Graph based Translation Memory for Neural Machine Translation

Mengzhou Xia∗
Carnegie Mellon University
mengzhox@andrew.cmu.edu

Guoping Huang
Tencent AI Lab
donkeyhuang@tencent.com

Lemao Liu
Tencent AI Lab
redmondliu@tencent.com

Shuming Shi
Tencent AI Lab
shumingshi@tencent.com

Abstract

A translation memory (TM) is proved to be helpful to improve neural machine translation (NMT). Existing approaches either pursue the decoding efficiency by merely accessing local information in a TM or encode the global information in a TM yet sacrificing efficiency due to redundancy. We propose an efficient approach to making use of the global information in a TM. The key idea is to pack a redundant TM into a compact graph and perform additional attention mechanisms over the packed graph for integrating the TM representation into the decoding network. We implement the model by extending the state-of-the-art NMT, Transformer. Extensive experiments on three language pairs show that the proposed approach is efficient in terms of running time and space occupation, and particularly it outperforms multiple strong baselines in terms of BLEU scores.

Introduction

A translation memory (TM) typically consists of bilingual sentence pairs that are most similar to the sentence to be translated (Robinson 2012). In computer-aided translation, professional translators are able to quickly figure out the faithful translation by standing on top of a TM instead of translating from scratch. In statistical machine translation (SMT), various research work have also been devoted to making use of a TM to improve translation quality (Simard and Isabelle 2009; Koehn and Senellart 2010; Ma et al. 2011).

Recent years have witnessed an evolutionary shift from SMT to neural machine translation (NMT), and many notable works investigate on how to integrate a TM into neural translation models (Li, Zhang, and Zong 2016; Zhang et al. 2018; Gu et al. 2018). For example, Zhang et al. (2018) propose a simple approach which defines a quantity over a TM to bias word selection for NMT decoding. The quantity is calculated based on n-gram matching between a TM and the sentence to be translated. This method indeed captures helpful local patterns from a TM while inescapably omitting significant global information such as long distance reordering in a TM. In addition, it introduces an additional hyperparameter to balance the contribution of a TM and the NMT model. The sensitive hyperparameter is required to be tuned manually for each specific task by selecting a suitable development set. On the other hand, Gu et al. (2018) propose an alternative approach to making full use of the global information in a TM. The method encodes all bilingual sentences in a TM into a key-value memory, in which each target word and its corresponding context vector represent a pair of value and key. Nevertheless, the integration of the memory leads to considerable inefficiency in terms of both computational time and space occupation due to redundant words in both source and target side of a TM.

In this paper, we propose a new approach to integrating a translation memory into NMT which utilizes global information within a TM in an efficient fashion. The key to the proposed approach is to represent a TM using a compact structure. Specifically, our approach retrieves a translation memory consisting of multiple bilingual sentences from our training corpus. As the source sentences in a TM are similar to the input sentence, we directly ignore the source side of a TM and only focus on representing the target side. To avoid redundancy in the target side of a TM, we further pack the target side into a directed graph, where each target sentence in a TM indicates a path from the beginning to the end. Then we incorporate this packed graph into the decoding network by performing a self-attention mechanism over the graph inspired by Veličković et al. (2018).

Our approach is more efficient while keeping global information because it performs an attention mechanism over fewer attentive nodes in the packed graph, compared with the large key-value memory in Gu et al. (2018). Furthermore, our approach is more robust in practice because it does not introduce additional sensitive hyperparameters to be manually tuned like Zhang et al. (2018).

To validate the effectiveness of the proposed approach, we implement our idea on top of the state-of-the-art model Transformer (Vaswani et al. 2017). Extensive experiments on six translation tasks demonstrate the following advantages of the proposed approach:

• It is more efficient than the approach in Gu et al. (2018) in terms of both running time and space occupation;
• It delivers BLEU improvements over strong baselines in-
including Transformer and both approaches in Zhang et al. (2018) and Gu et al. (2018).

