Guided Alignment Training for Topic-Aware Neural Machine Translation

Wenhu Chen hustchenwenhu@gmail.com
RWTH Aachen University, Ahornstr. 55, Aachen, Germany

Evgeny Matusov ematusov@ebay.com
eBay, Inc. Kasernenstr. 25, Aachen, Germany

Shahram Khadivi skhadivi@ebay.com
eBay, Inc. Kasernenstr. 25, Aachen, Germany

Jan-Thorsten Peter peter@cs.rwth-aachen.de
RWTH Aachen University, Ahornstr. 55 Aachen, Germany

Abstract

In this paper, we propose an effective way for biasing the attention mechanism of a sequence-to-sequence neural machine translation (NMT) model towards the well-studied statistical word alignment models. We show that our novel guided alignment training approach improves translation quality on real-life e-commerce texts consisting of product titles and descriptions, overcoming the problems posed by many unknown words and a large type/token ratio. We also show that meta-data associated with input texts such as topic or category information can significantly improve translation quality when used as an additional signal to the decoder part of the network. With both novel features, the BLEU score of the NMT system on a product title set improves from 18.6 to 21.3%. Even larger MT quality gains are obtained through domain adaptation of a general domain NMT system to e-commerce data. The developed NMT system also performs well on the IWSLT speech translation task, where an ensemble of four variant systems outperforms the phrase-based baseline by 2.1% BLEU absolute.

1 Introduction

NMT systems were shown to reach state-of-the-art translation quality on tasks established in MT research community such as IWSLT speech translation task Cettolo et al. (2012). In this paper, we also apply NMT approach to e-commerce data: user-generated product titles and descriptions for items put on sale. Such data are very different from newswire and other texts typically considered in the MT research community. Titles in particular are short (usually fewer than 15 words), contain many brand names which often do not have to be translated, but also product feature values and specific abbreviations and jargon. Also, the vocabulary size is very large due to the large variety of product types, and many words are observed in the training data only once. At the same time, these data are provided with additional meta-information about the item (e.g. product category such as clothing or electronics), which can be used as context to perform topic/domain adaptation for improved translation quality.

At first glance, established phrase-based statistical MT approaches are well-suited for e-commerce data translation. In a phrase-based approach, singleton, but unambiguous words and phrases are usually translated correctly. Also, since the alignment between source and target words is available, it is possible to transfer certain entities from the source sentence to the generated target sentence "in-context" without translating them. Such entities can include numbers, product specifications such as "5S" or brand names such as "Samsung" or "Lenovo". In training, these entities can be replaced with placeholders to reduce the vocabulary size.

However, NMT approaches are more powerful at capturing context beyond phrase boundaries and were shown to better exploit available training data. They also successfully adapt themselves to a domain, for which only a limited amount of parallel training data is available Luong and Manning (2015). Also, previous research Mathur et al. (2015) has shown that it is difficult to obtain translation quality improvements with topic adaptation in phrase-based SMT because of data sparseness and a large number of topics (e. g. corresponding to product categories), which may or may not be relevant for disambiguating between alternative translations or solving other known MT problems. In contrast, we expected NMT to better solve the topic adaptation problem by using the additional meta-information as an extra signal in the neural network. To the best of our knowledge, this is the first work where the additional information about the text topic is embedded into the vector space and used to directly influence NMT decisions.

In an NMT system, the attention mechanism introduced in Luong et al. (2014) is important both for decoding as well as for restoration of placeholder content and insertion of unknown words in the right positions in the target sentence. To improve the estimation of the soft alignment, we propose to use the Viterbi alignments of the IBM model 4 Brown et al. (1993) as an additional source of knowledge during NMT training. The additional alignment information helps the current system to bias the attention mechanism towards the Viterbi alignment.

This paper is structured as follows. After an overview of related NMT work in Section 2, we propose a novel approach in Section 3 on using statistical word alignemt to bias the training of neural MT attention mechanism, we call it *guided alignment training*. In Section 4, we describe in more detail how topic information can benefit NMT. Section 5 and Section 6 describes our domain adaptation approach. Experimental results are presented in Section 7. The paper is concluded with a discussion and outlook in Section 8.

