Modeling Term Translation for Document-informed Machine Translation

Fandong Meng[1, 2] **Deyi Xiong**[3] **Wenbin Jiang**[1, 2] **Qun Liu**[4, 1]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{mengfandong,jiangwenbin,liuqun}@ict.ac.cn
[3]School of Computer Science and Technology, Soochow University
dyxiong@suda.edu.cn
[4]Centre for Next Generation Localisation, Dublin City University

Abstract

Term translation is of great importance for statistical machine translation (SMT), especially document-informed SMT. In this paper, we investigate three issues of term translation in the context of document-informed SMT and propose three corresponding models: (a) a term translation disambiguation model which selects desirable translations for terms in the source language with domain information, (b) a term translation consistency model that encourages consistent translations for terms with a high strength of translation consistency throughout a document, and (c) a term bracketing model that rewards translation hypotheses where bracketable source terms are translated as a whole unit. We integrate the three models into hierarchical phrase-based SMT and evaluate their effectiveness on NIST Chinese-English translation tasks with large-scale training data. Experiment results show that all three models can achieve significant improvements over the baseline. Additionally, we can obtain a further improvement when combining the three models.

1 Introduction

A term is a linguistic expression that is used as the designation of a defined concept in a language (ISO 1087). As terms convey concepts of a text, term translation becomes crucial when the text is translated from its original language to another language. The translations of terms are often affected by the domain in which terms are used and the context that surrounds terms (Vasconcellos et al., 2001). In this paper, we study domain-specific and context-sensitive term translation for SMT.

In order to achieve this goal, we focus on three issues of term translation: 1) translation ambiguity, 2) translation consistency and 3) bracketing. First, term translation ambiguity is related to translations of the same term in different domains. A source language term may have different translations when it occurs in different domains. Second, translation consistency is about consistent translations for terms that occur in the same document. Usually, it is undesirable to translate the same term in different ways as it occurs in different parts of a document. Finally, bracketing concerns whether a multi-word term is bracketable during translation. Normally, a multi-word term is translated as a whole unit into a contiguous target string.

We study these three issues in the context of document-informed SMT. We use document-informed information to disambiguate term translations in different documents and maintain consistent translations for terms that occur in the same document. We propose three different models for term translation that attempt to address the three issues mentioned above. In particular,

- *Term Translation Disambiguation Model:* In this model, we condition the translations of terms in different documents on corresponding per-document topic distributions. In doing so, we enable the decoder to favor translation hypotheses with domain-specific term translations.
- *Term Translation Consistency Model:* This model encourages the same terms with a high strength of translation consistency that occur in different parts of a document to be translated in a consistent fashion. We calculate the translation consistency strength of a term based on the topic distribution of the documents where the term occurs in this model.
- *Term Bracketing Model:* We use the bracketing model to reward translation hypothe-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 546–556,
October 25-29, 2014, Doha, Qatar.

ses where bracketable multi-word terms are translated as a whole unit.

We integrate the three models into hierarchical phrase-based SMT (Chiang, 2007). Large-scale experiment results show that they are all able to achieve significant improvements of up to 0.89 BLEU points over the baseline. When simultaneously integrating the three models into SMT, we can gain a further improvement, which outperforms the baseline by up to 1.16 BLEU points.

In the remainder of this paper, we begin with a brief overview of related work in Section 2, and bilingual term extraction in Section 3. We then elaborate the proposed three models for term translation in Section 4. Next, we conduct experiments to validate the effectiveness of the proposed models in Section 5. Finally, we conclude and provide directions for future work in Section 6.

2 Related Work

In this section, we briefly introduce related work and highlight the differences between our work and previous studies.

As we approach term translation disambiguation and consistency via topic modeling, our models are related to previous work that explores the topic model (Blei et al., 2003) for machine translation (Zhao and Xing, 2006; Su et al., 2012; Xiao et al., 2012; Eidelman et al., 2012). Zhao and Xing (2006) employ three models that enable word alignment process to leverage topical contents of document-pairs with topic model. Su et al. (2012) establish the relationship between out-of-domain bilingual corpus and in-domain monolingual corpora via topic mapping and phrase-topic distribution probability estimation for translation model adaptation. Xiao et al. (2012) propose a topic similarity model for rule selection. Eidelman et al. (2012) use topic models to adapt lexical weighting probabilities dynamically during translation. In these studies, the topic model is not used to address the issues of term translation mentioned in Section 1.

