Linguistic Knowledge-Aware Neural Machine Translation

Qiang Li†, Derek F. Wong†, *Senior Member, IEEE*, Lidia S. Chao, *Member, IEEE*, Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang, *Member, IEEE*

***Abstract*—Recently, researchers have shown an increasing interest in incorporating linguistic knowledge into neural machine translation (NMT). To this end, previous works choose either to alter the architecture of NMT encoder to incorporate syntactic information into the translation model, or to generalize the embedding layer of the encoder to encode additional linguistic features. The former approach mainly focuses on injecting the syntactic structure of the source sentence into the encoding process, leading to a complicated model that lacks the flexibility to incorporate other types of knowledge. The latter extends word embeddings by considering additional linguistic knowledge as features to enrich the word representation. It thus does not explicitly balance the contribution from word embeddings and the contribution from additional linguistic knowledge. To address these limitations, this paper proposes a knowledge-aware NMT approach that models additional linguistic features in parallel to the word feature. The core idea is that we propose modeling a series of linguistic features at the word level (knowledge block) using a recurrent neural network (RNN). And in sentence level, those word-corresponding feature blocks are further encoded using a RNN encoder. In decoding, we propose a knowledge gate and an attention gate to dynamically control the proportions of information contributing to the generation of target words from different sources. Extensive experiments show that our approach is capable of better accounting for importance of additional linguistic, and we observe significant improvements from 1.0 to 2.3 BLEU points on Chinese↔English and English→German translation tasks.**

Manuscript received January 1, 2018; revised May 10, 2018; revised July 11, 2018. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61672555, 61432013 and 61732005), the Fundamental Research Funds for the Central Universities, the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ), and the Multi-year Research Grant from the University of Macau (Grant Nos. MYRG2017-00087-FST, MYRG2015-00175-FST and MYRG2015-00188-FST).

Qiang Li is with the Natural Language Processing Laboratory, School of Computer Science and Engineering, Northeastern University, Shenyang, China, and also with Alibaba Inc, Hangzhou, China (e-mail: liqiangneu@gmail.com).

Derek F. Wong and Lidia S. Chao are with the Natural Language Processing & Portuguese-Chinese Machine Translation (NLP^{2}CT) Laboratory, University of Macau, Macau, China (e-mail: derekfw@umac.mo; lidiasc@umac.mo)

Muhua Zhu is with Alibaba Inc, Hangzhou, China (e-mail: muhua.zmh@alibaba-inc.com)

Tong Xiao and Jingbo Zhu are with the Natural Language Processing Laboratory, School of Computer Science and Engineering, Northeastern University, Shenyang, China, and also with the Shenyang Yatrans Network Technology Co., Ltd., Shenyang, China (e-mail:xiaotong@mail.neu.edu.cn; zhujingbo@mail.neu.edu.cn)

Min Zhang is with the Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University, Suzhou, China (e-mail:minzhang@suda.edu.cn)

This work was done while the first author was at NLP^{2}CT Laboratory. † indicates equal contribution.

TABLE I
TRANSLATIONS PRODUCED BY THE NMT BASELINES IN CHINESE↔ENGLISH TRANSLATIONS. THE PROBLEMATIC CONTEXTS ARE HIGHLIGHTED. NMT$_{+\text{POS}}$ AND NMT$_{+\text{SYN}}$ INDICATE THE PROPOSED LINGUISTIC KNOWLEDGE-AWARE NMT MODELS THAT INCORPORATED WITH ADDITIONAL POS AND SYNTACTIC FEATURE, RESPECTIVELY

	Chinese → English Translation	
1	Chinese	美 欲 重启 中东 和谈
	Reference	US wants to reopen Middle East peace talks
	NMT	US to resume Middle East peace talks
	NMT$_{+\text{POS}}$	US wants to restart Middle East peace talks
	English → Chinese Translation	
2	English	No one is saying that in the media
	Reference	没有 人 在 媒体 中 那么 说
	NMT	没有 人 会 说 媒体
	NMT$_{+\text{Syn}}$	没有 人 在 媒体 上 说

***Index Terms*—Attention gate, knowledge block, knowledge gate, neural machine translation (NMT).**

I. INTRODUCTION

NOWADAYS, neural machine translation (NMT) has been the most prominent approach to machine translation (MT) [1]–[6], due to its simplicity, generality, and effectiveness. The principle of NMT is to directly maximize the conditional probabilities of target sentences given source sentences in an end-to-end paradigm [7]. One of the most widely used neural model follows the encoder-decoder framework [1]. It encodes the source sentence into a dense context representation by using a recurrent neural network (RNN), and then feeds the resulting vector to a RNN-based decoder to produce the target translation. By exploiting the gating [2] and attention mechanisms [3], NMT models have been shown to surpass the performance of previously dominant statistical machine translation (SMT) on many well-established translation tasks [4], [8], [9].

Unlike SMT, NMT does not rely on sub-modules and explicit linguistic features in crafting the translation [10], [11]. Instead, it learns the translation knowledge directly from parallel sentences without resorting to additional linguistic analysis. Despite its advantage in improving fluency in translation results, NMT is deemed an approach that often generates inadequate translations [12]. Similar observations have also been found and reported in the studies of Bentivogli et la. [13] and Toral and Sánchez-Cartagena [14]. For example, NMT systems tend to omit the translations of source words when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

they produce target sentences (aka *untranslated* or *under-translation*) [14]. Table I-(1) shows an example of Chinese to English translation, where a vanilla attention-based NMT system generates an inadequate translation with the translation of the Chinese word "欲 (want)" being omitted. However, when we introduce the Part-of-Speech (POS) information into the NMT (described in Section IV and V), the result is much more faithful to the source sentence that the verb "欲" can be properly translated into "wants". The observation implies that vanilla attention-based NMT models have difficulties in learning and interpreting the latent linguistic information in the source sentence, i.e. the verb complementation "欲 重启 (wants to reopen)" in this case. Another problem that often exists in NMT stems from the translation of complex syntactic structure that contains prepositional phrases [13], [15], [16]. As illustrated in the example of Table I-(2), the conventional NMT, in English to Chinese translation, misinterprets "the media" as the direct object of the verb "saying", and fails to attend to the preposition "in", leading to an incorrect translation "没有 人 会 说 媒体 (No one blames the media)". This again illustrates the limitation of NMT in capturing the syntactic information of a sentence, due to the fact that it does not account for the syntactic interpretations of sentence structure. By incorporating additional syntactic information into the model, the translation produced by the $\text{NMT}_{+\text{Syn}}$ is more adequate to the meaning as well as faithful to the syntactic structure of the source sentence. Therefore, we believe, in many aspects, linguistic knowledge can help to explain the complex sentence phenomena, and to some extent, contribute to the improvement in the translation.

To incorporate linguistic information as prior knowledge into the translation models has far been demonstrated to be valuable and effective approaches in SMT [11], [17], [18]. Recent studies in NMT also show that incorporating explicit linguistic information tends to improve the translation quality [19]–[22]. Much of the previous work is based on the idea of either altering the architecture of encoder to incorporate syntactic structure into the translation model [21], [22] or generalizing the embedding layer of the encoder to support modeling of prior knowledge [20], [23]. The former focuses on injecting the syntactic structure into the translation model, to capture dependencies between words that influence the word order with a consequent impact on the sentence meaning. The idea behind those proposals is to modify the architecture of the models, i.e. network topology, using the syntactic tree to guide recurrence and attention model, to better capture the short and long range dependencies and attachments of words [22]. However, the model itself does not account for the interpretation of the syntactic categories. Furthermore, it lacks flexibility of the inclusion of additional explicit linguistics, such as the POS tags, named entity information, chunk tags, dependency relations, etc. On the other hand, the latter approach employs a simple strategy to include linguistic knowledge as features, in addition to the word feature, into the word embedding [20]. The underlying intuition of this approach is to generalize the embedding layers of NMT, either the embedding layer of the source encoder [24] or the target decoder [23], aiming at enriching the word representations. In practice, the original embedding vector of word downsizes to make room for accommodating those additional linguistic features. In other word, the size of word embedding is reduced when compared to the vanilla model, and the sizes of feature embeddings vary depending on the types of linguistic features that need to be defined manually. In fact, this simple approach, that tightly couples the word and linguistic features, lacks of an effective mechanism to control the influence of the word and that of the linguistic features. Instead, we prefer a novel model that gives a manageable way to model the sequence of words and linguistic data separately, and balance the amount of information used in predicting the target words, as well as be able to dynamically incorporate an arbitrary number of linguistic factors.

In this paper, we propose a novel architecture, knowledge-aware NMT model (KaNMT), which extends to incorporate various linguistic knowledge as features into the NMT model. In contrast to the conventional NMT [3], on the source side, we use an additional RNN encoder (Knowledge Encoder) to encode the sequence of linguistic features in parallel to the encoder for word sequence. In the model, we define a knowledge block (KB) to represent a variable number of linguistic features of a word. In addition, we propose a gating mechanism, knowledge gate (KG), to balance the information between the word feature and the additional linguistic features that is best suited for inducing the source sentence representation. To effectively leverage the knowledge representation in predicting the target words inspired by Wang *et al.* [25], we propose a weighted variant attention mechanism, attention gate (AG), in which a time-dependent gating scalar is adopted to control the ratio of conditional information between the word and KB vectors. To investigate the effectiveness of different linguistic features, we respectively incorporate the POS tags, named entity information, chunk tags, and dependency relations of words, as prior knowledge, into the proposed knowledge-aware model. Empirical results for the Chinese↔English and English→German translations reveal that the proposed model outperforms the vanilla NMT model [3] and linguistic NMT model proposed by Sennrich and Haddow [20].

II. Related Work

Exploiting additional linguistic knowledge for NMT has attracted intensive attention in recent years. A variety of models and approaches have been proposed to incorporate explicit linguistic information as prior knowledge into the neural networks. These efforts could be categorized into three directions of research.