Preliminary

Suppose \( x = \langle x_1, \ldots, x_{|x|} \rangle \) is a source sentence with length \( |x| \) and \( y = \langle y_1, \ldots, y_{|y|} \rangle \) is the corresponding target sentence of \( x \) with length \( |y| \). Generally, for a given \( x \), the neural machine translation baseline, Transformer, aims to generate a translation \( y \) according to the conditional probability \( P(y \mid x) \) defined by neural networks:

\[
P(y \mid x) = \prod_{i=1}^{|y|} P(y_i \mid y_{<i}, x),
\]

where \( y_{<i} = \langle y_1, \ldots, y_{i-1} \rangle \) denotes a prefix of \( y \) with length \( i - 1 \). To expand each factor \( P(y_i \mid y_{<i}, x) \) in Eq.(1), Transformer adopts the encoder-decoder framework similar to the standard sequence-to-sequence learning in (Bahdanau, Cho, and Bengio 2014).

In encoding \( x \), an encoder employs \( L \) layers of neural networks, each layer consisting of different sub-layers, i.e. multi-head attention, residual connection, layer normalization, and feed-forward network as mentioned in (Vaswani et al. 2017). The output of the \( l \)th layer is obtained from the output of the \((l - 1)\)th layer via four sub-layers as follows:

\[
h_j^{E,l} = RL \circ F \circ RL \circ MH(h_j^{E,l-1}, h_j^{E,l-1})
\]

where \( h_j^{E,l} \) indicates the \( j \)th hidden unit with dimension \( d \) after passing \( l \) layers during the encoding phase and particularly \( h_j^{E,0} \) denotes the word embedding plus positional encoding of \( x_j \); \( h_j^{E,l} = \langle h_j^{E,1}, \ldots, h_j^{E,l} \rangle \) denotes an \((|x|, d)\)-matrix, i.e. a sequence of hidden units with dimension of \( d \); \( \circ \) denotes the composition operator between two functions; \( MH, RL \) and \( F \) respectively denote the functions corresponding to the sub-layers of multi-head attention, residual connection plus layer normalization, and feed-forward network with parameters omitted for easier understanding.

Note that \( RL \) and \( F \) maps a vector to another vector and we refer enthusiastic readers to Vaswani et al. (2017) for detailed definitions. \( MH(h_j, h) \) maps a vector \( h_j \) and a matrix \( h \) into another vector as follows:

\[
MH(h_j, h) = \varphi(\text{head}_1, \ldots, \text{head}_K)
\]

where \( \varphi \) is a linear projection function, and

\[
\text{head}_k = \text{softmax} \left( h_jW^1_k (h_kW^2_k)\top \right) W^3_k
\]

where \( W^1_k, W^2_k \) and \( W^3_k \) are transformation matrices for head \( k \) and \( d \) is the dimension of vector \( h_j \).

During decoding phase, Transformer similarly employs \( L \) layers of neural networks, yet each layer consisting of six sub-layers as follows:

\[
h_i^{D,l} = RL \circ F \circ RL \circ MH(RL \circ MH(h_i^{D,l-1}, h_i^{D,l-1}), h^{E,L})
\]

where \( h_i^{D,l} \) indicates the \( i \)th hidden unit at \( l \)th layer in the decoding phase, and in particular \( h_i^{D,0} \) denotes the word embedding of \( y_{i-1} \) plus its positioning encoding for the first layer; all other functions are defined as in Eq.(2). Note that Eq.(5) differs from Eq.(2) in an additional multi-head attention mechanism over the hidden unit sequence \( h^{E,L} \) derived from the encoding phase.

Finally, the factor \( P(y_i \mid y_{<i}, x) \) in Eq.(1) is defined as follows:

\[
P(y_i \mid y_{<i}, x) = \text{softmax} \left( \phi(h_i^{D,L}) \right)
\]

where \( \phi \) is a linear network to project the hidden unit to a vector with dimension of the target vocabulary size.

Graph Representation of TM

For a source sentence \( x \) to be translated, we firstly use the off-the-shelf search engine Lucene to retrieve a set of sentences \( \{ x^i \mid i \in [1, M] \} \) together with its translations \( \{ y^i \mid i \in [1, M] \} \). Following Gu et al. (2018) and Zhang et al. (2018), we prune the retrieved set to an \( N \)-best list \( (M \gg N) \) by the fine-grained similarity score based on the edit-distance:

\[
1 - \frac{\text{dist}(x, x^i)}{\max(|x|, |x^i|)}
\]

where \( \text{dist} \) denotes the edit-distance. This \( N \)-best list is referred to as a translation memory in this paper, which is represented by \( \{ (x^i, y^i) \mid i \in [1, N] \} \).

