Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

A Topic-Based Coherence Model for Statistical Machine Translation

Deyi Xiong[1,2] and **Min Zhang**[1,2*]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China 215006
[2]Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore138632
{*dyxiong, minzhang*}*@suda.edu.cn*

Abstract

Coherence that ties sentences of a text into a meaningfully connected structure is of great importance to text generation and translation. In this paper, we propose a topic-based coherence model to produce coherence for document translation, in terms of the continuity of sentence topics in a text. We automatically extract a coherence chain for each source text to be translated. Based on the extracted source coherence chain, we adopt a maximum entropy classifier to predict the target coherence chain that defines a linear topic structure for the target document. The proposed topic-based coherence model then uses the predicted target coherence chain to help decoder select coherent word/phrase translations. Our experiments show that incorporating the topic-based coherence model into machine translation achieves substantial improvement over both the baseline and previous methods that integrate document topics rather than coherence chains into machine translation.

1 Introduction

During the last two decades, statistical machine translation (SMT) has made significant progresses on modeling translation of individual sentences under the independence assumption that a text can be translated sentence by sentence. However, just as words within a sentence are logically and syntactically related to each other, sentences in a text are also semantically connected. Hence, it is necessary for SMT to advance from sentence translation to document translation. One of the most important problems that we have to deal with in document translation is how to generate a coherent target text from a coherent source text where sentences are meaningfully connected. In other words, how do we produce **coherence** in target documents?

Coherence is a property of well-formed texts that establishes links in meaning between sentences to make texts easy to read and understand (rather than randomly selected sentences). Linguists (de Beaugrande and Dressler 1981) define the foundation of coherence as a "continuity of senses". In this work, we specialize and confine the sense continuity to a continuous sentence topic transition. In order to keep a continuous flow of senses in a coherent text, sentences within the text should have the same or similar topics and topic changes in adjacent sentences should also be smooth. This explanation of coherence is similar to the concept of *content* adopted by Barzilay and Lee (2004), who propose HMMs to model sentence topics and topic shifts in a text in order to capture coherence.

Given a coherent document, we can assign a topic for each sentence in the document. The coherent document can be therefore characterized as a sentence topic sequence in which topics are connected and topic changes are continuous. We refer to such a sentence topic sequence as the **coherence chain** of the document.

Based on the document coherence chain, we propose a framework to produce coherence in document translation. In the framework, we predict the coherence chain for a target document according to the coherence chain of its corresponding source document. Then we build a topic-based coherence model on the predicted chain to capture the internal connectedness of the target document at the level of sentence-to-sentence topic transitions. In particular, we

- *Generate a coherence chain for each source document before we translate it (Section 3).* We train a topic model (Gruber, Rosen-zvi, and Weiss 2007) on our training data and then use the trained topic model to infer sentence topics in each source document to be translated.
- *Predict the coherence chain of target document given the source coherence chain (Section 4).* In order to obtain the target document coherence chain, we project source sentence topics onto target sentences under the assumption that each source sentence is translated to only one target sentence and vice versa[1]. Therefore the coherence chain prediction can be recast as a sequence labeling problem. We use a maximum entropy model to predict the target coherence chain from the source document sentence topic sequence.
- *Incorporate the predicted target coherence chain into document translation (Section 5).* We present two topic-based coherence models using the predicted target coherence chain. The two models are integrated into decoder to help it select appropriate target words/phrases that are related to the estimated topics of target sentences in which

*Corresponding author

[1]This assumption is reasonable as we use sentence-aligned bilingual corpus.

these words/phrases occur. In doing so, we want the decoder to produce coherent translations throughout the target document.

We investigate the effectiveness of our topic-based coherence models on NIST Chinese-to-English translation tasks. The topic-based coherence models can be constructed either at the word level or at the phrase level. Our best performing method uses a Markov model of order 2 to predict target coherence chains and builds a topic-based coherence model at the phrase level. Experiment results show that the word level coherence model is able to improve the performance by 0.53 BLEU (Papineni et al. 2002) points and the phrase level model 0.61 BLEU points.

We also compare our model against document topic based model which uses the topic of a document for all sentences within the document. Previous work (Xiao et al. 2012; Su et al. 2012) that explores topic model (Blei et al. 2003) for SMT uses only document topic for translations. They do not distinguish sentences of a document in terms of their topics. Although many sentences share the same topic with the document where they occur, we observe that there are sentence topic changes within a document and that a lot of sentences actually do have topics different from those of their documents in our training data. Experiment results also suggest that our sentence topic based coherence model is better than the document topic based model.

