Learning Topic Representation for SMT with Neural Networks*

Lei Cui[1], **Dongdong Zhang**[2], **Shujie Liu**[2], **Qiming Chen**[3], **Mu Li**[2], **Ming Zhou**[2], and **Muyun Yang**[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, P.R. China
leicui@hit.edu.cn, ymy@mtlab.hit.edu.cn
[2]Microsoft Research, Beijing, P.R. China
{dozhang,shujliu,muli,mingzhou}@microsoft.com
[3]Shanghai Jiao Tong University, Shanghai, P.R. China
simoncqm@gmail.com

Abstract

Statistical Machine Translation (SMT) usually utilizes contextual information to disambiguate translation candidates. However, it is often limited to contexts within sentence boundaries, hence broader topical information cannot be leveraged. In this paper, we propose a novel approach to learning topic representation for parallel data using a neural network architecture, where abundant topical contexts are embedded via topic relevant monolingual data. By associating each translation rule with the topic representation, topic relevant rules are selected according to the distributional similarity with the source text during SMT decoding. Experimental results show that our method significantly improves translation accuracy in the NIST Chinese-to-English translation task compared to a state-of-the-art baseline.

1 Introduction

Making translation decisions is a difficult task in many Statistical Machine Translation (SMT) systems. Current translation modeling approaches usually use context dependent information to disambiguate translation candidates. For example, translation sense disambiguation approaches (Carpuat and Wu, 2005; Carpuat and Wu, 2007) are proposed for phrase-based SMT systems. Meanwhile, for hierarchical phrase-based or syntax-based SMT systems, there is also much work involving rich contexts to guide rule selection (He et al., 2008; Liu et al., 2008; Marton and Resnik, 2008; Xiong et al., 2009). Although these methods are effective and proven successful in many SMT systems, they only leverage within-sentence contexts which are insufficient in exploring broader information. For example, the word *driver* often means "the operator of a motor vehicle" in common texts. But in the sentence "Finally, we write the user response to the buffer, i.e., pass it to our driver", we understand that *driver* means "computer program". In this case, people understand the meaning because of the IT topical context which goes beyond sentence-level analysis and requires more relevant knowledge. Therefore, it is important to leverage topic information to learn smarter translation models and achieve better translation performance.

Topic modeling is a useful mechanism for discovering and characterizing various semantic concepts embedded in a collection of documents. Attempts on topic-based translation modeling include topic-specific lexicon translation models (Zhao and Xing, 2006; Zhao and Xing, 2007), topic similarity models for synchronous rules (Xiao et al., 2012), and document-level translation with topic coherence (Xiong and Zhang, 2013). In addition, topic-based approaches have been used in domain adaptation for SMT (Tam et al., 2007; Su et al., 2012), where they view different topics as different domains. One typical property of these approaches in common is that they only utilize parallel data where document boundaries are explicitly given. In this way, the topic of a sentence can be inferred with document-level information using off-the-shelf topic modeling toolkits such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Hidden Topic Markov Model (HTMM) (Gruber et al., 2007). Most of them also assume that the input must be in document level. However, this situation does not always happen since there is considerable amount of parallel data which does not have document boundaries. In addition, contemporary SMT systems often works on sentence level rather than document level due to the efficiency. Although we can easily apply LDA at the

*This work was done while the first and fourth authors were visiting Microsoft Research.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 133–143,
Baltimore, Maryland, USA, June 23-25 2014.

sentence level, it is quite difficult to infer the topic accurately with only a few words in the sentence. This makes previous approaches inefficient when applied them in real-world commercial SMT systems. Therefore, we need to devise a systematical approach to enriching the sentence and inferring its topic more accurately.

In this paper, we propose a novel approach to learning topic representations for sentences. Since the information within the sentence is insufficient for topic modeling, we first enrich sentence contexts via Information Retrieval (IR) methods using content words in the sentence as queries, so that topic-related monolingual documents can be collected. These topic-related documents are utilized to learn a specific topic representation for each sentence using a neural network based approach. Neural network is an effective technique for learning different levels of data representations. The levels inferred from neural network correspond to distinct levels of concepts, where high-level representations are obtained from low-level bag-of-words input. It is able to detect correlations among any subset of input features through non-linear transformations, which demonstrates the superiority of eliminating the effect of noisy words which are irrelevant to the topic. Our problem fits well into the neural network framework and we expect that it can further improve inferring the topic representations for sentences.

To incorporate topic representations as translation knowledge into SMT, our neural network based approach directly optimizes similarities between the source language and target language in a compact topic space. This underlying topic space is learned from sentence-level parallel data in order to share topic information across the source and target languages as much as possible. Additionally, our model can be discriminatively trained with a large number of training instances, without expensive sampling methods such as in LDA or HTMM, thus it is more practicable and scalable. Finally, we associate the learned representation to each bilingual translation rule. Topic-related rules are selected according to distributional similarity with the source text, which helps hypotheses generation in SMT decoding. We integrate topic similarity features in the log-linear model and evaluate the performance on the NIST Chinese-to-English translation task. Experimental results demonstrate that our model significantly improves translation accuracy over a state-of-the-art baseline.