2 Related Work

Neural machine translation is mainly based on using recurrent neural networks to grasp long term dependencies in natural language. An NMT system is trained on end-to-end basis to maximize the conditional probability of a correct translation given a source sentence Sutskever et al. (2014), Bahdanau et al. (2014), Cho et al. (2014b). When using attention mechanism, large vocabularies Jean et al. (2014), and some other techniques, NMT is reported to achieve comparable translation quality to state-of-art phrase-based translation systems. Most NMT approaches are based on the encoder-decoder architecture Cho et al. (2014a), in which the input sentence is first encoded into a fixed-length representation, from which the recurrent neural network decoder generates the sequence of target words. Since fixed-length representation cannot give enough information for decoding, a more sophisticated approach using attention mechanism is proposed by Bahdanau et al. (2014). In this approach, the neural network learns to attend to different parts of source sentence to improve translation quality. Since the source and target language vocabularies for a neural network have to be limited, the rare words problem deteriorates translation quality significantly. The rare word replacement technique using soft alignment proposed by Luong et al. (2014) gives a promising solution for the problem. Both encoder-decoder architecture and insertion of unknown words into NMT output highly rely on the quality of the attention mechanism, thus it becomes the crucial part of NMT. Some research has been done to refine it by Luong et al. (2015), who proposed global and local attention-based models, and Cohn et al. (2016), who used biases, fertility and symmetric bilingual structure to improve the attention model mechanism. The most recent work done by Mi et al. (2016) is highly parallel with our *guided alignment training*, Section 3. They use statistical alignment to supervise the NMT in a similar fashion as we do, the difference is that they smooth the statistical alignment and apply Euclidean distance directly to the objective function, while we try with different divergence function and also re-weight it before adding to the overall objective function.

Research on topic adaptation most closely related to our work was performed by Hasler et al. (2014), but the features proposed there were added to the log-linear model of a phrase-based system. Here, we use the topic information as part of the input to the NMT system. Another difference is that we primarily work with human-labeled topics, whereas in Hasler et al. (2014) the topic distribution is inferred automatically from data.

When translating e-commerce content, we are faced with a situation when only a few product titles and descriptions were manually translated, resulting in a small in-domain parallel corpus, but a large general-domain parallel corpus is available. In such situations, domain adaption techniques have been used both in phrase-based systems Koehn and Schroeder (2007) and NMT Luong and Manning (2015). In addition, while diverse NMT models using different features and techniques are trained, an ensemble decoder can be used to combine them together to make a more robust model. This approach was used by Luong et al. (2015) to outperform the state-of-art phrase-based system with their NMT approach in the WMT 2015 evaluation.

3 Guided Alignment Training

When using the attention-based NMT Bahdanau et al. (2014), we observed that the attention mechanism sometimes fails to yield appropriate soft alignments, especially with increasing length of the input sentence and many out-of-vocabulary words or placeholders. In translation, this can lead to disordered output and word repetition.

In contrast to a statistical phrase-based system, the NMT decoder does not have explicit information about the candidates of the current word, so at each recurrent step, the attention weights only rely on the previously generated word and decoder/encoder state, as depicted in Figure 1. The target word itself is not used to compute its attention weights. If the previous word is an out-of-vocabulary (OOV) or a placeholder, then the information it provides for calculating the attention weights for the current word is neither sufficient nor reliable anymore. This leads to incorrect target word prediction, and the error propagates to the future steps due to feedback loop. The problem is even larger in the case of e-commerce data where the number of OOVs and placeholders is considerably higher.

To improve the estimation of the soft alignment, we propose to use the Viterbi alignments of the IBM model 4 as an additional source of knowledge during the NMT training. Therefore, we first extract Viterbi alignments using GIZA++ toolkit Och and Ney (2003), then we use them to bias the attention mechanism. Our approach is to optimize on both the decoder cost and the divergence between the attention weights and the alignment connections generated by statistical alignments. The multi-objective optimization task is then expressed as a single-objective function, which is a linear combination of two loss functions: original and new guided-alignment.

3.1 Decoder Cost

NMT proposed by Bahdanau et al. (2014) maximizes the conditional log-likelihood of target sentence $y_1, \ldots, y_T$ given the source sentence $x_1, \ldots, x'_T$:

$$H_D(y, x) = -\frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y_n | x_n) \tag{1}$$

where (y_n, x_n) refers to n_{th} training sentence pair, and N denotes the total number of sentence pairs in the training corpus. In the paper, we name the negative log-likelihood as decoder cost to distinguish from alignment cost. When using encoder-decoder architecture by Cho et al. (2014b), the conditional probability can be written as:

$$p(y_1 \dots y_T | x_1 \dots x_{T'}) = \prod_{t=1}^{T} p(y_t | y_{t-1} \cdots y_1, c) \tag{2}$$

with $p(y_t|y_{t-1} \cdots y_1, c) = g(s_t, y_{t-1}, c)$, where T is the length of the target sentence and T' is the length of source sentence, c is a fixed-length vector to encode source sentence, s_t is a hidden state of RNN at time step t, and $g(\cdot)$ is a non-linear function to approximate word probability. If attention mechanism is used, the vector c in each sentence is replaced by time-variant representation c_t that is a weighted summary over a sequence of annotations $(h_1, \cdots, h_{T'})$, and h_i contains information about the whole input sentence, but with a strong focus on the parts surrounding the i_{th} word Bahdanau et al. (2014). Then, the context vector can be defined as:

$$c_t = \sum_{i}^{T'} \alpha_{ti} h_i \quad \text{where} \quad \alpha_{ti} = \frac{exp(e_{ti})}{\sum_{k=1}^{T'} exp(e_{tk})}. \tag{3}$$