2 Related Work

Coherence model for text analysis Although coherence is rarely explored in SMT, it is widely studied in text analysis. Various coherence models are proposed in the context of document summarization and generation, e.g., entity-based local coherence model (Barzilay and Lapata 2008), content-based global coherence models (Barzilay and Lee 2004; Fung and Ngai 2006) and syntax-based coherence model (Louis and Nenkova 2012).

Our definition of coherence is partly inspired by the content model (Barzilay and Lee 2004) as mentioned in Section 1. We also infer topics for sentences in each document. But our key interest is to project source sentence topics and topic shifts onto sentences of target texts and then use the projected topics for target word/phrase selection during translation. Therefore our model can be considered as a bilingual coherence model.

Inter-sentence dependencies for document translation Recently SMT researchers have proposed models to explore inter-sentence dependencies for document translation, such as cache-based language models (Tiedemann 2010; Gong, Zhang, and Zhou 2011). Hardmeier et al. (2012) introduce a document-wide phrase-based decoder and integrate a semantic language model into the decoder. These studies normally focus on lexical cohesion (e.g., word repetitions in adjacent sentences) rather than coherence which deals with underlying sense connectedness within a document.

Topic model for SMT Our model is also related to previous approaches that employ topic model for SMT (Zhao and Xing 2006; Xiao et al. 2012; Su et al. 2012), especially the topic similarity model (Xiao et al. 2012) which explores document topics for hierarchical phrase selection. However, our coherence model is significantly different from the topic similarity model in two key aspects. Firstly, we use sentence topics instead of document topics to select words/phrase for document translation. We observe in training data that a great number of sentences do have a topic which is different from their document topic. We therefore propose a coherence chain prediction model to estimate target sentence topics. Secondly, we build a coherence model based on topic-related probabilities rather than a similarity model on the rule-topic distribution. Although using the rule-topic distribution is able to include all possible topics in the similarity model, the model is significantly larger as the number of topics increases. Additionally, the distribution-based similarity model can not differentiate topic-insensitive phrases (Xiao et al. 2012).

3 Source Coherence Chain Generation

Given a source document D_s that consists of sentences s_1^n, we want to obtain topics not only for the document itself (z_{D_s}) but also for all sentences in the document (z_1^n). Currently the most popular Latent Dirichlet Allocation (LDA) (Blei et al. 2003) model only generates topics for words and documents, ignoring sentence topics. We therefore resort to the Hidden Topic Markov Model (HTMM) (Gruber, Rosenzvi, and Weiss 2007) which assumes that all words in the same sentence have the same topic and hence is able to learn topics for sentences within documents.

We adopt the HTMM open-source toolkit[2] to train an HTMM on our training data where document boundaries are explicitly given. HTMM parameters are estimated iteratively via EM algorithm. The trained HTMM is then used to infer Viterbi sentence topic sequence for each document. Table 1 shows an example of source document in Chinese. We do not list all sentences of the document for the sake of saving space. The listed sentences are labeled with topics generated by the HTMM. The first 5 sentences have the same topic 123 which is related to *government and military* while the 6th sentence has a different topic 46 which is about *love*. Although the majority of sentences of the document have the same topic 123, we observe a topic change between sentence 5 and 6.

Once topics for all sentences in a source document are obtained, we can generate the coherence chain of the document by simply extracting the sequence of sentence topics. For example, the coherence chain of the document shown in Table 1 is "123 123 123 123 123 46 ...".

The way that the HTMM captures topic transitions between sentences is similar to that of the content model (Barzilay and Lee 2004). Both of them employ Hidden Markov Models (HMM). Integrating Markovian relations, the HTMM is able to drop the "bag-of-words" assumption that topics for words are independently learned. But still like the LDA model, the HTMM organizes all parameters via a hierarchical generative model. The learned conditional probability $p(w_j|z_i)$ for a word w_j under its hidden topic z_i will

[2]Available at: http://code.google.com/p/openhtmm/.

Table 1: Sentence topics inferred by the HTMM on a source document (written in Chinese Pinyin followed by English translations). SID indicates the sentence ID.

