Topic-Based Coherence Modeling for Statistical Machine Translation

Deyi Xiong, Min Zhang, *Member, IEEE*, and Xing Wang

***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 topic-based coherence models 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. We build two topic-based coherence models on the predicted target coherence chain: 1) a word level coherence model that helps the decoder select coherent word translations and 2) a phrase level coherence model that guides the decoder to select coherent phrase translations. We integrate the two models into a state-of-the-art phrase-based machine translation system. Experiments on large-scale training data show that our coherence models achieve substantial improvements over both the baseline and models that are built on either document topics or sentence topics obtained under the assumption of direct topic correspondence between the source and target side. Additionally, further evaluations on translation outputs suggest that target translations generated by our coherence models are more coherent and similar to reference translations than those generated by the baseline.**

***Index Terms*—Text coherence, text analysis, coherence chain, topic modeling, statistical machine translation (SMT), natural language processing.**

I. Introduction

UNDER an assumption that sentences of a text can be translated independently of each other, statistical machine translation (SMT) has made substantial progresses on modeling sentence-level translation over the last two decades. However, just as words within a sentence are logically and syntactically related to each other, sentences in a text are also semantically connected. The neglect of such inter-sentence semantic connectedness will hurt the coherence of the target document generated from a coherent source document where sentences are meaningfully connected.

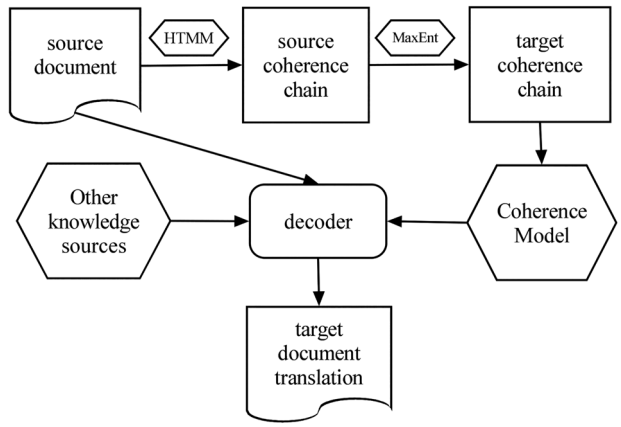

Fig. 1. Architecture of SMT system with the topic-based coherence model.

Coherence, establishing links in meaning between sentences, is an important property of well-formed texts. It makes texts cohesive and easy to read and understand, rather than a random group of sentences. Linguists de Beaugrande and Dressler [1] define the foundation of coherence as a "continuity of senses". In this article, 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 [2], who propose HMMs to model sentence topics and topic shifts in a text in order to capture coherence.

We can assign a topic for each sentence in a coherent 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 capture coherence for statistical machine translation.

Since the corresponding target document of a coherent source document is yet to be generated, we need to predict the coherence chain for the target document according to the coherence chain of its source document in our framework. Once we have the predicted target document coherence chain, we can build topic-based coherence models on it. Our key interest is to capture the internal connectedness of the target document at the level of sentence-to-sentence topic transitions for translation. The whole framework is visualized in Fig. 1.

Manuscript received July 31, 2014; revised October 19, 2014; accepted January 19, 2015. Date of current version February 26, 2015. This work was supported in part by the National Natural Science Foundation of China under Grants 61403269, 61432013, and 61333018, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20140355, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Helen Meng. *(Corresponding author: M. Zhang.)*

D. Xiong and M. Zhang were with the Institute for Infocomm Research, Singapore 138632. They are now with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China (e-mail: dyxiong@suda.edu.cn; minzhang@suda.edu.cn).

X. Wang is with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China (e-mail: xingwsuda@gmail.com).

Digital Object Identifier 10.1109/TASLP.2015.2395254

Specifically, first, we train a sentence topic model HTMM[1] [3] on our training data. The trained topic model is able to infer topics for individual sentences in a text. We use this topic model to produce a coherence chain for each source document to be translated. Second, we predict the target document coherence chain given the source document coherence chain. As each source sentence is translated to only one target sentence and vice versa in our training data, the coherence chain prediction can be recast as a sequence labeling problem. We use a maximum entropy (MaxEnt) model to project source sentence topics in the generated source coherence chain onto target sentences. Finally, we incorporate the predicted target coherence chain into document translation via two proposed topic-based coherence models. The two models are integrated into the decoder to help it select appropriate target words/phrases that are related to the estimated topics of target sentences in which these words/phrases occur. In doing so, we want the decoder to produce coherent translations throughout target documents.

We investigate the effectiveness of our coherence models on NIST Chinese-to-English translation. The coherence model 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 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 [4] points and the phrase level model 0.61 BLEU points.

We also compare our models against a document topic based translation model which uses the topic of a document for all sentences within the document. Previous work [5], [6] that explores topic model [7] 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 (1) that there are sentence topic changes within a document and (2) that a lot of sentences actually do have topics different from those of their documents in our training data.[2] Experiment results also suggest that our topic-based coherence model using sentence topics is better than the document topic based translation model.

The topic-based coherence model has been presented in our previous paper [8]. In this article, we make the following significant extensions to our previous work.