2 Background: Deep Learning

Deep learning is an active topic in recent years which has triumphed in many machine learning research areas. This technique began raising public awareness in the mid-2000s after researchers showed how a multi-layer feed-forward neural network can be effectively trained. The training procedure often involves two phases: a layer-wise unsupervised pre-training phase and a supervised fine-tuning phase. For pre-training, Restricted Boltzmann Machine (RBM) (Hinton et al., 2006), auto-encoding (Bengio et al., 2006) and sparse coding (Lee et al., 2006) are most frequently used. Unsupervised pre-training trains the network one layer at a time and helps guide the parameters of the layer towards better regions in parameter space (Bengio, 2009). Followed by fine-tuning in this parameter region, deep learning is able to achieve state-of-the-art performance in various research areas, including breakthrough results on the ImageNet dataset for objective recognition (Krizhevsky et al., 2012), significant error reduction in speech recognition (Dahl et al., 2012), etc.

Deep learning has also been successfully applied in a variety of NLP tasks such as part-of-speech tagging, chunking, named entity recognition, semantic role labeling (Collobert et al., 2011), parsing (Socher et al., 2011a), sentiment analysis (Socher et al., 2011b), etc. Most NLP research converts a high-dimensional and sparse binary representation into a low-dimensional and real-valued representation. This low-dimensional representation is usually learned from huge amount of monolingual texts in the pre-training phase, and then fine-tuned towards task-specific criterion. Inspired by previous successful research, we first learn sentence representations using topic-related monolingual texts in the pre-training phase, and then optimize the bilingual similarity by leveraging sentence-level parallel data in the fine-tuning phase.

3 Topic Similarity Model with Neural Network

In this section, we explain our neural network based topic similarity model in detail, as well as how to incorporate the topic similarity features into SMT decoding procedure. Figure 1 sketches the high-level overview which illustrates how to

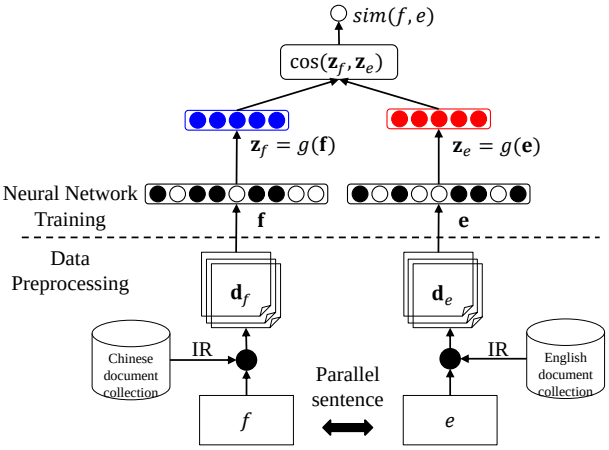


Figure 1: Overview of neural network based topic similarity model.

learn topic representations using sentence-level parallel data. Given a parallel sentence pair $\langle f, e \rangle$, the first step is to treat f and e as queries, and use IR methods to retrieve relevant documents to enrich contextual information for them. Specifically, the ranking model we used is a Vector Space Model (VSM), where the query and document are converted into tf-idf weighted vectors. The most relevant N documents $\mathbf{d}_f$ and $\mathbf{d}_e$ are retrieved and converted to a high-dimensional, bag-of-words input $\mathbf{f}$ and $\mathbf{e}$ for the representation learning[1].

There are two phases in our neural network training process: pre-training and fine-tuning. In the pre-training phase (Section 3.1), we build two neural networks with the same structure but different parameters to learn a low-dimensional representation for sentences in two different languages. Then, in the fine-tuning phase (Section 3.2), our model directly optimizes the similarity of two low-dimensional representations, so that it highly correlates to SMT decoding. Finally, the learned representation is used to calculate similarities which are integrated as features in SMT decoding procedure (Section 3.3).

3.1 Pre-training using denoising auto-encoder

In the pre-training phase, we leverage neural network structures to transform high-dimensional sparse vectors to low-dimensional dense vectors. The topic similarity is calculated on top of the learned dense vectors. This dense representation should preserve the information from the bag-of-words input, meanwhile alleviate data sparse problem. Therefore, we use a specially designed mechanism called auto-encoder to solve this problem. Auto-encoder (Bengio et al., 2006) is one of the basic building blocks of deep learning. Assuming that the input is a n-of-V binary vector $\mathbf{x}$ representing the bag-of-words (V is the vocabulary size), an auto-encoder consists of an encoding process $g(\mathbf{x})$ and a decoding process $h(g(\mathbf{x}))$. The objective of the auto-encoder is to minimize the reconstruction error $\mathcal{L}(h(g(\mathbf{x})), \mathbf{x})$. Our goal is to learn a low-dimensional vector which can preserve information from the original n-of-V vector.

