Translating with Bilingual Topic Knowledge for Neural Machine Translation

Xiangpeng Wei,[1,2] **Yue Hu,**[1,2] **Luxi Xing,**[1,2] **Yipeng Wang,**[1,2] **Li Gao**[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Platform & Content Group, Tencent, Beijing, China
{weixiangpeng, huyue, xingluxi, wangyipeng}@iie.ac.cn, leolgao@tencent.com

Abstract

The dominant neural machine translation (NMT) models that based on the encoder-decoder architecture have recently achieved the state-of-the-art performance. Traditionally, the NMT models only depend on the representations learned during training for mapping a source sentence into the target domain. However, the learned representations often suffer from implicit and inadequately informed properties. In this paper, we propose a novel bilingual topic enhanced NMT (BLT-NMT) model to improve translation performance by incorporating bilingual topic knowledge into NMT. Specifically, the bilingual topic knowledge is included into the hidden states of both encoder and decoder, as well as the attention mechanism. With this new setting, the proposed BLT-NMT has access to the background knowledge implied in bilingual topics which is beyond the sequential context, and enables the attention mechanism to attend to topic-level attentions for generating accurate target words during translation. Experimental results show that the proposed model consistently outperforms the traditional RNNsearch and the previous topic-informed NMT on Chinese-English and English-German translation tasks. We also introduce the bilingual topic knowledge into the newly emerged Transformer base model on English-German translation and achieve a notable improvement.

Introduction

Neural Machine Translation (NMT) (Sutskever, Vinyals, and V. Le 2014; Cho et al. 2014; Bahdanau, Cho, and Bengio 2015) is a novel approach to machine translation that has shown remarkable superiority over conventional statistical machine translation (SMT) across a variety of language pairs (Junczys-Dowmunt, Dwojak, and Hoang 2016), which directly models the entire translation process through training an encoder-decoder network in end-to-end style. The success of NMT depends on its capacity of using representation to bridge the source and target languages. However, this representation, a sequence of fixed-dimensional vectors learned from sequential context, suffers from implicit and inadequately informed properties.

Currently, many methods have been proposed to enrich the representations produced by NMT model, such as coverage mechanism (Tu et al. 2016), fertility constraint (Cohn et al. 2016), posterior regularization (Zhang et al. 2017a) and linguistic knowledge based methods (Sennrich and Haddow 2016; Eriguchi, Hashimoto, and Tsuruoka 2016; Chen et al. 2017a; 2017b; Wu, Zhou, and Zhang 2017; Wu et al. 2017). Although (Zhang et al. 2016) had proposed a topic-informed NMT model that can increase the likelihood of selecting words from the same topic or domain by conveying topic knowledge during translation, the topic-informed NMT suffers from the gap between the two topic distributions of source and target languages that are respectively constructed by two independent LDA models.

In this paper, we take bilingual topic information as prior knowledge and incorporate it into NMT for further improving translation performance. More concretely, (1) we use bilingual LDA to jointly learn the topic distribution of each source and target word by taking comparable documents from Wikipedia in two languages and mapping them into a shared topic space which represented by a group of universal topics. (2) We consider the topic distribution learned by bilingual LDA of each word as its topic embedding, and propose a novel bilingual topic enhanced NMT (BLT-NMT) model that takes topic embeddings of source and target words as additional inputs.

For encoding, we generate representations containing information both from literal words and from latent topics for each source sentence by introducing a joint-encoder network. During decoding, we convey target-side topic knowledge by developing a joint-decoder, and augment the existing attention mechanism with topic-level attentions by developing a joint-attention network. Through this way, each target word is generated according to both the literal relevance and the topical relevance with source words. Furthermore, we modify the softmax layer by adding a separate generation probability at each prediction timestep, which can bias the probability distribution over target vocabulary and contributes to generating quality translations.

We evaluate the proposed model on Chinese-English and English-German translation tasks. Experimental results show that the proposed model consistently outperforms the traditional RNNsearch and the previous topic-informed NMT. In addition, we also introduce the bilingual topic knowledge into the newly emerged Transformer on English-German translation and achieve a notable improvement.

Compared with the previous topic-informed NMT, the

Sample topics	Representative words
Software Topic	bǎnběn, yònghù, gōngnéng, ... Windows,users,version, ...
Music Topic	zhuānjí, yuèduì, gēqǔ, ... song,album,band,music, ...

Table 1: Sample of universal topics.

novelties of our method are in three folds:

- We use bilingual LDA to represent both the source and target languages into a shared topic space using a group of universal topics, as shown in table 1. By doing so, our BLT-NMT model can directly connect source and target words through specific topic dimensions.
- We develop a joint-attention network to model the relevance between source and target words at both word-level and topic-level, which makes words aligned more accurately.
- We introduce a biased-softmax mechanism to assign more generation probabilities to the target words that are relevant with the topics of the source sentence.

Neural Machine Translation

Neural Machine Translation is a neural network, which is implemented as an encoder-decoder framework with recurrent neural networks and attention mechanism (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015). We follow (Bahdanau, Cho, and Bengio 2015), and give a brief summarization here.