The popular NMT models rely on sequential encoder and decoder architecture [1], [3] without any explicit modeling of the syntactic structure of language. The first line of research attempts to exploit additional syntactic information of sentence to improve the translation. Eriguchi *et al.* [21] proposed a tree-to-sequence NMT model by incorporating phrase structure of the source sentence into the neural encoder. They employed a forward Long Short-Term Memory (LSTM) RNN [26] to encode the lexical nodes and a tree-LSTM [27] to recursively generate the phrase representations upwards. To some extent,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the encoded representation is deemed to take advantage of syntax in interpreting the meaning of a sentence. The model is further extended by Yang *et al.* [22] to encode every node of the syntactic tree with both the local and global context. Their encoder adopts a bidirectional encoding mechanism both at the lexical (leaf nodes) and phrase (tree nodes) level. The vector representations of the sentence, phrases as well as words, are therefore encoded with global meaning and the syntactic information of the sentence rather than local information. Apart from incorporating syntactic structure into the encoder, Stahlberg *et al.* [28] used the translation hypotheses produced by a syntactic SMT model, i.e. Hiero [29], to guide the NMT decoder in predicting the target words. In order to utilize syntactic structure for the NMT models, these approaches require redesigning the neural architecture of the encoder or decoder. As a result, these models are too rigid in modeling the syntax of language, leaving out other linguistic knowledge, such as morphological features and semantic labels.

In SMT, Koehn and Hoang [30] proposed factored translation models by allowing to integrate different morphological features into the translation process, where the representation of word is factored into a vector of morphological features. Following this principle, a recent line of research attempts to enrich the vector representation of words by integrating useful linguistic features in either the source encoder or target decoder. Sennrich and Haddow [20] proposed to generalize the embedding layer of an encoder in NMT by using a combination of morphological and syntactic features, such as word lemma, morphological attributes, POS and dependency labels. In their model, the input word embeddings are simply the concatenation of feature embedding vectors, while the other parts of the NMT model remain unchanged. On the contrary, García-Martínez *et al.* [24] factored the word into morphological (lemma) and syntactic (factors) features at the output decoder of NMT. With a heuristic morphological synthesizer, given the predicted lemma and factors, their model is able to generate unseen word forms, in the meanwhile, to deal with large vocabulary and the out-of-vocabulary (OOV) problem in translation. Unfortunately, their method did not yield any gain in the experiments. The above approaches adopt a tight coupling of the word and linguistic information, and primarily focus on modeling the additional linguistic knowledge at the lexical level. Different from the above work, this paper proposes encoding the additional linguistic features separately using a sequential RNN model, to enhance the model as a whole. Furthermore, due to the decoupling of components, it offers advantages to dynamically control the use of linguistic information in producing the target translation.

The third line of research attempts to incorporate external knowledge through multi-source or multi-task learning. The additional source can be considered as a kind of different but related and useful information for the learning of a translation model, which has shown to be quite effective in a variety of settings [31], [32]. Calixto *et al.* [33] introduced a multi-model NMT by incorporating visual information into the translation process. They employed two attentional models to independently attend the source sentence and the additional features of an image respectively. Firat *et al.* [34] proposed a multi-way,

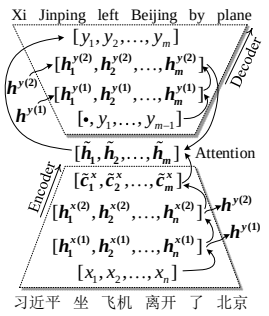


Fig. 1. The graphical illustration of the attentional neural machine translation.

multilingual NMT using n encoders and m decoders for the translation between n and m different languages, by using a single attentional model to share across all the language pairs. A similar idea was further explored to improve the translation of low-resource languages [31], [35]. In the context of incorporating syntactic information using multi-source model, Li *et al.* [36] proposed to linearize the syntactic structure of the source sentences into structural label sequence and feed it as an additional source to a multi-source NMT model. Chen *et al.* [37] modeled the dependency information of source sentence using a convolutional neural network (CNN) yielding a hybrid encoder that consists of a convolutional and a recurrent neural network, for dependency information and the source words respectively. While Wang *et al.* [25] employed a gating mechanism to consider broader context beyond the sentence. The present study, however, is motivated from a different perspective. Instead of using extrinsic information (i.e. other languages or visual information) [31], [32], [35], we focus on modeling the intrinsic linguistic information of source language. Our work is also different from the works of Li *et al.* [36] and Chen *et al.* [37] that we propose to model arbitrary linguistic knowledge spanning from lexical to syntactic, even semantic.

III. Neural Machine Translation

Our proposed KaNMT is built on an attentional NMT system [4], which simultaneously conducts dynamic alignments and generation of each target words, as illustrated in Fig. 1. Given a source sentence, $x = \{x_1, \ldots, x_n\}$, and its translation, $y = \{y_1, \ldots, y_m\}$, an attentional NMT model tries to model the translation probability $p(y|x)$:

$$\log\, p(y|x) = \sum_{t=1}^{m} \log\, p(y_t|y_{<t}, \boldsymbol{s^x}) \tag{1}$$

where $\boldsymbol{s^x}$ is the representation of source sentence x, m is the length of target sentence, t is current timestep. To compute $p(y_t|y_{<t}, \boldsymbol{s^x})$, softmax can be used as activation function:

$$p(y_t|y_{<t}, \boldsymbol{s^x}) = \mathrm{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_{\boldsymbol{t}}) \tag{2}$$

where $\tilde{\boldsymbol{h}}_{\boldsymbol{t}}$ is an attentional hidden state and $\boldsymbol{W_s}$ is the learned parameter. Given the target hidden state $\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}$ from top RNN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

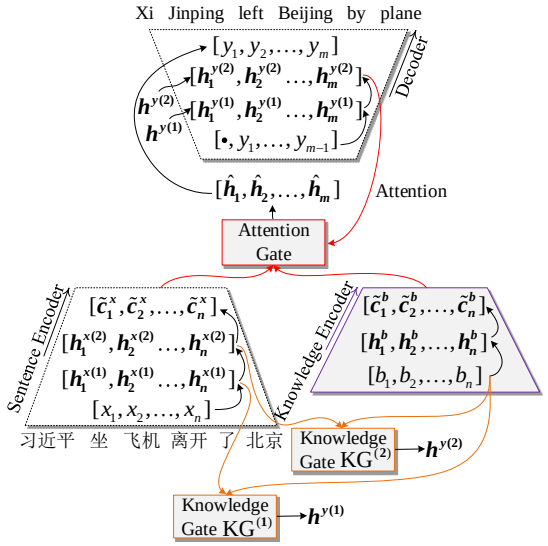


Fig. 2. The architecture of KaNMT which extends the attentional NMT with an additional RNN encoder to modeling the corresponding set of lexical features, knowledge block b_j, for a source word x_j.

layer and source context vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}$ for timestep t, $\tilde{\boldsymbol{h}}_{\boldsymbol{t}}$ is calculated as follows:

$$\tilde{\boldsymbol{h}}_{\boldsymbol{t}} = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}; \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}]) \tag{3}$$

where $\boldsymbol{W}_{\boldsymbol{c}}$ is the learned parameter and $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}$ is computed as the weighted average over all the source hidden states $\boldsymbol{h}_{\leq \boldsymbol{n}}^{\boldsymbol{x(2)}}$ [4]. Target hidden state $\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(l)}}$ is computed by previous hidden state $\boldsymbol{h}_{\boldsymbol{t-1}}^{\boldsymbol{y(l)}}$ and cell state $\boldsymbol{c}_{\boldsymbol{t-1}}^{\boldsymbol{y(l)}}$ as our RNN is based on LSTM unit [1], [26], here l is the layer number in deep RNN architecture. The last hidden state $\boldsymbol{h}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ and cell state $\boldsymbol{c}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ from the encoder are used as the initial values of the input hidden state $\boldsymbol{h}^{\boldsymbol{y(l)}}$ and cell state $\boldsymbol{c}^{\boldsymbol{y(l)}}$ at the decoder. Then the learning objective is to seek the optimal model parameters θ^*, that maximize the likelihood of the training data $\mathbb{D}$:

$$\theta^* = \arg\max_{\theta} \sum_{(x,y)\in\mathbb{D}} \log\, p(y|x;\theta). \tag{4}$$

IV. Knowledge-Aware NMT

In this section, we describe the KaNMT model in detail. Fig. 2 demonstrates the architecture of our proposed KaNMT. In addition to the conventional model as shown in Fig. 1, we use additional RNN encoder to model the external knowledge as prior knowledge. More formally, given a source sentence $x = \{x_1, \ldots, x_n\}$ and its corresponding knowledge blocks $b = \{b_1, \ldots, b_n\}$, the conditional probability of the target sentence $y = \{y_1, \ldots, y_m\}$ is now expressed as:

$$\log\, p(y|x,b) = \sum_{t=1}^{m} log\; p(y_t|y_{<t}, \boldsymbol{s}^{\boldsymbol{x}}, \boldsymbol{s}^{\boldsymbol{b}}) \tag{5}$$

where $\boldsymbol{s}^{\boldsymbol{b}}$ denotes the representation of prior knowledge b. We explain this in Section IV-A. The conditional probability of the

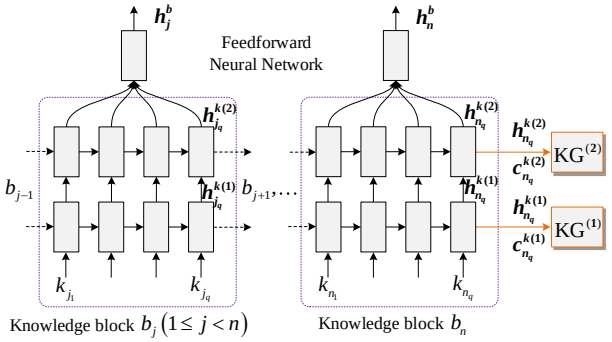


Fig. 3. Knowledge blocks $b_j = \{k_{j_1}, \ldots, k_{j_q}\}$ and b_n for source words x_j and x_n. The size of each knowledge block is q, which means that there are q types of linguistic knowledge for each source word. The number of layers of RNN in knowledge block and sentence encoder is the same as they are used to initialize the decoder.

t-th target word y_t is calculated using a non-linear function softmax, as follows:

$$p(y_t|y_{<t}, \boldsymbol{s}^{\boldsymbol{x}}, \boldsymbol{s}^{\boldsymbol{b}}) = \text{softmax}(\hat{\boldsymbol{W}}_{\boldsymbol{s}}\hat{\boldsymbol{h}}_{\boldsymbol{t}}) \tag{6}$$

where $\hat{\boldsymbol{W}}_{\boldsymbol{s}}$ represents the weight matrix, $\hat{\boldsymbol{h}}_{\boldsymbol{t}}$ is the new attentional hidden state that is based on source sentence context vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}$, source knowledge context vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$, and target hidden state $\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}$. The computation of $\hat{\boldsymbol{h}}_{\boldsymbol{t}}$ will be described in Section IV-B, and the computation of $\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}$ will be described in Section IV-C.