Our work is also related to document-level SMT in that we use document-informed information for term translation. Tiedemann (2010) propose cache-based language and translation models, which are built on recently translated sentences. Gong et al. (2011) extend this by further introducing two additional caches. They employ a static cache to store bilingual phrases extracted from documents in training data that are similar to the document being translated and a topic cache with target language topic words. Recently we have also witnessed efforts that model lexical cohesion (Hardmeier et al., 2012; Wong and Kit, 2012; Xiong et al., 2013a; Xiong et al., 2013b) as well as coherence (Xiong and Zhang, 2013) for document-level SMT. Hasler et al. (2014a) use topic models to learn document-level translation probabilities. Hasler et al. (2014b) use topic-adapted model to improve lexical selection. The significant difference between our work and these studies is that term translation has not been investigated in these document-level SMT models.

Itagaki and Aikawa (2008) employ bilingual term bank as a dictionary for machine-aided translation. Ren et al. (2009) propose a binary feature to indicate whether a bilingual phrase contains a term pair. Pinis and Skadins (2012) investigate that bilingual terms are important for domain adaptation of machine translation. These studies do not focus on the three issues of term translation as discussed in Section 1. Furthermore, domain and document-informed information is not used to assist term translation.

Itagaki et al. (2007) propose a statistical method to calculate translation consistency for terms with explicit domain information. Partially inspired by their study, we introduce a term translation consistency metric with document-informed information. Furthermore, we integrate the proposed term translation consistency model into an actual SMT system, which has not been done by Itagaki et al. (2007). Ture et al. (2012) use IR-inspired tf-idf scores to encourage consistent translation choice. Guillou (2013) investigates what kind of words should be translated consistently. Term translation consistency has not been investigated in these studies.

Our term bracketing model is also related to Xiong et al. (2009)'s syntax-driven bracketing model for phrase-based translation, which predicts whether a phrase is bracketable or not using rich syntactic constraints. The difference is that we construct the model with automatically created bilingual term bank and do not depend on any syntactic knowledge.

3 Bilingual Term Extraction

Bilingual term extraction is to extract terms from two languages with the purpose of creating or ex-

tending a bilingual term bank, which in turn can be used to improve other tasks such as information retrieval and machine translation. In this paper, we want to automatically build a bilingual term bank so that we can model term translation to improve translation quality of SMT. Our interest is to extract multi-word terms.

Currently, there are mainly two strategies to conduct bilingual term extraction from parallel corpora. One of them is to extract term candidates separately for each language according to monolingual term metrics, such as C-value/NC-value (Frantzi et al., 1998; Vu et al., 2008), or other common cooccurrence measures such as Log-Likelihood Ratio, Dice coefficient and Pointwise Mutual Information (Daille, 1996; Piao et al., 2006). The extracted monolingual terms are then paired together (Hjelm, 2007; Fan et al., 2009; Ren et al., 2009). The other strategy is to align words and word sequences that are translation equivalents in parallel corpora and then classify them into terms and non-terms (Merkel and Foo, 2007; Lefever et al., 2009; Bouamor et al., 2012). In this paper, we adopt the first strategy. In particular, for each sentence pair, we collect all source phrases which are terms and find aligned target phrases for them via word alignments. If the target side is also a term, we store the source and target term as a term pair.

We conduct monolingual term extraction using the C-value/NC-value metric and Log-Likelihood Ratio (LLR) measure respectively. We then combine terms extracted according to the two metrics mentioned above. For the C-value/NC-value metric based term extraction, we implement it in the same way as described in Frantzi et al. (1998). This extraction method recognizes linguistic patterns (mainly noun phrases) listed as follows.

$$((Adj|Noun)^{+}|((Adj|Noun)^{*}$$
$$(NounPrep)^{?})(Adj|Noun)^{*})Noun$$

It captures the linguistic structures of terms. For the LLR metric based term extraction, we implement it according to Daille (1996), who estimate the propensity of two words to appear together as a multi-word expression. We then adopt LLR-based hierarchical reducing algorithm proposed by Ren et al. (2009) to extract terms with arbitrary lengths. Since the C-value/NC-value metric based extraction method can obtain terms in strict linguistic patterns while the LLR measure based method extracts more flexible terms, these two methods are complementary to each other. Therefore, we use these two methods to extract monolingual multi-word terms and then combine the extracted terms.

4 Models

This section presents the three models of term translation. They are the term translation disambiguation model, term translation consistency model and term bracketing model respectively.