SID	Topic	Sentence
1	123	balin gongzhu xia jia meidabing jing shi hunyin wu nian xuangao polie // Bahraini Princess Marries US Soldier, Astonishing 5-Year Bond Comes to End
...	...	...
5	123	tamen liang ren zai yijiujiujiunian xiangyu, dangshi, qiangsheng hai shi zhiye junren, paizhu zai balin. // The pair met in 1999 when career military man Johnson was stationed in Bahrain.
6	46	ta renshi zhege doukou nianhua de xiao gongzhu hou, liang ren cha chu ai de huohua, ta de shengming yiner chuxian jubian. // But his life changed dramatically when he met the beautiful teenage princess and the pair fell in love.
...	...	...

be used in our topic-based coherence model (Section 5).

4 Target Coherence Chain Prediction

When we translate a coherent source document D_s, we want the generated target document D_t to be coherent too. In order to produce coherence in D_t, we can use the coherence chain of D_t to help decoder select words and phrases that are coherent. However, we can not directly infer the target coherence chain via the HTMM as the target document D_t is yet to be generated.

Fortunately, we can obtain the source coherence chain as described in the last section. It is widely accepted that the target document translation should be meaningfully faithful to the source document. Thus, corresponding sentences between the source and target document should have equivalent topics. If a topic change happens in the source coherence chain, a similar topic shift should also occur in the target coherence chain. This suggests that we can predict the target coherence chain based on its counterpart on the source side. We further assume a one-to-one mapping between sentence topics in the source/target coherence chain. Therefore the target coherence chain prediction is actually a sequence labeling problem, in which the source coherence chain is the observation sequence while the target chain is the hidden state sequence to be predicted.

Yet another way to generate the target coherence chain is to learn a bilingual topic model (Mimno et al. 2009). Since only source-side documents are available during decoding, marginalization is required to infer monolingual topics for source-side documents and sentences using the learned bilingual topic model. We do not adopt this method due to this computation-intensive marginalization, which is explained in detail by Xiao et al. (2012).

In this section, we introduce our projection method, including the prediction model, features used in the model and the training procedure.

4.1 Prediction Model

Given a source coherence chain $z_1^n = z_1, ..., z_n$ along with the source document topic z_{D_s}, we choose the target coherence chain $\mathbf{z}_1^n = \mathbf{z}_1, ..., \mathbf{z}_n$ with the highest probability among all possible chains.

$$\hat{\mathbf{z}}_1^n = \underset{\mathbf{z}_1^n}{\operatorname{argmax}}\, Pr(\mathbf{z}_1^n | z_1^n, z_{D_s}) \tag{1}$$

Note that a source topic (value of z_i) may align to different target topics (value of $\mathbf{z}_i$) and vice versa in the training data (Xiao et al. 2012). The posterior probability $Pr(\mathbf{z}_1^n | z_1^n, z_{D_s})$ is factorized and modeled under a Markov assumption as follows.

$$Pr(\mathbf{z}_1^n | z_1^n, z_{D_s}) \approx \prod_{i=1}^{n} p(\mathbf{z}_i | \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) \tag{2}$$

That is, we determine the hidden state $\mathbf{z}_i$ according to its preceding k states $\mathbf{z}_{i-k}^{i-1}$, a 5-sentence window z_{i-2}^{i+2} centered at the current observed source sentence topic z_i and the source document topic z_{D_s}. We set k to 0/1/2 and the model is referred to as the prediction model of order 0/1/2 correspondingly.

We use a maximum entropy classifier to estimate the probability $p(\mathbf{z}_i | \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$, which is calculated as follows.

$$\begin{aligned} & p(\mathbf{z}_i | \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) \\ = & \frac{exp(\sum_m \theta_m h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}))}{\sum_{\mathbf{z}_i'} exp(\sum_m \theta_m h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}))} \end{aligned} \tag{3}$$

where $h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$ are binary valued feature functions and θ_m are weights for these feature functions.

4.2 Features

We have integrated the following features into the prediction model.

Source sentence topic features: Source sentence topics are used in the features formulated as follows.

$$\begin{aligned} & h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) \\ = & \begin{cases} 1, & \text{if } z_{i+d} = z_s \text{ and } \mathbf{z}_i = z_t \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where $d \in \{-2, ..., 2\}$. The features will be fired if the source sentence topic z_{i+d} is z_s and the prediction for the current target sentence topic equals z_t. Note that z_s is not necessarily the same as z_t. Even if they are equal to each other, they may represent different topics as we infer sentence topics on the source and target side separately (see the next subsection).