Finally, our training objective is formulated as follows:

$$\hat{\theta}^* = \arg\max_{\hat{\theta}} \sum_{(x,b,y)\in\hat{\mathbb{D}}} \log\, p(y|x,b;\hat{\theta}) \tag{7}$$

where $\hat{\mathbb{D}}$ is our parallel training corpus with their corresponding prior knowledge, may it be POS tags, chunk tags, named entity information, and dependency relations, or using them at the same time. The learning objective is to seek the optimal parameters $\hat{\theta}^*$.

A. Knowledge Block

Our proposed knowledge block uses a RNN to model a series of linguistic features at the word level, and those word-corresponding feature blocks are further encoded using a RNN encoder in sentence level. There is a knowledge block $b_j = \{k_{j_1}, \ldots, k_{j_q}\}$ for each source word x_j in x, q is the size of each knowledge block[1] and $\{k_{j_1}, \ldots, k_{j_q}\}$ represents the set of linguistic features of a word. Fig. 3 demonstrates the architecture of our proposed knowledge block, first, a LSTM-based RNN encodes linguistic knowledge $\{k_{j_1}, \ldots, k_{j_q}\}$ into hidden states $\{\boldsymbol{h}_{\boldsymbol{j_1}}^{\boldsymbol{k(l)}}, \ldots, \boldsymbol{h}_{\boldsymbol{j_q}}^{\boldsymbol{k(l)}}\}$ for knowledge block b_j, the number of layers of RNN in knowledge block and sentence encoder is the same as they are used to initialize the decoder, here l is the layer number. Then, a feedforward neural network converts the hidden states $\{\boldsymbol{h}_{\boldsymbol{j_1}}^{\boldsymbol{k(2)}}, \ldots, \boldsymbol{h}_{\boldsymbol{j_q}}^{\boldsymbol{k(2)}}\}$ from top layer

[1] q is a variable number and we can use it to dynamically incorporate an arbitrary number of linguistic factors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

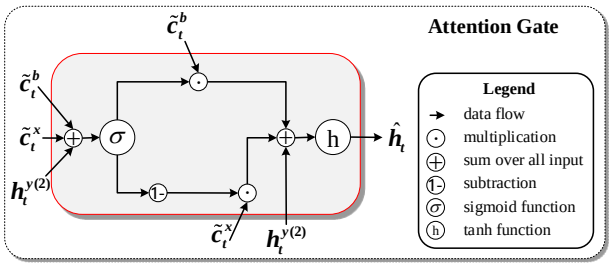


Fig. 4. Our proposed attention gate that is based on the context vectors $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}$ and $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$ from sentence encoder and knowledge encoder, and current hidden state $\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}$ from the top layer of the decoder.

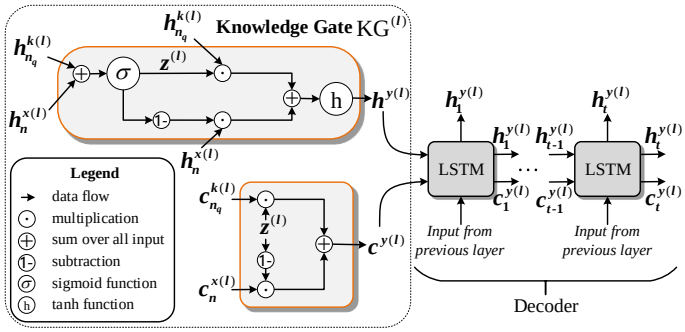


Fig. 5. Our proposed knowledge gate that is computed by the last hidden states and last cell states from sentence encoder and knowledge encoder.

of RNN into a fixed-dimensional vector $\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}} \in \mathbb{R}^d$, where $\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}$ is computed as follows:

$$\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}} = \tanh(\sum_{i=1}^{q} \boldsymbol{W}_{\boldsymbol{i}} \boldsymbol{h}_{\boldsymbol{j}_i}^{\boldsymbol{k(2)}}) \tag{8}$$

where $\boldsymbol{W}_{\boldsymbol{i}} \in \mathbb{R}^{d \times d}$ are learned parameters. The last hidden state $\boldsymbol{h}_{\boldsymbol{n_q}}^{\boldsymbol{k(l)}}$ and cell state $\boldsymbol{c}_{\boldsymbol{n_q}}^{\boldsymbol{k(l)}}$ from the last knowledge block b_n along with the last hidden state $\boldsymbol{h}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ and cell state $\boldsymbol{c}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ from the sentence encoder are used to initialize the decoder, which will be illustrated in Section IV-C.

Then we can compute the context vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$ based on $\boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}$ for timestep t. To compute $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$, the local-$p$ attention model [4] works as follows. First, a position to look at the knowledge encoder is predicted by equation:

$$p_t = n \cdot \sigma(\boldsymbol{v}_{\boldsymbol{p}}^{\top} \tanh(\boldsymbol{W}_{\boldsymbol{p}} \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}})) \tag{9}$$

where σ is a logistic sigmoid function, n is the number of knowledge blocks, and $\boldsymbol{v}_{\boldsymbol{p}}$ and $\boldsymbol{W}_{\boldsymbol{p}}$ are learned parameters. It is worth noting that p_t is a real number. After p_t is computed, a window of size $2w+1$ is looked at in the feedforward neural network layer of the knowledge encoder centered around p_t. For each hidden state in this window, we compute an alignment score $\boldsymbol{a}_{\boldsymbol{t}}(j)$:

$$\boldsymbol{a}_{\boldsymbol{t}}(j) = \text{align}(\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}, \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}) \exp(\frac{-(j-p_t)^2}{2\delta^2}) \tag{10}$$

$$\text{align}(\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}, \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}) = \frac{\exp(\text{score}(\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}, \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}))}{\sum_{j'} \exp(\text{score}(\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}, \boldsymbol{h}_{\boldsymbol{j'}}^{\boldsymbol{b}}))} \tag{11}$$

$$\text{score}(\boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}, \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}}) = \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}\top} \boldsymbol{W}_{\boldsymbol{o}} \boldsymbol{h}_{\boldsymbol{j}}^{\boldsymbol{b}} \tag{12}$$

where δ is set to $w/2$ and j is the index for that knowledge hidden state. After computing $\boldsymbol{a}_{\boldsymbol{t}}(j)$, $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$ is created by taking a weighted sum of all knowledge hidden states $\boldsymbol{h}_{\boldsymbol{\leq n}}^{\boldsymbol{b}}$.

B. Attention Gate

In our proposed KaNMT model, one straightforward method to calculate $\hat{\boldsymbol{h}}_{\boldsymbol{t}}$ in equation 6 for timestep t works as follows:

$$\hat{\boldsymbol{h}}_{\boldsymbol{t}} = \tanh(\hat{\boldsymbol{W}}_{\boldsymbol{c}}[\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}; \tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}; \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}]) \tag{13}$$

where $\hat{\boldsymbol{W}}_{\boldsymbol{c}} \in \mathbb{R}^{d \times 3d}$ is learned parameter, $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}}$ and $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}}$ are context vectors of sentence encoder and knowledge encoder for timestep t.

To our knowledge, if a source word is a function word, we prefer more information from external knowledge than from its surface as we do not want to translate it, otherwise we prefer more information from its surface as we must obtain the translation for it. Inspired by this intuition, to effectively leverage the knowledge representation in predicting the target words, we propose a weighted variant attention mechanism, attention gate, in which a time-dependent gating scalar $\boldsymbol{e}_{\boldsymbol{t}} \in \mathbb{R}^d$ is adopted to control the ratio of conditional information between the source sentence and external knowledge, as shown in Fig. 4. Formally, an attention gate consists of a sigmoid neural network layer and an element-wise multiplication operation. At timestep t, $\boldsymbol{e}_{\boldsymbol{t}}$ is calculated as follows:

$$\boldsymbol{e}_{\boldsymbol{t}} = \sigma(\boldsymbol{W}_{\boldsymbol{e}} \tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}} + \boldsymbol{M}_{\boldsymbol{e}} \tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}} + \boldsymbol{U}_{\boldsymbol{e}} \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}) \tag{14}$$

where $\boldsymbol{W}_{\boldsymbol{e}}$, $\boldsymbol{M}_{\boldsymbol{e}}$, and $\boldsymbol{U}_{\boldsymbol{e}} \in \mathbb{R}^{d \times d}$ are learned parameters. Then $\hat{\boldsymbol{h}}_{\boldsymbol{t}}$ is computed as follows:

$$\hat{\boldsymbol{h}}_{\boldsymbol{t}} = \tanh((1 - \boldsymbol{e}_{\boldsymbol{t}}) \odot \boldsymbol{W}_{\boldsymbol{a}} \tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{x}} + \boldsymbol{e}_{\boldsymbol{t}} \odot \boldsymbol{M}_{\boldsymbol{a}} \tilde{\boldsymbol{c}}_{\boldsymbol{t}}^{\boldsymbol{b}} + \boldsymbol{U}_{\boldsymbol{a}} \boldsymbol{h}_{\boldsymbol{t}}^{\boldsymbol{y(2)}}) \tag{15}$$

where $\boldsymbol{W}_{\boldsymbol{a}}, \boldsymbol{M}_{\boldsymbol{a}}, \boldsymbol{U}_{\boldsymbol{a}} \in \mathbb{R}^{d \times d}$ are learned parameters.