Gu et al. (2018) encode each word in both the source and target sides of TM as a neural memory, and repeatedly access the memory during decoding. Unfortunately, since a word even a phrase may appear repeatedly in both source and target side of a TM, redundant words are encoded multiple times, leading to a large key-value memory, where the key value pairs are determined by target words. This large memory makes the attention computation inefficient in speed, and also consumes a considerable GPU memory, which is expensive for machines in practice.

To address these issues, we propose an effective approach to representing a TM in a compact way by the following two steps in this paper.

Ignoring source side

It is observed that most source words in the TM also appear in the input sentence \( x \) and have already been represented by the encoder. In addition, we believe that those words in the TM yet beyond \( x \) may not be informative to translate the sentence \( x \) itself. Therefore, in our proposed model, we directly ignore the source sentences of the TM and only represent the target side. In other words, we expect the target side of the TM to directly capture relevant parts in the sentence \( x \) without bothering to access its source side counterpart. Since the GPU memory consumed by TM encoding is proportional to the total number of words in a TM, this trick is able to reduce memory consumption greatly. Note that this trick can only improve the computational speed to a limited extent, because it still maintains the same number of target words corresponding to key value pairs.
Esta enmienda no afecta la versión inglesa.

Figure 1: The architecture of the proposed NMT with graph based TM. 1) Graph representation - The part in the dashed box is a concrete example of the graph representation of a TM. The source and target languages are Spanish and English respectively. Src is the source sentence and Ref is the corresponding target sentence. Note that only 2 TM pairs are used in this example for simplicity. 2) Model architecture - The part outside of the dashed box shows the core components of the model architecture. A light blue box consists of a multi-head attention mechanism and a residual connection plus a layer normalization. A light yellow box consists of one more free-forward network and a residual connection plus a layer normalization. Specifically, the graph attention operation is presented with three selected nodes. L is the number of repetition of each part.

**Packing target side**

Secondly, instead of sequentially encoding target sentences in a TM, we pack them into a compact graph such that some words in different sentences may correspond to the same node in the graph, which is inspired by the notion of lattice or hypergraph in statistical machine translation (Koehn 2009). To this end, we convert the target side in a TM into a confusion network by using the algorithm proposed by Mango, Brill, and Stolcke (1999) and Mango, Brill, and Stolcke (2000). The basic idea is to cluster each word into an equivalent class, and organize all the classes according to an order defined over classes. Finally, we can obtain a graph via computing the dual of the confusion network, by treating edges as nodes (West and others 2001).

Figure 1 shows an example of a resulting graph for a TM (including two sentences for simplicity), where each node represents a word. In the graph, the nodes representing the words "This", "amendment", "the", "Danish", and "." appear in both sentences, but only appear once in the graph. In terms of both space and speed, it is more efficient to encode the graph, compared with the representation of all source and target words of a TM in Gu et al. (2018).

**NMT with Graph based TM**

Suppose \( G = (v, e) \) is the directed graph obtained from a translation memory, where \( v = \langle v_1, \cdots, v_{|v|} \rangle \) is a sequence of nodes, and \( e = \langle e_1, \cdots, e_{|e|} \rangle \) is a sequence of edges. In addition, \( n_i = \langle n_{i1}, \cdots, n_{i|n_i|} \rangle \) denotes a sequence of first-order neighborhood nodes of \( v_i \) in graph \( G \) including \( v_i \) itself. Nodes are arranged in topological order. For example, as illustrated in Figure 1, \( v_1 = \text{This}, n_1 = \langle \text{This, amendment} \rangle \) and \( v_2 = \text{amendment}, n_2 = \langle \text{This, amendment, does, only} \rangle \).

Generally, the enhanced Transformer shares the similar architecture as the baseline but with two major differences in encoding and decoding phases. The model structure is illustrated in figure 1.