Given a source sentence $\mathbf{x} = x_1, x_2, ..., x_L$ and a target sentence $\mathbf{y} = y_1, y_2, ..., y_{L'}$, where L and L respectively indicates the length of $\mathbf{x}$ and $\mathbf{y}$. NMT models the translation probability as

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{L'} p(y_i|y_{<i}, \mathbf{x}; \Theta) \quad (1)$$

where Θ is the model parameters and $y_{<i} = y_1, y_2, ..., y_{i-1}$ is partial translation. The generation probability of y_i is

$$p(y_i|y_{<i}, \mathbf{x}) \propto \exp\{g(y_{i-1}, s_i, \mathbf{c}_i)\} \quad (2)$$

where $g(\cdot)$ is a non-linear activation function, y_{i-1} is the previous translated target word and

$$s_i = f(y_{i-1}, s_{i-1}, \mathbf{c}_i) \quad (3)$$

is the i-th hidden state of decoder, $f(\cdot)$ is a non-linear transformation. The attention $\mathbf{c}_i$ is the context vector that denotes the relevance with source words for generating y_i and is calculated by an attention model. As

$$\mathbf{c}_i = \sum_{j=1}^{L} \alpha_{ij} h_j \quad (4)$$

where α_{ij} is given by

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{m} exp(e_{ik})} \quad (5)$$
$$e_{ij} = a(s_{i-1}, h_j)$$

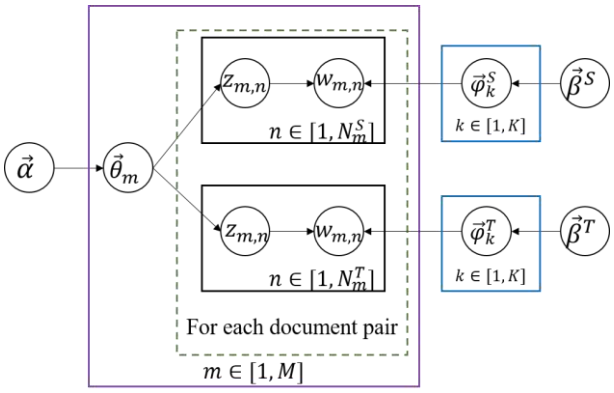

Figure 1: Graphic model of bilingual LDA. $w_{m,n}$ indicates the n-th word in m-th document, $z_{m,n}$ indicates the topic assigned to $w_{m,n}$, M indicates the number of documents and N_m indicates the length of the m-th document. The letter S denotes the source language, and the letter T denotes the target language. $\vec{\alpha}$, $\vec{\beta}^S$ and $\vec{\beta}^T$ are hyper-parameters of Dirichlet prior distributions.

where a is parametrized as a feedforward neural network, h_j is the j-th hidden state of encoder.

Given a set of training examples $\mathrm{D} = \{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{|\mathrm{D}|}$, the log-likelihood of the training data is

$$\mathcal{L}(\mathrm{D}; \Theta) = \sum_{s=1}^{|\mathrm{D}|} \log p(\mathbf{y}^{(s)}|\mathbf{x}^{(s)}; \Theta) \quad (6)$$

The whole model, consisting of the encoder, decoder and attention model, is then tuned end-to-end to maximize the log-likelihood.

Bilingual Topic Knowledge Acquisition

Instead of letting the NMT model rely solely on the implicit representation it learns during training, we improve its performance by augmenting it with bilingual topic knowledge. In this section, we describe our method on bilingual topic knowledge acquisition.

We use bilingual LDA (BL-LDA) (Ni et al. 2009) to mine bilingual topics by taking comparable documents from Wikipedia in two different languages and mapping them into a shared topic space. BL-LDA adapts Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) to model bilingual topics, and assumes that every two documents of a concept unit in Wikipedia, although in different languages, share identical topic distribution.

Referring to (Ni et al. 2009), we implement a BL-LDA model, as illustrated in Figure 1. In our experiments, bilingual document-aligned texts are used to represent the mixture of topics. We train the BL-LDA model using the Wikipedia comparable corpora, and estimate the parameters of the BL-LDA by Gibbs Sampling algorithm (Gregor 2005). When training is finished, we can obtain the topic