- We carry out new experiments to investigate whether our MaxEnt-based topic projection from the source to the target side is better than a bilingual topic model based on the assumption of direct topic correspondence between the source and target side. Such a bilingual topic model does not require the MaxEnt-based topic projection, i.e., the second step in our framework in Fig. 1.
- We conduct an in-depth analysis to disclose how the coherence models improve translation quality through intrinsic and extrinsic evaluations on translation outputs.
- We provide more details, such as the accuracy of the MaxEnt-based prediction model as well as the model size and decoding speed of the proposed coherence models. These details will help us look inside the topic-based coherence models.

The remainder of this article proceeds as follows. Section II elaborates the two topic-based coherence models. Section III presents the prediction model that we use to predict target coherence chains from generated source coherence chains, as well as features that are used in the prediction model. Section IV introduces how we generate coherence chains for source documents. Section V evaluates the topic-based coherence models with large-scale training data on Chinese-to-English translation. Section VI deeply investigates how the incorporated coherence model improves translation quality. Section VII introduces related work and highlights the differences of our coherence models from previous approaches. Finally we conclude in Section VIII with future directions.

II. Topic-Based Coherence Model

In this section, we describe our topic-based coherence models. 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 the decoder select words and phrases that are coherent. Let us first suppose that we have already predicted the target coherence chain $\underline{z}_1^n = \{\underline{z}_1, \ldots, \underline{z}_n\}$ for the target document D_t from the coherence chain of its source document. We can use this coherence chain to provide constraints for the target document translation.

Our goal is to make the target document D_t as coherent as possible. We use the conditional probability $Pr(D_t|\underline{z}_1^n)$ 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:

$$Pr(D_t|\underline{z}_1^n) \approx \prod_{i=1}^{n} p(D_t^i|\underline{z}_i) \tag{1}$$

where D_t^i is the ith sentence in the target document.

The probability $p(D_t^i|\underline{z}_i)$ estimates the relatedness between the sentence translation D_t^i and its corresponding topic $\underline{z}_i$ in the continuous sense chain of the target document. We can further factorize this probability by decomposing the sentence translation into words or phrases. Correspondingly, we propose two topic-based coherence models: word and phrase level coherence model.

Word level coherence model (WCM). The probability $p(D_t^i|\underline{z}_i)$ is further factorized into topic probabilities over words as follows:

$$Pr(D_t|\underline{z}_1^n) \approx \prod_{i=1}^{n} p(D_t^i|\underline{z}_i) \approx \prod_{i=1}^{n} \prod_{j} p(w_j|\underline{z}_i) \tag{2}$$

where w_j are words in D_t^i. The topic-word probability $p(w_j|\underline{z}_i)$ can be directly obtained from the outputs of the trained topic model (see Section IV). As we discard all stop words when training our topic model[3], stop words occurring in the sentence translation D_t^i are therefore ignored.

[1]See Section IV for more details.

[2]For more details, see Section V-E

[3]English stop words are from http://snowball.tartarus.org/algorithms/english/stop.txt, As for Chinese stop words, we obtain them from our training data according to word frequency.

Phrase level coherence model (PCM). We can also factorize $p(D_t^i|\underline{z}_i)$ at the phrase level as follows:

$$Pr(D_t|\underline{z}_1^n) \approx \prod_{i=1}^{n} p(D_t^i|\underline{z}_i) \approx \prod_{i=1}^{n}\prod_{j} p(r_j|\underline{z}_i) \quad (3)$$

where r_j are target phrases that are used to generate translation D_t^i. Since the number of phrases is much larger than that of words, we have to consider data sparseness problem when estimating the probability distribution of topic $\underline{z}_i$ over phrases r_j. Instead of directly estimating $p(r_j|\underline{z}_i)$ in our phrase level coherence model, we actually calculate the probability $p(\underline{z}_i|r_j)$. This is reasonable as both $p(\underline{z}_i|r_j)$ and $p(r_j|\underline{z}_i)$ measure the relatedness of phrase r_j to topic $\underline{z}_i$.

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

$$p(\underline{z}_i|r_j) = \frac{f(r_j, \underline{z}_i) + 1}{f(r_j) + K} \quad (4)$$

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 [10], as shown in Fig. 1.

III. Target Coherence Chain Prediction

In the previous section, we assume that the coherence chain of a target document is available before the decoder generates the target document. So how can we obtain the target document coherence chain before decoding? We cannot directly infer the target coherence chain via topic models trained on texts of the target language as the target document D_t is yet to be generated.

However, we can obtain the coherence chain of its corresponding source document D_s with trained topic models. 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 comparable 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 sentences in the source/target document[4]. 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 which is similar to that proposed in [9]. This bilingual model learns topics for the source and target language in the same topic space, where aligned source and target sentences share the same topics. We can approximate this bilingual topic model with a "bilingual trained" topic model by running our monolingual topic model on modified bilingual training data. In Section V-C, we will empirically compare our target coherence chain generation strategy against this approximate bilingual topic model based coherence chain generation.

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

A. Prediction Model

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

$$\underline{z}_1^n = \arg\max_{\underline{z}_1^n} Pr(\underline{z}_1^n|z_1^n, z_{D_s}) \quad (5)$$

Note that a source sentence topic (value of z_i) may align to different target topics (value of $\underline{z}_i$) and vice versa in training data [5]. This is because we separately train two sentence topic models on the source and target language: one is use to infer source coherence chains z_1^n and the other for target coherence chains $\underline{z}_1^n$. Therefore there is no direct one-to-one topic correspondence between topics inferred by these two separately trained topic models. However, these topics are connected to each other by document and sentence alignments. Using these alignments, we can project topics from the source topic space to the target topic space.