This means, α_{ti} for each annotation h_i is computed by normalizing the score function with the softmax. Also, $e_{ti} = a(s_{t-1}, h_i)$ is the function to calculate the score of t-th target word aligning to i-th word in the source sentence. The alignment model $a(\cdot, \cdot)$ is used to calculate similarity between previous state s_{t-1} and bi-directional state h_i. In our experiments, we took the idea of the dot global attention model of Luong et al. (2015), but we still keep the order $h_{t-1} \to a_t \to c_t \to h_t$ as proposed by Bahdanau et al. (2014). We calculate the dot product of encoder state h_i with the last decoder state s_{t-1} instead of the current decoder state. We observe that this dot attention model (Equation 4) works better than concatenation in our experiments.

$$a(s_{t-1}, h_i) = (W_s s_{t-1})^T (W_h h_i) \tag{4}$$

3.2 Alignment Cost

We introduce alignment cost to penalize attention mechanism when it is not consistent with statistical word alignment. We represent the pre-trained statistical alignments by a matrix A, where A_{ti} refers to the probability of the t_{th} word in the target sentence of being aligned to the i_{th} word in the source sentence. In case of multiple source words aligning to the same target word, we normalize to make sure $\sum_i A_{ti} = 1$, in the case of non-aligned target words, we do not add any penalty. In attention-based NMT, the matrix of attention weights α has the same shape and semantics as A. We propose to penalize NMT based on the divergence of the two matrices during the training, the divergence function can e. g. be cross entropy G_{ce} or mean square error G_{mse} as in Equation 5. As shown in Figure 1, A comes from statistical alignment and is fed into our guided-alignment NMT as an additional input to penalize the attention mechanism.

$$G_{ce}(A, \alpha) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{T'} A_{ti} \log \alpha_{ti} \qquad G_{mse}(A, \alpha) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{T'} (A_{ti} - \alpha_{ti})^2 \tag{5}$$

We combine decoder cost and alignment cost to build the new loss function $H(y, x, A, \alpha)$:

$$H(y, x, A, \alpha) = w_1 G(A, \alpha) + w_2 H_D(y, x) \tag{6}$$

During training, we optimize the new compound loss function $H(y, x, A, \alpha)$ with regard to the same parameters as before. The guided-alignment training influences the attention mechanism to generate alignment closer to Viterbi alignment and has the advantage of unchanged parameter space and model complexity. When training is done, we assume that NMT can generate robust alignment by itself, so there is no need to feed an alignment matrix as input during evaluation. As indicated in Equation 6, we set w_1 and w_2 for weights of decoder cost and alignment cost to balance their weight ratio. We performed further experiments (see section 7) to analyze the impact of different weight settings on translation quality.

4 Topic-aware Machine Translation

In the e-commerce domain, the information on the product category (e.g., “mens’ clothing”, “mobile phones“, “kitchen appliances”) often accompanies the product title and description and can be used as an additional source of information both in the training of a MT system and during translation. In particular, such meta-information can help to disambiguate between alternative translations of the same word that have different meaning. The choice of the right translation often depends on the category. For example, the word “skin” has to be translated differently in the categories “mobile phone accessories” and “make-up”. Outside of the e-commerce world, similar topic information is available in the form of e.g. tags and keywords for a given document (on-line article, blog post, patent, etc.) and can also be used for word sense disambiguation and topic adaptation. In general, a document may belong to multiple topics.

Here, we propose to feed such meta-information into the recurrent neural network to help generate words which are appropriate given a particular category or topic.

4.1 Topic Representation

The idea is to represent topic information in a D-dimensional vector l, where D is the number of topics. Since one sentence can belong to multiple topics (possibly with different probabilities/weights), we normalize the topic vector so that the sum of its elements is 1. It is fed into the decoder to influence the proposed target word distribution. The conditional probability given the topic membership vector can be written as (cf. Equations 2 and 3):

$$p(y_t|y_{<t-1}, c_t, s_{t-1}, l) = p(y_t|y_{t-1}, c_t, s_{t-1}, l) \approx g(y_{t-1}, s_{t-1}, c_t, l)$$

where $g(\cdot)$ is used to approximate the probability distribution. In our implementation, we introduce an intermediate readout layer to build the function $g(\cdot)$, which is a feed-forward network as depicted in Figure 2.