One problem with auto-encoder is that it treats all words in the same way, making no distinguishment between function words and content words. The representation learned by auto-encoders tends to be influenced by the function words, thereby it is not robust. To alleviate this problem, Vincent et al. (2008) proposed the Denoising Auto-Encoder (DAE), which aims to reconstruct a clean, "repaired" input from a corrupted, partially destroyed vector. This is done by corrupting the initial input $\mathbf{x}$ to get a partially destroyed version $\tilde{\mathbf{x}}$. DAE is capable of capturing the global structure of the input while ignoring the noise. In our task, for each sentence, we treat the retrieved N relevant documents as a single large document and convert it to a bag-of-words vector $\mathbf{x}$ in Figure 2. With DAE, the input $\mathbf{x}$ is manually corrupted by applying masking noise (randomly mask 1 to 0) and getting $\tilde{\mathbf{x}}$. Denoising training is considered as "filling in the blanks" (Vincent et al., 2010), which means the masking components can be recovered from the non-corrupted components. For example, in IT related texts, if the word *driver* is masked, it should be predicted through hidden units in neural networks by active signals such as "buffer", "user response", etc.

In our case, the encoding process transforms the corrupted input $\tilde{\mathbf{x}}$ into $g(\tilde{\mathbf{x}})$ with two layers: a linear layer connected with a non-linear layer. Assuming that the dimension of the $g(\tilde{\mathbf{x}})$ is L, the linear layer forms a $L \times V$ matrix W which projects the n-of-V vector to a L-dimensional hidden layer. After the bag-of-words input has been transformed, they are fed into a subsequent layer to model the highly non-linear relations among words:

$$\mathbf{z} = f(W\tilde{\mathbf{x}} + \mathbf{b}) \quad (1)$$

where $\mathbf{z}$ is the output of the non-linear layer, $\mathbf{b}$ is a

[1]We use $\mathbf{f}$ and $\mathbf{e}$ to denote the n-of-V vector converted from the retrieved documents.

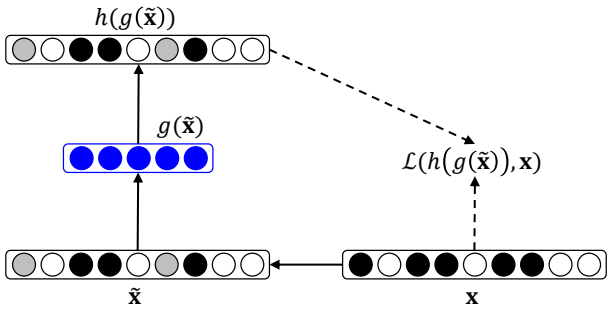


Figure 2: Denoising auto-encoder with a bag-of-words input.

L-length bias vector. $f(\cdot)$ is a non-linear function, where common choices include sigmoid function, hyperbolic function, "hard" hyperbolic function, rectifier function, etc. In this work, we use the rectifier function as our non-linear function due to its efficiency and better performance (Glorot et al., 2011):

$$rec(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The decoding process consists of a linear layer and a non-linear layer with similar network structures, but different parameters. It transforms the L-dimensional vector $g(\tilde{\mathbf{x}})$ to a V-dimensional vector $h(g(\tilde{\mathbf{x}}))$. To minimize reconstruction error with respect to $\tilde{\mathbf{x}}$, we define the loss function as the L2-norm of the difference between the uncorrupted input and reconstructed input:

$$\mathcal{L}(h(g(\tilde{\mathbf{x}})), \mathbf{x}) = \|h(g(\tilde{\mathbf{x}})) - \mathbf{x}\|_2 \quad (3)$$

Multi-layer neural networks are trained with the standard back-propagation algorithm (Rumelhart et al., 1988). The gradient of the loss function is calculated and back-propagated to the previous layer to update its parameters. Training neural networks involves many factors such as the learning rate and the length of hidden layers. We will discuss the optimization of these parameters in Section 4.

3.2 Fine-tuning with parallel data

In the fine-tuning phase, we stack another layer on top of the two low-dimensional vectors to maximize the similarity between source and target languages. The similarity scores are integrated into the standard log-linear model for making translation decisions. Since the vectors from DAE are trained using information from monolingual training data independently, these vectors may be inadequate to measure bilingual topic similarity due to their different topic spaces. Therefore, in this stage, parallel sentence pairs are used to help connecting the vectors from different languages because they express the same topic. In fact, the objective of fine-tuning is to discover a latent topic space which is shared by both languages as much as possible. This shared topic space is particularly useful when the SMT decoder tries to match the source texts and translation candidates in the target language.

