN translation, the average improvement is up to 2.23 BLEU points (34.94 vs. 32.71). In EN-JA translation, the improvement can reach 1.96 BLEU points (27.95 vs. 25.99). It indicates that incorporating a phrase table into NMT can substantially improve NMT's translation quality.

In Fig. 3, we show an illustrative example of CH-EN translation. In this example, our method is able to obtain a correct translation while the baseline is not. Specifically, baseline NMT system mistranslates “jinkou dafu xiahua (the sharp decline in imports)” into “import of imports”, and incorrectly translates “maoyi shuncha (trade surplus)” into “trade”. But these two mistakes are fixed by our method, because there are two phrase translation pairs (“jinkou dafu xiahua” to “the sharp decline in imports” and “maoyi shuncha” to “trade surplus”) in the phrase table, and the correct translations are obtained due to our recommendation method.

Lexicon vs. Phrase

A natural question arises that whether it is more beneficial to incorporate a phrase translation table than the translation lexicons. From Table 1, we can conclude that both translation lexicons and phrase

translation table can improve NMT system's translation quality. In CH-EN translation, Arthur improves the baseline NMT system with 0.81 BLEU points, while our method improves the baseline NMT system with 2.23 BLEU points. In EN-JA translation, Arthur improves the baseline NMT system with 0.73 BLEU points, while our method improves the baseline NMT system with 1.96 BLEU points. Therefore, it is very obvious that phrase information is more effective than lexicon information when we use them to improve the NMT system.

Fig. 4 shows an illustrative example. In this example, baseline NMT mistranslates “dianli (electricity) anquan (safe)” into “coal”. Arthur partially fixes this error and it can correctly translate “dianli (electrical)” into “electrical”, but the source word “anquan (safe)” is still missed. Fortunately, this mistake is fixed by our proposed method. The reason behind this is that Arthur uses information from translation lexicons, which makes the system only fix the translation mistake of an individual lexicon (in this example, it is “dianli (electrical)”), while our method uses the information from phrases, which makes the system can not only obtain the correct translation of the individual lexicon but also capture local lexicon reordering and fixed collocation etc.

Besides the BLEU score, we also conduct a subjective evaluation to validate the benefit of incorporating a phrase table in NMT. The subjective evaluation is conducted on CH-EN translation. As our method tries to solve the problem that NMT system cannot reflect the true meaning of the source sentence, the criterion of the subjective evaluation is the faithfulness of translation results. Specifically, five human evaluators, who are native Chinese and expert in English, are asked to evaluate the translations of 500 source sentences randomly sampled from the test sets without knowing which system a translation is selected from. The score ranges from 0 to 5. For a translation result, the higher its score is, the more faithful it is. Table 2 shows the average results of five subjective evaluations on CH-EN translation. As shown in Table 2，the faithfulness of translation results produced by our method is better than Arthur and baseline NMT system.

Different Methods to Construct Recommendation Set

When constructing the word recommendation set, our current methods are adding the next word of the match part into recommendation set. In order to test the validity of this strategy, we compare the current strategy with another system, in which, we can add all words in candidate target phrase set into recommendation set without matching. We denote this system by system(no matching), whose results are reported in line 5 in Table 1. From the results, we can conclude that in both CH-EN translation and EN-JA translation, system(no matching) can boost the baseline system, while the improvements are much smaller than our methods. It indicates that the matching between the phrase and partial translation is quite necessary for our methods.

As we discussed in Section 3.1, we allow the prefix and suffix to be an empty string to make first word of each phrase into the word recommendation set. To show effectiveness of this setting, we also implement another system as a comparison. In the system, the first words of each phrase are not included in the recommendation set (we denote the system by system(no first)). The results of this system are reported in line 6 in Table 1. As shown in Table 1, our methods performs better than system(no first)) on both CH-EN translation and EN-JA translation. This result shows that the first word of the target phrase is also important for our method and is worthy to recommend.

Translation Results on Large Data

We also conduct another experiment to find out whether or not our methods are still effective when much more sentence pairs are available. Therefore, the CH-EN experiments on millions of sentence pairs are conducted and Table 3 reports the results. We can conclude from Table 3 that our model can also improve the NMT translation quality on all of the test sets and the average improvement is up to 1.83 BLEU points.

Related Work

In this work, we focus on integrating the phrase translation table of SMT into NMT. And there have been several effective works to combine SMT and NMT.

Using coverage mechanism. BIBREF9 and BIBREF10 improved the over-translation and under-translation problems in NMT inspired by the coverage mechanism in SMT.

Extending beam search. BIBREF25 extended the beam search method with SMT hypotheses. BIBREF13 improved the beam search by using the SMT lattices.

Combining SMT features and results. BIBREF12 presented a log-linear model to integrate SMT features (translation model and the language model) into NMT. BIBREF26 and BIBREF27 proposed a supervised attention model for NMT to minimize the alignment disagreement between NMT and SMT. BIBREF11 proposed a method that incorporates the translations of SMT into NMT with an auxiliary classifier and a gating function. BIBREF28 proposed a neural combination model to fuse the NMT translation results and SMT translation results.

Incorporating translation lexicons. BIBREF6 , BIBREF17 attempted to integrate NMT with the probabilistic translation lexicons. BIBREF16 moved forward further by incorporating a bilingual dictionaries in NMT.

In above works, integrating the phrase translation table of SMT into NMT has not been fully studied.

Translating phrase in NMT. The most related works are BIBREF18 and BIBREF19 . Both methods attempted to explore the possibility of translating phrases as a whole in NMT. In their models, NMT can generate a target phrase in phrase memory or a word in vocabulary by using a gate. However, their

“phrases” are different from that are used in phrase-based SMT. BIBREF18 's models only support a unique translation for a source phrase. In BIBREF19 's models, the source side of a phrase pair must be a chunk. Different from above two methods, our model can use any phrase pair in the phrase translation table and promising results can be achieved.

Conclusions and Future Work

In this paper, we have proposed a method to incorporate a phrase translation table as recommendation memory into NMT systems to alleviate the problem that the NMT system is opt to generate fluent but unfaithful translations.

Given a source sentence and a phrase translation table, we first construct a word recommendation set at each decoding step by using a matching method. Then we calculate a bonus value for each recommendable word. Finally we integrate the bonus value into NMT. The extensive experiments show that our method achieved substantial increases in both Chinese-English and English-Japanese translation tasks.

In the future, we plan to design more effective methods to calculate accurate bonus values.

Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303 and the Natural Science Foundation of China under Grant No. 61333018 and 61673380. The research work in this paper also has been supported by Beijing Advanced Innovation Center for Language Resources.