4.2 Topic-aware Decoder

In the NMT decoder, we feed the topic membership vector to the readout layer in each recurrent step to enhance word selection. As shown in Figure 1 and Figure 2, topic membership vector l is fed into the NMT decoder as an additional input besides source and target sentences:

$$p(y_t|y_{<t-1}, c_t, s_{t-1}, l) = p(y_t|r_t) \quad \text{where} \quad r_t = W_r[c_t; f_{t-1}; s_{t-1}; l] + b_r \tag{7}$$

Here, W_r is the concatenation of original transformation matrix and l, r_t is the output from readout layer and f_t is the embedding of the last target word y_{t-1}; s_{t-1} refers the last decoder state. W_r and b_r are weights and bias for the linear transformation, respectively. We can rearrange the formula as:

$$\begin{aligned} r_t &= [W_r', W_c][c_t; f_{t-1}; s_{t-1}; l] + b_r \\ &= [W_r'[c_t; f_{t-1}; s_{t-1}] + b_r] + W_c l \\ &= r_t' + E_c \end{aligned} \tag{8}$$

where W_r is concatenation of original transformation matrix W_r' and topic transformation matrix W_c. Then adding topic into readout layer input is equivalent to adding an additional topic

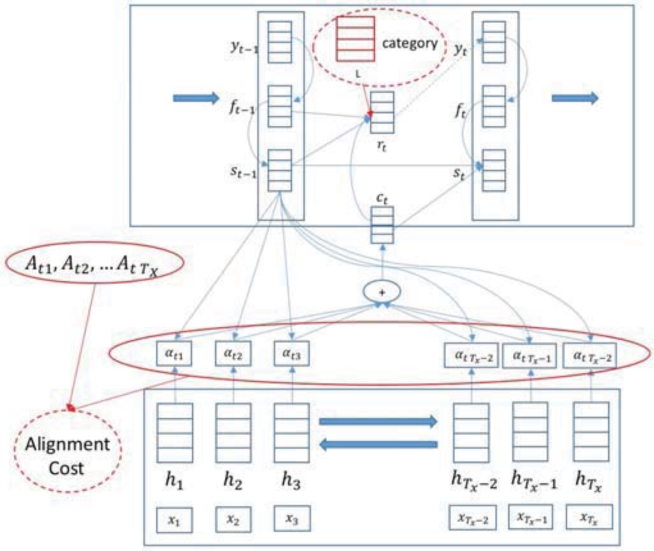

Figure 1: Topic-aware, alignment-guided encoder-decoder model. Topic information l is added to the decoder as an additional input, influencing every decoding step; statistical alignment A is added to the attention layer to supervise the learning of the attention mechanism.

vector E_c into the original readout layer output. Assuming l is a one-hot category vector, then $W_c l$ is equivalent to retrieving a specific column from the matrix W_c. Hence, we can name this additional vector E_c as topic embedding, regarded as a vector representation of topic information. It is quite similar to word embedding by Mikolov et al. (2013), we will further analyze the similarity between different topics in Figure 3.

The readout layer depicted in Figure 2 merges information from the last state s_{t-1}, previous word embedding f_{t-1} (coming from word index y_{t-1}, which is sampled w.r.t. the proposed word distribution), as well as the current context c_t to generate output. It can be seen as a shallow network, which consists of a max-out layer Goodfellow et al. (2013), a fully-connected layer, and a softmax layer.

5 Bootstrapping

When trained on small amounts of data, the attention-based neural network approach does not always produce reliable soft alignment. The problem gets worse when the sentence pairs available for training are getting longer. To solve this problem, we extracted bilingual sub-sentence units from existing sentence pairs to be used as additional training data. These units are exclusively aligned to each other, i. e. all words within the source sub-sentence are aligned only to the words within the corresponding target sub-sentence and vice versa. The alignment is determined with the standard approach (IBM Model 4 alignment trained with the GIZA++ toolkit Och and Ney (2003)). As boundaries for sub-sentence units, we used punctuation marks, including period, comma, semicolon, colon, dash, etc. To simplify bilingual sentence splitting, we used the standard phrase pair extraction algorithm for phrase-based SMT, but set the minimum/maximum source phrase length to 8 and 30 tokens, respectively. From all such long phrase pairs extracted by the algorithm, we only kept those which are started or ended with a punctuation mark or started/ended a sentence; both on the source and on the target side.