C. Knowledge Gate

Attention mechanism is one method that allows information flow from encoder to decoder. Meanwhile, the last hidden states and cell states from sentence encoder and knowledge encoder are used to initialize the decoder, which is another method that allows information flow from encoder to decoder. In our proposed KaNMT, source sentence and prior knowledge have their own encoders. As we have two encoders, the question is how to combine the last hidden states $\boldsymbol{h}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ and $\boldsymbol{h}_{\boldsymbol{n_q}}^{\boldsymbol{k(l)}}$, and last cell states $\boldsymbol{c}_{\boldsymbol{n}}^{\boldsymbol{x(l)}}$ and $\boldsymbol{c}_{\boldsymbol{n_q}}^{\boldsymbol{k(l)}}$ from each encoder, to pass on to the decoder, on which the computations of $\boldsymbol{h}^{\boldsymbol{y(l)}}$ and $\boldsymbol{c}^{\boldsymbol{y(l)}}$ is based.

One straightforward method to combine two hidden states works by concatenating the two hidden states from the source encoders, applying a linear transformation $\boldsymbol{W}_{\boldsymbol{r}}^{(\boldsymbol{l})}$, then sending its output through a tanh non-linearity. The new cell state is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

simply the average of the two cell states from each encoder. These operations are represented by the following equations:

$$\boldsymbol{h^{y(l)}} = \tanh(\boldsymbol{W_r^{(l)}}[\boldsymbol{h_n^{x(l)}}; \boldsymbol{h_{n_q}^{k(l)}}]) \tag{16}$$

$$\boldsymbol{c^{y(l)}} = \frac{\boldsymbol{c_n^{x(l)}} + \boldsymbol{c_{n_q}^{k(l)}}}{2} \tag{17}$$

where $\boldsymbol{W_r^{(l)}} \in \mathbb{R}^{d\times 2d}$ is learned parameter, $\boldsymbol{h^{y(l)}} \in \mathbb{R}^d$ and $\boldsymbol{c^{y(l)}} \in \mathbb{R}^d$ are the initial hidden state and cell state for the RNN decoder in the l layer.

Different from the simple method shown above, we propose a gating mechanism, knowledge gate $\boldsymbol{z^{(l)}} \in \mathbb{R}^d$, to balance the information between the word feature and the additional linguistic features that is best suited for inducing the source sentence representation, as shown in Fig. 5. The knowledge gate assigns an element-wise weight $\boldsymbol{z^{(l)}}$ to the input signals of the decoder, computed by:

$$\boldsymbol{z^{(l)}} = \sigma(\boldsymbol{W_z^{(l)}}\boldsymbol{h_n^{x(l)}} + \boldsymbol{M_z^{(l)}}\boldsymbol{h_{n_q}^{k(l)}}) \tag{18}$$

where $\boldsymbol{W_z^{(l)}}, \boldsymbol{M_z^{(l)}} \in \mathbb{R}^{d\times d}$ are learned parameters and l is the layer number. Next, we use the knowledge gate to initialize the decoder as follows:

$$\boldsymbol{h^{y(l)}} = \tanh((1-\boldsymbol{z^{(l)}}) \odot \boldsymbol{W_k^{(l)}}\boldsymbol{h_n^{x(l)}} + \boldsymbol{z^{(l)}} \odot \boldsymbol{M_k^{(l)}}\boldsymbol{h_{n_q}^{k(l)}}) \tag{19}$$

$$\boldsymbol{c^{y(l)}} = (1-\boldsymbol{z^{(l)}}) \odot \boldsymbol{c_n^{x(l)}} + \boldsymbol{z^{l}} \odot \boldsymbol{c_{n_q}^{k(l)}} \tag{20}$$

where $\boldsymbol{W_k^{(l)}}, \boldsymbol{M_k^{(l)}} \in \mathbb{R}^{d\times d}$ are learned parameters. After $\boldsymbol{h^{y(l)}}$ and $\boldsymbol{c^{y(l)}}$ are computed, $\boldsymbol{h_1^{y(l)}}$ and $\boldsymbol{c_1^{y(l)}}$ can be updated with LSTM unit, and then target hidden state $\boldsymbol{h_t^{y(l)}}$ can be updated with LSTM-based RNN and has the information from source sentence and external knowledge, as shown in the right part of Fig. 5.

V. Incorporating Linguistic Knowledge

To evaluate the effectiveness of our proposed KaNMT model, knowledge block b_j contains four types of linguistic knowledge for a surface word x_j, that is, q is 4 in equation 8. Here, b_j is summarized as follows:

$$\begin{aligned} b_j &= \{k_{j_1}, k_{j_2}, k_{j_3}, k_{j_4}\} \\ &= \{g_{\text{pos}}(x_j), g_{\text{ne}}(x_j), g_{\text{chk}}(x_j), g_{\text{dp}}(x_j)\} \end{aligned} \tag{21}$$

where k is the prior knowledge, *pos* is part-of-speech tag, *ne* is named entity tag, *chk* is chunk information, and *dp* is dependency relation. Given x_j, to compute $g_*(x_j)$, first we should transform POS tags, named entity tags, chunk tags, and dependency relations into sequence, and the transformed sequence has the same length with source input sentence, then $g_*(x_j)$ returns the annotation at position j.

A. Part-of-Speech Tags

The row *POS Tags* in Fig. 6 shows an example of POS tags, Chinese word "了" is a *function* word with "AS" annotation, "北京 (Beijing)" is a *noun* with "NR" annotation, and "坐 (travel by)" is a *verb* with "VV" annotation. What we want to learn from POS tags for the KaNMT system is that function words can be omitted and content words should be translated during decoding. Therefore, incorporation of POS tags into KaNMT can help in disambiguation and alleviating meaningful word missing. Given x_j, the definition of $g_{\text{pos}}(x_j)$ is as follows:

- $g_{\text{pos}}(x_j)$: we arrange POS tags in order, and each source word x_j has its corresponding POS tag. $g_{\text{pos}}(x_j)$ returns the POS tag at position j.

For example, we annotate the Chinese source sentence "今天 上午 十点 习近平 坐 飞机 离开 了 北京 (Jinping Xi left Beijing by plane at ten this morning)" with "NT NT NT NR VV NN VV AS NR", and g_{pos}(上午) returns "NT".

B. Named Entity and Chunk Tags

Time, person, and location are often rare words and tend to produce incorrect translations during decoding [38]. These kinds of rare words can be recognized by named entity recognition and we incorporate these named entity tags into our KaNMT model to improve the translation performance for sentences containing these words. The row *NE Tags* in Fig. 6 shows an example of named entity, Chinese word "今天 (today)", "习近平 (Jingping Xi)", and "北京 (Beijing)" are "DATE", "PSN", and "LOC" types, respectively. Chinese phrase "上午 十点 (ten this morning)" is a "TIME" type. For an input source sentence and its corresponding named entities, we transform the named entities into sequence according to the following rules:

- A single word x_i with annotation *T*, output *T*.
- Sequence words $x_i, x_{i+1}, \ldots, x_j\,(i < j \leq n)$ with annotation *T*, output *T*-B, *T*-M, ..., *T*-M, where the meaning of B and M is begin and middle, respectively.
- Otherwise, output N/A.

Here, *T* is a variable. With the above rules, we can transform named entities into sequence for an input sentence and each source word x_j has its corresponding annotation. And then $g_{\text{ne}}(x_j)$ returns the annotation at position j, for example, g_{ne}(上午) is "TIME-B".

POS tagger can assign parts of speech to each word, further more, chunking which is also called shallow parsing can classify a phrase into verb phrase or noun phrase. The row *Chunk Tags* in Fig. 6 shows the result of chunking. Here, Chinese phrases "今天 上午 十点 (ten this morning)", "飞机 (plane)", and "北京 (Beijing)" are "NP" phrases, "习近平 坐 (Jinping Xi travel by)" and "离开 了 (left)" are "VP" phrases. The approach that we use to transform chunks into sequence is the same with the approach for named entities, for example, g_{chk}(上午) is "NP-M".

