Encoding Gated Translation Memory into Neural Machine Translation

Qian Cao and **Deyi Xiong***
School of Computer Science and Technology, Soochow University, Suzhou, China
20174227009@stu.suda.edu.cn; dyxiong@suda.edu.cn

Abstract

Translation memories (TM) facilitate human translators to reuse existing repetitive translation fragments. In this paper, we propose a novel method to combine the strengths of both TM and neural machine translation (NMT) for high-quality translation. We treat the target translation of a TM match as an additional reference input and encode it into NMT with an extra encoder. A gating mechanism is further used to balance the impact of the TM match on the NMT decoder. Experiment results on the UN corpus demonstrate that when fuzzy matches are higher than 50%, the quality of NMT translation can be significantly improved by over 10 BLEU points.

1 Introduction

Neural machine translation, an emerging machine translation (MT) technology, has made remarkable progress in the past few years (Cho et al., 2014; Sutskever et al., 2014), which strongly encourages many translation agencies to embrace it for product deployment. A natural question during this deployment is how the strengths of both the traditional TM and new NMT technologies can be combined together for professional high-quality translation.

Such attempts to the TM and MT combination have been already conducted in the context of statistical machine translation (SMT). A variety of efforts have been made to incorporate matched translation segments from TM into SMT (Koehn and Senellart, 2010). Partially inspired by these efforts, we aim at combining TM and NMT in this paper.

Different from TM and SMT, both of which use symbolic fragments to construct translations, NMT induces translations from a real-valued continuous space. Furthermore, NMT is trained in an end-to-end fashion, which makes it not easy to be amenable to external intervention. Therefore, incorporating TM as external knowledge into NMT is challenging.

In this paper, we propose a novel and effective method to address this issue in the combination of TM and NMT. The key idea behind this method is to mimic human translators in translating a source sentence given a similar source sentence with a translation. We treat the matched TM translation as an additional signal and try to encode it with a new encoder to guide the NMT decoder to translate the current sentence. Specifically, we first find the sentence that is most similar to the current source sentence from TM by calculating their semantic similarity based on sentence embeddings. In order to prevent the TM matched translation from dominating the decoding process, we introduce a gate mechanism to balance the TM translation signal and the current source sentence which are encoded separately by two different encoders.

A series of experiments on the Chinese-English UN corpus demonstrate that when fuzzy matches are over 50%, the proposed method can significantly improve NMT with the gated TM signal. We also conduct an in-depth analysis on the TM gate, which shows that the gate can indeed regulate the information flow from TM to the NMT decoder.

2 Encoding Gated TM into NMT

In this section, we elaborate our proposed method that encodes translation memories into neural machine translation with a gating mechanism. We refer to our method as NMT-GTM, which consists of three essential components: i) coupled encoders that encodes both the source sentence and matched TM translation separately, ii) a TM gating network that controls the encoded signal from the TM matched translation and iii) a TM-guided decoder

* Corresponding author

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3042–3047
Brussels, Belgium, October 31 - November 4, 2018.

that incorporates the gated TM signal into the decoding. The diagram of NMT-GTM is shown in Figure 1.

For each source sentence src, we retrieve TM to find the most similar sentence to it. Different from the combination of TM and SMT, we define the best TM match as the sentence with the highest cosine similarity which is calculated based on sentence embeddings (Le and Mikolov, 2014), instead of being selected based on fuzzy match score. This is consistent with NMT that performs in an embedding-defined semantic space. But we display our results in experiments according to fuzzy match scores for easy understanding. We use tm_s to denote the most semantically similar sentence to src from TM and tm_t its translation.

2.1 Coupled Encoders

We use a pair of encoders to separately encode the source sentence src and its matched TM translation tm_t. Both encoders are running independently of each other with bidirectional GRU recurrent neural networks[1] (Chung et al., 2014). Accordingly, two separate attention networks are employed to obtain context representations for both src and tm_t, which we denote as c^{src} and c^{tm_t} respectively. The attention network for the TM matched translation is able to help detect matched translation segments from tm_t for the decoder.

2.2 TM Gating Network

When we translate a source sentence, in addition to the input of the sentence itself, we also have a TM matched translation (tm_t) semantically similar to the sentence as an additional input. We want the additional input to act as a translation example for providing positive guide to target word prediction. In order to balance the information flow from the two inputs (src and tm_t) into the decoder, we further introduce a TM gating network to control the respective proportions of tm_t and src, partially inspired by Tu et al. (2017) who propose a gating mechanism to combine source and target contexts. We formulate the TM gating network as follows:

$$g^{tm} = f(s_{t-1}, y_{t-1}, c^{src}, c^{tm\_t})$$

where s_{t-1} is the previous hidden state, y_{t-1} is the previously predicted target word, and f is a logistic sigmoid function.