Given a parallel sentence pair $\langle f, e \rangle$, the DAE learns representations for f and e respectively, as $\mathbf{z}_f = g(\mathbf{f})$ and $\mathbf{z}_e = g(\mathbf{e})$ in Figure 1. We then take two vectors as the input to calculate their similarity. Consequently, the whole neural network can be fine-tuned towards the supervised criteria with the help of parallel data. The similarity score of the representation pair $\langle \mathbf{z}_f, \mathbf{z}_e \rangle$ is defined as the cosine similarity of the two vectors:

$$\begin{aligned} sim(f, e) &= \cos(\mathbf{z}_f, \mathbf{z}_e) \\ &= \frac{\mathbf{z}_f \cdot \mathbf{z}_e}{\|\mathbf{z}_f\| \|\mathbf{z}_e\|} \end{aligned} \quad (4)$$

Since a parallel sentence pair should have the same topic, our goal is to maximize the similarity score between the source sentence and target sentence. Inspired by the contrastive estimation method (Smith and Eisner, 2005), for each parallel sentence pair $\langle f, e \rangle$ as a positive instance, we select another sentence pair $\langle f', e' \rangle$ from the training data and treat $\langle f, e' \rangle$ as a negative instance. To make the similarity of the positive instance larger than the negative instance by some margin η, we utilize the following pairwise ranking loss:

$$\mathcal{L}(f, e) = \max\{0, \eta - sim(f, e) + sim(f, e')\} \quad (5)$$

where $\eta = \frac{1}{2} - sim(f, f')$. The rationale behind this criterion is, the smaller $sim(f, f')$ is, the more we should penalize negative instances.

To effectively train the model in this task, negative instances must be selected carefully. Since different sentences may have very similar topic distributions, we select negative instances that are dissimilar with the positive instances based on the following criteria:

1. For each positive instance $\langle f, e \rangle$, we select e' which contains at least 30% different content words from e.

2. If we cannot find such e', remove $\langle f, e\rangle$ from the training instances for network learning.

The model minimizes the pairwise ranking loss across all training instances:

$$\mathcal{L} = \sum_{\langle f,e\rangle} \mathcal{L}(f, e) \tag{6}$$

We used standard back-propagation algorithm to further fine-tune the neural network parameters W and $\mathbf{b}$ in Equation (1). The learned neural networks are used to obtain sentence topic representations, which will be further leveraged to infer topic representations of bilingual translation rules.

3.3 Integration into SMT decoding

We incorporate the learned topic similarity scores into the standard log-linear framework for SMT. When a synchronous rule $\langle \alpha, \gamma\rangle$ is extracted from a sentence pair $\langle f, e\rangle$, a triple instance $\mathcal{I} = (\langle \alpha, \gamma\rangle, \langle f, e\rangle, c)$ is collected for inferring the topic representation of $\langle \alpha, \gamma\rangle$, where c is the count of rule occurrence. Following (Chiang, 2007), we give a count of one for each phrase pair occurrence and a fractional count for each hierarchical phrase pair. The topic representation of $\langle \alpha, \gamma\rangle$ is then calculated as the weighted average:

$$\mathbf{z}_\alpha = \frac{\sum_{(\langle \alpha,\gamma\rangle,\langle f,e\rangle,c)\in\mathcal{T}} \{c \times \mathbf{z}_f\}}{\sum_{(\langle \alpha,\gamma\rangle,\langle f,e\rangle,c)\in\mathcal{T}} \{c\}} \tag{7}$$

$$\mathbf{z}_\gamma = \frac{\sum_{(\langle \alpha,\gamma\rangle,\langle f,e\rangle,c)\in\mathcal{T}} \{c \times \mathbf{z}_e\}}{\sum_{(\langle \alpha,\gamma\rangle,\langle f,e\rangle,c)\in\mathcal{T}} \{c\}} \tag{8}$$

where $\mathcal{T}$ denotes all instances for the rule $\langle \alpha, \gamma\rangle$, $\mathbf{z}_\alpha$ and $\mathbf{z}_\gamma$ are the source-side and target-side topic vectors respectively.

By measuring the similarity between the source texts and bilingual translation rules, the SMT decoder is able to encourage topic relevant translation candidates and penalize topic irrelevant candidates. Therefore, it helps to train a smarter translation model with the embedded topic information. Given a source sentence s to be translated, we define the similarity as follows:

$$Sim(\mathbf{z}_s, \mathbf{z}_\alpha) = \cos(\mathbf{z}_s, \mathbf{z}_\alpha) \tag{9}$$

$$Sim(\mathbf{z}_s, \mathbf{z}_\gamma) = \cos(\mathbf{z}_s, \mathbf{z}_\gamma) \tag{10}$$

where $\mathbf{z}_s$ is the topic representation of s. The similarity calculated against $\mathbf{z}_\alpha$ or $\mathbf{z}_\gamma$ denotes the source-to-source or the source-to-target similarity.

We also consider the topic sensitivity estimation since general rules have flatter distributions while topic-specific rules have sharper distributions. A standard entropy metric is used to measure the sensitivity of the source-side of $\langle \alpha, \gamma\rangle$ as:

$$Sen(\alpha) = -\sum_{i=1}^{|\mathbf{z}_\alpha|} z_{\alpha i} \times \log z_{\alpha i} \tag{11}$$

where $z_{\alpha i}$ is a component in the vector $\mathbf{z}_\alpha$. The target-side sensitivity $Sen(\gamma)$ can be calculated in a similar way. The larger the sensitivity is, the more topic-specific the rule manifests.