C. Dependency Relations

In machine translation, dependency parse tree expresses the structure of a source sentence, and can help in generating correct sentence structure if our translation model uses it during translation inference. The row *DP Relations* in Fig. 6 presents an example parse tree used in our approach. In our proposed approach, we transform dependency tree into sequence as we arrange the dependency relations in order.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

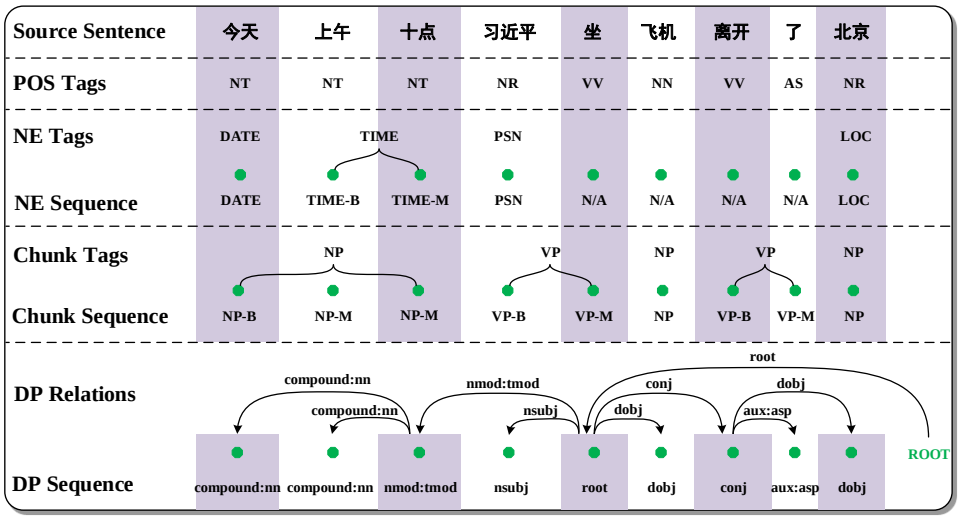


Fig. 6. Incorporating POS tags, named entity tags, chunk tags, and syntactic relations into our proposed KaNMT system. Here, we show the method that we use to transform these prior knowledge into sequence.

TABLE II
STATISTICS OF TRAINING AND VALIDATION CORPORA ON CHINESE↔ENGLISH TRANSLATION TASKS. THE SENTENCES ARE LIMITED UP TO 20 WORDS LONG

Data	# Sentences	# Chn tokens	# Eng tokens
Training	2.85 M	31.29 M	33.75 M
Validation	2,492	27,511	32,284

TABLE III
STATISTICS OF VOCABULARY SIZE FOR 4 TYPES OF PRIOR KNOWLEDGE AS WELL AS SURFACE WORDS ON CHINESE↔ENGLISH TRANSLATION TASKS

Feature	Chn vocab size	Eng vocab size
POS tags	32	45
NE tags	36	7
Chunk tags	23	19
Dependency relations	11	553
All	102	624
Concatenate	6,743	10,213
Words	148,688	148,479

The dependency relation is that it connects a word to its syntactic head, or "root" if the word has no syntactic head. Here we just focus on the syntactic head as previous work has found that both local and global syntactic information about source sentences is captured by the encoder [16], and we enhance the learning ability of the source encoder with syntactic head. For example, we annotate the Chinese sentence with "compound:nn compound:nn nmod:tmod nsubj root dobj conj aux:asp dobj" in Fig. 6, and each Chinese word has an annotation. $g_{\text{dp}}(x_j)$ returns the annotation at position j, for example, $g_{\text{dp}}(\text{上午})$ returns "compound:nn".

VI. EVALUATION

In our work, we incorporate four kinds of prior knowledge shown in Section V into our proposed KaNMT model and carry out experiments on Chinese↔English and English→German translations.

A. Training and Validation Corpora

On Chinese↔English translations, our bilingual training corpora consist of 2.85M sentence pairs selected from the NIST portion of the bilingual data of NIST MT 2008 Evaluation. We randomly select 2,492 parallel sentences from the training dataset as our validation dataset, and remove them from the training dataset. For more details about training and validation datasets, please refer to Table II. We use NIST 2006, NIST 2008, and NIST 2008 progress MT evaluation sets as our test datasets for Chinese→English translation. For English→Chinese translation task, first, we concatenate CWMT2009 and CWMT2011 English→Chinese evaluation sets into CWMT2009-2011, and then we use CWMT2009-2011 and NIST 2008 MT English→Chinese evaluation sets as our test datasets. All of our test datasets have 4 references on both translation tasks.

B. Experiment Setup

In our experiments, we use NiuParser[2] [39] to annotate Chinese input sentences with POS tags, named entity tags, chunk tags, and syntactic relations on Chinese→English translation. The column *Chn vocab size* in Table III shows the statistics of vocabulary size for 4 types of linguistic knowledge as well as surface words for Chinese. We use Stanford CoreNLP[3] to annotate English input sentences with POS tags [40], named entity tags [41], and syntactic relations [42] on English→Chinese translation. For English chunking, we use YamCha[4] [43] that is trained on CoNLL-2000 training data for chunking. For English, the column *Eng vocab size* in Table III shows the

[2]http://www.niuparser.com/index.en.html
[3]https://stanfordnlp.github.io/CoreNLP/
[4]http://chasen.org/~taku/software/yamcha/

statistics of vocabulary size for 4 types of linguistic knowledge as well as surface words. For Chinese and English source sentence, as we use different toolkits to obtain their prior knowledge, so the number of linguistic tags is different for the same knowledge. In our experiments, we also concatenate all 4 types of prior knowledge into one single token, for example, the concatenated prior knowledge for "今天 (today)" in Fig. 6 is "NT|DATE|NP-B|compound:nn", and we name these systems as Knowledge Concatenate on Chinese↔English translations. The row *Concatenate* shows the statistics of vocabulary size for this method. The results are evaluated in case-insensitive BLEU using the `mteval-v13a.pl` script.[5]

C. Compared Systems

First of all, we compare our proposed KaNMT systems with conventional NMT baseline systems on both translation tasks. In this part, we incorporate POS tags, named entity tags, chunk tags, and dependency relations into our KaNMT systems. Then, we compare KaNMT systems with our methods that do not contain attention and knowledge gates on both translation tasks. Third, we compare KaNMT systems with our model that uses simple concatenation method for external linguistic knowledge. Finally, we compare our KaNMT systems with the extended embedding NMT systems [20] and Transformer [44] on both translation tasks. More detail about our compared systems are shown below:

- **NMT baseline**: our NMT systems have a deep LSTM network with 2 encoder and 2 decoder layers with local attention model and feed-input model [4]. Our encoder-decoder with LSTM units [26] is trained for maximum likelihood with back-propagation through time (BPTT) [45]. All of the models use $1,024$ LSTM nodes per encoder and decoder layers. The size of source and target word embeddings is $1,024$. The size of source and target vocabularies is 50K with the subword method proposed by Sennrich [46]. We use minibatches of size 80 and a maximum sentence length of 20 words. We clip the gradient norm to 5.0. Our parameters are uniformly initialized in $[-0.08, 0.08]$. We train for 10 epochs[6] using stochastic gradient descent (SGD), start with a learning rate of 0.70, and begin to halve the learning rate every epoch after 5 epochs. We set dropout rate to 0.2 for our LSTMs [47]. In our experiments, we present the source sentence in reverse order as suggested by Sutskever, Vinyals, and Le [1].
- **KaNMT**: the proposed knowledge-aware NMT has an additional knowledge encoder, 1 attention gate, and 2 knowledge gates associated with it. The size of knowledge embedding is set to $1,024$. The other settings are the same with NMT baseline. We incorporate POS tags, named entity tags, chunk tags, and dependency relations into the KaNMT with the method shown in Fig. 2. The row *All* in Table III shows the vocabulary size of all prior knowledge.
- **KaNMT w/o KG & AG**: we use Eqs. (13), (16) and (17) to compute $\hat{\boldsymbol{h}}_t$, $\boldsymbol{h}^{\boldsymbol{y}(\boldsymbol{l})}$, and $\boldsymbol{c}^{\boldsymbol{y}(\boldsymbol{l})}$ instead of using the attention gate and knowledge gates in KaNMT systems, the other settings are the same with our proposed KaNMT systems.
- **KaNMT w/o KG**: to investigate the effectiveness of the knowledge gates, we remove the knowledge gates from the KaNMT systems while the other settings remain the same.
- **KaNMT Concatenate**: for one source surface word, we concatenate its 4 types of linguistic knowledge into one single token. Then, we use the source words with their corresponding concatenated knowledge to train and test our KaNMT systems. The other settings are the same with our KaNMT. The row *Concatenate* in Table III shows the vocabulary size of the prior knowledge.
- **Extended Embedding**: to compare the method proposed by Sennrich and Haddow [20] with our KaNMT, we generalize the embedding layer of the encoder to support the inclusion of POS tags, named entity tags, chunk tags, and dependency relations in our baseline system. The embeddings of POS tags, named entity tags, chunk tags, and dependency relations are set to 20, and the word embedding is set to 944, totally $1,024$. The other settings are the same as the NMT baseline.
- **Transformer**:[7] transformer [44] is a state-of-the-art neural machine translation model and we use it as another contrast system. In transformer, the number of attention heads is 8, both the attention key and value dimensions are 64. The size of minibatch is $2,048$, the dimension of hidden state is 512, and the number of layers is 6. We use the Adam [48] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. Dropout [49] is applied to the output of each sub-layer and the dropout rate is set to 0.1. We train transformer for $100,000$ steps.

D. Translation Performance

Table IV shows the BLEU scores of the NMT baseline systems and our KaNMT systems on Chinese↔English translation. We use the paired bootstrap resampling for testing the statistical significance of the difference between two systems [50]. On Chinese→English translation, we can see that our KaNMT systems improve the translation performance by incorporating 4 types of external linguistic knowledge into it. Compared with baseline system, our method with POS tag (the row *KaNMT w/ POS*) achieved 0.2, 1.4, and 1.2 BLEU points increase on NIST MT 2006, 2008, and 2008 progress evaluation datasets, respectively. NE achieves 0.3, 0.8, and 1.1 BLEU points increase and Chunk achieves 0.1, 0.7, and 1.7 BLEU points increase, respectively. The BLEU points increase of our approach with dependency relations is 1.0, 1.2, and 0.5, respectively. Finally, we use 4 types of linguist knowledge together (the row *KaNMT*) and the BLEU points increase is 1.0, 2.2, and 2.3, respectively, which are significantly better than the NMT baseline system. We can see that BLEU points increase is consistently higher for the system trained with all

[5]http://www.itl.nist.gov/iad/mig/tests/mt/2009/

[6]We do not observe any noticeable improvement in BLEU scores after 10 epochs.