[1] In this paper, we use GRU encoders and decoders. However, our method can be applicable to other encoders and decoders.

	train	dev	test
#Sentences	1, 117, 452	804	1, 614
Average FMS	0.1890	0.5493	0.5392

Table 1: Statistics of the training data, development and test set. FMS: fuzzy match score.

2.3 TM-Guided Decoder

In the TM-guided decoder, we integrate the gated TM information into the decoding process and use the context representations of src and tm_t to predict the hidden state of the decoder in each time step. The decoder hidden state s_t is computed as follows:

$$s_t = GRU(s_{t-1}, y_{t-1}, c^{src} * (1 - g^{tm}), c^{tm\_t} * g^{tm})$$

where * is an element-wise multiplication.

The conditional probability of the next word y_t is calculated as follows:

$$p(y_t | y_{<t}, src) = g(f(s_t, y_{t-1}, c^{src}))$$

Please notice that we only incorporate the gated TM into the hidden state of the decoder, rather than the prediction of the next word. Our goal is to correctly translate the source sentence with reference to the translation of the TM match tm_t. In other words, tm_t only plays a supporting role in translation. We don't want too much information from TM to affect the translation of the source sentence. Therefore, we incorporate the gated TM in a way that it can only indirectly influence the target generation via hidden states. In our experiments, we observe that this helps our proposed model to faithfully translate a source sentence, instead of copying all information from the TM matched translation, especially for source sentences with slight differences (e.g., dates or numbers) from TM matches.

3 Experiments

We conducted a series of experiments on Chinese-English corpus to evaluate the effectiveness of the proposed NMT-GTM and analyzed the TM gate.

3.1 Experimental Settings

Our data come from the Chinese-English United Nations Parallel Corpus (Rafalovitch et al., 2009), which consists of official records and other parliamentary documents. Since large-scale public

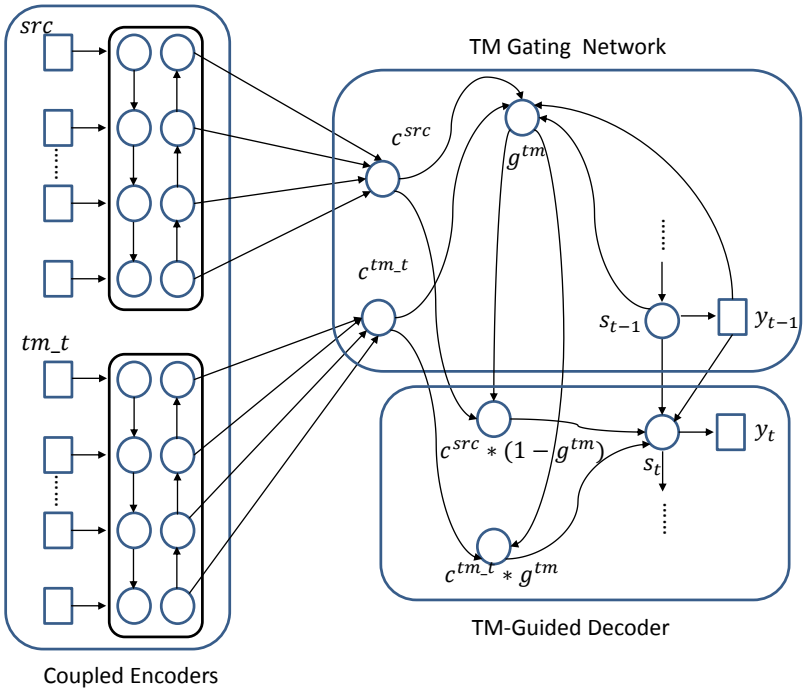

Figure 1: Model Architecture of NMT-GTM

FMS	#Sentences
$[0.9, 1.0)$	171
$[0.8, 0.9)$	182
$[0.7, 0.8)$	178
$[0.6, 0.7)$	179
$[0.5, 0.6)$	181
$[0.4, 0.5)$	177
$[0.3, 0.4)$	180
$[0.2, 0.3)$	185
$(0.0, 0.2)$	181