In addition to traditional SMT features, we add new topic-related features into the standard log-linear framework. For the SMT system, the best translation candidate $\hat{e}$ is given by:

$$\hat{e} = \arg\max_e P(e|f) \tag{12}$$

where the translation probability is given by:

$$\begin{aligned} P(e|f) &\propto \sum_i w_i \cdot \log \phi_i(f, e) \\ &= \underbrace{\sum_j w_j \cdot \log \phi_j(f, e)}_{\text{Standard}} + \underbrace{\sum_k w_k \cdot \log \phi_k(f, e)}_{\text{Topic related}} \end{aligned} \tag{13}$$

where $\phi_j(f, e)$ is the standard feature function and w_j is the corresponding feature weight. $\phi_k(f, e)$ is the topic-related feature function and w_k is the feature weight. The detailed feature description is as follows:

Standard features: Translation model, including translation probabilities and lexical weights for both directions (4 features), 5-gram language model (1 feature), word count (1 feature), phrase count (1 feature), NULL penalty (1 feature), number of hierarchical rules used (1 feature).

Topic-related features: rule similarity scores (2 features), rule sensitivity scores (2 features).

4 Experiments

4.1 Setup

We evaluate the performance of our neural network based topic similarity model on a Chinese-to-English machine translation task. In neural network training, a large number of monolingual documents are collected in both source and target languages. The documents are mainly from two domains: news and weblog. We use Chinese and

English Gigaword corpus (Version 5) which are mainly from news domain. In addition, we also collect weblog documents with a variety of topics from the web. The total data statistics are presented in Table 1. These documents are built in the format of inverted index using Lucene[2], which can be efficiently retrieved by the parallel sentence pairs. The most relevant N documents are collected, where we experiment with $N = \{1, 5, 10, 20, 50\}$.

Domain	Chinese		English	
	Docs	Words	Docs	Words
News	5.7M	5.4B	9.9M	25.6B
Weblog	2.1M	8B	1.2M	2.9B
Total	7.8M	13.4B	11.1M	28.5B

Table 1: Statistics of monolingual data, in numbers of documents and words (main content). "M" refers to million and "B" refers to billion.

We implement a distributed framework to speed up the training process of neural networks. The network is learned with mini-batch asynchronous gradient descent with the adaptive learning rate procedure called AdaGrad (Duchi et al., 2011). We use 32 model replicas in each iteration during the training. The model parameters are averaged after each iteration and sent to each replica for the next iteration. The vocabulary size for the input layer is 100,000, and we choose different lengths for the hidden layer as $L = \{100, 300, 600, 1000\}$ in the experiments. In the pre-training phase, all parallel data is fed into two neural networks respectively for DAE training, where network parameters W and $\mathbf{b}$ are randomly initialized. In the fine-tuning phase, for each parallel sentence pair, we randomly select other ten sentence pairs which satisfy the criterion as negative instances. These training instances are leveraged to optimize the similarity of two vectors.

In SMT training, an in-house hierarchical phrase-based SMT decoder is implemented for our experiments. The CKY decoding algorithm is used and cube pruning is performed with the same default parameter settings as in Chiang (2007). The parallel data we use is released by LDC[3]. In total, the datasets contain nearly 1.1 million sentence pairs. Translation models are trained over the parallel data that is automatically word-aligned using GIZA++ in both directions, and the diag-grow-final heuristic is used to refine symmetric word alignment. An in-house language modeling toolkit is used to train the 5-gram language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995). The English monolingual data used for language modeling is the same as in Table 1. The NIST 2003 dataset is the development data. The testing data consists of NIST 2004, 2005, 2006 and 2008 datasets. The evaluation metric for the overall translation quality is case-insensitive BLEU4 (Papineni et al., 2002). The reported BLEU scores are averaged over 5 times of running MERT (Och, 2003). A statistical significance test is performed using the bootstrap re-sampling method (Koehn, 2004).

[2]http://lucene.apache.org/

[3]LDC2003E14, LDC2002E18, LDC2003E07, LDC2005T06, LDC2005T10, LDC2005E83, LDC2006E34, LDC2006E85, LDC2006E92, LDC2006E26, LDC2007T09

4.2 Baseline

The baseline is a re-implementation of the Hiero system (Chiang, 2007). The phrase pairs that appear only once in the parallel data are discarded because most of them are noisy. We also use the fix-discount method in Foster et al. (2006) for phrase table smoothing. This implementation makes the system perform much better and the translation model size is much smaller.