[7]https://github.com/tensorflow/tensor2tensor

TABLE IV
BLEU SCORES ON CHINESE↔ENGLISH TRANSLATION TASKS WITH OUR KANMT THAT INCORPORATES POS TAGS, NAMED ENTITY TAGS, CHUNK TAGS, AND DEPENDENCY RELATIONS INTO IT. BOLD FONT INDICATES HIGHEST BLEU POINTS. HERE, * INDICATES A SCORE SIGNIFICANTLY BETTER THAN BASELINE AT $p < 0.05$

Systems	Chinese→English			English→Chinese	
	MT06	MT08	MT08-pro	CWMT2009-2011	MT08
Baseline	35.8	27.7	23.6	26.5	27.1
KaNMT w/ POS	36.0	29.1*	24.8*	27.6*	27.4
KaNMT w/ NE	36.1	28.5*	24.7*	27.4*	27.5*
KaNMT w/ Chunk	35.9	28.4*	25.3*	27.3*	27.3
KaNMT w/ DP	**36.8***	28.9*	24.1*	27.7*	27.9*
KaNMT	**36.8***	**29.9***	**25.9***	**28.3***	**28.4***

TABLE V
COMPARING OUR PROPOSED KANMT WITH KANMT W/O KG & AG, KANMT W/O KG, KANMT CONCATENATE, EXTENDED EMBEDDING, AND TRANSFORMER ON CHINESE↔ENGLISH TRANSLATIONS. $\Diamond$, $+$, †, AND ‡ INDICATE THAT OUR KANMT IS SIGNIFICANTLY BETTER THAN KANMT W/O KG & AG, KANMT CONCATENATE, EXTENDED EMBEDDING, AND TRANSFORMER AT $p < 0.05$, RESPECTIVELY. HERE, WE INCORPORATE POS TAGS, NAMED ENTITY TAGS, CHUNK TAGS, DEPENDENCY RELATIONS INTO OUR PROPOSED MODEL AT THE SAME TIME. BOLD FONT INDICATES HIGHEST BLEU POINTS

Systems	Chinese→English				English→Chinese		
	MT06	MT08	MT08-pro	Avg.	CWMT2009-2011	MT08	Avg.
Baseline	35.8	27.7	23.6	29.0	26.5	27.1	26.8
KaNMT w/o KG & AG	36.0	28.6	24.8	29.8	26.7	27.4	27.1
KaNMT w/o KG	36.6	29.6	25.7	30.6	28.0	28.2	28.1
KaNMT Concatenate	36.6	29.4	25.2	30.4	27.7	28.0	27.9
Extended Embedding	36.4	28.3	24.6	29.8	27.6	27.8	27.7
KaNMT	**36.8**$^{\Diamond\dagger\ddagger}$	**29.9**$^{\Diamond+\dagger\ddagger}$	**25.9**$^{\Diamond+\dagger}$	**30.9**	**28.3**$^{\Diamond+\dagger\ddagger}$	**28.4**$^{\Diamond+\dagger}$	**28.4**
Transformer [44]	36.5	29.1	25.7	30.4	27.9	28.2	28.1

linguistic features compared with other single systems on all test datasets. On English→Chinese translation, our approach with POS tag achieves 1.1 and 0.3 BLEU points increase on CWMT2009-2011 and NIST MT 2008 evaluation datasets compared with baseline system, respectively. NE achieves 0.9 and 0.4 BLEU points increase and Chunk achieves 0.8 and 0.2 BLEU points increase. Our approach with dependency relations achieves 1.4 and 0.8 BLEU points increase, respectively. With all external linguistic knowledge in the row *KaNMT*, the BLEU points increase is 1.8 and 1.3, which are significantly better than baseline system.

In Table V, first, we compare our KaNMT systems (the row *KaNMT*) with our method without knowledge and attention gates (the row *KaNMT w/o KG & AG*) on Chinese→English and English→Chinese translations. The method without gates is computed by Eqs. (13), (16) and (17). From these two tables we can see that our proposed KaNMT systems perform significantly better than our method without gates on all test datasets. As our proposed attention and knowledge gates can control the proportions of information flowing from sentence encoder and knowledge encoder to the decoder. Then, we compare our KaNMT systems with the KaNMT w/o KG. We can see that our proposed knowledge gates can obtain 0.3 BLEU points of improvement on both translation tasks. Then, we compare our KaNMT systems with the KaNMT Concatenate on Chinese↔English translations. From these two tables we can see that our proposed KaNMT systems always perform significantly better than KaNMT Concatenate systems on all test datasets, except for MT06 Chinese→English test set. Then we compare our KaNMT systems with Extended Embedding systems [20] on Chinese↔English translation tasks. We can see that our KaNMT systems always perform significantly better than Extended Embedding systems that incorporate 4 types of external knowledge. The main reason is that this straightforward approach of extending embedding is overly simplistic and lacks explicit mechanism to balance the context that contributed from the word embedding and those of additional linguistic knowledge, however our proposed model can do it. Finally, our proposed KaNMT systems are significantly better than transformer systems on MT06, MT08, and CWMT2009-2011 test datasets on Chinese↔English translation tasks.

The last columns of Table V show the average BLEU scores of different MT systems on Chinese↔English translation tasks. We can see that our proposed KaNMT systems always perform much better than other NMT systems. On Chinese→English translation, our KaNMT system achieves 1.9, 1.1, 0.3, 0.5, 1.1, and 0.5 BLEU points increase compared with NMT baseline, KaNMT w/o KG & AG, KaNMT w/o KG, KaNMT Concatenate, Extended Embedding, and transformer, respectively. On English→Chinese translation, the BLEU points increase is 1.6, 1.3, 0.3, 0.5, 0.7, and 0.3, respectively.

E. Perplexities

Figure 7 shows the perplexities of validation datasets for the NMT baseline systems, KaNMT without KG & AG systems, KaNMT concatenate systems, and our KaNMT systems on Chinese↔English translation tasks, all KaNMT systems

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

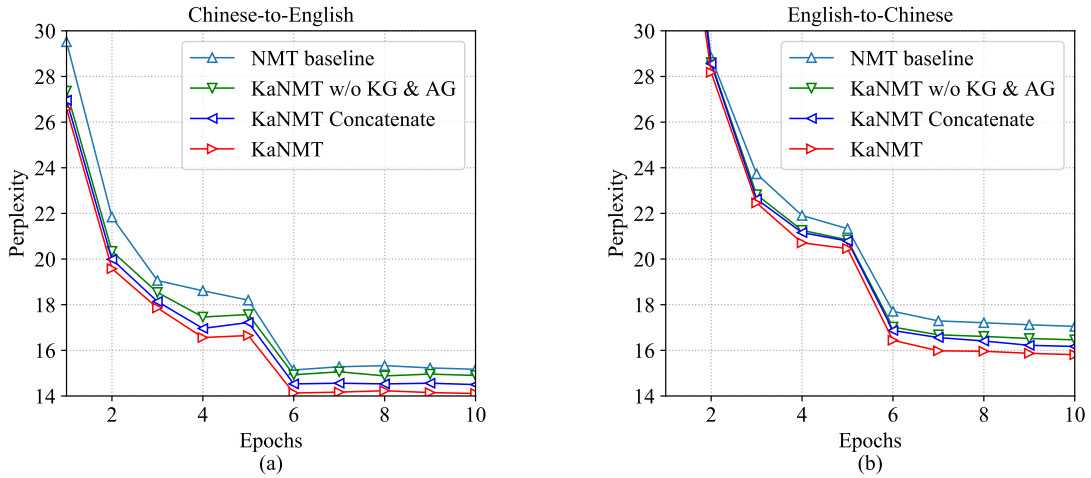


Fig. 7. Perplexities (y-axis) as a function of epochs (x-axis) on validation dataset for the NMT baseline systems, KaNMT without KG & AG systems, KaNMT concatenate systems, and our KaNMT systems on Chinese↔English translation tasks.

TABLE VI
SAMPLE TRANSLATIONS OF THE PROPOSED KANMT SYSTEM AND THE BASELINE SYSTEM ON CHINESE→ENGLISH TRANSLATION. WORDS WITH GRAY BACKGROUND INDICATE TRANSLATION ERRORS IN THE BASELINE OUTPUT BUT CORRECT IN THE PROPOSED MODEL'S OUTPUT. BLUE WORDS ARE THE CORRESPONDING LINGUISTIC KNOWLEDGE FOR EACH SOURCE PROBLEMATIC WORD

		Chinese → English Translation
1	Source	美 欲$_{VV}$ 重启 中东 和谈
	Reference	US wants to reopen Middle East peace talks
	Baseline	US to resume Middle East peace talks
	KaNMT w/ POS	US wants to restart Middle East peace talks
2	Source	联合国 因 德黑兰$_{LOC}$ 未能 冻结 铀 浓缩 作业 , 已 对 伊朗 实施 两 套 制裁 。
	Reference	The UN has already imposed two sets of sanctions on Iran because Tehran failed to freeze its uranium enrichment operation.
	Baseline	The United Nations has imposed two sanctions against Iran since it failed to freeze uranium enrichment operations.
	KaNMT w/ NE	The United Nations has imposed two sets of sanctions against Iran because of Tehran failed to freeze uranium enrichment operations .
3	Source	英国 警方 将 调查 范围$_{NP}$ 转向 印度 和 澳大利亚 。
	Reference	British police has switched its scope of investigation to India and Australia.
	Baseline	British police now turn the investigation into India and Australia.
	KaNMT w/ Chunk	British police have shifted the survey scope to India and Australia.
4	Source	朝 韩 就 开展$_{ROOT}$ 轻工业 和 地下 资源 开发$_{compound:nn}$ 合作 达成 协议
	Reference	North-South Koreas reach cooperation agreement on developing light industry and tapping underground resources
	Baseline	South Korea reaches agreement on light industry and underground resources developing
	KaNMT w/ DP	DPRK, South Korea reach agreement on developing light industry and underground resources

incorporate 4 types of prior knowledge. From this two figures we can see that perplexities are consistently lower for our proposed KaNMT systems trained with 4 types of linguistic features compared with the NMT baseline systems, the KaNMT without KG & AG, and the KaNMT concatenate, and the results shown here track the results of BLEU scores in Table IV and V.