**Enhanced Graph Encoding**

Besides encoding the input sequence \( x \), the proposed model also encodes \( v \) from \( G \) by using \( L \) layers of networks in a similar fashion to the encoding of the input \( x \). Suppose \( h_i^{G,l} \) denotes the hidden units at \( l_{th} \) layer regarding the node \( v_i \). Inspired by the graph attention in Velicković et al. (2018), \( h_i^{G,l} \) is calculated by the hidden units at \( l - 1_{th} \) layer as follows:

\[
    h_i^{G,l} = RL \circ F \circ RL \circ MH(h_i^{G,l-1}, h_{G,l-1}^{G,l-1}[n_i])
\]

where \( RL \), \( F \) and \( MH \) are functions corresponding to sublayers of residual connection plus layer normalization, feed-forward network and multi-head attention mechanism defined in Eq.(2). \( h_{G,l-1}^{G,l-1}[n_i] \) is a matrix obtained by taking the slice over the matrix \( h_{G,l-1}^{G,l-1} \) along the first axis via \( n_i \).

Based on the encoding schema elaborated above, there are different variations of the graph encoding in real practice. We can choose to fix the graph encoding after \( L \) layers’ computation and introduce the fixed encoding to each decoding layer, which is the same as the encoding of the source side \( x \). We can also introduce a flexible encoding to each decoding layer, which is calculated as follows:

\[
    h_i^{G,0} = RL \circ MH_i(h_i^{G,0}, h^{G,0}[n_i])
\]

where \( h^{G,0} \) denotes the word embedding plus positional encoding of the translation memory sequence. When flexible
encoding is applied, each decoding layer gets its own unique graph encoding. Besides, the outermost components $RL$ and $F$ are removed from flexible encoding to reduce complexity.

Because the directed graph $G$ roughly records the global order of nodes, which somehow captures the reordering in the original TM, it would be helpful to encode the directional information for translating the input $x$. In this paper, we index the nodes $v$ in topological order and integrate additional positional encoding signal before the graph attention similar to the encoding of $x$ as in the standard Transformer architecture. In detail, the first layer of graph encoding is the sum of the word embedding of $v_i$ and position encoding of $i$ (Vaswani et al. 2017).

**Enhanced Decoding**

Suppose $h^{E,L}$ is the hidden unit sequence encoded from $x$ and $h^{G,L}$ is that encoded from $G$. As mentioned in the last subsection, $h^{G,L}$ could be a fixed graph encoding or a flexible one. Similar to Transformer, the proposed model employs $L$ layers of networks in the decoding phase, but each layer includes two additional sub-layers ($MH$ and $RL$) to incorporate $h^{G,L}$, besides other six sub-layers. We place the extra two layers nearest to the output of the decoder network in order to let the graph encoding fully influence the decoding process. Formally, at the $l$th layer in the decoding phase, the hidden unit $h^{D,l}_{i}$ is computed from $h^{D,l-1}_{i}$, $h^{E,L}$ and $h^{G,L}$ as follows:

$$h^{D,l}_{i} = RL \circ F \circ RL \circ MH \left( RL \circ MH \left( h^{D,l-1}_{i}, h^{G,L}_{i+1}, h^{E,L}, h^{G,L} \right) \right)$$

This equation Eq.(10) is similar to Eq.(5) except two extra sub-layers including $MH$ over $h^{G,L}$ and its immediate $RL$. By applying these additional layers, the model is able to dynamically extract favorable semantic and syntactic information from translation memory for each time step.

Note that in encoding $v_i$, the sequence of $n^i$ only includes the nodes which are directly adjacent to $v_i$, one might argue that the hidden unit $h^{G,l}_{i}$ ignores information far away from the node $v_i$. Indeed, we tried to reset $n^i$ to include higher order neighbors of $v_i$, but we did not observe additional gains in translation quality. The possible reason is that the long distance information is captured by the multi-layer structure.

**Related Work**

This section reviews the mostly related works, according to the following three research lines.