Source document topic feature: We also use the source document topic to predict the target document coherence chain as follows.

$$\begin{aligned} & h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) \\ = & \begin{cases} 1, & \text{if } z_{D_s} = z_D \text{ and } \mathbf{z}_i = z_t \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Target sentence topic transition features: We use these features to capture the dependence on preceding target sentence topics.

$$h_m(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) = \begin{cases} 1, & \text{if } \mathbf{z}_{i-k} = z_{t'} \text{ and } \mathbf{z}_i = z_t \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

If $k = 0$, the features will be not used. That means that the topic of the current target sentence is estimated independent of topics of preceding target sentences.

4.3 Training

In order to train the maximum entropy classifier in equation (3), we need to collect training events $(\mathbf{z}_i, \mathbf{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$ from aligned source/target coherence chains. We first infer all sentence topics on source documents in our bilingual training data using the HTMM as described in Section 3. Similarly, we also train an HTMM on target documents and use the trained HTMM to infer sentence topics of target documents in our training data. Once we complete the sentence topic inference on both source and target documents, we can extract coherence chains for all aligned source/target documents. From these extracted coherence chain pairs, we generate the features as introduced in the last section. Finally, we train the maximum entropy classifier via the off-the-shelf MaxEnt toolkit[3].

5 Topic-Based Coherence Model

Once we predict the target coherence chain $\mathbf{z}_1^n$ for target document D_t, we can use the coherence chain to provide constraints for the target document translation. Our key interest is to make the target document translation T_{D_t} as coherent as possible. We use the conditional probability $Pr(T_{D_t}|\mathbf{z}_1^n)$ to measure the coherence of the target document translation. As we define the coherence as a continuous sense transition over sentences within a document, the probability is factorized as follows.

$$Pr(T_{D_t}|\mathbf{z}_1^n) \approx \prod_{i=1}^{n} p(T_i|\mathbf{z}_i) \quad (7)$$

where T_i is the translation of the ith sentence in the target document.

The probability $p(T_i|\mathbf{z}_i)$ estimates the relatedness between the sentence translation and its corresponding topic $\mathbf{z}_i$ in the continuous sense chain of the target document. We can further factorized this probability by decomposing the sentence translation into words or phrases.

Word level coherence model (WCM): The probability $p(T_i|\mathbf{z}_i)$ is further factorized into the topic probability over words as follows.

$$Pr(T_{D_t}|\mathbf{z}_1^n) \approx \prod_{i=1}^{n} p(T_i|\mathbf{z}_i) \approx \prod_{i=1}^{n} \prod_{j} p(w_j|\mathbf{z}_i) \quad (8)$$

The topic-word probability $p(w_j|\mathbf{z}_i)$ can be directly obtained from the outputs of the HTMM. As we discard all stop words when training the HTMM, stop words occurring in the sentence translation T_i are therefore ignored.

[3]Available at: http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

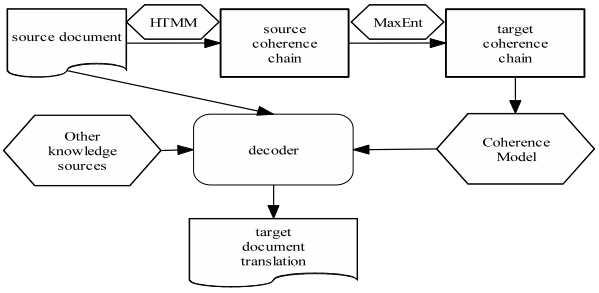

Figure 1: Architecture of SMT system with the topic-based coherence model.

Phrase level coherence model (PCM): We can also factorize $p(T_i|\mathbf{z}_i)$ at the phrase level.

$$Pr(T_{D_t}|\mathbf{z}_1^n) \approx \prod_{i=1}^{n} p(T_i|\mathbf{z}_i) \approx \prod_{i=1}^{n} \prod_{j} p(r_j|\mathbf{z}_i) \quad (9)$$

where r_j are phrases. Since the number of phrases is much larger than that of words, data sparseness becomes a problem when estimating the probability distribution of topic $\mathbf{z}_i$ over phrases r_j. We actually use the probability $p(\mathbf{z}_i|r_j)$ instead of $p(r_j|\mathbf{z}_i)$ in our phrase level coherence model. This is reasonable as both $p(\mathbf{z}_i|r_j)$ and $p(r_j|\mathbf{z}_i)$ measure the relatedness of phrase r_j to topic $\mathbf{z}_i$.