F. Sample Translation

Table VI shows four sample translations of 1) our KaNMT systems, i.e. KaNMT w/ POS, KaNMT w/ NE, KaNMT w/ Chunk and KaNMT w/ DP, which is respectively trained with each type of linguistic features, and 2) the baseline system on the task of Chinese→English translation. We want to know how each category of prior knowledge helps in solving the problems that appear in the translation of the baseline system. For the first sample in Table VI-(1), our approach with POS tags obtains translation "wants" for the verb "欲" whereas the translation for the word in the baseline system is missing. From this sample example we can see that POS tags have ability in alleviating absence of meaningful words. One possible reason is that NMT systems with POS tags can distinguish words from function words. In Table VI-(2), the baseline system fails to translate source word "德黑兰" whose correct translation is "Tehran". Named entity recognition system knows the word is a location and our KaNMT system with such knowledge can translate it correctly. In a similar way, our NMT system with NE knowledge manages to translate the source word "套" into "sets", whereas the baseline system omits it. In Table VI-(3), source phrase "调查 范围 (scope of investigation)" is a noun phrase in the chunking results. The baseline system fails to translate the source word "范围" whereas our KaNMT system with chunk knowledge obtains the correct translation "scope" for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE VII
SAMPLE TRANSLATIONS OF THE PROPOSED KANMT SYSTEM AND THE BASELINE SYSTEM ON ENGLISH→CHINESE TRANSLATION. WORDS WITH GRAY BACKGROUND INDICATE TRANSLATION ERRORS IN THE BASELINE OUTPUT BUT CORRECT IN THE PROPOSED MODEL'S OUTPUT. BLUE WORDS ARE THE CORRESPONDING LINGUISTIC KNOWLEDGE FOR EACH SOURCE PROBLEMATIC WORD

		English → Chinese Translation
1	Source	The spokeswoman said the price$_{NN}$ agreed for the stake was 1.7 billion dollars (1.25 billion euros) .
	Reference	这 位 发言人 说 , 商定 的 股份 价格 为 17亿 美元 (合 12.5亿 欧元) 。
	Baseline	这 位 发言人 说 , 为 17亿 美元 (12.5亿 欧元) 。
	KaNMT w/ POS	这 位 发言人 说 , 所 涉 价格 为 17亿 美元 (12.5亿 欧元) 。
2	Source	Pakistan$_{LOC}$ cleric says would rather die than surrender
	Reference	巴基斯坦 教士 声称 宁 死 不 降
	Baseline	巴 米 韦 说 不如 投降
	KaNMT w/ NE	巴基斯坦 教士 说 , 他 将 死亡
3	Source	"We have decided that we$_{NP}$ can be martyred$_{VP}$ but we will not surrender .
	Reference	" 我们 已 决定 我们 可以 牺牲 , 但 我们 不会 投降 。
	Baseline	" 我们 已 决定 可以 牺牲 , 但 我们 不会 投降 。
	KaNMT w/ Chunk	" 我们 已 决定 我们 可以 牺牲 , 但 我们 不会 投降 。
4	Source	No one is saying$_{ROOT}$ that$_{dobj}$ in$_{case}$ the$_{det}$ media$_{nmod:in}$.
	Reference	没有 人 在 媒体 中 那么 说 。
	Baseline	没有 人 会 说 媒体 。
	KaNMT w/ DP	没有 人 在 媒体 上 说 。

it. In Table VI-(4), our approach obtains the perfect sentence structure like "on developing . . .", whereas the baseline system obtains the sentence structure like "on . . . developing". The sample show us that the KaNMT system has better ability in generating correct sentence structures.

Table VII shows four sample translations on English→Chinese translation. In Table VII-(1), source word "price" gets perfect translation "价格" in our KaNMT system with POS knowledge, whereas the translation for it in the baseline system is missing. In Table VII-(2), source word "Pakistan" is a location and our KaNMT system with NE knowledge gets correct translation "巴基斯坦", whereas the baseline system incorrectly translate it to "巴 米 韦". At the same time, our KaNMT system can express the meaning of the source phrase "would rather die than surrender" with "将 死亡", where the baseline system obtains the opposite meaning "不如 投降" that means "would rather surrender". In Table VII-(3), "we" and "can be martyred" are NP and VP phrases in chunking, respectively, and our KaNMT system with chunk knowledge gets correct translation "我们 可以 牺牲" for source phrase "we can be martyred", whereas the baseline system translates it to "可以 牺牲 (can be martyred)" that the pronoun "we" in source side fails to translate. In Table VII-(4), the meaning of translation in the baseline system and the source English sentence is different because of incorrect sentence structure "说 媒体 (blame the media)". The translation of the baseline system has incorrect sentence structure and incorrect meaning according to the source English sentence, whereas our approach with dependency parsing relations generates correct sentence structure "在 媒体 . . . 说 (say in the media)" and obtains correct translation result.

From Table VI and Table VII we can see that our proposed KaNMT systems can obtain better translation results with correct word selection and sentence structure compared with conventional NMT systems. This is the reason why the BLEU scores of our KaNMT systems can outperform the baseline systems on Chinese↔English translations.

TABLE VIII
COMPARING OUR PROPOSED KANMT CONCATENATE SYSTEM WITH THE BASELINE, BYTENET, GNMT, CONVS2S, MOE, AND TRANSFORMER ON THE ENGLISH→GERMAN NEWSTEST2014 TEST SET

Systems	**newstest2014**
ByteNet [51]	23.75
Baseline	24.09
GNMT [52]	24.61
ConvS2S [53]	25.16
KaNMT Con.	25.34
MoE [54]	26.03
Transformer [44]	27.30

TABLE IX
BLEU SCORES OVER SENTENCE LENGTHS ON ENGLISH→GERMAN NEWSTEST2014 TEST SET

Systems	**Sentence Length**		
	$[2, 16]_{11.6}$	$[17, 26]_{21.1}$	$[27, 91]_{36.1}$
Baseline	23.88	23.42	24.37
KaNMT Con.	$24.64_{\uparrow 3.2\%}$	$24.79_{\uparrow \mathbf{5.8}\%}$	$25.68_{\uparrow 5.4\%}$

Note: $[min, max]_{avg}$, where min and max indicate the minimum and maximum length of the sentences in the group, and avg is the average length.

G. English→German Translation

We also evaluate our model on the WMT English→German translation task. For these experiments, the training set contains 4.56M sentence pairs. We compute case-sensitive tokenized BLEU (`multi-bleu.perl`[8]) to compare against previous work [44], [51]–[54]. The newstest2013 and the newstest2014 are used as the validation set and the test set,

[8]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

respectively. In this experiment, we extend to use the data up to 91 English words and 75 German words of sentences in length. To prevent from running out of memory due to the long training sentences, our MT system is trained using the KaNMT Concatenate method. The settings of the baseline system and our KaNMT concatenate system are same as that of the Chinese↔English translation. We use a source and target vocabulary with 32K BPE types. Table VIII reports the BLEU scores of our baseline system, our KaNMT Concatenate system, and the compared models on the newstest2014 test set. From the results, we can see that our KaNMT model outperforms the ByteNet, Baseline, GNMT, and ConvS2S, by 1.59, 1.25, 0.73 and 0.18 BLEU points, respectively. While the BLEU score is slightly lower than that of the MoE and Transformer.

Table IX reports the BLEU scores with respect to the varying lengths of the source sentences on the English→German newstest2014 test set. We sort the source sentences from short to long, then divide the sentences into three disjoint groups. Each group has $1,001$ sentences. From the translation results, we observe that our KaNMT model yields an improvement of 5.8% in the group of $[17, 26]$, and consistently outperforms the baseline on the other two groups. In addition, we find that, as the sentences getting longer, our KaNMT model tends to achieve a steady improvement over the previous groups of the shorter sentences, showing that our proposed model is better for long sentences.

VII. Conclusion

In this paper, we proposed a linguistic knowledge-aware NMT that independently models additional linguistic features in parallel to the word feature. In our proposed KaNMT, first, we use attention and knowledge gates to dynamically control the proportions of information at which word and linguistic knowledge contributes to the generation of target words. Then, we incorporate four kinds of external linguistic knowledge, including POS tags, named entity tags, chunk tags, and dependency relations, into the knowledge block b_j for each source word x_j in our proposed KaNMT system to demonstrate the effectiveness of our method. Extensive experiments show that our approach is capable of better using additional linguistic knowledge, and we observe significant improvements of BLEU scores from 1.0 to 2.3 on the Chinese↔English and English→German translation tasks.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks

[2] K. Cho, B. van Merrienboer, 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1724–1734. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1179.pdf

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

[4] 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1166.pdf

[5] J. Zhang and C. Zong, "Deep neural networks in machine translation: An overview," *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 16–25, 2015. [Online]. Available: https://doi.org/10.1109/MIS.2015.69

[6] B. Zhang, D. Xiong, J. Su, and H. Duan, "A context-aware recurrent encoder for neural machine translation," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 12, pp. 2424–2432, 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2017.2751420

[7] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2013, pp. 1700–1709. [Online]. Available: http://aclweb.org/anthology/D/D13/D13-1176.pdf

[8] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for WMT 16," in *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, 2016, pp. 371–376. [Online]. Available: http://aclweb.org/anthology/W/W16/W16-2323.pdf

[9] R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. M. Barone, and P. Williams, "The university of edinburgh's neural MT systems for WMT17," in *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, 2017, pp. 389–399. [Online]. Available: http://aclanthology.info/papers/W17-4739/w17-4739

[10] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, 2003, pp. 48–54. [Online]. Available: http://aclweb.org/anthology/N/N03/N03-1017.pdf

[11] P. Koehn and H. Hoang, "Factored translation models," in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, 2007, pp. 868–876. [Online]. Available: http://www.aclweb.org/anthology/D07-1091

[12] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context gates for neural machine translation," *TACL*, vol. 5, pp. 87–99, 2017. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/948

[13] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: a case study," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 257–267. [Online]. Available: http://aclweb.org/anthology/D/D16/D16-1025.pdf

[14] A. Toral and V. M. Sánchez-Cartagena, "A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 1063–1073. [Online]. Available: http://www.aclweb.org/anthology/E17-1100

[15] M. Popović, "Comparing language related issues for nmt and pbmt between german and english," *The Prague Bulletin of Mathematical Linguistics*, vol. 108, no. 1, pp. 209–220, 2017.