4.1 Term Translation Disambiguation Model

The most straightforward way to disambiguate term translations in different domains is to calculate the conditional translation probability of a term given domain information. We use the topic distribution of a document obtained by a topic model to represent the domain information of the document. Since Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the most widely-used topic model, we exploit it for inferring topic distributions of documents. Xiao et al. (2012) proposed a topic similarity model for rule selection. Different from their work, we take an easier strategy that estimates topic-conditioned term translation probabilities rather than rule-topic distributions. This makes our model easily scalable on large training data.

With the bilingual term bank created from the training data, we calculate the source-to-target term translation probability for each term pair conditioned on the topic distribution of the source document where the source term occurs. We maintain a K-dimension (K is the number of topics) vector for each term pair. The k-th component $p(t_e|t_f, z = k)$ measures the conditional translation probability from source term t_f to target term t_e given the topic k.

We calculate $p(t_e|t_f, z = k)$ via maximum likelihood estimation with counts from training data. When the source part of a bilingual term pair occurs in a document D with topic distribution $p(z|D)$ estimated via LDA tool, we collect an instance $(t_f, t_e, p(z|D), c)$, where c is the fraction count of the instance as described in Chiang (2007). After collection, we get a set of instances $I = \{(t_f, t_e, p(z|D), c)\}$ with different document-topic distributions for each bilingual term pair. Using these instances, we calculate the probability

$p(t_e|t_f, z = k)$ as follows:

$$p(t_e|t_f, z = k) = \frac{\sum_{i \in I, i.t_f = t_f, i.t_e = t_e} i.c * p(z = k|D)}{\sum_{i \in I, i.t_f = t_f} i.c * p(z = k|D)} \quad (1)$$

We associate each extracted term pair in our bilingual term bank with its corresponding topic-conditioned translation probabilities estimated in the Eq. (1). When translating sentences of document $D^{'}$, we first get the topic distribution of $D^{'}$ using LDA tool. Given a sentence which contains T terms $\{t_{f_i}\}_1^T$ in $D^{'}$, our term translation disambiguation model $TermDis$ can be denoted as

$$TermDis = \prod_{i=1}^{T} P_d(t_{e_i}|t_{f_i}, D^{'}) \quad (2)$$

where the conditional source-to-target term translation probability $P_d(t_{e_i}|t_{f_i}, D^{'})$ given the document $D^{'}$ is formulated as follows:

$$P_d(t_{e_i}|t_{f_i}, D^{'}) = \sum_{k=1}^{K} p(t_{e_i}|t_{f_i}, z = k) * p(z = k|D^{'}) \quad (3)$$

Whenever a source term t_{f_i} is translated into t_{e_i}, we check whether the pair of t_{f_i} and its translation t_{e_i} can be found in our bilingual term bank. If it can be found, we calculate the conditional translation probability from t_{f_i} to t_{e_i} given the document $D^{'}$ according to Eq. (3).

The term translation disambiguation model is integrated into the log-linear model of SMT as a feature. Its weight is tuned via minimum error rate training (MERT) (Och, 2003). Through the feature, we can enable the decoder to favor translation hypotheses that contain target term translations appropriate for the domain represented by the topic distribution of the corresponding document.

4.2 Term Translation Consistency Model

The term translation disambiguation model helps the decoder select appropriate translations for terms that are in accord with their domains. Yet another translation issue related to the domain-specific term translation is to what extent a term should be translated consistently given the domain where it occurs. Term translation consistency indicates the translation stability that a source term is translated into the same target term (Itagaki et al., 2007). When translating a source term, if the translation consistency strength of the source term is high, we should take the corresponding target term as the translation for it. Otherwise, we may need to create a new translation for it according to its context. In particular, we want to enable the decoder to choose between: 1) translating a given source term into the extracted corresponding target term or 2) translating it in another way according to the strength of its translation consistency. In doing so, we can encourage consistent translations for terms with a high translation consistency strength throughout a document.

Our term translation consistency model can exactly measure the strength of term translation consistency in a document. Since the essential component of our term translation consistency model is the translation consistency strength of the source term estimated under the topic distribution, we describe how to calculate it before introducing the whole model.

With the bilingual term bank created from training data, we first group each source term and all its corresponding target terms into a 2-tuple $G\langle t_f, Set(t_e)\rangle$, where t_f is the source term and $Set(t_e)$ is the set of t_f's corresponding target terms. We maintain a K-dimension (K is the number of topics) vector for each 2-tuple $G\langle t_f, Set(t_e)\rangle$. The k-th component measures the translation consistency strength $cons(t_f, k)$ of the source term t_f given the topic k.