The posterior probability $Pr(\underline{z}_1^n|z_1^n, z_{D_s})$ is factorized and modeled under a Markov assumption as follows:

$$Pr(\underline{z}_1^n|z_1^n, z_{D_s}) \approx \prod_{i=1}^{n} p(\underline{z}_i|\underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) \quad (6)$$

That is, we determine the hidden state $\underline{z}_i$ according to its preceding k states $\underline{z}_{i-k}^{i-1}$, 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 $p(\underline{z}_i|\underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$, which is calculated as follows:

$$p(\underline{z}_i|\underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) = \frac{exp(\sum_m \theta_m h_m(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}))}{\sum_{\underline{z}_i'} exp(\sum_m \theta_m h_m(\underline{z}_i', \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}))} \quad (7)$$

where $\underline{z}_i'$ represents a possible topic of the ith target sentence, $h_m(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$ are binary valued feature functions which will be introduced in the next subsection, and θ_m are weights for these feature functions.

B. Features

We have integrated the following features into the prediction model.

Source sentence topic features. Source sentence topics are used to create features formulated as follows:

$$h_m(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{Ds}) = \begin{cases} 1, & \text{if } z_{i+d} = a \text{ and } \underline{z}_i = b \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

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

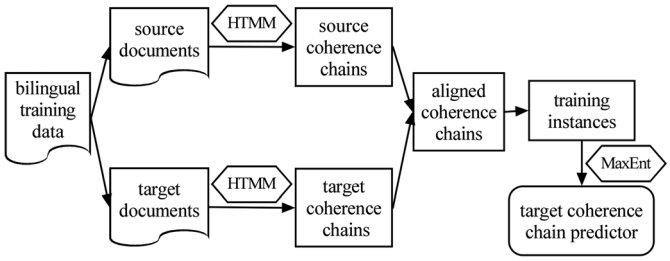

Fig. 2. The training process of the target coherence chain predictor.

where $d \in \{-2, \ldots, 2\}$, a and b are a specific topic ID of the source and target language, respectively. The feature will be fired if the source sentence topic z_{i+d} is a and the prediction for the current target sentence topic equals b. Note that a is not necessarily the same as b. 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 z_{D_s} to predict the target document coherence chain as follows:

$$h_m(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) = \begin{cases} 1, & \text{if } z_{D_s} = a \text{ and } \underline{z}_i = b \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The feature will be fired if the source document topic z_{D_s} equals a and the prediction for $\underline{z}_i$ is b.

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

$$h_m(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s}) = \begin{cases} 1, & \text{if } \underline{z}_{i-k} = a \text{ and } \underline{z}_i = b \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

This feature will be fired if the topics of the $(i-k)$th and ith target sentences are a and b respectively. If $k = 0$, the feature will be not used. It means that the topic of the current target sentence is estimated independent of topics of preceding target sentences.

C. Training

The process of training the maximum entropy based predictor shown in equation (7) is visualized in Fig. 2. In order to train the predictor, we need to collect training instances $(\underline{z}_i, \underline{z}_{i-k}^{i-1}, z_{i-2}^{i+2}, z_{D_s})$ from aligned source/target coherence chains. Particularly,

- We first train a source sentence topic model HTMM on the source language and use the trained source topic model to infer all sentence topics on source documents in our bilingual training data. The details of this procedure will be described in the next section.
- 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.

TABLE I
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.
...	...	...

- From these extracted coherence chain pairs, we collect training instances and generate the features as described in the last subsection.
- Finally, we train the maximum entropy classifier via the off-the-shelf MaxEnt toolkit[5].

IV. SOURCE COHERENCE CHAIN GENERATION

The last question about our proposed topic-based coherence framework is how we generate coherence chains for source documents. Given a source document D_s that consists of sentences $\{D_s^i\}_{i=1}^n$, 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) [7] model only generates topics for words and documents, ignoring sentence topics. We therefore resort to the Hidden Topic Markov Model (HTMM) [3] 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[6] to train an HTMM on our training data where document boundaries are explicitly given. HTMM parameters are estimated iteratively via the EM algorithm. The trained HTMM is then used to infer Viterbi sentence topic sequence for each document. Table I 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 I 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 [2]. Both of

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

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

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 given its hidden topic z_i is used in our word level coherence model (Section II).

V. Experiments

In this section, we conducted a series of experiments to evaluate the proposed topic-based coherence models on NIST Chinese-English translation trained with large-scale data. In particular, we aim at the following tasks.

- Measuring the impact of two parameters on our coherence models: the number of topics K and the Markov order k of the prediction model (See Section III).
- Comparing the coherence models built on target sentence topics learned by our prediction model against that on topics learned by an approximate bilingual topic model.
- Investigating the effect of the word and phrase level coherence models.
- Comparing our coherence models against the document topic based model.

We also investigated the size of the proposed coherence models and the decoding speed of systems with the coherence models.

A. Setup

Our baseline system is a state-of-the-art BTG-based phrasal system which adopts Bracketing Transduction Grammars (BTG) [11] for phrasal translation [12]. We integrate the proposed 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 [13] with modified Kneser-Ney smoothing.

In order to train the HTMM and the coherence chain prediction model, we selected 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 and News). In total, our training data for the coherence chain prediction model contain 103,236 documents with 2.80M sentences. When we train the HTMM with the EM algorithm, we set the hyper parameters $\alpha = 1 + 50/K$ and $\eta = 1.01$ according to the values used by Gruber *et al.* [3]. 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 III).