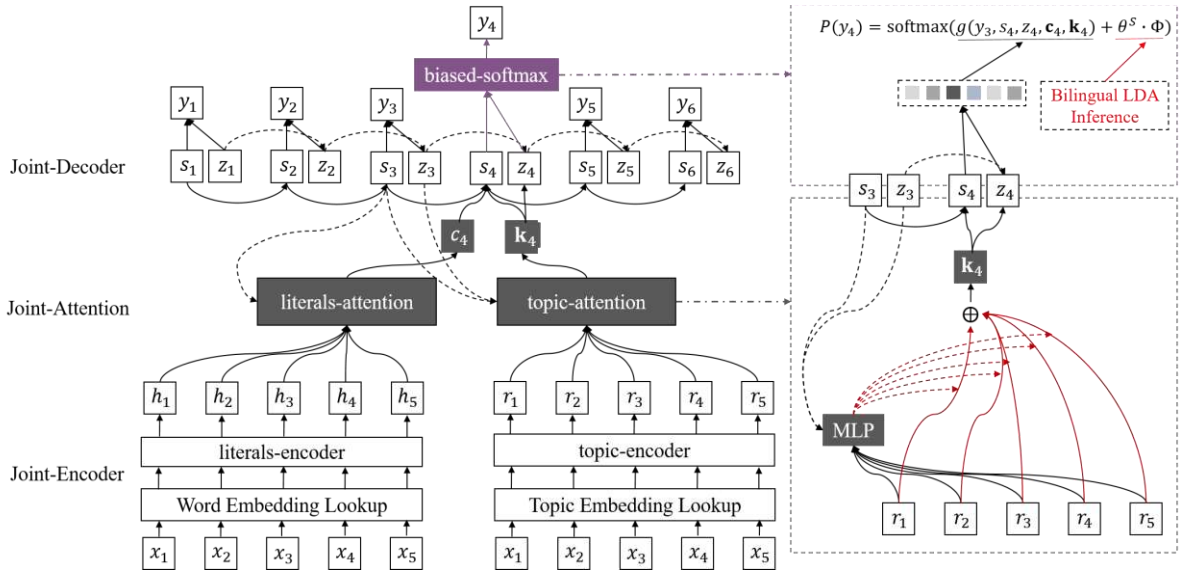

Figure 2: Graphic structure of bilingual topic enhanced NMT model.

distribution φ_t^S of the t-th source word w_t^S as

$$\varphi_t^S = [\phi_{1t}^S, ..., \phi_{kt}^S, ..., \phi_{\mathbf{K}t}^S]$$
$$\phi_{kt}^S = \frac{C_{kt}^S + \beta^S}{\sum_{k'=1}^{\mathbf{K}} (C_{k't}^S + \beta^S)} \quad (7)$$

and the topic distribution φ_t^T of the t-th target word w_t^T as

$$\varphi_t^T = [\phi_{1t}^T, ..., \phi_{kt}^T, ..., \phi_{\mathbf{K}t}^T]$$
$$\phi_{kt}^T = \frac{C_{kt}^T + \beta^T}{\sum_{k'=1}^{\mathbf{K}} (C_{k't}^T + \beta^T)} \quad (8)$$

where $\mathbf{K}$ is the number of the universal topics, C_{kt}^S and C_{kt}^T denote the number of times that w_t^S and w_t^T are assigned to the shared topic k, respectively. We take the topic distribution of each word as its topic embedding. Thus, the BL-LDA maps the two languages into a shared topic space, and we can build direct connections between all words of the two languages according to specific topic dimensions.

In inference phase, we can use the BL-LDA model to assign topics to a new document by Gibbs Sampling, and calculate the topic distribution of the new document as

$$\vartheta = [\theta_1, ..., \theta_k, ..., \theta_{\mathbf{K}}]$$
$$\theta_k = \frac{C_k + \alpha}{\sum_{k'=1}^{\mathbf{K}} (C_{k'} + \alpha)} \quad (9)$$

where C_k is the number of times that topic k is assigned to the new document.

BLT-NMT: Bilingual Topic enhanced Neural Machine Translation

We aim to enhance NMT through incorporating bilingual topic knowledge into neural machine translation. Figure 2 gives the graphical illustration of the proposed BLT-NMT model, which consists of three parts: **Joint-Encoder**, **Joint-Attention** and **Joint-Decoder**.

Joint-Encoder

In encoding phase, we convert the sequence of words into a sequence of word embeddings and another sequence of topic embeddings. The word embeddings are obtained by looking up the word embedding table that is randomly initialized and updated during training, and the topic embeddings of source words are pre-calculated according to Equation 7 that are kept fixed during translation process. We introduce a bidirectional GRU (Cho et al. 2014) as the literals-encoder that represents a sequence of word embeddings as a sequence of hidden vectors $\{h_1, h_2, ..., h_L\}$, where h_j is the concatenation of the outputs of the forward and backward GRUs. And we additionally develop a topic-encoder which maps a sequence of topic embeddings of the source sentence into a sequence of hidden topic-vectors $\{r_1, r_2, ..., r_L\}$. We also utilize a bidirectional GRU as topic-encoder, and the j-th hidden topic-vector r_j is calculated as

$$r_j = BiGRU(\varphi_j^S, r_{j-1}) \quad (10)$$

where φ_j^S is the topic distribution of the j-th source word. Thus, the joint-encoder generates representations containing information both from literal words and from latent topics.

Joint-Attention

We augment the existing attention mechanism with topic-level attentions by developing a joint-attention network to improve word alignment quality, which consists of a literals-attention module and a topic-attention module.

Literals-attention The literals-attention is used to summarize $\{h_j\}_{j=1}^L$ as the sequential context vector $\mathbf{c}_i$, which calculated according to Equations 4 and 5.