For the bootstrapped training, we merged the original training data with the extracted sub-

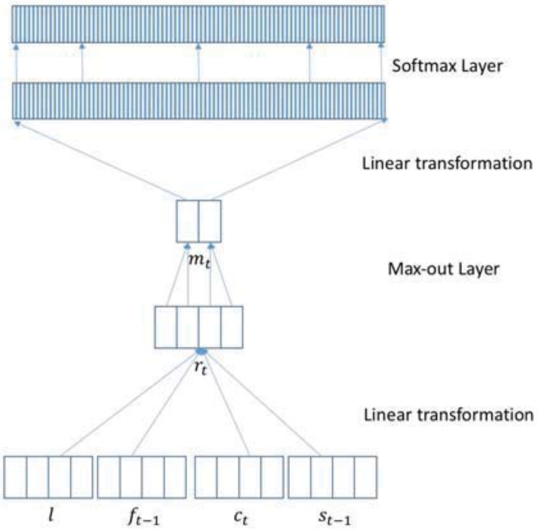

Figure 2: Topic-aware readout layer. The topic information vector, l, is fed to the readout layer in each recurrent step to influence target word selection.

sentence units and ran the neural training algorithm on this extended training set. Since the extracted bilingual sub-sentence units generally showed good correspondence between source and target due to the constraints described above, the expectation was that having such units repeated in the training data as stand-alone training instances would guide the attention mechanism to become more robust and make it easier for the neural training algorithm to find better correspondences between more difficult source/target sentence parts. Also, having both short and long training instances was expected to make neural translation quality less dependent on the input length.

6 E-commerce Domain Adaptation

For the e-commerce English-to-French translation task, we only have a limited amount of in-domain parallel training data (item titles and descriptions). To benefit from large amounts of general-domain training data, we follow the method described by Luong and Manning (2015). We first train a baseline NMT model on English-French WMT data (common-crawl, europarl v7, and news commentary corpora) for two epochs to get the best result on a development set, and then we continue training the same model on the in-domain training set for a few more epochs. In contrast to Luong and Manning (2015), however, we use the vocabularies of the most frequent 52K source/target words in the in-domain data (instead of the out-of-domain data vocabularies). This causes NMT to focus on translation of the most relevant in-domain words.

7 Experimental Results

7.1 Data Sets and Preprocessing

We performed MT experiments on the German-to-English IWLST 2015 speech translation task Cettolo et al. (2012) and on an in-house English-to-French e-commerce translation task. As part of data preprocessing, we tokenized and lowercased the corpora, as well as replaced numbers, product specifications, and other special symbols with placeholders such as $num. We only keep these placeholders in training, but preserve their content as XML markups in the dev/test sets, which we try to restore using attention mechanism. This content is inserted for the generated placeholders on the target side based on the attention mechanism (see Luong et al. (2014)). In the beam search for the best translation, we make sure that each placeholder content is used only once. Using the same mechanism, we also pass OOV words to the target side "as is" (without using any special unknown word symbol).

On both tasks, we evaluate all systems and system variants using case-insensitive BLEU Papineni et al. (2002) and TER Snover et al. (2006) scores on held-out development and test data using a single human reference translation.

Data-set		IWSLT		e-commerce	
Language		German	English	English	French
Training	Sentences	165 201		516 000	
	Running words	3 873 816	3 656 038	2 592 202	2 895 089
	Full vocabulary	103 390	45 068	119 607	129 848
Dev	Sentences	567		910	
	Running words	9 812	10 695	10 339	11 283
Test	Sentences	1100		910	
	Running words	19 019	22 895	10 817	11 016
Source OOV rate w.r.t. full/NMT Voc.		5.16/6.12 %		2.56/5.76 %	

Table 1: Corpus statistics for the IWSLT and e-commerce translation tasks. OOV rate is calculated after preprocessing, placeholders like \$num, \$url, etc. largely decrease the OOV rate in the e-commerce dev and test sets.

7.1.1 IWSLT TED Talk Data

For the IWSLT German-to-English task (translation of transcribed TED talks), we map the topic keywords of each TED talk in the 2015 training/dev/test evaluation campaign release to ten general topics such as politics, environment, education, and others. All sentences in the same talk share the same topic, and one talk can belong to several topics. Instead of using the official IWSLT dev/test data, we set aside 81/159 talks for development/test set, respectively. Out of these talks, we used 567 dev and 1100 test sentences which have the highest probability of relating to a particular topic (bag-of-words classification using the remaining 1365 talks as the training data). The corpus statistics of the data sets obtained this way are given in Table 1[1].

7.1.2 E-commerce Data

For the e-commerce English-to-French task, we used the product category such as "fashion" or "electronics" as topic information (a total of 80 most widely used categories plus the category "other" that combined all the less frequent categories). The training set contained both product titles and product descriptions, while dev and test set only contained product titles. Each title or description sentence was assigned to only one category. The statistics of the e-commerce data sets are given in Table 1.