We compare our method with the LDA-based approach proposed by Xiao et al. (2012). In (Xiao et al., 2012), the topic of each sentence pair is exactly the same as the document it belongs to. Since some of our parallel data does not have document-level information, we rely on the IR method to retrieve the most relevant document and simulate this approach. The PLDA toolkit (Liu et al., 2011) is used to infer topic distributions, which takes 34.5 hours to finish.

4.3 Effect of retrieved documents and length of hidden layers

We illustrate the relationship among translation accuracy (BLEU), the number of retrieved documents (N) and the length of hidden layers (L) on different testing datasets. The results are shown in Figure 3. The best translation accuracy is achieved when N=10 for most settings. This confirms that enriching the source text with topic-related documents is very useful in determining topic representations, thereby help to guide the synchronous rule selection. However, we find that as N becomes larger in the experiments, e.g. N=50, the translation accuracy drops drastically. As more documents are retrieved, less relevant information

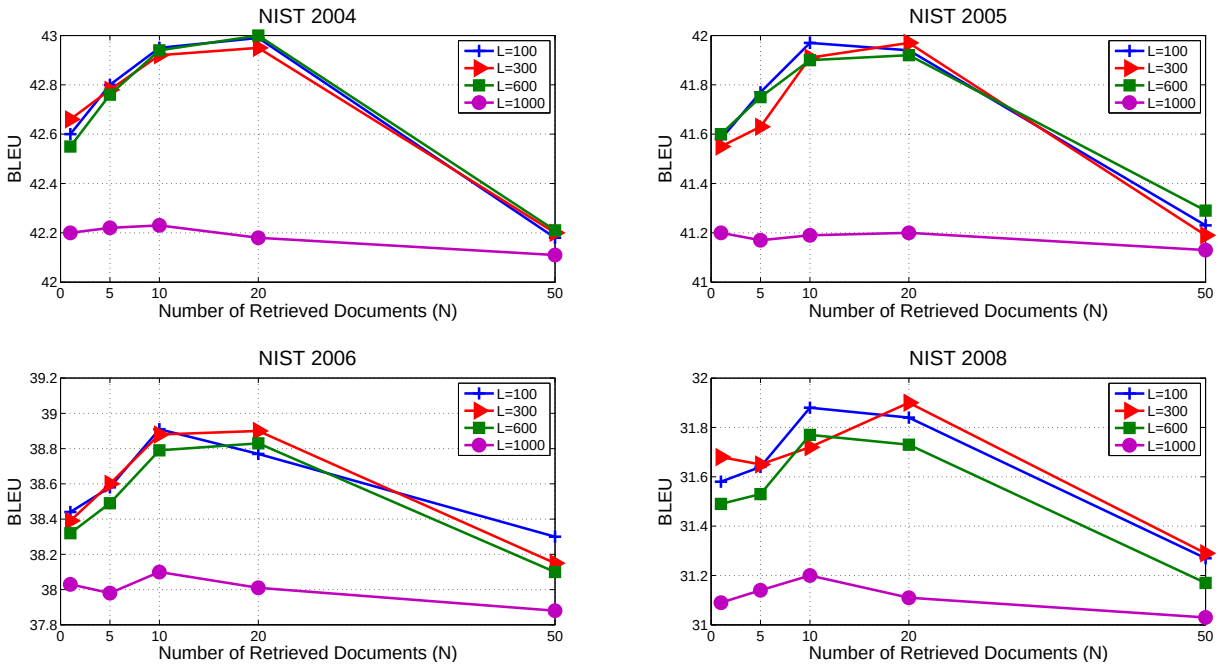


Figure 3: End-to-end translation results (BLEU%) using all standard and topic-related features, with different settings on the number of retrieved documents N and the length of hidden layers L.

is also used to train the neural networks. Irrelevant documents bring so many unrelated topic words hence degrade neural network learning performance.

Another important factor is the length of hidden layers L in the network. In deep learning, this parameter is often empirically tuned with human efforts. As shown in Figure 3, the translation accuracy is better when L is relatively small. Actually, there is no obvious distinction of the performance when L is less than 600. However, when L equals 1,000, the translation accuracy is inferior to other settings. The main reason is that parameters in the neural networks are too many to be effectively trained. As we know when L=1000, there are a total of $100,000 \times 1,000$ parameters between the linear and non-linear layers in the network. Limited training data prevents the model from getting close to the global optimum. Therefore, the model is likely to fall in local optima and lead to unacceptable representations.

4.4 Effect of topic related features

We evaluate the performance of adding new topic-related features to the log-linear model and compare the translation accuracy with the method in (Xiao et al., 2012). To make different methods comparable, we set the dimension of topic representation as 100 for all settings. This takes 10 hours in pre-training phase and 22 hours in fine-tuning phase. Table 2 shows how the accuracy is improved with more features added. The results confirm that topic information is indispensable for SMT since both (Xiao et al., 2012) and our neural network based method significantly outperforms the baseline system. Our method improves 0.86 BLEU points at most and 0.76 BLEU points on average over the baseline. We observe that source-side similarity is more effective than target-side similarity, but their contributions are cumulative. This proves that bilingually induced topic representation with neural network helps the SMT system disambiguate translation candidates. Furthermore, rule sensitivity features improve SMT performance compared with only using similarity features. Because topic-specific rules usually have a larger sensitivity score, they can beat general rules when they obtain the same similarity score against the input sentence. Finally, when all new features are integrated, the performance is the best, preforming substantially better than (Xiao et al., 2012) with 0.39 BLEU points on average.