Data sparseness in the estimation of $p(\mathbf{z}_i|r_j)$ is under control as the number of topics is normally smaller than 1000. In order to calculate $p(\mathbf{z}_i|r_j)$, we annotate phrases with topic $\mathbf{z}_i$ when these phrases are extracted from sentence T_i. The probability $p(\mathbf{z}_i|r_j)$ is estimated using smoothed counts:

$$p(\mathbf{z}_i|r_j) = \frac{f(r_j, \mathbf{z}_i) + 1}{f(r_j) + K} \quad (10)$$

where $f(\cdot)$ denotes the frequency and K is the total number of topics.

After the factorization at the word/phrase level, the topic-based coherence model can be directly integrated into SMT decoder just like the lexical/phrasal translation probability model in phrase-based SMT (Koehn, Och, and Marcu 2003). Figure 1 shows the architecture of SMT system with the topic-based coherence model.

6 Experiments

In this section, we conducted a series of experiments to validate the effectiveness of our topic-based coherence models on NIST Chinese-English translation tasks trained with large-scale data. In particular, we aim at: 1) Measuring the impacts of two parameters on our coherence models: the number of topics K and the Markov order k of the prediction model (Section 4), 2) Investigating the effects of the word and phrase level coherence model, and 3) Comparing our coherence model against the document topic based model.

6.1 Setup

Our baseline system is a state-of-the-art BTG-based phrasal system which adopts Bracketing Transduction Grammars (BTG) (Wu 1997) for phrasal translation and a maximum entropy based reordering model for phrasal reordering (Xiong, Liu, and Lin 2006). We integrate the proposed word level and phrase level coherence model into this system.

Our training data (including LDC2002E18, LDC2003E07, LDC2003E14, LDC2004E12, LDC2004T07, LDC2004T08 (Only Hong Kong News), LDC2005T06 and LDC2005T10) consists of 3.8M sentence pairs with 96.9M Chinese words and 109.5M English words. We used a 5-gram language model which was trained on the Xinhua section of the English Gigaword corpus (306 million words) using the SRILM toolkit (Stolcke 2002) with modified Kneser-Ney smoothing.

In order to train the HTMM and the coherence chain prediction model, we selected corpora LDC2003E14, LDC2004T07, LDC2005T06 and LDC2005T10 from our bilingual training data, where document boundaries are explicitly provided. We also used all data from the corpus LDC2004T08 (Hong Kong Hansards/Laws/News). In total, our training data for the coherence chain prediction model contain 103,236 documents with 2.80M sentences. When training the HTMM by EM algorithm, we set the hyper parameters $\alpha = 1 + 50/K$ and $\eta = 1.01$ according to the values used by Gruber et al. (Gruber, Rosen-zvi, and Weiss 2007). We performed 100 iterations of the L-BFGS algorithm implemented in the MaxEnt toolkit with both Gaussian prior and event cutoff set to 1 to train the prediction model (Section 4).

We adopted the NIST MT03 evaluation test data as our development set, and the NIST MT05 as the test set. The numbers of documents in MT03/05 are 100/100 respectively. We used the case-insensitive BLEU-4 (Papineni et al. 2002) and NIST (Doddington 2002) to evaluate translation quality. In order to alleviate the impact of MERT (Och 2003) instability, we followed the suggestion of Clark et al. (Clark et al. 2011) to run MERT three times and report average BLEU/NIST scores over the three runs for all our experiments.

6.2 Impacts of the Number of Topics and the Markov Order of the Prediction Model

Our first group of experiments were carried out to study the impacts of two important parameters on our coherence model: the number of topics K and the Markov order k. The former parameter determines the granularity of senses and sense changes that are allowed in document translation. The latter specifies whether to capture the dependencies on the topics of preceding sentences in the coherence chain prediction. We evaluated the impacts of both parameters on the word level coherence model by setting $K \in \{100, 150, 200\}$ and $k \in \{0, 1, 2\}$. The results are shown in Table 2. From the table, we can observe that

- When we increase the number of topics K from 100 to 150, the average BLEU/NIST scores on the three different Markov order settings are improved from 0.3429/9.4121 to 0.3457/9.4227. However, when K is further increased to 200, the average BLEU/NIST scores drop to 0.3441/9.4113 respectively. The reason may be that the probability distribution of topic transitions is becoming sparse when the number of topics K is large.
- As we increase the Markov order k from 0 to 1, the performance of the word level coherence model first drops. However, when k is set to 2, both BLEU and NIST scores rise and are higher on average than those scores when k is 0. This indicates that capturing topic dependencies helps the coherence chain prediction model, which in turn benefits the topic-based coherence model.