We calculate $cons(t_f, k)$ for each $G\langle t_f, Set(t_e)\rangle$ with counts from training data as follows:

$$cons(t_f, k) = \sum_{m=1}^{M} \sum_{n=1}^{N_m} (\frac{q_{mn} * p(k|m)}{Q_k})^2 \quad (4)$$

$$Q_k = \sum_{m=1}^{M} \sum_{n=1}^{N_m} q_{mn} * p(k|m) \quad (5)$$

where M is the number of documents in which the source term t_f occurs, N_m is the number of unique corresponding term translations of t_f in the mth document, q_{mn} is the frequency of the nth translation of t_f in the mth document, $p(k|m)$ is the conditional probability of the mth document over topic k, and Q_k is the normalization factor. All translations of t_f are from $Set(t_e)$. We adapt Itagaki et al. (2007)'s translation consistency metric for terms to our topic-based translation consistency measure in the Eq. (4). This equation calculates the translation consistency strength of the source term t_f given the topic k according to the distribution of t_f's translations in each document

where they occur. According to Eq. (4), the translation consistency strength is a score between 0 and 1. If a source term only occurs in a document and all its translations are the same, the translation consistency strength of this term is 1.

We reorganize our bilingual term bank into a list of 2-tuples $G\langle t_f, Set(t_e)\rangle$s, each of which is associated with a K-dimension vector storing the topic-conditioned translation consistency strength calculated in the Eq. (4). When translating sentences of document D, we first get the topic distribution of D via LDA tool. Given a sentence which contains T terms $\{t_{f_i}\}_1^T$ in D, our term translation consistency model $TermCons$ can be denoted as

$$TermCons = \prod_{i=1}^{T} exp(S_c(t_{f_i}|D)) \quad (6)$$

where the strength of translation consistency for t_{f_i} given the document D is formulated as follows:

$$S_c(t_{f_i}|D) = log(\sum_{k=1}^{K} cons(t_{f_i}, k) * p(k|D)) \quad (7)$$

During decoding, whenever a hypothesis just translates a source term t_{f_i} into t_e, we check whether the translation t_e can be found in $Set(t_e)$ of t_{f_i} from the reorganized bilingual term bank. If it can be found, we calculate the strength of translation consistency for t_{f_i} given the document D according to Eq. (7) and take it as a soft constraint. If the $S_c(t_{f_i}|D)$ of t_{f_i} is high, the decoder should translate t_{f_i} into the extracted corresponding target terms. Otherwise, the decoder will select translations from outside of $Set(t_e)$ for t_{f_i}. In doing so, we encourage terms to be translated in a topic-dependent consistency pattern in the test data similar to that in the training data so that we can control the translation consistency of terms in the test data.

The term translation consistency model is also integrated into the log-linear model of SMT as a feature. Through the feature, we can enable the decoder to translate terms with a high translation consistency in a document into corresponding target terms from our bilingual term bank rather than other translations in a consistent fashion.

4.3 Term Bracketing Model

The term translation disambiguation model and consistency model concern the term translation accuracy with domain information. We further propose a term bracketing model to guarantee the integrality of term translation. Xiong et al. (2009) proposed a syntax-driven bracketing model for phrase-based translation, which predicts whether a phrase is bracketable or not using rich syntactic constraints. If a source phrase remains contiguous after translation, they refer to this type of phrase as bracketable phrase, otherwise unbracketable phrase. For multi-word terms, it is also desirable to be bracketable since a source term should be translated as a whole unit and its translation should be contiguous.

In this paper, we adapt Xiong et al. (2009)'s bracketing approach to term translation and build a classifier to measure the probability that a source term should be translated in a bracketable manner. For all source parts of the extracted bilingual term bank, we find their target counterparts in the word-aligned training data. If the corresponding target counterpart remains contiguous, we take the source term as a bracketable instance, otherwise an unbracketable instance. With these bracketable and unbracketable instances, we train a maximum entropy binary classifier to predict bracketable (b) probability of a given source term t_f within particular contexts $c(t_f)$. The binary classifier is formulated as follows:

$$P_b(b|c(t_f)) = \frac{exp(\sum_j \theta_j h_j(b, c(t_f)))}{\sum_{b'} exp(\sum_j \theta_j h_j(b', c(t_f)))} \quad (8)$$

where $h_j \in \{0, 1\}$ is a binary feature function and θ_j is the weight of h_j. We use the following features: 1) the word sequence of the source term, 2) the first word of the source term, 3) the last word of the source term, 4) the preceding word of the first word of the source term, 5) the succeeding word of the last word of the source term, and 6) the number of words in the source term.