**TM based SMT**

Many research works are devoted to integrating a translation memory into the statistical machine translation paradigm. For example, Koehn and Senellart (2010) extract bilingual segments from a TM which matches the source sentence to be translated, and constrain SMT to decode for those unmatched parts of the source sentence. Unlike Koehn and Senellart (2010) employing a heuristic score to decide whether the extracted segments should be used as decoding constraints or not, Ma et al. (2011) design a fine-grained classifier to predict the score for making more reliable decisions. In addition, Simard and Isabelle (2009), Wang, Zong, and Su (2013) and Wang, Zong, and Su (2014) add the extracted bilingual segments to the translation table of SMT, and then decode the source sentence with the augmented translation table, which bias the decoder in a soft constraint manner instead of a hard one as in Koehn and Senellart (2010) and Ma et al. (2011).

**Graph based NMT**

There are many works involving graph structures to improve NMT. For instance, Stahlberg et al. (2016) implicitly employs a translation graph (i.e. lattice) generated by SMT by rewarding the words which have high posterior probabilities over the lattice. Meanwhile, Khayrallah et al. (2017) explicitly use a lattice generated by SMT to constrain the decoding space and rerank to select the best translation from this constrained lattice. Both approaches obviously differ from ours in that they do not encode the discrete graph into the continuous vectors from the point view of neural networks. Ma et al. (2018) and Su et al. (2017) encode a packed parse forest or a word segmentation lattice of a source sentence. However, the focus of our work is oriented to a translation memory instead of a parse tree or word segmentation. Furthermore, we integrate the encoded vectors into the strong Transformer, which leads to nontrivial implementations compared with attention operations in recurrent networks.

**Experiments**

In this section, we demonstrate, by experiments, the advantages of the proposed model: it yields better translation than
Zhang et al. (2018) with the help of global information from translation memory; and it is more efficient than the model in Gu et al. (2018) in terms of running time and memory consumption mainly because of the compact graph representation of translation memory.

**Settings**

**Data Preparation** Following the previous works investigating on incorporating TM into NMT models, we use the JRC-Acquis corpus for training and evaluating our proposed model. The JRC-Acquis corpus is a collection of parallel legislative text of European union Law applicable in the EU member states. The highly related text in the corpus is suitable for us to make evaluations. To fully explore the effectiveness of our proposed model, we conduct translation experiments on three language pair bidirectionally, namely, en-fr, en-es, and en-de.

We manage to obtain preprocessed datasets from Gu et al. (2018). For each language pair, we randomly select 3000 samples to form a development and a test set respectively. The rest of the pairs are used as the training set. Sentences longer than 80 and 100 are removed from the training and development/test set. The technique of Byte-pair Encoding (Sennrich, Haddow, and Birch 2016) is applied and the vocabulary size is set to be 20K for all the experiments.

**Baseline Systems** The proposed TM graph based model is built on transformer (Vaswani et al. 2017), and it is denoted by G-TFM. We also propose and implement another simpler TM representation based transformer, SEQ-TFM which sequentially encodes all target sentences in a TM as one of the baseline models. Specifically, each target sentence in TM goes through a multi-head mechanism and an immediate residual connection plus layer normalization in $l_{th}$ layer, which is calculated as follows:

$$h_{k,i}^{l} = RL \circ MH_{l}(h_{k,i}^{0}, h_{k}^{0})$$

where $h_{k,i}^{l}$ is the $i_{th}$ hidden unit of the $k_{th}$ sentence in the TM; $h_{k,i}^{0}$ is the word embedding plus positional encoding of $k_{th}$ sentence in TM. The derived representations for these sentences are then concatenated to form the representation of the translation memory $h_{3,i}^{k}$, which can be utilized flexibly in $l_{th}$ decoding layer.

Besides, as the proposed model is directly built upon the Transformer architecture, Transformer itself provides a natural baseline, which is referred to as TFM in this paper. In addition, following Zhang et al. (2018) and Gu et al. (2018), we implement two translation memory based systems on top of Transformer for fair comparison and we refer them to P-TFM and SEG-TFM, respectively.\(^1\) For the broader comparison, we reproduce the model (denoted by P-RNN) in Zhang et al. (2018) on top of our in-house RNNsearch system (denoted by RNN).\(^2\)

**Training systems** For each sentence, we retrieve 100 translation pairs from the training set by using Apache Lucene. We score the source side of each retrieved pair against the source sentence $x$ with fuzzy matching score in Eq. (7) and select top $N = 5$ translation sentence pairs as a translation memory for the sentence $x$ to be translated. Sentences from the target side in the translation memory are used to form a graph, with each word represented as a node and the connection between adjacent words in a sentence represented as an undirected edge.