Table 2: BLEU and NIST scores of the word level coherence model on the development set with topic number K varying from 100 to 200 and the Markov order k of the prediction model from 0 to 2.

	$k=0$	$k=1$	$k=2$	Avg
$K=100$	0.3431 9.4787	0.3390 9.3481	0.3466 9.4094	0.3429 9.4121
$K=150$	0.3461 9.3817	0.3443 9.4200	**0.3466** **9.4665**	0.3457 9.4227
$K=200$	0.3456 9.4529	0.3422 9.3359	0.3444 9.4452	0.3441 9.4113

The best performance is obtained when we set $K = 150$ and $k = 2$. This setting is used for our coherence models in all experiments thereafter.

6.3 Effects of the Word and Phrase Level Coherence Model

The second group of experiments aims at investigating and comparing the effects of the word and phrase level coherence models. Table 3 presents the results. The word level coherence model outperforms the baseline by an absolute 0.53 BLEU points while the phrase level achieves a larger improvement of 0.61 BLEU points over the baseline on the test set. NIST scores obtained by the two coherence mdoels are also much higher than that of the baseline. These suggest that the proposed topic-based coherence models are able to improve document translation quality by selecting coherent word/phrase translations that are related to their corresponding sentence topics.

Table 4 gives an example showing how the topic-based coherence model improves document translation quality. The bold Chinese word "dongzuo" has two different meanings (original meaning and derived sense), which can be translated into English word *movement* (of body) and *action* respectively. According to the meaning of the word in this given sentence, *action* is a better translation for it. The topic of this sentence in the predicted target coherence chain is 19, whose probability distribution over words is shown in Table 5. Clearly, the distribution probability over word *action* is much higher than that of word *movement*. Therefore our coherence model is able to select the translation *action* for the source word instead of the translation *movement*.

Table 3: BLEU and NIST scores of the word/phrase level coherence models on the test set. WCM/PCM ($\mathbf{z}_1^n$): the word/phrase level coherence model based on the target document coherence chain $\mathbf{z}_1^n$; WCM/PCM (z_{D_t}): the degenerated word/phrase level coherence model only using the target document topic z_{D_t}.

	BLEU	NIST
Base	0.3393	9.1639
WCM ($\mathbf{z}_1^n$)	0.3446	9.3699
PCM ($\mathbf{z}_1^n$)	0.3454	9.3746
WCM (z_{D_t})	0.3387	9.3023
PCM (z_{D_t})	0.3404	9.3368

Table 4: A Chinese (shown in pinyin) to English translation example showing the difference between the baseline translation (Base) and the translation generated by the system enhanced with our coherence model (WCM ($\mathbf{z}_1^n$)).

src	zhunbei gongzuo jiang hui jinxing dao qiyue, ranhou zai zhankai zhengzhi **dongzuo**
Base	preparatory work will be carried out until July , and then launched a political **movement**
WCM ($\mathbf{z}_1^n$)	preparatory work will be carried out until July , then a political **action**
ref	preparations would take place until July, after which political **action** will begin

6.4 Coherence Chain vs. Document Topic

In the last group of experiments, we investigated whether it is necessary to use sentence topic sequences (coherence chains) instead of document topics in our coherence model. We observe that 40.86% of sentences in our development/test sets have topics that are different from topics of documents where these sentences belong.

In order to study the impact of these sentences with topics different from their document topics, we design a model which only uses the topic of target document z_{D_t} rather than the target coherence chain $\mathbf{z}_1^n$ to select translations for words/phrases. The new model can be considered as a degenerated variation of our proposed coherence model. It can be formulated and factorized as follows.

$$Pr(T_{D_t}|z_{D_t}) \approx \prod_{i=1}^{n} p(T_i|z_{D_t}) \tag{11}$$

The probability $p(T_i|z_{D_t})$ is further factorized at the word and phrase level, similarly to equation (8) and (9)

We still use a maximum entropy classifier to predict the target document topic given its source document topic with the following feature:

$$h_m(z_{D_t}, z_{D_s}) = \begin{cases} 1, & \text{if } z_{D_s} = z_D \text{ and } z_{D_t} = z'_D \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

The results are shown in the last two rows in Table 3. We can clearly observe that BLEU/NIST scores of both the word and phrase level coherence models significantly drop on the test set when using the target document topic for all sentences. This suggests that the coherence chain based model is better than document topic based model.