Topic-attention The topic-attention is used to synthetise the topical context vector $\mathbf{k}_i$ from $\{r_j\}_{j=1}^{L}$. More concretely, when generating the i-th target word, the topic-attention takes the hidden topic-vectors of the topic-encoder $\{r_j\}_{j=1}^{L}$, the previous hidden state of the literals-decoder s_{i-1}, and the previous hidden state of the topic-decoder z_{i-1} as inputs, and predicts topical relevance between the i-th target word and all words in source sentence. That is,

$$\tilde{\alpha}_{ij} = \frac{exp(\tilde{e}_{ij})}{\sum_{o=1}^{m} exp(\tilde{e}_{io})} \qquad (11)$$
$$\tilde{e}_{ij} = \tilde{a}(s_{i-1}, z_{i-1}, r_j)$$

where $\tilde{a}$ is parametrized as a feedforward neural network, and the topic context vector $\mathbf{k}_i$ is synthetised as:

$$\mathbf{k}_i = \sum_{j=1}^{L} \tilde{\alpha}_{ij} t_j \qquad (12)$$

Following (Xing et al. 2017), there is an extra input (i.e., s_{i-1}) in topic-attention. With this strategy, the topical relevance between target and source words can be calculated under the guidance of sequential context.

Joint-Decoder

In decoding phase, we convey target-side topic knowledge during translation by developing a joint-decoder. The joint-decoder consists of a GRU based literals-decoder and a GRU based topic-decoder, $\{s_i\}_{i=1}^{L}$ and $\{z_i\}_{i=1}^{L}$ are hidden states of the two decoders, respectively. At time step i, the literals-decoder is used to generate the target word y_i by considering both the literal relevance and the topical relevance with source words. Accordingly, the i-th hidden state of the literals-decoder is updated as:

$$s_i = f(y_{i-1}, s_{i-1}, \mathbf{c}_i, \mathbf{k}_i) \qquad (13)$$

where $\mathbf{c}_i$ is the sequential context vector, $\mathbf{k}_i$ is the topical context vector, both of them are synthetised by the joint-attention network. And the target-side topic knowledge is maintained as:

$$z_i = f'(\varphi^{T}_{i-1}, z_{i-1}, \mathbf{k}_i) \qquad (14)$$

where φ^{T}_{i-1} is the topic embedding of y_{i-1}, which calculated according to Equation 8 and is kept fixed during translation process. f denotes the topic-decoder unit.

In softmax layer, we assign more generation probabilities to the target words that are relevant with the topics of the source sentence through introducing a biased-softmax mechanism. At i-th time step, the generation probability $p(y_i)$ is calculated as

$$p(y_i) = \text{softmax}\{g(y_{i-1}, s_i, z_i, \mathbf{c}_i, \mathbf{k}_i) + \mathbf{b}\} \qquad (15)$$
$$\mathbf{b} = \vartheta^{S} \cdot \Phi$$

where $g(\cdot)$ is a non-linear activation function, $\vartheta^{S} \in \mathbb{R}^{\mathbf{K}}$ is the topic representation of the source sentence and is calculated by Equation 9 (i.e., the inference mode of bilingual LDA). $\Phi \in \mathbb{R}^{\mathbf{K} \times V_y}$ is the topic embedding matrix of the target language that established according to Equation 8, and the v-th column of which denotes the topic embedding of the v-th word in target vocabulary. Adding such an extra generation probability can bias the target word distribution, that is, target words that are relevant with the topics of the source sentence are more likely to be generated.

Experiments

We mainly evaluate the proposed model on Chinese-English and English-German translation tasks.

Setup

Datasets Bilingual topic modeling: We have built two document-aligned corpora from Wikipedia comparable corpora[1] for Chinese-English and English-German language pairs, respectively. For Chinese-English, the comparable corpus consist of 405574 document-pairs, with 338.4M English words and 156.8M Chinese words. And for English-German, the comparable corpus consist of 852363 document-pairs, with 544.4M English words and 405.0M German words.

Chinese-English: The parallel training data consists of 1.25M sentence pairs extracted from LDC corpora[2], with 27.9M Chinese words and 34.5M English words respectively. We choose NIST 2002 (NIST02) as development set, and the NIST 2003 (NIST03), NIST 2004 (NIST04), NIST 2005 (NIST05), NIST 2006 (NIST06) datasets as our test sets. The Stanford Chinese word segmenter (Tseng et al. 2005) is used to segment the Chinese training data. The English side of the corpora is tokenized.

English-German: As previous work (Luong, Pham, and Manning 2015; Jean et al. 2015; Wu et al. 2016), we use the same subset of WMT 2014 training corpus that contains 4.5M sentence pairs with 91M English words and 87M German words. The concatenation of *newstest2012* and *newstest2013* is used as the development set and *newstest2014* is used as the test set.

We apply the script *mteval-v11b.pl* to evaluate the Chinese-English translation and utilize the script *multi-belu.pl* for English-German translation. The metrics are exactly the same as in the previous literatures.

Training Details For all experiments of bilingual topic modeling, we set the hyper parameters as $\grave{\alpha} = 0.5/\mathbf{K}$, $\grave{\beta}^{S} = 0.1$ and $\grave{\beta}^{T} = 0.1$, where $\mathbf{K}$ is the number of universal topics ranging from 50 to 500, with 50 as the step size. For each value of $\mathbf{K}$, the model is estimated using 200 Gibbs Sampling iterations.