7.2 Model Training

We implemented our neural translation model in Python using the Blocks deep learning library van Merriënboer et al. (2015) based on the open-source MILA translation project. We compared our implementation of NMT baseline system with Bahdanau et al. (2014) on the WMT 2014 English-to-French machine translation task and obtained a similar BLEU score on the official test set as they reported in Bahdanau et al. (2014). Then we implemented the topic-aware

[1] This data set with topic labels is publicly available at https://github.com/wenhuchen/iwslt-2015-de-en-topics.

E-commerce En→Fr	BLEU [%]	TER [%]
Baseline NMT	18.6	68.5
+prefixed human-labeled categs	18.3	69.3
+readout human-labeled categs	19.7	65.3
+readout LDA topics	14.5	74.9

Table 2: Comparison of different approaches for topic-aware NMT.

algorithm (section 4), guided alignment training (section 3), and the bootstrapped training (section 5) into the NMT model. We trained separate models with various feature combinations. We also created an ensemble of different models to obtain the best NMT translation results.

In our experiments, we set the dimension of both source and target word embeddings to 620 and use a bi-directional GRU encoder and attention-based GRU decoder, the cell dimension of both are set to 1000. We selected the 50k most frequent German words and top 30k English words as vocabularies for the IWSLT task, and most frequent 52k English/French words for the e-commerce task. The optimization of the objective function was performed by using AdaDelta algorithm Zeiler (2012). We set the beam size to 10 for dev/test set beam search translation.

For training implementation, we use stochastic gradient descent with batch size of 100, saving model parameters after a certain number of epochs. We saved around 30 consecutive model parameters. We selected the best parameter set according to the sum of the established MT evaluation measures BLEU Papineni et al. (2002) and 1-TER Snover et al. (2006) on the development set. After model selection, we evaluated the best model on the test set. We report the test set BLEU and TER scores in Table 5 and Table 7.

We use TITAN X GPUs with 12GB of RAM to run experiments on Ubuntu Linux 14.04. The training converges in less than 24 hours on the IWSLT talk task and around 30 hours on the e-commerce task. The beam search on the test set for both tasks takes around 10 minutes, the exact time depends on the vocabulary size and beam size.

7.3 Effect of Topic-aware NMT

We tested different approaches to find out where topic information fits best into NMT, since topic information can affect alignment, word selection, etc. The most naive approach is to insert a pseudo topic word in the beginning of a sentence to bias the context of the sentence to a certain topic. We also tried topic vectors of different origin in the read-out layer of the network. We used both topics predicted automatically with the Latent Dirichlet Analysis (LDA) and human-labeled topics to feed into the network as shown in Figure 1.

The results on the e-commerce task in Table 2 show that category information as a pseudo topic word does not carry enough semantic and syntactic meaning in comparison to real source words to have a positive effect on the target words predicted in the decoder. The BLEU score of such system (18.3%) is even below the baseline (18.6%). In contrast, the human-labeled categories are more reliable and are able to positively influence word selection in the NMT decoder, significantly (19.7% BLEU) outperforming the baseline.

Replacing the human-labeled topic one-hot vectors of size 80 with the LDA-predicted topic distribution vectors of the same dimension in the read-out layer of the neural network deteriorated the BLEU and TER scores significantly. We attribute this to data sparseness problems when training the LDA of dimension 80 on product titles.

On the German-to-English task, we also observed MT quality improvements when using human-labeled topic information as described in Figure 1. Here, we extracted the topic embedding E_c from different experiments and show their cosine distance in Figure 3. It's straightforward that in different experiments, the same topic tends to share similar representation in

SRC	ich möchte Ihnen heute Morgen gerne von meinem Projekt, Kunst Aufräumen, erzählen.
NMT	I want to clean you this morning, from my project, to say Art.
+topics	I would like to talk to you today by my project, Art clean.
REF	I would like to talk to you this morning about my project, Tidying Up Art.
SRC	. . . unsere Kollegen an Tufts verbinden Modelle wie diese mit durch Tissue Engineering erzeugten Knochen, um zu sehen, wie Krebs sich von einem Teil des Körpers zum nächsten verbreiten könnte.
NMT	. . . our NOAA colleagues combined models of models like this with tissue generated bones from bones to see how cancer could spread from one part of the body, to the next distribution.
+topics	. . . our colleagues at Tufts are using models like this with tissue-based engineered bones to see how cancer could spread from a part of the body to the next part.
REF	. . . our colleagues at Tufts are mixing models like these with tissue-engineered bone to see how cancer might spread from one part of the body to the next.

Table 3: Example of improved translation quality when topic information is used as input in the neural MT system (German-to-English IWSLT test set).