Table 5: Ten most probable words for topic 19. We also show the probability of the topic 19 over word *action* and *movement*. $p = p(w|z_i = 19)$.

Word	p	Word	p
united	0.0209182	russia	0.00637757
states	0.0203053	security	0.00617798
china	0.00922345	international	0.00601291
countries	0.00842481	...	...
military	0.00749308	action	0.000886684
defense	0.00702691	...	...
bush	0.00658136	movement	0.000151846

7 Conclusions

We have presented a topic-based coherence model for statistical machine translation at the document level. Our method uses a Markovian topic model to generate a coherence chain for a source document and projects the source coherence chain onto the corresponding target document by a MaxEnt-based prediction model. The projected coherence chain captures topic-related constraints on word/phrase selection for the target document translation. Integration of the topic-based coherence model into phrase-based machine translation yields significant improvement over the baseline.

We have also observed 1) that the phrase level coherence model is marginally better than the word level coherence model and 2) that our coherence models significantly outperform the degenerated coherence model which only uses target document topic to constrain word/phrase translations.

We address the text coherence for document translation from the lexical and topical perspective. There exists yet another dimension of coherence: intentional structure that is concerned with the purpose of discourse. Louis and Nenkova (2012) find that syntactic patterns shared by a sequence of sentences in a text are able to capture intentional structure. Therefore an important future direction lies in studying and modeling the intentional structure dimension of coherence for syntax-based machine translation (Galley et al. 2006) that uses syntactical rules to generate translations. By automatically learning syntactic patterns and intentional coherence embedded in these patterns from large-scale training data with parse trees, we may be able to select syntactic translation rules in a more efficient and appropriate fashion.

We only model sentence topics and their changes in the content structure of a text. There are many other important relations, such as rhetorical relations (Lin, Ng, and Kan 2011), which should also be considered when translating a text. Furthermore, the discourse structure is frequently modeled hierarchically in the literature. Therefor we also plan to incorporate more hierarchical discourse information into phrase/syntax-based machine translation at the document level in the future.

References

Barzilay, R., and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1):1–34.

Barzilay, R., and Lee, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In Susan Dumais, D. M., and Roukos, S., eds., *HLT-NAACL 2004: Main Proceedings*, 113–120.

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and Lafferty, J. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Clark, J. H.; Dyer, C.; Lavie, A.; and Smith, N. A. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 176–181.

de Beaugrande, R.-A., and Dressler, W. 1981. *Introduction to Text Linguistics*.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, 138–145.

Fung, P., and Ngai, G. 2006. One story, one flow: Hidden markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing* 3(2):1–16.

Galley, M.; Graehl, J.; Knight, K.; Marcu, D.; DeNeefe, S.; Wang, W.; and Thayer, I. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 961–968.

Gong, Z.; Zhang, M.; and Zhou, G. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 909–919.

Gruber, A.; Rosen-zvi, M.; and Weiss, Y. 2007. Hidden topic markov models. In *In Proceedings of Artificial Intelligence and Statistics*.

Hardmeier, C.; Nivre, J.; and Tiedemann, J. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1179–1190.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 58–54.

Lin, Z.; Ng, H. T.; and Kan, M.-Y. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 997–1006.

Louis, A., and Nenkova, A. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1157–1168.

Mimno, D.; Wallach, H. M.; Naradowsky, J.; Smith, D. A.; and McCallum, A. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 880–889.

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

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Stolcke, A. 2002. Srilm–an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 901–904.

Su, J.; Wu, H.; Wang, H.; Chen, Y.; Shi, X.; Dong, H.; and Liu, Q. 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)*, 459–468.

Tiedemann, J. 2010. To cache or not to cache? experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 189–194.

Wu, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3):377–403.

Xiao, X.; Xiong, D.; Zhang, M.; Liu, Q.; and Lin, S. 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)*, 750–758.

Xiong, D.; Liu, Q.; and Lin, S. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 521–528.

Zhao, B., and Xing, E. P. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 969–976.