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 [4] and NIST [14] to evaluate translation quality. In order to alleviate the impact of MERT [15] instability, we followed the suggestion of Clark *et al.* [16] to run MERT three times and report average BLEU/NIST scores over the three runs for all our experiments.

TABLE II
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	0.3390	0.3466	0.3429
	9.4787	9.3481	9.4094	9.4121
$K=150$	0.3461	0.3443	**0.3466**	0.3457
	9.3817	9.4200	**9.4665**	9.4227
$K=200$	0.3456	0.3422	0.3444	0.3441
	9.4529	9.3359	9.4452	9.4113

B. The Number of Topics and the Markov Order of the Prediction Model

We first investigated the impact of two important parameters: 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 impact 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 II. 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 sparser when the number of topics K is larger.
- 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 when K is less than 200, which in turn benefits the coherence model.

These findings are consistent with the training accuracies of the MaxEnt-based coherence chain prediction model as shown in Table III. If we do not use any topic dependency information, the prediction accuracy drops as the topic number K ranges from 100 to 200. If we use topic information from previous sentences, the prediction accuracy increases when K changes from 100 to 150 and then drops when K is 200. This is reasonable since richer information (e.g., previous sentence topics) will improve prediction accuracy while more classes (topics) to be predicted will tend to decrease the prediction accuracy.

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.

TABLE III
TRAINING ACCURACY (%) OF THE MAXENT-BASED COHERENCE CHAIN PREDICTION MODEL

	$k=0$	$k=1$	$k=2$
$K=100$	52.90	73.88	74.11
$K=150$	52.45	73.97	**74.21**
$K=200$	51.90	72.10	72.34

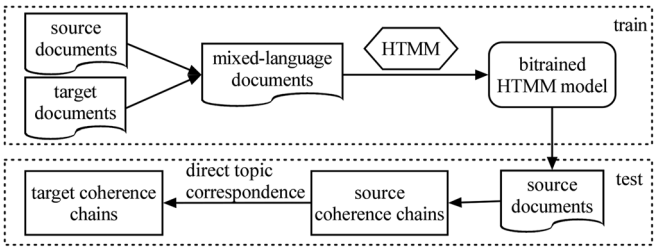

Fig. 3. The training and testing process of the bitrained HTMM model.

C. Sentence Topic Projection vs. Direct Topic Correspondence

One may wonder why we do not assume a direct topic correspondence between source and target sentences so that we do not need to train a MaxEnt-based prediction model for sentence topic projection. In order to investigate this question, we carried out experiments to compare these two strategies (sentence topic projection vs. direct topic correspondence) using the word level coherence model. Based on the assumption of direct topic correspondence between source and target sentences, one can easily build a pseudo bilingual topic model to infer topics for sentence pairs by concatenating each source sentence and its aligned target sentence into one mixed-language sentence. After concatenation, one can train a bilingual topic model on the concatenated corpus using the same topic tool HTMM without any changes. We refer to this topic model as "bilingual trained" or "bitrained" for short HTMM model. The bitrained HTMM model is then used to infer topics for source sentences on the test set. The inferred source sentence topics are used as topics for corresponding target sentences generated by SMT system based on the direct topic correspondence assumption. We visualize the training and testing process of the bitrained HTMM model in Fig. 3. Using these target sentence topics and topic-word probabilities learned by the HTMM model on the concatenated corpus, we can build a word level coherence model following equation (2).

The experiment results of our projection method vs. this approximate bilingual topic model are shown in Table IV. We can observe that the word level coherence model built on projected topics is better than that built on bilingual topics obtained in the way mentioned above. This finding is similar to that of Zhang *et al.* [17], which shows that many-to-many topic projection between the source and target side is better than one-to-one topic mapping for translation rule selection. This may be due to the more serious problem of sparsity in the bitrained HTMM model as its vocabulary size is about twice as large as that of the monolingual HTMM model.[7]

D. Word Level vs. Phrase Level Coherence Model

We investigated and compared the effect of the word and phrase level coherence models. Table V presents the results, where the last two rows will be discussed in the next subsection. 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 models are also much higher than that of the baseline. These suggest that the proposed coherence models are able to improve document translation quality by selecting coherent word/phrase translations that are related to their corresponding sentence topics.

TABLE IV
BLEU AND NIST SCORES OF THE WORD LEVEL COHERENCE MODEL BUILT ON OUR PROJECTION METHOD VS. THE APPROXIMATE BILINGUAL TOPIC MODEL ON THE TEST SET. WCM (BILINGUAL) REFERS TO THE WORD LEVEL COHERENCE MODEL BUILT ON SENTENCE TOPICS INFERRED BY THE BITRAINED HTMM MODEL WHILE WCM (PROJECTION) IS THE COHERENCE MODEL BUILT ON PREDICTED TARGET COHERENCE CHAINS VIA TOPIC PROJECTION

	BLEU	NIST
Base	0.3393	9.1639
Base+WCM (bilingual)	0.3420	9.1873
Base+WCM (projection)	0.3446	9.3699

E. Coherence Chain vs. Document Topic

In this section, we investigated whether it is necessary to use sentence topic sequences (coherence chains) instead of document topics in our coherence models. 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.