[16] X. Shi, I. Padhi, and K. Knight, "Does string-based neural mt learn source syntax?" in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 1526–1534. [Online]. Available: http://aclweb.org/anthology/D/D16/D16-1159.pdf

[17] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string alignment template for statistical machine translation," in *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006. [Online]. Available: http://aclweb.org/anthology/P06-1077

[18] T. Xiao, D. F. Wong, and J. Zhu, "A loss-augmented approach to training syntactic machine translation systems," *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, vol. 24, no. 11, pp. 2069–2083, Nov 2016.

[19] M. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[20] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, 2016, pp. 83–91. [Online]. Available: http://aclweb.org/anthology/W/W16/W16-2209.pdf

[21] A. Eriguchi, K. Hashimoto, and Y. Tsuruoka, "Tree-to-sequence attentional neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016, pp. 823–833. [Online]. Available: http://aclweb.org/anthology/P/P16/P16-1078.pdf

[22] B. Yang, D. F. Wong, T. Xiao, L. S. Chao, and J. Zhu, "Towards bidirectional hierarchical representations for attention-based neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1443–1452. [Online]. Available: http://aclweb.org/anthology/D17-1151

[23] M. Nadejde, S. Reddy, R. Sennrich, T. Dwojak, M. Junczys-Dowmunt, P. Koehn, and A. Birch, "Syntax-aware neural machine translation using ccg," *CoRR*, vol. abs/1702.01147, 2017. [Online]. Available: http://arxiv.org/abs/1702.01147

[24] M. García-Martínez, L. Barrault, and F. Bougares, "Factored neural machine translation," *CoRR*, vol. abs/1609.04621, 2016. [Online]. Available: http://arxiv.org/abs/1609.04621

[25] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting cross-sentence context for neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 2826–2831.

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

[27] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015, pp. 1556–1566. [Online]. Available: http://aclweb.org/anthology/P/P15/P15-1150.pdf

[28] F. Stahlberg, E. Hasler, A. Waite, and B. Byrne, "Syntactically guided neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*, 2016. [Online]. Available: http://aclweb.org/anthology/P/P16/P16-2049.pdf

[29] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007. [Online]. Available: http://dx.doi.org/10.1162/coli.2007.33.2.201

[30] P. Koehn and H. Hoang, "Factored translation models," in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, 2007, pp. 868–876. [Online]. Available: http://www.aclweb.org/anthology/D07-1091

[31] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *CoRR*, vol. abs/1611.04558, 2016. [Online]. Available: http://arxiv.org/abs/1611.04558

[32] B. Zoph and K. Knight, "Multi-source neural translation," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 30–34. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1004.pdf

[33] I. Calixto, Q. Liu, and N. Campbell, "Doubly-attentive decoder for multi-modal neural machine translation," *CoRR*, vol. abs/1702.01287, 2017. [Online]. Available: http://arxiv.org/abs/1702.01287

[34] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 866–875. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1101.pdf

[35] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Yarman-Vural, and K. Cho, "Zero-resource translation with multi-lingual neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 268–277. [Online]. Available: http://aclweb.org/anthology/D/D16/D16-1026.pdf

[36] J. Li, D. Xiong, Z. Tu, M. Zhu, M. Zhang, and G. Zhou, "Modeling source syntax for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 688–697. [Online]. Available: http://aclanthology.coli.uni-saarland.de/pdf/P/P17/P17-1064.pdf

[37] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao, "Neural machine translation with source dependency representation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2017, pp. 2836–2842. [Online]. Available: http://aclanthology.info/papers/D17-1303/d17-1303

[38] T. Luong, I. Sutskever, Q. V. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015, pp. 11–19. [Online]. Available: http://aclweb.org/anthology/P/P15/P15-1002.pdf

[39] J. Zhu, M. Zhu, Q. Wang, and T. Xiao, "Niuparser: A chinese syntactic and semantic parsing toolkit," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, System Demonstrations*, 2015, pp. 145–150. [Online]. Available: http://aclweb.org/anthology/P/P15/P15-4025.pdf

[40] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, 2003, pp. 173–180. [Online]. Available: http://aclweb.org/anthology/N/N03/N03-1033.pdf

[41] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, 2005, pp. 363–370. [Online]. Available: http://aclweb.org/anthology/P/P05/P05-1045.pdf

[42] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 740–750. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1082.pdf

[43] T. Kudo and Y. Matsumoto, "Fast methods for kernel-based text analysis," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, 2003, pp. 24–31. [Online]. Available: http://aclweb.org/anthology/P/P03/P03-1004.pdf

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 6000–6010. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need

[45] P. J. Werbos, "Bachpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[46] 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016. [Online]. Available: http://aclweb.org/anthology/P/P16/P16-1162.pdf

[47] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2864648, IEEE/ACM Transactions on Audio, Speech, and Language Processing

optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[49] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

[50] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, 2004, pp. 388–395. [Online]. Available: http://www.aclweb.org/anthology/W04-3250

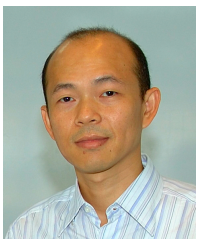
[51] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *CoRR*, vol. abs/1610.10099, 2016. [Online]. Available: http://arxiv.org/abs/1610.10099

[52] 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

[53] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 1243–1252. [Online]. Available: http://proceedings.mlr.press/v70/gehring17a.html

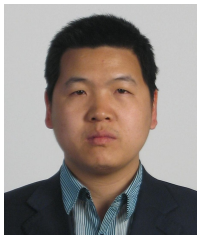
[54] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *CoRR*, vol. abs/1701.06538, 2017. [Online]. Available: http://arxiv.org/abs/1701.06538

Qiang Li is currently a Ph.D. candidate in the School of Computer Science and Engineering, Northeastern University, Shenyang, China. He received his M.Sc. with honors in the institute of computer software and theory from Northeastern University in 2013, from where he received his B.Sc. in 2011. Moreover, he had internships at Microsoft Research Asia (Beijing) in 2015, Information Sciences Institute, University of Southern California (USC/ISI) in 2016, NLP^{2}CT Lab, University of Macau in 2017.



Derek F. Wong received the Ph.D. degree in Automation from Tsinghua University in 2005. He is currently an Associate Professor in the Department of Computer and Information Science at the University of Macau, with a secondary appointment as a project manager in the Instituto de Engenharia de Sistemas e Computadores de Macau during 2003–2013. His active and diverse research interests span the areas of natural language processing and machine translation. He is the leader of the Natural Language Processing & Portuguese–Chinese Machine Translation (NLP^{2}CT) research group and the founder of the NLP^{2}CT laboratory.

Lidia S. Chao received the Ph.D. degree in Software Engineering from the University of Macau in 2008. Since 1996, she has been with the Department of Computer and Information Science at the University of Macau, being currently an Assistant Professor. Her current research focuses are data mining and machine learning technology, and knowledge acquisition in language and bioinformatics.



Muhua Zhu is currently a researcher in Alibaba Inc. He finished his Ph.D study at Northeastern University, Shenyang, China in 2013. He received his B.Sc. in 2003 and received his M.Sc. in 2006 from the same university.

Tong Xiao received the Bachelor's, Master's, and Ph.D. degrees in computer science all from Northeastern University, Shenyang, China, in 2005, 2008, and 2012, respectively. In 2014, he was elected as a candidate of the Excellent Ph.D. Dissertation Award established by Chinese Information Processing Society of China. He is currently an Associate Professor in the College of Computer Science and Engineering, Northeastern University. He is a Co-PI of the NiuTrans MT project. His current research interests include machine translation and language modeling.

Jingbo Zhu received the Ph.D. degree in computer science from Northeastern University, Shenyang, China, in 1999. He has been with Northeastern University, since 1999. He is currently a Full Professor with the College of Computer Science and Engineering and is in charge of research activities within the Natural Language Processing Laboratory. He has published more than 180 papers and holds four US patents. His current research interests include syntactic parsing, machine translation, and machine learning for natural language processing.

Min Zhang received the Bachelor's and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1991 and 1997, respectively. In 2013, he joined Soochow University, Suzhou, China, where he is currently a Distinguished Professor. From 1997 to 1999, he was a Postdoctoral Research Fellow with the Korean Advanced Institute of Science and Technology, South Korea. He began his academic and industrial career as a Researcher at Lernout & Hauspie Asia Pacific (Singapore) in 1999. He joined Infotalk Technology (Singapore) as a Researcher in 2001 and became a Senior Research Manager in 2002. He joined the Institute for Infocomm Research (Singapore) in 2003. He has authored 150 papers in leading journals and conferences. His current research interests include machine translation, natural language processing, information extraction, large-scale text processing, intelligent computing, and machine learning. He is the Vice-President of COLIPS, a steering committee member of PACLIC, an executive member of AFNLP, and a member of ACL.