Table 2: The numbers of sentences of the test set in each fuzzy match score group.

translation memories are not easily available, we built a translation memory from the UN corpus. Specifically, we divided the Chinese-English UN corpus into two parts UN_a and UN_b with equal size. For each source sentence s_a from UN_a, we chose the source sentence s_b from UN_b that has the highest semantically similarity to s_a, computed in the way described in the last section. In doing so, we built a corpus with matched pairs $(s_a/t_a, s_b/t_b)$ where $t_{a/b}$ are translations corresponding to $s_{a/b}$. Then we computed the fuzzy match score for each pair of source sentences as follows:

$$FMS(s_a, s_b) = 1 - \frac{Levenshtein(s_a, s_b)}{max(|s_a|, |s_b|)}$$

where $Levenshtein(s_a, s_b)$ is the word-based Levenshtein Distance between s_a and s_b. The fuzzy match score can also be calculated with other methods, e.g., the method introduced in (Bloodgood and Strauss, 2015). We leave FMS estimated with different methods to our future work. We selected all pairs $(s_a/t_a, s_b/t_b)$ with a fuzzy match score $FMS >= 0.5$. From those pairs with $FMS < 0.5$, we randomly selected 20% of them. These selected pairs were then divided into 9 groups according to their fuzzy match scores (e.g., $FMS \in [0.5, 0.6)$). We randomly chose approximately the same number of sentences from each group to create a development set and test set. The remaining data were used to create the training data (i.e., $\{(s_a, t_b, t_a)_{selected}\}$) and translation memory (i.e., $\{(s_b, t_b)_{selected}\}$). Statistics of the training data, development and test set are shown in Table 1. The numbers of sentences of the test set in each fuzzy match score group are presented in Table 2.

We used RNNSearch as our NMT baseline. We set the maximum sentence length of training corpus to 50 words both for the Chinese and English sides. The sizes of vocabularies of both sides were

FMS	RNNSearch	NMT-GTM	TM
[0.9, 1.0)	43.97	77.67	94.23
[0.8, 0.9)	47.32	79.78	79.84
[0.7, 0.8)	50.95	**71.53**	67.11
[0.6, 0.7)	56.12	**65.39**	58.93
[0.5, 0.6)	65.01	**66.46**	46.99
[0.4, 0.5)	67.83	66.30	34.67
[0.3, 0.4)	58.51	56.83	22.93
[0.2, 0.3)	46.12	44.42	9.72
(0.0, 0.2)	31.41	29.83	1.18
(0.0, 1.0)	51.11	**61.43**	47.16

Table 3: BLEU scores for translations from RNNSearch, NMT-GTM and TM.

FMS	ref_as_TM	TM_ave_gate	ref_ave_gate
[0.9, 1.0)	81.51	0.6712	0.6735
[0.8, 0.9)	85.94	0.6543	0.6582
[0.7, 0.8)	85.87	0.6385	0.6477
[0.6, 0.7)	83.13	0.6075	0.6267
[0.5, 0.6)	84.55	0.5995	0.6218
[0.4, 0.5)	85.13	0.5755	0.6035
[0.3, 0.4)	78.63	0.5721	0.6083
[0.2, 0.3)	76.78	0.5652	0.6409
(0.0, 0.2)	70.89	0.5633	0.6699
(0.0, 1.0)	81.04	0.6047	0.6388

Table 4: Changes of the TM gate. The second column shows the BLEU scores with reference translations being used as additional TM inputs. The third column represents the average gate values of the standard setting, while the last column represents the average gate values when references are used as additional TM inputs.

set to 30k. For those words that are not in the vocabulary, we replaced them with a special token UNK. We set the dropout to 0.5. All the other settings were the same as those described by Bahdanau et al. (2014). We used the stochastic gradient descent algorithm with Adam (Kingma and Ba, 2014) to train NMT models. The learning rate was set to 0.0004. The size of mini-batch was set to 80 sentences. The beam size was set to 10 during decoding.

For the proposed NMT-GTM model, we used tuples (src, tm_t, tgt) as input. The rest of the parameter settings were consistent with the baseline model. To calculate the cosine similarity, we used the fasttext tool [2] with the dimension of 100 to obtain sentence embeddings.