Given a source sentence which contains T terms $\{t_{f_i}\}_1^T$, our term bracketing model $TermBrack$ can be denoted as

$$TermBrack = \prod_{i=1}^{T} P_b(b|c(t_{f_i})) \quad (9)$$

Whenever a hypothesis just covers a source term t_{f_i}, we calculate the bracketable probability of t_{f_i} according to Eq. (8).

The term bracketing model is integrated into the log-linear model of SMT as a feature. Through the feature, we want the decoder to translate source terms with a high bracketable probability as a whole unit.

Source	Target	D	M
Fángyù Xìtǒng	defence mechanisms	470	56
Fángyù Xìtǒng	defence systems		
Fángyù Xìtǒng	defense programmes		
Fángyù Xìtǒng	prevention systems		
...	...		
Zhànluè Dǎodàn Fángyù Xìtǒng	strategic missile defense system	7	0

Table 1: Examples of bilingual terms extracted from the training data. "D" means the total number of documents in which the corresponding source term occurs and "M" denotes the number of documents in which the corresponding source term is translated into different target terms. The source side is Chinese Pinyin. To save space, we do not list all the 23 different translations of the source term "Fángyù Xìtǒng".

5 Experiments

In this section, we conducted experiments to answer the following three questions.

1. Are our term translation disambiguation, consistency and bracketing models able to improve translation quality in BLEU?
2. Does the combination of the three models provide further improvements?
3. To what extent do the proposed models affect the translations of test sets?

5.1 Setup

Our training data consist of 4.28M sentence pairs extracted from LDC[1] data with document boundaries explicitly provided. The bilingual training data contain 67,752 documents, 124.8M Chinese words and 140.3M English words. We chose NIST MT05 as the MERT (Och, 2003) tuning set, NIST MT06 as the development test set, and NIST MT08 as the final test set. The numbers of documents/sentences in NIST MT05, MT06 and MT08 are 100/1082, 79/1664 and 109/1357 respectively.

The word alignments were obtained by running GIZA++ (Och and Ney, 2003) on the corpora in both directions and using the "grow-diag-final-and" balance strategy (Koehn et al., 2003). We adopted SRI Language Modeling Toolkit (Stolcke and others, 2002) to train a 4-gram language model with modified Kneser-Ney smoothing on the Xinhua portion of the English Gigaword corpus. For the topic model, we used the open source LDA tool GibbsLDA++[2] with the default setting for training and inference. We performed 100 iterations of the L-BFGS algorithm implemented in the MaxEnt toolkit[3] with both Gaussian prior and event cutoff set to 1 to train the term bracketing prediction model (Section 4.3).

We performed part-of-speech tagging for monolingual term extraction (C-value/NC-vaule method in Section 3) of the source and target languages with the Stanford NLP toolkit[4]. The bilingual term bank was extracted based on the following parameter settings of term extraction methods. Empirically, we set the maximum length of a term to 6 words[5]. For both the C-value/NC-value and LLR-based extraction methods, we set the context window size to 5 words, which is a widely-used setting in previous work. And we set C-value/NC-value score threshold to 0 and LLR score threshold to 10 according to the training corpora.

We used the case-insensitive 4-gram BLEU[6] as our evaluation metric. In order to alleviate the impact of the instability of MERT (Och, 2003), we ran it three times for all our experiments and presented the average BLEU scores on the three runs following the suggestion by Clark et al. (2011).

We used an in-house hierarchical phrase-based decoder to verify our proposed models. Although the decoder translates a document in a sentence-by-sentence fashion, it incorporates document-informed information for sentence translation via the proposed term translation models trained on documents.

[1]The corpora include LDC2003E07, LDC2003E14, LDC2004T07, LDC2004E12, LDC2005E83, LDC2005T06, LDC2005T10, LDC2006E24, LDC2006E34, LDC2006E85, LDC2006E92, LDC2007E87, LDC2007E101, LDC2008E40, LDC2008E56, LDC2009E16 and LDC2009E95.

[2]http://sourceforge.net/projects/gibbslda/

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

[4]http://nlp.stanford.edu/software/tagger.shtml

[5]We determine the maximum length of a term by testing {5, 6, 7, 8} in our preliminary experiments. We find that length 6 produces a slightly better performance than other values.