We further investigate how varying sentence topics within documents (or coherence chains) are in the training data. We calculate the percentages of documents with m different sentence topics in both the source and target part of training data. The results are displayed in Table VI. From the table, we can see that documents with only one topic for all sentences within them accounting for less than 16%. About 48% coherence chains have more than 5 different topics for sentences within them. Note that the average length of a coherence chain is 28.4 sentences.

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 $\underline{z}_1^n$ to select translations for words/phrases. The new model can be considered as a degenerated variation of our proposed coherence models. It can be formulated and factorized as follows:

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

The probability $p(D_t^i|z_{D_t})$ is further factorized at the word and phrase level, similarly to equation (2) and (3)

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

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

[7]Thanks to an anonymous reviewer for pointing this out.

The results are shown in the last two rows in Table V. 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.

F. Coherence Model Size and Decoding Speed

In this section we discuss the size of our coherence models and decoding speed of systems enhanced with these coherence models. This is directly related to the implementation details about additional memory usage and runtime incurred by the integration of the coherence models into the decoder. Table VII shows the information of model size and decoding speed for the word and phrase level coherence model vs. the baseline system. We set the aggregate size of all models (including the language model, translation model and reordering model) and the decoding speed of the baseline system as the reference (i.e., values are set to 1) in order to have a clear comparison. We then compute the ratios of model size and decoding speed of the systems with coherence models against those of the baseline system.

From the table, we can observe that

- The size of Base + WCM[8] is marginally larger than that of Base as we only integrate topic-word probabilities learned by the HTMM model into the decoder.
- The size of Base + PCM is about 60% larger than that of Base. This is because the number of phrases is much larger than that of words as we discussed in II.
- The decoding speeds of Base + WCM and Base + PCM are slower than that of Base by 17.6% and 22.0% respectively.

From the perspective of implementation, the memory usage and decoding time of Base + WCM and Base + PCM can be largely reduced. For example, we can postpone the integration of topic probabilities that are related to specific words and phrases into the decoder until these words/phrases are to be translated. More specifically, given a source sentence, we first obtain all possible target words and phrases that will be used to generate translation hypotheses according to bilingual phrases collected from our phrase table, which match to the source sentence on the source side. Then we load topic probabilities of these target words and phrases. In doing so, we can not only reduce the extra memory usage but also decrease the time cost of finding these probabilities at runtime.

VI. Analysis

In this section, we will investigate more details of our topic-based coherence models by looking at the differences that they make on target documents and individual translation hypotheses. We conduct two types of evaluations on target documents generated by the baseline and our coherence model: 1) intrinsic evaluation that measures the degree of semantic relatedness between sentences in target documents, and 2) extrinsic evaluation that calculates the ratio of target sentences whose topics inferred by HTMM match to those of sentences in reference translations. Section VI-A will provide the intrinsic evaluation results while Section VI-B the extrinsic evaluation results. We also provide examples in Section VI-C to give a further look into differences on translation hypotheses.

[8]WCM/PCM refer to the word/phrase level coherence model built on the predicted target coherence chain unless otherwise specified.

TABLE V
BLEU and NIST Scores of the Word/Phrase Level Coherence Models on the Test Set. WCM/PCM (z_1^n): the Word/Phrase Level Coherence Model Based on the Target Document Coherence Chain 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
Base+WCM (z_1^n)	0.3446	9.3699
Base+PCM (z_1^n)	0.3454	9.3746
Base+WCM (z_{D_t})	0.3387	9.3023
Base+PCM (z_{D_t})	0.3404	9.3368

TABLE VI
Percentages (%) of Documents with m Different Sentence Topics in the Source and Target Part of the Training Data

m	Perc. in the source part	Perc. in the target part
1	14.3	15.6
2	12.2	12.5
3	11.0	8.8
4	7.6	8.5
5	6.8	6.2
>5	48.1	48.4

TABLE VII
Coherence Model Size and Decoding Speed

Model	Model Size	Decoding Speed
Base	1	1
Base+WCM	1.10	0.85
Base+PCM	1.61	0.82

A. Intrinsic Evaluation: Degree of Semantic Relatedness

In order to quantitatively evaluate how coherent target documents generated by the baseline system and the enhanced systems (Base + WCM or Base + PCM) are, we follow Lapata and Barzilay [18] to measure the coherence of a document as the degree of semantic relatedness between sentences by calculating word-based similarities of these sentences. Formally, the coherence of a document D with n sentences is computed as follows.

$$coherence(D) = \frac{\sum_{i=1}^{n-1} sim_w(s_i, s_{i+1})}{n-1} \tag{13}$$

where the word-based similarity $sim_w(s_i, s_{i+1})$ is calculated as the ratio of word overlap between two sentences.

$$sim_w(s_i, s_{i+1}) = \frac{2|w(s_i) \bigcap w(s_{i+1})|}{|w(s_i)| + |w(s_{i+1})|} \tag{14}$$

Here $w(s)$ represents the set of words in sentence s.