It is worth mentioning that the performance of (Xiao et al., 2012) is similar to the settings with N=1 and L=100 in Figure 3. This is not simply coincidence since we can interpret their approach as a special case in our neural network method: when a parallel sentence pair has

Settings	NIST 2004	NIST 2005	NIST 2006	NIST 2008	Average
Baseline	42.25	41.21	38.05	31.16	38.17
(Xiao et al., 2012)	42.58	41.61	38.39	31.58	38.54
Sim(Src)	42.51	41.55	38.53	31.57	38.54
Sim(Trg)	42.43	41.48	38.4	31.49	38.45
Sim(Src+Trg)	42.7	41.66	38.66	31.66	38.67
Sim(Src+Trg)+Sen(Src)	42.77	41.81	38.85	31.73	38.79
Sim(Src+Trg)+Sen(Trg)	42.85	41.79	38.76	31.7	38.78
Sim(Src+Trg)+Sen(Src+Trg)	**42.95**	**41.97**	**38.91**	**31.88**	**38.93**

Table 2: Effectiveness of different features in BLEU% ($p < 0.05$), with N=10 and L=100. "Sim" denotes the rule similarity feature and "Sen" denotes rule sensitivity feature. "Src" and "Trg" means utilizing source-side/target-side rule topic vectors to calculate similarity or sensitivity, respectively. The "Average" setting is the averaged result of four datasets.

document-level information, that document will be retrieved for training; otherwise, the most relevant document will be retrieved from the monolingual data. Therefore, our method can be viewed as a more general framework than previous LDA-based approaches.

4.5 Discussion

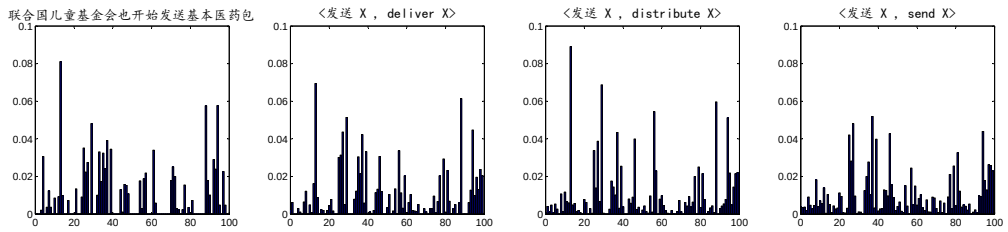
In this section, we give a case study to explain why our method works. An example of translation rule disambiguation for a sentence from the NIST 2005 dataset is shown in Figure 4. We find that the topic of this sentence is about "rescue after a natural disaster". Under this topic, the Chinese rule "发送 X" should be translated to "deliver X" or "distribute X". However, the baseline system prefers "send X" rather than those two candidates. Although the translation probability of "send X" is much higher, it is inappropriate in this context since it is usually used in IT texts. For example, ⟨发送邮件, send emails⟩, ⟨发送信息, send messages⟩ and ⟨发送数据, send data⟩. In contrast, with our neural network based approach, the learned topic distributions of "deliver X" or "distribute X" are more similar with the input sentence than "send X", which is shown in Figure 4. The similarity scores indicate that "deliver X" and "distribute X" are more appropriate to translate the sentence. Therefore, adding topic-related features is able to keep the topic consistency and substantially improve the translation accuracy.

5 Related Work

Topic modeling was first leveraged to improve SMT performance in (Zhao and Xing, 2006; Zhao and Xing, 2007). They proposed a bilingual topical admixture approach for word alignment and assumed that each word-pair follows a topic-specific model. They reported extensive empirical analysis and improved word alignment accuracy as well as translation quality. Following this work, (Xiao et al., 2012) extended topic-specific lexicon translation models to hierarchical phrase-based translation models, where the topic information of synchronous rules was directly inferred with the help of document-level information. Experiments show that their approach not only achieved better translation performance but also provided a faster decoding speed compared with previous lexicon-based LDA methods.

Another direction of approaches leveraged topic modeling techniques for domain adaptation. Tam et al. (2007) used bilingual LSA to learn latent topic distributions across different languages and enforce one-to-one topic correspondence during model training. They incorporated the bilingual topic information into language model adaptation and lexicon translation model adaptation, achieving significant improvements in the large-scale evaluation. (Su et al., 2012) investigated the relationship between out-of-domain bilingual data and in-domain monolingual data via topic mapping using HTMM methods. They estimated phrase-topic distributions in translation model adaptation and generated better translation quality. Recently, Chen et al. (2013) proposed using vector space model for adaptation where genre resemblance is leveraged to improve translation accuracy. We also investigated multi-domain adaptation where explicit topic information is used to train domain specific models (Cui et al., 2013).