For training all systems, we maintain the same hyperparameters as shown in Table 1 for comparison. Besides, we adopt the same training algorithm to learn the models as follows. We use a customized leaning rate decay paradigm following Tensor2Tensor (Vaswani et al. 2018) package. The learning rate increases linearly on early stages for a certain number of steps, known as warm-up steps, and decay exponentially later on. We set the warm-up step to be 5 epochs and we early stop the model after training 20 epochs, typically the time when the development performance varies insignificantly.

Furthermore, since there is a hyperparameter in the system P-TFM of Zhang et al. (2018) which is sensitive to the specific translation task, we tune it carefully on the development set for all translation tasks. Its optimized value is 0.7 for es and de tasks while it is 0.8 for fr task.\(^3\)

**Results and analysis on es-en task**

**Translation accuracy** Table 2 shows the experiment results of all the systems on the es-en task in terms of BLEU. Several observations can be made from the results. First, the baseline TFM achieves substantial gains over RNN and even outperforms P-RNN by around 1 BLEU point on the test set. Compared with the strongest baseline P-TFM, the proposed SEQ-TFM and G-TFM are able to obtain some gains up to 1.9 BLEU points on the test set. This result verifies that our

\(^{1}\)Note that due to the architecture divergence between RNN-based NMT and Transformer, SEG-TFG differs from the RNN-based counterpart in that two quantities $c_{l}$ and $z_{i}$ in Gu et al. (2018) are replaced by the hidden units obtained from the multi-head attention over the encoding units and the decoding hidden state units before the softmax operator.

\(^{2}\)We did not reimplement the model in Gu et al. (2018) on top of our in-house RNN, because it was clearly demonstrated in Zhang et al. (2018) that P-RNN works comparable to or even better than Gu et al. (2018) on the same JRC-Acquis corpora as conducted in our experiments.

\(^{3}\)We run all 6 tasks with hyperparameters among [0.5, 1.5] with scale of 0.1, and manually pick the optimized value according to its performance on the development set.

<table>
<thead>
<tr>
<th>Word embedding</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layers</td>
<td>6</td>
</tr>
<tr>
<td>TM dropout</td>
<td>0.6</td>
</tr>
<tr>
<td>Other dropout</td>
<td>0.1</td>
</tr>
<tr>
<td>Beam size</td>
<td>5</td>
</tr>
<tr>
<td>Label smoothing</td>
<td>0.1</td>
</tr>
<tr>
<td>Batch size (tokens)</td>
<td>8192</td>
</tr>
</tbody>
</table>

Table 1: Hyper-parameters for training baselines and the proposed model on top of the transformer architecture.
Graphs due to temporary variables. Ideally, the total memory consumption can be calculated.

Running time Since the retrieval time can be neglected compared with the decoding time as found in Zhang et al. (2018), we thereby eliminate the retrieval time and directly compare running time for neural models as shown in Table 3.

Memory consumption Ideally, the total memory consumed by a system is proportional to the size of the computational tensor graph. However, it is impossible to exactly compare the complexity of the computational graph, because it is dependent on the coding implementation. As a result, we roughly evaluate the memory consumption in terms of the number of words encoded by the models, which corresponds to a part of the nodes in the computational graph and can be calculated.

Table 3 depicts the total number of source and target words encoded by the corresponding model for each test sentence on average. It is observed that SEG-TFM needs to encode approximately 3 times and SEQ-TFM encodes approximately 2 times the number of words of our proposed model, G-TFM. There’s no wonder that TFM takes the fewest words to encode because no extra TM is included. These statistics indicate that under the scenario of incorporating TM in NMT, our model requires the least memory.