The coherence of a whole test corpus $\mathcal{C}$ is calculated as the mean of coherence degrees of all documents (m documents) in the corpus.

$$coherence(\mathcal{C}) = \frac{\sum\limits_{D \in \mathcal{C}} coherence(D)}{m} \tag{15}$$

TABLE VIII
COHERENCE DEGREES OF TARGET DOCUMENTS GENERATED BY THE BASELINE SYSTEM AND THE ENHANCED SYSTEM WITH OUR COHERENCE MODELS ON THE TEST SET

System	Coherence Degree (%)
Base	17.8
Base+WCM	19.7
Base+PCM	20.3

TABLE IX
RATIOS (%) OF TARGET SENTENCES GENERATED BY BASE AND BASE + WCM/PCM ON THE TEST SET, WHERE THE TOPICS OF THESE TARGET SENTENCES MATCH TO THOSE OF REFERENCE TRANSLATIONS

System	ref_1	ref_2	ref_3	ref_4	avg
Base	66.4	73.2	73.0	72.4	71.3
Base+WCM	68.7	74.9	73.9	74.1	72.9
Base+PCM	68.8	74.4	73.5	74.0	72.7

Table VIII shows the intrinsic evaluation results for Base and Base+WCM/PCM. We can observe that target documents generated by our topic-based coherence models are more coherent than those generated by the baseline in terms of word-based sentence similarity (0.197/0.203 vs. 0.178).

B. Extrinsic Evaluation: Ratio of Topic Matches

We can also evaluate target translations generated by Base and Base + WCM/PCM in an extrinsic manner: quantifying the similarity of system translations to reference translations in terms of sentence topic matches. Specifically, we use the HTMM model trained on the target side of the training data to infer sentence topics for reference translations, translations generated by Base, and translations generated by Base + WCM/PCM. We then calculate the topic match ratio R of sentences whose topics match to those of reference translations as follows.

$$R = \frac{\sum_i \delta(z_{r_i}, z_{s_i})}{N} \tag{16}$$

where $\delta(x, y)$ is the Kronecker function that is 1 if $x = y$ and 0 otherwise, z_{r_i} and z_{s_i} are sentence topics inferred by the HTMM for the reference translation r_i and system translation s_i respectively, and N is the total number of sentences in the set to be evaluated.

Since there are 4 reference translations for each source sentences on the test set, we can calculate 4 topic match ratios, one per set of reference translations ($ref_1 - ref_4$). Table IX shows the topic match ratios to reference translations for system translations generated by Base, Base + WCM and Base + PCM. Obviously, all the topic match ratios to 4 different reference translations of Base + WCM/PCM are higher than those of Base. This suggests that Base + WCM/PCM translations are more similar to reference translations than Base translations in terms of sentence topic matches.

We can also observe that the topic match ratios of Base + WCM are higher than those of Base + PCM for most reference translation sets although Base+PCM is better than Base+WCM in terms of BLEU score. The reason is not quite clear. But we conjecture that this may be due to the fact that HTMM models are trained at the word level rather than the phrase level. Base + PCM generates translations using phrases with topics similar to corresponding reference translations. But it can not guarantee that words in these phrases have similar topics too.

C. Differences on Translation Hypotheses

Table X gives several examples to further shed light on how the topic-based coherence model improves translation quality. In Eg. 1, 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 XI. 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*.

Similarly, in Eg. 2, the Chinese word "jianshe" can be translated into *building*, *construction*, *development* and so on. Given the target sentence topic 106, the topic-word probability of *construction* is higher than that of *building*: $p(construction|106) = 0.00578426$ vs. $p(building|106) = 0.00158552$. This is the reason why Base + WCM selects *construction* rather than *building*.

In the third example, the baseline translates the Chinese phrase "gaoyuan fanying" in a word-by-word manner into a target translation *plateau reaction*. If this translation is also a candidate translation for Base+PCM, its coherence score will be computed as follows by the phrase level coherence model according to equation (4).

$$p(44|\text{plateau}) \times p(44|\text{reaction}) = 0.0113 \times 0.0134 = 1.51\text{E}{-4}$$

where 44 is the topic of the target sentence in this example. The coherence score for the translation *altitude sickness*, however, is

$$p(44|\text{altitudesickness}) = 2.53\text{E}{-2}$$

This is much larger than the score of *plateau reaction*. We further find that the target translation *altitude sickness* mainly distribute over three topics: 44, 100 and 132, among which the topic 44 has the largest distribution probability. Even if we take the impact of phrase penalty feature [10] into account, this example suggests that the topic-based coherence model is able to help the decoder select appropriate translations.

VII. RELATED WORK

We roughly divide previous work related to our topic-based coherence modeling framework into three categories: 1) coherence models for text analysis, 2) inter-sentence dependencies for document translation and 3) topic models for SMT. We will introduce these related models and approaches and highlight the differences of our coherence model from them in this section.

A. Coherence Models for Text Analysis

Although coherence is rarely explored in SMT, it is widely studied in text analysis. Various coherence models are proposed

TABLE X
CHINESE (SHOWN IN PINYIN) TO ENGLISH TRANSLATION EXAMPLES SHOWING THE DIFFERENCE BETWEEN THE BASELINE TRANSLATION (BASE) AND THE TRANSLATION GENERATED BY THE SYSTEM ENHANCED WITH OUR COHERENCE MODELS (Base + WCM/PCM)

Eg. 1	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**
	Base+WCM	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
Eg. 2	src	yi shi yao jiakuai tuijin jinrong zichan guanli gongsi youguan fagui **jianshe**
	Base	a **building** of relevant laws and regulations to speed up financial assets management companies
	Base+WCM	**construction** of a relevant laws and regulations to speed up financial assets management companies
	ref	The first will be to accelerate the **construction** process of relevant laws and regulations against asset management companies
Eg. 3	src	ling yige kunnan shi **gaoyuan fanying**
	Base	Another difficulty is **plateau reaction**
	Base+PCM	Another difficulty was **altitude sickness**
	ref	Another difficulty was **altitude sickness**

TABLE XI
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

in the context of document summarization and generation, e.g., entity-based local coherence model [19], content-based global coherence models [2], [20] and syntax-based coherence model [21].