[6]ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

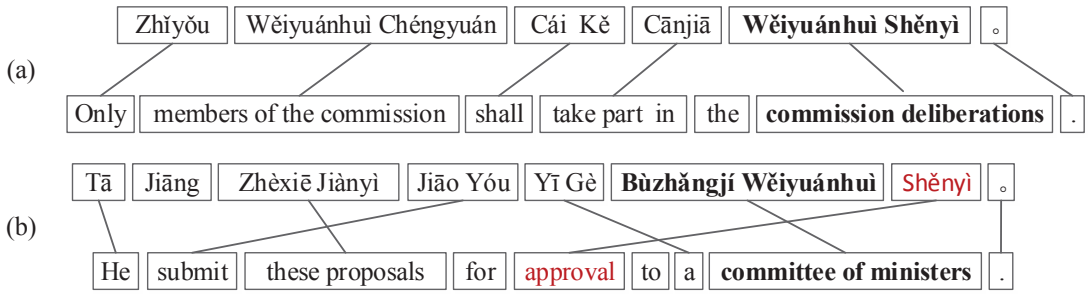

Figure 1: An example of unbracketable source term in the training data. In (a), "Wěiyuánhuì Shěnyì" is bracketable while in (b) it is unbracketable. The solid lines connect bilingual phrases. The source side is Chinese Pinyin.

5.2 Bilingual Term Bank

Before reporting the results of the proposed models, we provide some statistics of the bilingual term bank extracted from the training data.

According to our statistics, about 1.29M bilingual terms are extracted from the training data. 65.07% of the sentence pairs contain bilingual terms in the training data. And on average, a source term has about 1.70 different translations. These statistics indicate that terms are frequently used in real-world data and that a source term can be translated into different target terms.

We also present some examples of bilingual terms extracted from the training data in Table 1. Accordingly, we show the total number of documents in which the corresponding source term occurs and the number of documents in which the corresponding source term is translated into different target terms. The source term "Fángyù Xìtǒng" has 23 different translations in total. They are distributed in 470 documents in the training data. In 414 documents, "Fángyù Xìtǒng" has only one single translation. However, in the other 56 documents it has different translations. This indicates that "Fángyù Xìtǒng" is not consistently translated in these 56 documents. Different from this, the source term "Zhànluè Dǎodàn Fángyù Xìtǒng" only has one translation. And it is translated consistently in all 7 documents where it occurs. In fact, according to our statistics, there are about 5.19% source terms whose translations are not consistent even in the same document.

These examples and statistics suggest 1) that source terms have domain-specific translations and 2) that source terms are not necessarily translated in a consistent manner even in the same document. These are exactly the reasons why we propose the term translation disambiguation and consistency model based on domain information represented by topic distributions.

Actually, 36.13% of the source terms are not necessarily translated into target strings as a whole unit. We show an example of such terms in Figure 1. In Figure 1-(a), "Wěiyuánhuì Shěnyì" is a term, and is translated into "commission deliberations" as a whole unit. Therefore "Wěiyuánhuì Shěnyì" is bracketable in this sentence. However, in Figure 1-(b), "Wěiyuánhuì" and "Shěnyì" are translated separately. Therefore "Wěiyuánhuì Shěnyì" is an unbracketable term in this sentence. This is the reason why we propose a bracketing model to predict whether a source term is bracketable or not.

5.3 Effect of the Proposed Models

In this section, we validate the effectiveness of the proposed term translation disambiguation model, consistency model and bracketing model respectively. In addition to the traditional hiero (Chiang, 2007) system, we also compare against the "CountFeat" method in Ren et al. (2009) who use a binary feature to indicate whether a bilingual phrase contains a term pair. Although Ren et al. (2009)'s experiments are conducted in a phrase-based system, the idea can be easily applied to a hierarchical phrase-based system.

We carried out experiments to investigate the effect of the term translation disambiguation model (Dis-Model) and report the results in Table 2. In order to find the topic number setting with which our model has the best performance, we ran experiments using the MT06 as the development test set. From Table 2, we observe that the Dis-Model obtains steady improvements over the baseline and "CountFeat" method with the topic number K

Models		MT06	MT08	Avg
Baseline		32.43	24.14	28.29
CountFeat		32.77	24.29	28.53
Dis-Model	$K = 50$	32.94*	24.53	28.74
	$K = 100$	33.10*	24.57	28.84
	$K = 150$	**33.16***	**24.67***	**28.92**
	$K = 200$	33.08*	24.55	28.81
Cons-Model	$K = 50$	33.09*	24.59	28.84
	$K = 100$	33.13*	24.74*	28.94
	$K = 150$	**33.32*+**	**24.84*+**	**29.08**
	$K = 200$	33.02*	24.73*	28.88
Brack-Model		**33.09***	**24.66***	**28.88**
Combined-Model		**33.59*+**	**24.99*+**	**29.29**