To efficiently train NMT models, we remove sentences longer than 50 words for Chinese-English and sentences longer than 100 words for English-German. Besides, we use both source and target vocabularies with 30K most frequent words for Chinese-English translation, and 32K sub-word tokens based on byte-pair encoding (Sennrich, Haddow, and

[1] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

[2] LDC2002E18, LDC2003E07, LDC2003E14, the Hansards portion of LDC2004T07, LDC2004T08, and LDC2005T06

Birch 2016) for English-German translation. All the out-of-vocabulary words are mapped to a special token UNK. Finally, such vocabularies contained 98.43% Chinese words and 99.40% English words of the Chinese-English corpus, and almost 100% English and German words of the English-German corpus. For all experiments, we set the following hyper-parameters: word embedding dimension as 512, hidden layer size as 1024, batch size as 80, gradient norm as 5.0, dropout rate as 0.3 and beam width as 10. Inspired by (Wu et al. 2016), we initialize all trainable parameters uniformly between [-0.04, 0.04]. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ and follow the same learning rate schedule in (Vaswani et al. 2017).

Models for Comparison We compare the proposed model against the following three models:

- **DL4MT** is an open source NMT toolkit[3].
- **RNNsearch** is a re-implementation of (Bahdanau, Cho, and Bengio 2015), and we assemble it with some advanced techniques, such as using the output of forward RNN as the input of backward RNN, training with dynamic learning rate and initializing all word embeddings with word2vec[4].
- **Topic-Informed NMT** is an in-house implementation of (Zhang et al. 2016) based on RNNsearch. For fair comparison, instead of the translation training data used in (Zhang et al. 2016), we use the additional Wikipedia comparable corpora[5] introduced in this paper to independently learn the topic distributions of the source and target languages.

Parameters & Speed RNNsearch and Topic-Informed NMT models have 79.9M and 80.4M parameters, respectively. By contrast, the parameter size of our proposed model is about 90.7M. We train our BLT-NMT on single NVIDIA Titan X GPU. For Chinese-English translation, each training step takes about 0.5 seconds, the model is trained about 150,000 steps (21 hours) and is saved at each 1,000 updates. For English-German translation, each training step takes about 0.8 seconds, the model is trained about 300,000 steps (2.8 days) and is saved at each 2,000 updates.

Effect of universal-topic number

We first investigate the impact of the number of universal topics on the development set for Chinese-English and English-German translation tasks. To this end, we gradually varied $\mathbf{K}$ from 50 to 500 with 50 as step size. As shown in Figure 3, we find that our model achieves the best performance when $\mathbf{K} = 100$ and $\mathbf{K} = 150$ for Chinese-English and English-German translations respectively. Therefore, we set $\mathbf{K} = 100$ for Chinese-English translation, and set $\mathbf{K} = 150$ for English-German translation. In addition, we find that the BLEU scores of both the two translation tasks

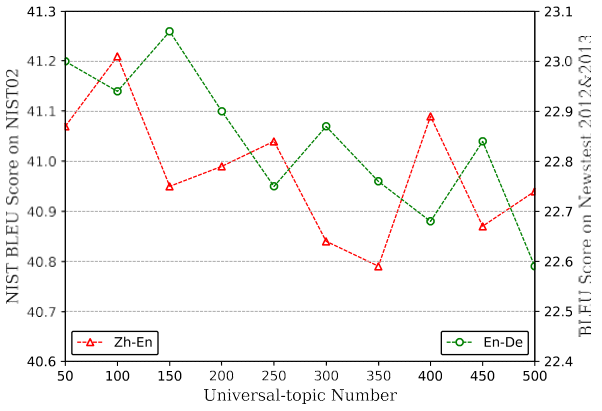

Figure 3: Experimental results on the development sets of various settings of universal topic number for Chinese-English (the left *y*-axis + the red line) and English-German (the right *y*-axis + the green line) translation tasks.

tend down with the number of universal topics increasing. This phenomenon had also emerged in (Zhang et al. 2016). One possible reason we conjecture is that the number of universal topics is too large for overfitting, and another is that the information is too fragmentary to represent a topic.

Results on Chinese-English

The experimental results on Chinese-English translation are depicted in Table 2. At first, compared to the DL4MT, our basic RNNsearch achieves a significant improvement by +4.39 BLEU points. Although our RNNsearch is a basic attentional NMT model, we assemble it with some advanced techniques, such as using the output of forward RNN as the input of backward RNN, training with dynamic learning rate and initializing all word embeddings with word2vec. Whatever, we give the comparison between our basic RNNsearch and DL4MT is to prove that our baseline is strong enough, and all improvements over baseline model are reliable.

Clearly BLT-NMT leads to a remarkable improvement over its competitors. Compared to RNNsearch and Topic-Informed NMT, BLT-NMT is +3.56 and +1.68 BLEU scores higher respectively, showing the modeling power gained from the bilingual topic knowledge. The reason is that our proposed model has access to the background knowledge which is beyond the sequential context.

To further prove the effectiveness of BLT-NMT, we also make a comparison with some dominant individual models such as COVERAGE, NMT_{IA}, MemDec, $\text{NMT}_{H\ Distortion}$ and DeepLAU. Our best single model outperforms both a coverage model (COVERAGE) as well as a memory enhanced NMT model (MemeDec) by +4.60 and +3.13 BLEU points on the same data set respectively. Even compared with DeepLAU, the BLT-NMT also achieves a notable improvement by +1.13 BLEU points.