Our definition of coherence is partly inspired by the content model [2] as mentioned in Section I. 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.

B. 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 [22], [23]. Hardmeier *et al.* [24] 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.

C. Topic Models for SMT

Our model is also related to previous approaches that employ topic model for SMT [25], [5], [6], especially the topic similarity model [5] which explores document topics for hierarchical phrase selection. However, our coherence model is significantly different from the topic similarity model in two key aspects. First, 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. Second, 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 size of the model is becoming larger and larger as we increase the number of topics. Additionally, the distribution-based similarity model cannot differentiate topic-insensitive phrases [5].

Finally our work is also related to most recent work that uses domain information to help lexical and phrasal selection [26], [27] since domain information can be considered as corpus-level information that is beyond sentence boundaries, just like our coherence chains defined at the document level.

VIII. 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 models into phrase-based machine translation yields significant improvements over the baseline.

We have also observed that

- The coherence model built on topics projected from source sentences by the MaxEnt-based prediction model is better than that built on topics inferred by the bitrained HTMM model as described in Section V-C.
- The phrase level coherence model is marginally better than the word level coherence model.
- Our coherence models significantly outperform the degenerated coherence model which only uses target document topics to constrain word/phrase translations.
- Target translations generated by the proposed coherence models are more coherent and similar to reference translations than those generated by the baseline system according to the intrinsic and extrinsic evaluations in Section VI.

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 [21] 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 [28], [29], [30], [31] 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.

In the future, we want to build new coherence models on multiple coherence chains for each document in order to reduce error propagation from HTMM models. Specifically, we would like to output the n-best topic sequences from HTMM models for each document, rather than only the best coherence chain, and use them to train our topic projection model. Additionally, 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 [32], which should also be considered when translating a text. Finally, the discourse structure is frequently modeled hierarchically in the literature. Therefor we also plan to incorporate more hierarchical discourse information (e.g., discourse connectives [33]) into phrase/syntax-based machine translation at the document level in the future.

ACKNOWLEDGMENT

We would like to thank anonymous reviewers for their insightful comments.

REFERENCES

[1] R.-A. de Beaugrande and W. Dressler, *Introduction to Text Linguistics*. New York, NY, USA: Longman London, 1981.

[2] R. Barzilay, L. Lee, D. M. Susan, Dumais, S. Roukos, and B. Massachusetts, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. Assoc. Comput. Linguist.*, May 2–7, 2004, pp. 113–120.

[3] A. Gruber, M. Rosen-zvi, and Y. Weiss, "Hidden topic Markov models," in *Proc. Artif. Intell. Statist.*, 2007.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, Philadelphia, PA, USA, Jul. 2002, 2007, pp. 311–318 [Online]. Available: http://www.aclweb.org/anthology/P02-1040

[5] X. Xiao, D. Xiong, M. Zhang, Q. Liu, and S. Lin, "A topic similarity model for hierarchical phrase-based translation," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguist. (Vol. 1: Long Papers)*, Jeju Island, Korea, Jul. 2012, pp. 750–758 [Online]. Available: http://www.aclweb.org/anthology/P12-1079

[6] J. Su, H. Wu, H. Wang, Y. Chen, X. Shi, H. Dong, and Q. Liu, "Translation model adaptation for statistical machine translation with monolingual topic information," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguist. (Volume 1: Long Papers)*, Jeju Island, Korea, Jul. 2012, pp. 459–468 [Online]. Available: http://www.aclweb.org/anthology/P12-1048

[7] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] D. Xiong and M. Zhang, "A topic-based coherence model for statistical machine translation," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013.

[9] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proc. Conf. Empir. Meth. Nat. Lang. Process.*, Singapore, Aug. 2009, pp. 880–889 [Online]. Available: http://www.aclweb.org/anthology/D/D09/D09-1092

[10] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. Assoc. Comput. Linguist.*, Edmonton, AB, Canada, May–Jun. 2003, pp. 58–54.

[11] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Comput. Linguist.*, vol. 23, no. 3, pp. 377–403, 1997.

[12] D. Xiong, Q. Liu, and S. Lin, "Maximum entropy based phrase reordering model for statistical machine translation," in *Proc. 21st Int. Conf. Comput. Linguist. 44th Annu. Meeting Assoc. Comput. Linguist.*, Sydney, Australia, Jul. 2006, pp. 521–528 [Online]. Available: http://www.aclweb.org/anthology/P06-1066

[13] A. Stolcke, "SRILM–an extensible language modeling toolkit," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, Sep. 2002, pp. 901–904.

[14] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. 2nd Int. Conf. Human Lang. Technol. Res.*, San Francisco, CA, USA, 2002, pp. 138–145 [Online]. Available: http://dl.acm.org/citation.cfm?id=1289189.1289273, ser. HLT '02

[15] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguist.*, Sapporo, Japan, Jul. 2003, pp. 160–167 [Online]. Available: http://www.aclweb.org/anthology/P03-1021, Assoc. Comput. Linguist.