Table 2: BLEU-4 scores (%) of the term translation disambiguation model (Dis-Model), the term translation consistency model (Cons-Model), the term bracketing model (Brack-Model), and the combination of the three models, on the development test set MT06 and the final test set MT08. $K \in \{50, 100, 150, 200\}$ which is the number of topics for the Dis-Model and the Cons-Model. "Combined-Model" is the combination of the three single modes with topic number 150 for the Dis-Model and the Cons-Model. "Baseline" is the traditional hierarchical phrase-based system. "CountFeat" is the method that adds a counting feature to reward translation hypotheses containing bilingual term pairs. The "*" and "+" denote that the results are significantly (Clark et al., 2011) better than those of the baseline system and the CountFeat method respectively ($p<0.01$).

ranging from 50 to 150. However, when we set K to 200, the performance drops. The highest BLEU scores 33.16 and 24.67 are obtained at the topic setting $K = 150$. In fact, our Dis-Model gains higher performance in BLEU than both the traditional hiero baseline and the "CountFeat" method with all topic settings. The "CountFeat" method rewards translation hypotheses containing bilingual term pairs. However it does not explore any domain information. Our Dis-Model incorporates domain information to conduct translation disambiguation and achieves higher performance. When the topic number is set to 150, we gain the highest BLEU score, which is higher than that of the baseline by 0.73 and 0.53 BLEU points on MT06 and MT08, respectively. The final gain over the baseline is on average 0.63 BLEU points.

We conducted the second group of experiments to study whether the term translation consistency model (Cons-Model) is able to improve the performance in BLEU, as well as to investigate the impact of different topic numbers on the Cons-Model. Results are shown in Table 2, from which we observe the similar phenomena to what we have found in the Dis-Model. Our Cons-Model gains higher BLEU scores than the baseline system and the "CountFeat" method with all topic settings. Setting topic number to 150 achieves the highest BLEU score, which is higher than baseline by 0.89 BLEU points and 0.70 BLEU points on MT06 and MT08 respectively, and on average 0.79 BLEU points.

We also conducted experiments to verify the effectiveness of the term bracketing model (Brack-Model), which conducts bracketing prediction for source terms. Results in Table 2 show that our Brack-Model gains higher BLEU scores than those of the baseline system and the "CountFeat" method. The final gain of Brack-Model over the baseline is 0.66 BLEU points and 0.52 points on MT06 and MT08 respectively, and on average 0.59 BLEU points.

5.4 Combination of the Three Models

As shown in the previous subsection, the term translation disambiguation model, consistency model and bracketing model substantially outperform the baseline. Now, we investigate whether using these three models simultaneously can lead to further improvements. The last row in Table 2 shows that the combination of the three models (Combined-Model) achieves higher BLEU score than all single models, when we set the topic number to 150 for the term translation disambiguation model and consistency model. The final gain

Models	MT06	MT08
Best-Dis-Model	30.89	30.14
Best-Cons-Model	38.04	36.70
Brack-Model	60.46	55.78
Combined-Model	54.39	50.85

Table 3: Percentage (%) of 1-best translations which are generated by the Combined-Model and the three single models with best settings on the development test set MT06 and the final test set MT08. The topic number is 150 for Best-Dis-Model and Best-Cons-Model.

of the Combined-Model over the baseline is 1.16 BLEU points and 0.85 points on MT06 and MT08 respectively, and on average 1.00 BLEU points.

5.5 Analysis

In this section, we investigate to what extent the proposed models affect the translations of test sets. In Table 3, we show the percentage of 1-best translations affected by the Combined-Model and the three single models with best settings on test sets MT06 and MT08. For single models, if the corresponding feature (disambiguation, consistency or bracketing) is activated in the 1-best derivation, the corresponding model has impact on the 1-best translation. For the Combined-Model, if any of the corresponding features is activated in the 1-best derivation, the Combined-Model affects the 1-best translation.

From Table 3, we can see that 1-best translations of source sentences affected by any of the proposed models account for a high proportion (30%∼60%) on both MT06 and MT08. This indicates that all proposed models play an important role in the translation of both test sets. Among the three proposed models, the Brack-Model is the one that affects the largest number of 1-best translations in both test sets. And the percentage is 60.46% and 55.78% on MT06 and MT08 respectively. The Brack-Model only considers source terms during decoding, while the Dis-Model and Cons-Model need to match both source and target terms. The Brack-Model is more likely to be activated. Hence the percentage of 1-best translations affected by this model is higher than those of the other two models. Since we only investigate the 1-best translations generated by the Combined-Model and single models, the translations generated by some single models (e.g., Brack-Model) may not be generated by the Combined-Model. Therefore it is hard to say that the numbers of 1-best translations affected by the Combined-Model must be greater than those of single models.