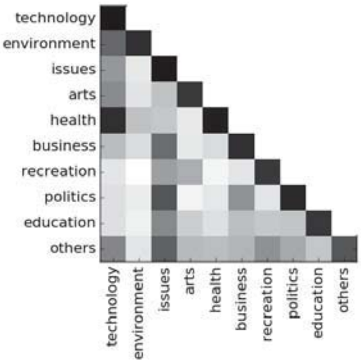

Figure 3: Topic embedding cosine distance.

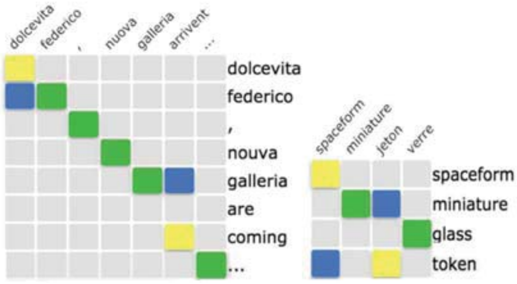

Figure 4: Refined alignment examples using guided-alignment learning (green blocks refer to the identical alignments from Baseline NMT and guided-alignment NMT, blue blocks refer to the alignment from baseline NMT, yellow blocks refer to guided-alignment NMT).

continuous embedding space. At the same time, closer topic pairs like "politics" and "issues" tend to have shorter distance from each other. Examples of improved German-to-English NMT translations when human-labeled topic information is used are shown in Table 3.

7.4 Implementation of Guided Alignment

To balance decoder cost and the attention weight cost, we experimented with different weights for these costs. We analyzed the relation between weight ratio and the final result in Table 4. Besides fixing the cost ratios during training, we also apply a heuristic to adjust the ratio as the training is progressing. One approach is to set a high value for the alignment cost in the beginning, then decay the weight to 90% after every epoch, finally eliminating the influence of the alignment after some time. This approach helps for the IWSLT task, but not on the e-commerce task. We assume that the alignment for the TED talk sentences seems to be easier for NMT to learn on its own than the alignment between product titles and their translations. We also analyzed the effect of using different loss functions for calculating alignment divergence (see Section 3.2). The difference between the squared error and cross-entropy is not so large as

En→Fr	BLEU %	TER %
Baseline NMT	18.6	68.3
+ce (decay)	20.5	65.8
+ce (1:2)	20.6	65.5
+ce (1:1)	20.2	65.0
+ce (2:1)	20.9	65.7
+mse (1:1)	20.8	64.5

Table 4: Comparison of different loss functions and weight ratios for guided alignment (cf. Equation 5).

En→Fr systems		BLEU %	TER %
1. NMT in-domain (ID)		18.6	68.5
2. 1) + topic vectors		19.7	65.3
3. 1) + bootstrapping		20.1	66.2
4. 1) + guided alignment		20.9	65.7
5. NMT with 2) and 4)		21.3	64.3
6. NMT with 2) and 3) and 4)		20.7	66.2
7. NMT out-of-domain (OOD)		13.8	77.4
8. 7) + guided alignment		16.3	74.5
9. 8) + domain adaptation		25.0	60.1
Ensemble NMT ID	system 4)	24.5	60.9
	system 5)		
	system 6)		
	NMT w. 3) and 4)		
Ensemble NMT OOD	system 9)	25.6	58.6
	9) with DW		
	9) w. topic vectors		

Table 5: Translation results on the En→Fr e-commerce task. (DW: decaying weight for the statistical alignment).

SRC	Vintage Ollech & Wajs Early Bird Diver watch, Excellent!
SMT	Vintage Ollech & Wajs début oiseau montre de plongée, excellent!
NMT	Montre de plongée vintage Ollech & Wajs early bird, excellent!
REF	Montre de Plongée Vintage Ollech & Wajs Early Bird, Excellent !

Table 6: Example of improved translation quality of the NMT ensemble system vs. phrase-based baseline system (English-to-French title test set).

shown in Table 4. Since the cross-entropy function has a consistent form as decoder cost, we decided to use it in further experiments. We extracted the NMT attention weights and marked the connection with the highest score as hard alignment for each word. We drew the alignment in Figure 4 to compare baseline NMT and alignment-guided NMT. It can be seen from the graph that the guided alignment training truly improves the alignment correspondence.

7.5 Overall Results

The overall results on the e-commerce translation task and IWSLT task are shown in Table 5 and Table 7, respectively. We observed consistency on both tasks, in a sense that a feature that improves BLEU/TER results on one task is also beneficial for the other.

For comparison, we trained phrase-based SMT models using the Moses toolkit Koehn et al. (2007) on both translation tasks. We used the standard Moses features, including a 4-gram LM trained on the target side of the bilingual data, word-level and phrase-level translation probabilities, as well as the distortion model with the maximum distortion of 6. Our stronger phrase-based baseline included additional 5 features of a 4-gram operation sequence model – OSM Durrani et al. (2015).