[3]https://github.com/nyu-dl/dl4mt-tutorial/

[4]https://code.google.com/archive/p/word2vec/

[5]The Wikipedia comparable corpora consist of over 41 million aligned articles for 253 language pairs, which can be generalized to many translation tasks especially to low-resource language pairs.

MODEL	NIST03	NIST04	NIST05	NIST06	Average
COVERAGE (Tu et al. 2016)	34.49	38.34	34.91	34.25	35.50
NMT$_{IA}$ (Meng et al. 2016)	35.69	39.24	35.74	35.10	36.44
MemDec (Wang et al. 2016)	36.16	39.81	35.91	35.98	36.97
NMT$_{H-Distortion}$ (Zhang et al. 2017b)	37.93	40.40	36.81	35.77	37.73
DeepLAU (Wang et al. 2017)	39.35	41.15	38.07	37.29	38.97
DL4MT	32.37	34.22	31.03	30.97	32.15
RNNsearch	36.25	39.69	35.51	34.70	36.54
Topic-Informed NMT	38.82	40.48	37.75	36.64	38.42
BLT-NMT	**40.41**$^{*++}$	**41.15**$^{**++}$	**39.88**$^{**++}$	**38.97**$^{**+}$	**40.10**

Table 2: Evaluation of the Chinese-English translation task using case-insensitive BLEU scores. We also displayed the experimental results of the five models reported in (Wang et al. 2017; Zhang et al. 2017b). COVERAGE is a basic NMT model with a coverage model. NMT$_{IA}$ exploits a interactive attention mechanism to keep track of interactive history in decoding. MemDec improves translation quality with external memory. NMT$_{H\ Distortion}$ incorporates word reordering knowledge into NMT. DeepLAU reduces the gradient propagation length inside the recurrent unit of RNN-based NMT. "$*$" significantly better than RNNsearch ($p < 0.05$); "$**$" significantly better than RNNsearch ($p < 0.01$); "+" significantly better than Topic-Informed NMT ($p < 0.05$); "++" significantly better than Topic-Informed NMT ($p < 0.01$).

MODEL	Voc.	BLEU
(Jean et al. 2015)	500K	19.40
(Luong, Pham, and Manning 2015)	50K	20.90
(Zhou et al. 2016)	160K	20.60
GNMT WMP-32K	40K	24.61
ConvS2S		25.16
DL4MT	32K	20.54
RNNsearch	32K	23.80
Topic-Informed NMT	32K	24.56
BLT-NMT	32K	**25.68**
Transformer (base)	–	27.30
Transformer (base) + BLT	–	**27.93**

Table 3: Evaluation of the WMT English-German translation using case-insensitive BLEU scores. We directly cited the experimental results of various existing state-of-the-art models, such as GNMT (Wu et al. 2016), ConvS2S (Gehring et al. 2017) and Transformer (Vaswani et al. 2017). "BLT" in the last row indicates "bilingual topic knowledge". All the improvements are statistically significant ($p < 0.05$).

Results on English-German

To enhance the persuasion of our model, we also report some experimental results of various existing state-of-the-art models on the same data set, including Deep RNN models (Jean et al. 2015; Luong, Pham, and Manning 2015; Zhou et al. 2016; Wu et al. 2016), Deep CNN model (Gehring et al. 2017) and Deep Attention model (Vaswani et al. 2017). For fair comparison, we just list the single model results reported in their papers.

Table 3 presents the results on WMT English-German translation. Our BLT-NMT still significantly outperforms RNNsearch and Topic-Informed NMT by +1.88 and +1.12 BLEU points, respectively. Compared to other RNN based and CNN based models, our BLT-NMT is very competitive, although it is a shallow model.

MODEL	Softmax	Average	Q
BLT-NMT	biased	40.10	-
BLT-NMT	vanilla	39.65	-0.45
RM-TEnc	biased	39.33	-0.77
RM-TAttn	biased	39.18	-0.92
RM-TEnc-TAttn	biased	39.04	-1.06
RM-BLT	vanilla	36.83	-3.27

Table 4: Averaged case-insensitive BLEU scores on Chinese-English translation for BLT-NMT with different settings. ("biased" indicates our biased-softmax mechanism, and "vanilla" indicates standard softmax operation. The "Q" column presents the drop of test compared to BLT-NMT.)

We also introduce the bilingual topic knowledge into the newly emerged state-of-the-art Transformer (Vaswani et al. 2017). More concretely, we just concatenate the word embedding and the topic embedding of each (source or target) word, and then feed this concatenation into the standard Transformer base model, the other settings are the same as described in (Vaswani et al. 2017). With this strategy, we achieve a notable improvement of +0.63 BLEU points over the standard Transformer base model. We believe that the bilingual topic knowledge introduced in this paper is helpful for NMT and can be applied to other architectures easily.

Analysis

We further look into the proposed BLT-NMT model and study the main factors that influence our results on Chinese-English translation task.