6 Conclusion and Future Work

We have studied the three issues of term translation and proposed three different term translation models for document-informed SMT. The term translation disambiguation model enables the decoder to favor the most suitable domain-specific translations with domain information for source terms. The term translation consistency model encourages the decoder to translate source terms with a high domain translation consistency strength into target terms rather than other new strings. Finally, the term bracketing model rewards hypotheses that translate bracketable terms into continuous target strings as a whole unit. We integrate the three models into a hierarchical phrase-based SMT system[7] and evaluate their effectiveness on the NIST Chinese-English translation task with large-scale training data. Experiment results show that all three models achieve significant improvements over the baseline. Additionally, combining the three models achieves a further improvement. For future work, we would like to evaluate our models on term translation across a range of different domains.

Acknowledgments

This work was supported by National Key Technology R&D Program (No. 2012BAH39B03) and CAS Action Plan for the Development of Western China (No. KGZD-EW-501). Deyi Xiong's work was supported by Natural Science Foundation of Jiangsu Province (Grant No. BK20140355). Qun Liu's work was partially supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the CNGL at Dublin City University. Sincere thanks to the anonymous reviewers for their thorough reviewing and valuable suggestions. The corresponding author of this paper, according to the meaning given to this role by University of Chinese Academy of Sciences and Soochow University, is Deyi Xiong.

[7] Our models are not limited to hierarchical phrase-based SMT. They can be easily applied to other SMT formalisms, such as phrase- and syntax-based SMT.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Houda Bouamor, Aurélien Max, and Anne Vilnat. 2012. Validation of sub-sentential paraphrases acquired from parallel monolingual corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 716–725. Association for Computational Linguistics.

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

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 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: short papers-Volume 2*, pages 176–181.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. *Journal of The balancing act: Combining symbolic and statistical approaches to language*, 1:49–66.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 115–119. Association for Computational Linguistics.

Xiaorong Fan, Nobuyuki Shimizu, and Hiroshi Nakagawa. 2009. Automatic extraction of bilingual terms from a chinese-japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 41–45. ACM.

Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Research and Advanced Technology for Digital Libraries*, pages 585–604. Springer.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 909–919.

Liane Guillou. 2013. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 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*, pages 1179–1190.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014a. Dynamic topic adaptation for phrase-based mt. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden*.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014b. Dynamic topic adaptation for smt using distributional profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 445–456, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Hans Hjelm. 2007. Identifying cross language term equivalents using statistical machine translation and distributional association measures. In *Proceedings of 16th Nordic Conference of Computational Linguistics Nodalida*, pages 97–104.

Masaki Itagaki and Takako Aikawa. 2008. Post-mt term swapper: Supplementing a statistical machine translation system with a user dictionary. In *LREC*.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. *Proceedings of MT summit XI*, pages 269–274.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54.

Els Lefever, Lieve Macken, and Veronique Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 496–504.

Magnus Merkel and Jody Foo. 2007. Terminology extraction and term ranking for standardizing term banks. In *Proceedings of 16th Nordic Conference of Computational Linguistics Nodalida*, pages 349–354.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167.

Scott SL Piao, Guangfan Sun, Paul Rayson, and Qi Yuan. 2006. Automatic extraction of chinese multiword expressions with a statistical tool. In *Workshop on Multi-word-expressions in a Multilingual Context held in conjunction with the 11th EACL, Trento, Italy*, pages 17–24.

Pinis and Skadins. 2012. Mt adaptation for under-resourced domains–what works and what not. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012*, volume 247, page 176. IOS Press.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904.

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

Jörg Tiedemann. 2010. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15.

Ferhan Ture, Douglas W Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426. Association for Computational Linguistics.

Muriel Vasconcellos, Brian Avey, Claudia Gdaniec, Laurie Gerber, Marjorie León, and Teruko Mitamura. 2001. Terminology and machine translation. *Handbook of Terminology Management*, 2:697–723.

Thuy Vu, Ai Ti Aw, and Min Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of the third international joint conference on natural language processing*.

Billy Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068.

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

Deyi Xiong and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13), Bellevue, Washington, USA, July*.

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

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü, and Qun Liu. 2013a. Modeling lexical cohesion for document-level machine translation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2183–2189. AAAI Press.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. 2013b. Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1563—1573.

Bing Zhao and Eric P Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 969–976.