On the e-commerce task, which is more challenging due to a high number of OOV words and placeholders, we observed that NMT translation output had many errors related to incorrect attention weights. To improve the attention mechanism, we applied guided alignment and

#	De→En systems		BLEU %	TER %
1	Phrase-based system		24.7	55.4
2	Phrase-based system + OSM		25.7	55.1
3	NMT		23.4	60.1
4	NMT + topic vectors		23.7	59.6
5	NMT + bootstrapping		24.1	58.6
6	NMT + guided alignment		23.8	60.8
7	NMT + topic vectors + bootstrapping		24.2	59.4
8	NMT + topic vectors + bootstrapping + guided alignment		24.6	57.7
9	Ensemble	NMT + topic vectors	27.8	55.4
		NMT + topic vectors + guided alignment		
		NMT + topic vectors + bootstrapping		
		NMT + topic v. + guided alignment + bootstrapping		

Table 7: Overview of the translation results on the German-to-English IWSLT task.

bootstrapping. Both boosted the translation performance. Adding topic information increased the BLEU score to 21.3%. We selected the four best model parameters from various experiments to make an ensemble system, this improved the BLEU score to 24.5%. For the following experiment, we had pre-trained a model on WMT15 parallel data with the guided alignment technique, and then continued training on the e-commerce data for several epochs as described in section 6, performing domain adaptation. This approach proved to be extremely helpful, giving an increase of over 3.0% absolute in BLEU. Finally, we also applied ensemble methods on variants of the domain-adapted models to further increase the BLEU score to 25.6, which is 7.0 BLEU higher than the NMT baseline system, and only 0.6% BLEU behind the BLEU score of 26.2% for the state-of-the-art phrase-based baseline. Table 6 shows examples where the ensemble NMT system is better than the phrase-based system despite the slightly lower corpus-level BLEU score. In fact, a more detailed analysis of the sentence-level BLEU scores showed that the NMT translation of 386 titles out of 910 was ranked higher than the SMT translation, the reverse was true for 460 titles. In particular, the word order of noun phrases was observed to be better in the NMT translations.

On the IWSLT task (Table 7), the baseline NMT was not as far behind the phrase-based system as on the e-commerce task, and thus the obtained improvements were smaller than for product title translations. We observed that topic information is less helpful than bootstrapping and guided alignment learning. When we combined them, we reached the same BLEU score as the phrase-based system (see Table 7). Finally, we combined four variant systems to create an ensemble, which resulted in the BLEU score of 27.8%, surpassing the phrase-based translation with the OSM model by 2.1% BLEU absolute.

8 Conclusion

We have presented a novel guided alignment training for a NMT model that utilizes IBM model 4 Viterbi alignments to guide the attention mechanism. This approach was shown experimentally to bring consistent improvements of translation quality on e-commerce and spoken language translation tasks. Also on both tasks, the proposed novel way of utilizing topic meta-information in NMT was shown to improve BLEU and TER scores. We also showed improvements when using domain adaptation by continuing training of an out-of-domain NMT system on in-domain parallel data. In the future, we would like to investigate how to effectively make use of the abundant monolingual data with human-labeled product category information that we have available for the envisioned e-commerce application.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cohn, T., Hoang, C. D. V., and Vymolova, E. (2016). Incorporating structural alignment biases into an attention neural translation model. *arXiv preprint arXiv:1601.01085*.

Durrani, N., Schmid, H., Fraser, A., Koehn, P., and Schütze, H. (2015). The operation sequence model-combining n-gram-based and phrase-based statistical machine translation. *Computational Linguistics*.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout networks. *arXiv preprint arXiv:1302.4389*.

Hasler, E., Blunsom, P., Koehn, P., and Haddow, B. (2014). Dynamic topic adaptation for phrase-based MT. In *Proceedings of EACL*, pages 328–337.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. *CoRR*, abs/1412.2007.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224–227. Association for Computational Linguistics.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domain.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Mathur, P., Federico, M., Köprü, S., Khadivi, S., and Sawaf, H. (2015). Topic adaptation for machine translation of e-commerce content. *Proceedings of MT Summit XV*, page 270.

Mi, H., Wang, Z., and Ittycheriah, A. (2016). Supervised attentions for neural machine translation. *arXiv preprint arXiv:1608.00112*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

van Merriënboer, B., Bahdanau, D., Dumoulin, V., Serdyuk, D., Warde-Farley, D., Chorowski, J., and Bengio, Y. (2015). Blocks and fuel: Frameworks for deep learning. *arXiv preprint arXiv:1506.00619*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.