Ablation Study To understand the importance of different components of the proposed model, we develop five variants of BLT-NMT. (1) BLT-NMT with vanilla softmax, which means we just remove the item $\mathbf{b}$ in Equation 15; (2) RM-TEnc: we remove the topic-encoder, which means that the topic-attention directly summarizes the sequence of

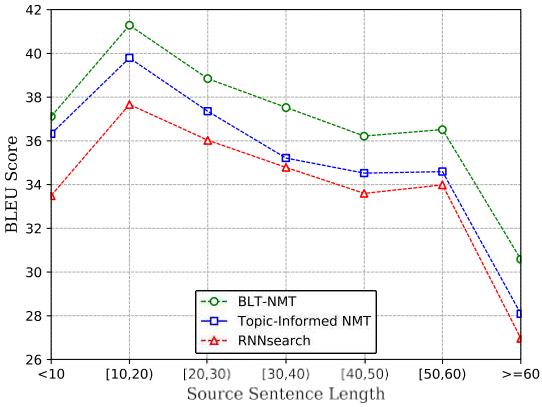

Figure 4: BLEU scores on different translation groups divided according to source sentence length.

topic embeddings of source sentence as context topic vector; (3) RM-TAttn: we remove the topic-attention module, that is, we concatenate the outputs of the literals-encoder and the topic-encoder, and then feed this concatenation into literals-attention; (4) RM-TEnc-TAttn: we remove both topic-encoder and topic-attention, which is structurally the same as (Zhang et al. 2016); (5) RM-BLT: for excluding the effect of additional model parameters, we just use the same architecture as illustrated in Figure 2 but feed the "topic" branch of this architecture with word embeddings rather than with bilingual topic embeddings.

Results are reported in Table 4. Firstly, by comparing row 1 to row 2, we can see that the biased-softmax demonstrates a remarkable improvement, as it enables the BLT-NMT assigns more generation probabilities to the target words that are relevant with the topics of the source sentence and consequently contributes to generating quality translation. Then, removing topic-encoder, topic-attention and both lead to 0.77, 0.92 and 1.06 BLEU drops respectively. We can conclude that both the topic-encoder and the topic-attention are essential for BLT-NMT. Finally, when we use the same model architecture as BLT-NMT but don't actually provide any topic information from the bilingual LDA, the BLEU score dramatically decrease by 3.27. Therefore, we can conclude that the bilingual topics provide useful background knowledge for improving translation performance.

Source Sentence Lengths We carry out a more detailed comparison between RNNsearch, Topic-Informed NMT and BLT-NMT, suggests the superior performance of our model. In particular, we plot BLEU scores with respect to the length of source sentences in Figure 4. We observe that our model achieves the best performance in all groups. The improvements become larger for sentences longer than 40 words, and this provides some evidence for the importance of the bilingual topic knowledge for long sentences. Intuitively, it makes sense that 1) the bilingual topic knowledge provides "global" relevance for any two words regardless of their distance in input sentence, thus is helpful in capturing long-range dependencies and 2) the proposed model can avoid prematurely producing EOS (end of sentence) symbol through attending to topic-level attentions, as the EOS symbol not favours to any topic.

Related work

Many studies have focused on using explicit prior knowledge to help learn sentence representations for NMT, such as (Tu et al. 2016; Cohn et al. 2016; Sennrich and Haddow 2016; Eriguchi, Hashimoto, and Tsuruoka 2016; Zhang et al. 2017a; Chen et al. 2017a; Wu et al. 2017; Chen et al. 2017b; Wu, Zhou, and Zhang 2017). (Tu et al. 2016) incorporate coverage mechanism to improve the adequacy of translation. (Cohn et al. 2016) add agreement constraints to training objectives and improve translation performance at both directions synchronously. (Zhang et al. 2017a) propose to use posterior regularization to provide a general framework for integrating prior knowledge into NMT. (Sennrich and Haddow 2016) incorporate linguistic features to improve the NMT performance by appending feature vectors to word embeddings. (Eriguchi, Hashimoto, and Tsuruoka 2016) and (Chen et al. 2017a) respectively introduce the source-side syntactic trees into NMT by developing a bottom-up tree encoder and a bidirectional tree encoder. (Wu, Zhou, and Zhang 2017) and (Chen et al. 2017b) enhance the NMT by enriching each encoder state with global source dependency structure. (Wu et al. 2017) jointly construct the target word sequence and its dependency structure to facilitate word generations. Differs from these previous works, in this paper we focus on improving NMT with bilingual topic knowledge.

Although topic modeling has shown its effectiveness in statistical machine translation (SMT) models (Chiang, Deneefe, and Pust 2011; Eidelman, Boyd-Graber, and Resnik 2012; Hasler, Haddow, and Koehn 2014; Hasler et al. 2014), most proposed NMT models (a notable exception being that of (Zhang et al. 2016)) do not explicitly exploit topic knowledge during translation. (Zhang et al. 2016) propose a topic-informed NMT model that makes use of source-side and target-side topics, which are separately learned by two independent LDA models from translation training data and consequently lack direct connections between source and target words under the same topic.