[2]Available at: https://fasttext.cc/

3.2 Experimental Results

Table 3 shows the results of different NMT systems measured by BLEU (Papineni et al., 2002). From the table, we can find that when fuzzy match scores are over 50%, the extra introduction of TM information can significantly help NMT to better translate. Even when fuzzy match scores are lower than 50%, the translation quality does not drop too much. On the entire test set, the proposed gated combination model of TM and NMT improves the translation quality by 10.32 BLEU points over the baseline.

In addition, in order to investigate how similar the matched TM translations tm_t are to the reference translations ref, we also measured the BLEU scores of the matched TM translations against the reference translations. The results are also shown in Table 3, indicated as TM.

3.3 Analysis

We further took a deep look into how the TM gate is varying when we incorporate TM matches with different fuzzy match scores. As a comparison, we used the reference translations as the matched TM translations and incorporated them into NMT-GTM to check the changes of the gate. The BLEU scores measured when we used reference translations as matched TM translations as well as average gate values are shown in Table 4. The results demonstrate that when the matched TM is semantically closer to the current source sentence, the TM gate is larger, indicating that more information from the matched TM translation is used to guide the decoder.

Table 5 shows an example from our test set. The highlighted fragments of the source sentence and the matched TM source sentence are not actually the same in terms of their surface forms. However, they are semantically close and can be translated into the same target translation. Our proposed NMT-GMT is able to successfully incorporate the translation of such a fragment into the decoder.

4 Related Work

Various strategies have been proposed to combine TM and SMT (Koehn and Senellart, 2010; He et al., 2010). Their key ideas are to integrate the translations of the same fragments from TM into SMT, and let SMT only translate those different parts. In order to better model this process, Wang et al. (2013, 2014) use different features to allow relevant TM information to guide SMT decoding.

src	主席 说 ， 津巴布韦 代表 根据 议事规则 43 条 要求 参加 该项 目的 讨论
ref	the chairman said that the representative of zimbabwe asked to participate in the discussion of the item in accordance with rule 43 of the rules of procedure .
tm_s	主席 说 ， 塞尔维亚 代表 请求 依据 议事规则 第 43 条 参与 讨论 项目
tm_t	the chairman said that the representative of serbia had asked to participate in the discussion of the item in accordance with rule 43 of the rules of procedure .
RNNSearch	the chairman said that the representative of zimbabwe, in accordance with rule 43, requested a discussion of the item .
NMT-GTM	the chairman said that the representative of zimbabwe had asked to participate in the discussion of the item in accordance with rule 43 of the rules of procedure .

Table 5: A translation example from the test set. Semantically similar fragments are highlighted with red color.

The related work on combining TM and NMT is quite limited. Gu et al. (2017) propose a TM-NMT model that first finds the most similar segments through search engines according to fuzzy match scores and saves them as key-value pairs in memory. In the subsequent decoding, the saved information is used to help decoding. Our work is significantly different from theirs in two aspects. First, we use semantic similarity based on sentence embeddings to detect the best TM matches rather than the fuzzy match score. Second, we encode the entire TM matched translation rather than segments into NMT with coupled encoders and a gating network.

Our work is also related to multi-source NMT (Zoph and Knight, 2016). The difference is that in our case, the multiple source inputs are just semantically similar, rather than identical. This is the reason that we use a gate to combine these inputs.

5 Conclusion and Future work

In this paper, we have presented a novel gated method to encode translation memory into NMT so as to convey the information of the matched TM translation into the NMT decoder. Extensive experiments verify that our method can indeed effectively improve translation quality, especially when fuzzy match scores are higher than 50%. Further analysis reveals that the proposed TM gate is able to vary according to the similarity between the matched TM translation and the current sentence.

Acknowledgments

The present research was supported by the National Natural Science Foundation of China (Grants No. 61622209 and 61861130364). We would like to thank three anonymous reviewers for their insightful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Michael Bloodgood and Benjamin Strauss. 2015. Translation memory retrieval methods. *Computer Science*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2017. Search Engine Guided Non-Parametric Neural Machine Translation. *arXiv preprint arXiv:1705.07267*.

Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of*

the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Alexandre Rafalovitch, Robert Dale, et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association of Computational Linguistics*, 5(1):87–99.

Kun Wang, Chengqing Zong, and Keh-Yih Su. 2013. Integrating translation memory into phrase-based machine translation during decoding. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 11–21.

Kun Wang, Chengqing Zong, and Keh-Yih Su. 2014. Dynamically integrating cross-domain translation memory into phrase-based machine translation during decoding. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 398–408.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of NAACL-HLT*, pages 30–34.