Influence on similarity Intuitively, the performance of our proposed model is supposed to vary on the basis of the similarity between the sentences to be translated and the retrieved translation memory. To verify the hypothesis, we divided the test set into bucketed subsets based on the averaged similarity of each sentence in the retrieved translation memory. From the results reported in Table 4, it is observed that the proposed G-TFM outperforms the baseline TF on all bucketed subsets in general. As the similarity score increases, G-TFM leads to more significant improvements than those of the baseline TFM. In detail, the improvements over TFM are up to 5 BLEU points when similarity is around 0.9; while the improvements are less than 1 BLEU as the similarity score is between 0.2 and 0.3. This fact shows that the translation memory indeed brings improvements of translation quality over all similarity buckets. And the more similar the TM is, the better it will do.

In addition, P-TFM is slightly better than the proposed G-TFM when the similarity is between 0.2 and 0.5. But G-TFM outperforms P-TFM with a substantial margin if the similarity is more than 0.5. In particular, as the similarity is above 0.9, G-TFM delivers a gain of 5 BLEU over P-TFM.

Table 4: Translation quality on es-en task for the divided subsets according to similarity. The averaged similarity for the entire test set is 0.48.

<table>
<thead>
<tr>
<th>Similarity</th>
<th>Sents Percent</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 0.1)</td>
<td></td>
<td>90.25</td>
</tr>
<tr>
<td>[0.1, 0.2]</td>
<td></td>
<td>89.96</td>
</tr>
<tr>
<td>[0.2, 0.3]</td>
<td></td>
<td>89.72</td>
</tr>
<tr>
<td>[0.3, 0.4]</td>
<td></td>
<td>89.48</td>
</tr>
<tr>
<td>[0.4, 0.5]</td>
<td></td>
<td>89.24</td>
</tr>
<tr>
<td>[0.5, 0.6]</td>
<td></td>
<td>88.99</td>
</tr>
<tr>
<td>[0.6, 0.7]</td>
<td></td>
<td>88.75</td>
</tr>
<tr>
<td>[0.7, 0.8]</td>
<td></td>
<td>88.51</td>
</tr>
<tr>
<td>[0.8, 0.9]</td>
<td></td>
<td>88.27</td>
</tr>
<tr>
<td>[0.9, 1)</td>
<td></td>
<td>88.03</td>
</tr>
</tbody>
</table>

Table 5: Translation results when adding reference to TM.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>TFM</th>
<th>P-TFM</th>
<th>G-TFM</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.68</td>
<td>65.00</td>
<td>90.25</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Running time and memory. Training time reports the time in seconds for training one epoch on average, and testing time reports the time in seconds for translating one sentence on average. Words (#) denotes the number of words encoded in the neural models on average.
Table 6: An example of effects on using global information from es-en test set. S and R respectively denote source and reference sentences, TM shows only one sentence in the given translation memory, T1, T2 and T3 represent TFM, P-TFM and G-TFM.

Table 7: Translation Results on both development and test sets across other 5 translation tasks.

Results on other tasks
We pick stronger baselines from the es-en task, i.e. TFM and P-TFM, and compare them with the proposed G-TFM model on other 5 translation tasks. Table 7 summarizes their results on both the development and test sets. From this table, we can see that on the test set, G-TFM steadily outperforms TFM by up to 3 BLEU points across all these 5 tasks. In addition, contrast to P-TFM, G-TFM demonstrates better performance by exceeding at least 1 BLEU point across all these tasks except the en-fr task. These results are consistent with the results on es-en task and further validates the effectiveness of integrating graph-based translation memory into the Transformer model.

Conclusion
We have proposed an approach to augmenting the neural machine translation by a translation memory. The proposed approach firstly packs the translation memory into a compact graph, where each node may correspond to multiple words for different sentences in a TM, and then it encodes the packed graph into a deep representation during the decoding phase. The proposed approach is able to make use of the global information from TM and is also efficient enough even if the size of TM increases. Extensive experiments on six tasks with three language pairs show that the proposed approach is effective when compared with the strong baseline Transformer: it not only gains over the baseline with a large margin, but also consistently outperforms the approaches in Zhang et al. (2018) and Gu et al. (2018).

Acknowledgments
We thank Jiatao Gu and Jingyi Zhang for their help on the datasets and reviewers for their valuable feedbacks on this paper.

References