Conclusion and Future Work

In this paper, we propose a novel encoder-decoder NMT model that makes use of bilingual topic knowledge to improve translation performance. The topics from the bilingual LDA provide useful background knowledge which enriches the representations produced by the encoder and decoder, and supervises the attention model to select more accurate words for translation. Experimental results on Chinese-English and English-German translations show that the proposed BLT-NMT model significantly outperforms the traditional RNNsearch and the previous topic-informed NMT.

For future work, we will focus on applying our method to Transformer model and validate the model on more language pairs. Second, we will study neural machine translation using the Wikipedia comparable corpora with other kinds of information, such as word embeddings trained on the external comparable documents.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Key Research and Development Program of China (No. 2017YFB0803003). Yue Hu is the corresponding author.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 993–1022.

Chen, H.; Huang, S.; Chiang, D.; and Chen, J. 2017a. Improved neural machine translation with a syntax–aware encoder and decoder. In *ACL 2017*, 1936–1945.

Chen, K.; Wang, R.; Utiyama, M.; Liu, L.; Tamura, A.; Sumita, E.; and Zhao, T. 2017b. Neural machine translation with source dependency representation. In *EMNLP 2017*, 2846–2852.

Chiang, D.; Deneefe, S.; and Pust, M. 2011. Two easy improvements to lexical weighting. In *ACL 2011*, 455–460.

Cho, K.; Merrienboer, B. V.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *arXiv:1409.1259*.

Cohn, T.; Cong, D. V. H.; Vymolova, E.; Yao, K.; Dyer, C.; and Haffari, G. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *NAACL-HLT 2016*, 876–885.

Eidelman, V.; Boyd-Graber, J.; and Resnik, P. 2012. Topic models for dynamic translation model adaptation. In *ACL 2012*, 115–119.

Eriguchi, A.; Hashimoto, K.; and Tsuruoka, Y. 2016. Tree–to–sequence attentional neural machine translation. In *ACL 2016*, 823–833.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *arXiv:1705.03122*.

Gregor, H. 2005. Parameter estimation for text analysis. In *Technical report*.

Hasler, E.; Blunsom, P.; Koehn, P.; and Haddow, B. 2014. Dynamic topic adaptation for phrase-based mt. In *EACL 2014*, 328–337.

Hasler, E.; Haddow, B.; and Koehn, P. 2014. Dynamic topic adaptation for smt using distributional profiles. In *Workshop on Statistical Machine Translation 2014*, 445–456.

Jean, S.; Cho, K.; Memisevic, R.; and Bengio, Y. 2015. On using very large target vocabulary for neural machine translation. In *ACL 2015*, 1–10.

Junczys-Dowmunt, M.; Dwojak, T.; and Hoang, H. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *arXiv: 1610.01108*.

Luong, M. T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention–based neural machine translation. In *EMNLP 2015*, 1412–1421.

Meng, F.; Lu, Z.; Li, H.; and Liu, Q. 2016. Interactive attention for neural machine translation. In *COLING 2016*, 2174–2185.

Ni, X.; Sun, J. T.; Hu, J.; and Chen, Z. 2009. Mining multilingual topics from wikipedia. In *WWW 2009*, 1155–1156.

Sennrich, R., and Haddow, B. 2016. Linguistic input features improve neural machine translation. In *Machine Translation 2016*, 83–91.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *ACL 2016*, 1715–1725.

Sutskever, I.; Vinyals, O.; and V. Le, Q. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014*, 3104–3112.

Tseng, H.; Chang, P.; Andrew, G.; Jurafsky, D.; and Manning, C. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *SIGHAN 2005*, 168–171.

Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; and Li, H. 2016. Modeling coverage for neural machine translation. In *ACL 2016*, 78–85.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *arXiv:1706.03762*.

Wang, M.; Lu, Z.; Li, H.; and Liu, Q. 2016. Memory–enhanced decoder for neural machine translation. In *EMNLP 2016*, 278–286.

Wang, M.; Lu, Z.; Zhou, J.; and Liu, Q. 2017. Deep neural machine translation with linear associative unit. In *ACL 2017*, 136–145.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; and Macherey, K. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *arXiv:1609.08144*.

Wu, S.; Zhang, D.; Yang, N.; Li, M.; and Zhou, M. 2017. Sequence–to–dependency neural machine translation. In *ACL 2017*, 698–707.

Wu, S.; Zhou, M.; and Zhang, D. 2017. Improved neural machine translation with source syntax. In *IJCAI 2017*, 4179–4185.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. Y. 2017. Topic aware neural response generation. In *AAAI 2017*.

Zhang, J.; Li, L.; Way, A.; and Liu, Q. 2016. Topic–informed neural machine translation. In *COLING 2016*, 1807–1817.

Zhang, J.; Liu, Y.; Luan, H.; Xu, J.; and Sun, M. 2017a. Prior knowledge integration for neural machine translation using posterior regularization. In *ACL 2017*, 1514–1523.

Zhang, J.; Wang, M.; Liu, Q.; and Zhou, J. 2017b. Incorporating word reordering knowledge into attention–based neural machine translation. In *ACL 2017*, 1524–1534.

Zhou, J.; Cao, Y.; Wang, X.; Li, P.; and Xu, W. 2016. Deep recurrent models with fast–forward connections for neural machine translation. In *TACL 2016*, 371–383.