Generally, most previous research has leveraged conventional topic modeling techniques such as LDA or HTMM. In our work, a novel neural network based approach is proposed to infer topic representations for parallel data. The advantage of



Rules	$P(\gamma\|\alpha)$	$Sim(\mathbf{z}_s, \mathbf{z}_\alpha)$
⟨发送 X, deliver X⟩	0.0237	0.8469
⟨发送 X, distribute X⟩	0.0546	0.8268
⟨发送 X, send X⟩	0.2464	0.6119

Src	联合国儿童基金会也开始发送基本医药包
Ref	(1) the united nations children's fund has also begun *delivering* basic medical kits (2) the unicef has also started to *distribute* basic medical kits (3) the united nations children's fund has also begun *distributing* basic medical kits (4) the united nations children's fund has begun *delivering* basic medical kits
Baseline	the united nations children's fund began to *send* basic medical kits
Ours	the united nations children's fund has begun to *distribute* basic medical kits

Figure 4: An example from the NIST 2005 dataset. We illustrate the normalized topic representations of the source sentence and three ambiguous synchronous rules. Details are explained in Section 4.5.

our method is that it is applicable to both sentence-level and document-level SMT, since we do not place any restrictions on the input. In addition, our method directly maximizes the similarity between parallel sentence pairs, which is ideal for SMT decoding. Compared to document-level topic modeling which uses the topic of a document for all sentences within the document (Xiao et al., 2012), our contributions are:

- We proposed a more general approach to leveraging topic information for SMT by using IR methods to get a collection of related documents, regardless of whether or not document boundaries are explicitly given.
- We used neural networks to learn topic representations more accurately, with more practicable and scalable modeling techniques.
- We directly optimized bilingual topic similarity in the deep learning framework with the help of sentence-level parallel data, so that the learned representation could be easily used in SMT decoding procedure.

6 Conclusion and Future Work

In this paper, we propose a neural network based approach to learning bilingual topic representation for SMT. We enrich contexts of parallel sentence pairs with topic related monolingual data and obtain a set of documents to represent sentences. These documents are converted to a bag-of-words input and fed into neural networks. The learned low-dimensional vector is used to obtain the topic representations of synchronous rules. In SMT decoding, appropriate rules are selected to best match source texts according to their similarity in the topic space. Experimental results show that our approach is promising for SMT systems to learn a better translation model. It is a significant improvement over the state-of-the-art Hiero system, as well as a conventional LDA-based method.

In the future research, we will extend our neural network methods to address document-level translation, where topic transition between sentences is a crucial problem to be solved. Since the translation of the current sentence is usually influenced by the topic of previous sentences, we plan to leverage recurrent neural networks to model this phenomenon, where the history translation information is naturally combined in the model.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments. We also thank Fei Huang (BBN), Nan Yang, Yajuan Duan, Hong Sun and Duyu Tang for the helpful discussions. This work is supported by the National Natural Science Foundation of China (Granted No. 61272384)

References

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2006. Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, Cambridge, MA.

Yoshua Bengio. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. *Proceedings of Machine Translation Summit XI*, pages 73–80.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1293, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Lei Cui, Xilun Chen, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Multi-domain adaptation for SMT using multi-task learning. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1055–1065, Seattle, Washington, USA, October. Association for Computational Linguistics.

George E. Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1):30–42, January.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 53–61, Sydney, Australia, July. Association for Computational Linguistics.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pages 315–323.

Amit Gruber, Michal Rosen-zvi, and Yair Weiss. 2007. Hidden topic markov models. In *In Proceedings of Artificial Intelligence and Statistics*.

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 321–328, Manchester, UK, August. Coling 2008 Organizing Committee.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. 2012. Imagenet classification with deep convolutional neural networks. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1106–1114.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. 2006. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, Cambridge, MA.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 89–97, Honolulu, Hawaii, October. Association for Computational Linguistics.

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and*

Technology, special issue on Large Scale Machine Learning. Software available at `http://code.google.com/p/plda`.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-08: HLT*, pages 1003–1011, Columbus, Ohio, June. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1988. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. 2011a. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011b. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 459–468, Jeju Island, Korea, July. Association for Computational Linguistics.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual lsa-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207, December.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA. ACM.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. 2012. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 750–758, Jeju Island, Korea, July. Association for Computational Linguistics.

Deyi Xiong and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *AAAI*.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2009. A syntax-driven bracketing model for phrase-based translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 315–323, Suntec, Singapore, August. Association for Computational Linguistics.

Bing Zhao and Eric P. Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 969–976, Sydney, Australia, July. Association for Computational Linguistics.

Bing Zhao and Eric P. Xing. 2007. Hm-bitam: Bilingual topic exploration, word alignment, and translation. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1689–1696. MIT Press, Cambridge, MA.