[16] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist.: Human Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 176–181 [Online]. Available: http://www.aclweb.org/anthology/P11-2031

[17] M. Zhang, X. Xiao, D. Xiong, and Q. Liu, "Topic-based dissimilarity and sensitivity models for translation rule selection," *J. Artif. Intell. Res.*, vol. 50, pp. 1–30, 2014.

[18] M. Lapata and R. Barzilay, "Automatic evaluation of text coherence: Models and representations," in *Proc. 19th Int. Joint Conf. Artif. Intell.*, 2005, pp. 1085–1090.

[19] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," *Comput. Linguist.*, vol. 34, no. 1, pp. 1–34, 2008.

[20] P. Fung and G. Ngai, "One story, one flow: Hidden markov story models for multilingual multidocument summarization," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–16, 2006.

[21] A. Louis and A. Nenkova, "A coherence model based on syntactic patterns," in *Proc. Joint Conf. Empir. Meth. Nat. Lang. Process. Comput. Nat. Lang. Learn.*, Jeju Island, Korea, Jul. 2012, pp. 1157–1168 [Online]. Available: http://www.aclweb.org/anthology/D12-1106

[22] J. Tiedemann, "To cache or not to cache? experiments with adaptive models in statistical machine translation," in *Proc. Joint 5th Workshop Statist. Mach. Transl. Metrics (MATR)*, Uppsala, Sweden, Jul. 2010, pp. 189–194 [Online]. Available: http://www.aclweb.org/anthology/W10-1728

[23] Z. Gong, M. Zhang, and G. Zhou, "Cache-based document-level statistical machine translation," in *Proc. 2011Conf. Empir. Meth. Nat. Lang. Process.*, Edinburgh, U.K., Jul. 2011, pp. 909–919 [Online]. Available: http://www.aclweb.org/anthology/D11-1084

[24] C. Hardmeier, J. Nivre, and J. Tiedemann, "Document-wide decoding for phrase-based statistical machine translation," in *Proc. Joint Conf. Empir. Meth. Nat. Lang. Process. Comput. Nat. Lang. Learn.*, Jeju Island, Korea, Jul. 2012, pp. 1179–1190 [Online]. Available: http://www.aclweb.org/anthology/D12-1108

[25] B. Zhao and E. P. Xing, "BiTAM: Bilingual topic admixture models for word alignment," in *Proc. 21st Int. Conf. Comput. Linguist. and 44th Annu. Meeting Assoc. Comput. Linguist.*, Sydney, Australia, Jul. 2006, pp. 969–976 [Online]. Available: http://www.aclweb.org/anthology/P/P06/P06-2124

[26] C. Hoang and K. Sima'an, "Latent domain translation models in mix-of-domains haystack," in *Proc. 25th Int. Conf. Comput. Linguist.: Technical Papers*, Dublin, Ireland, Aug. 2014, pp. 1928–1939 [Online]. Available: http://www.aclweb.org/anthology/C14-1182, Dublin City Univ. and Assoc. for Comput. Linguist.

[27] C. Hoang and K. Sima'an, "Latent domain phrase-based translation models for adaptation," in *Proc. Conf. Empir. Meth. Nat. Lang. Process.*, Doha, Qatar, Oct. 2014, pp. 1928–1939.

[28] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable inference and training of context-rich syntactic translation models," in *Proc. 21st Int. Conf. Comput. Linguist. and 44th Annu. Meeting Assoc. Comput. Linguist.*, Sydney, Australia, Jul. 2006, pp. 961–968 [Online]. Available: http://www.aclweb.org/anthology/P06-1121

[29] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string alignment template for statistical machine translation," in *Proc. 21st Int. Conf. Comput. Linguist. and the 44th Annu. Meeting Assoc. Comput. Linguist.*, Stroudsburg, PA, USA, 2006, pp. 609–616 [Online]. Available: http://dx.doi.org/10.3115/1220175.1220252, ser. ACL-44,Assoc. Comput. Linguist.

[30] H. Hassan, K. Sima'an, and A. Way, "Efficient accurate syntactic direct translation models: One tree at a time," *Mach. Translat.*, pp. 121–136, 2012.

[31] Y. Feng, D. Zhang, M. Li, and Q. Liu, "Hierarchical chunk-to-string translation," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguist. (Volume 1: Long Papers)*, Jeju Island, Korea, Jul. 2012, pp. 950–958 [Online]. Available: http://www.aclweb.org/anthology/P12-1100, Assoc. Comput. Linguist.

[32] Z. Lin, H. T. Ng, and M.-Y. Kan, "Automatically evaluating text coherence using discourse relations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist.: Human Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 997–1006 [Online]. Available: http://www.aclweb.org/anthology/P11-1100

[33] C. Hardmeier, "Discourse in statistical machine translation: A survey and a case study," *Discours - Revue de Linguistique, Psycholinguistique et Informatique*, vol. 11, 2012.

Deyi Xiong is a Professor at Soochow University. Previously, he was a Research Scientist at the Institute for Infocomm Research of Singapore from 2007–2013. He completed his Ph.D. in computer science at the Institute of Computing Technology of the Chinese Academy of Sciences in 2007. His research interests are in the area of natural language processing, including parsing and statistical machine translation.

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

Xing Wang is a Ph.D. candidate at Soochow University. He is supervised by Prof. Min Zhang and Prof. Deyi Xiong. His current research interests include statistical machine translation and machine learning.