Artificial Intelligence 232 (2016) 54–75

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Topic-based term translation models for statistical machine translation

Deyi Xiong [a,*], Fandong Meng [b], Qun Liu [b,c]

[a] Soochow University, Suzhou, China
[b] Institute of Computing Technology, China
[c] School of Computing, Dublin City University, Ireland

ARTICLE INFO

Article history:
Received 24 August 2014
Received in revised form 9 December 2015
Accepted 14 December 2015
Available online 18 December 2015

Keywords:
Term
Term translation disambiguation
Term translation consistency
Term unithood
Statistical machine translation

ABSTRACT

Term translation is of great importance for machine translation. In this article, we investigate three issues of term translation in the context of statistical machine translation 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 unithood model that rewards translation hypotheses where source terms are translated into target strings as a whole unit. We integrate the three models into hierarchical phrase-based SMT and evaluate their effectiveness on NIST Chinese–English translation with large-scale training data. Experiment results show that all three models can achieve substantial improvements over the baseline. Our analyses also suggest that the proposed models are capable of improving term translation.

1. Introduction

A term is a linguistic expression that is used as the designation of a defined concept in a language (ISO 1087). The following sentences provide several term examples (in Italic).

Cambodia and Vietnam jointly hold *commodity exhibition*.
Indonesia reiterated its opposition to *foreign military presence*.
Native Mandarin speakers teach you *Chinese as foreign language*.

As shown in these examples, terms are compound words that are composed of nouns, adjectives and prepositions in special linguistic patterns.

As terms convey concepts of a text, appropriately translating terms is 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 [1]. In this article, we study domain-specific and context-sensitive term translation in the context of statistical machine translation (SMT).

* Corresponding author.
E-mail addresses: dyxiong@suda.edu.cn (D. Xiong), mengfandong@ict.ac.cn (F. Meng), liuqun@ict.ac.cn (Q. Liu).

http://dx.doi.org/10.1016/j.artint.2015.12.002
0004-3702/

Table 1
Translation examples from the NIST MT02 Chinese-to-English test set. The underlined and underwaved words are source terms and their counterparts in baseline and reference translations, which highlight the three issues of term translation (ambiguity, consistency and unithood).

Eg. 1	Source	dan4 you2yu2 chang2gui1 sai4 yi3 lin2jin4 wei3sheng1, hua2sheng4dun4 qi2cai2 dui4 si4hu1 nan2yi3 yin1ci3 er2 chong1ji2 ji4 hou4 sai4
	Baseline	but because of conventional tournament is nearing an end, Washington Wizards team seems difficult to result after shocks quarter respectively
	Reference	However, as the regular season is approaching its end, it seems hard for Washington Wizards to impact the after season games as a result of this
Eg. 2	Source	yin4ni2 chong2shen1 fan3dui4 wai4guo2 jun1dui4 jin4zhu4... chong2shen1 fan3dui4 wai4guo2 jun1dui4 jin4zhu4 zhe4ge4 dao3guo2
	Baseline	Indonesia reiterates rejection of foreign military presence... reaffirming their opposition to foreign troops stationed in the island
	Reference	Indonesia Reiterated its Opposition to Foreign Military Presence... reiterated its opposition to foreign military presence in this island country

In order to achieve this goal, we focus on three issues of term translation: 1) ambiguity, 2) consistency and 3) unithood. First, term translation ambiguity is related to multiple translations of the same term in different domains. A source term may have different translations when it occurs in different domains. Second, term translation consistency is about consistent translations of 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, term unithood[1] concerns whether a multi-word term is still a unit after translation. Normally, a multi-word source term is translated as a whole unit into a contiguous target string.

Table 1 demonstrates the three issues of term translation with two Chinese-to-English translation examples. The first translation example (Eg. 1) visualizes two issues of term translation: ambiguity and unithood. In regard to the term translation ambiguity, the underlined source term "chang2gui1 sai4" can be translated into either "conventional tournament" or "regular season". The latter translation "regular season" is more widely used in the specific domain of NBA basketball games. Therefore given the domain of Eg. 1, "regular season" is a more appropriate translation for "chang2gui1 sai4" than "conventional tournament" that is chosen by the machine-generated baseline translation. As for the term unithood, the underwaved source term "ji4hou4 sai4" should be translated as a unit into target string "after season games". Unfortunately, the baseline translation violates the unithood constraint of this source term and translates it into an inconsecutive phrase that is interrupted by word "shocks".

The second translation example (Eg. 2) is related to term translation consistency. In this example, we display two sentences in the same text. The underlined source term "wai4guo2 jun1dui4 jin4zhu4" is not translated consistently in the baseline translations. It is translated as "foreign military presence" in the first sentence while "foreign troops stationed" in the second sentence (an undesirable translation).

In order to address these three issues of term translation, we propose a topic-based framework to model term translation for SMT. We capitalize on document-level topic information to disambiguate term translations in different documents and to maintain consistent translations for terms that occur in the same document. In particular, we propose the following three models.

- Term Translation Disambiguation Model: In this model, we condition the translations of source terms in different documents on the topic distributions of corresponding documents. In doing so, we enable the decoder to favor translation hypotheses with topic-specific term translations.
- Term Translation Consistency Model: We introduce a topic-dependent translation consistency metric for each source term to measure how consistently it is translated across documents in training data. With this metric, we encourage 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.
- Term Unithood Model: We explore rich contextual information in the term unithood model to calculate how likely a source term should remain contiguous after translation. We use this unithood model to reward translation hypotheses where multi-word terms are translated as a whole unit.

A bilingual term bank is required to build these three models. We construct this term bank from our bilingual training data via automatic term extraction methods. We use a hierarchical phrase-based SMT system [3] to validate the effectiveness of the three term translation models. Large-scale experiment results show that they are all able to achieve substantial improvements of up to 0.88 BLEU points over the baseline. When simultaneously integrating the three models into SMT, we can gain a further improvement. The combination of the three models outperforms the baseline by up to 1.27 BLEU points.

The three term translation models have been first presented in our previous paper [4]. In this article, we make significant extensions to our previous work. First, for the purpose of completeness, we provide a background introduction of SMT and topic modeling, more details about bilingual term extraction, especially how we pair monolingual terms into bilingual terms

[1] Term unithood is defined as "*the degree of strength or stability of syntagmatic combinations and collocations*" by Kageura and Umino [2]. In this article we are interested in the unithood property of a target translation of a term.

based on word alignments, as well as more details about how we calculate the proposed term translation consistency model. Second, we conduct new experiments to study the impact of the size of extracted bilingual term bank on the three models and the impact of topic information on the term translation consistency and unithood model. We also carry out experiments to compare different bilingual term extraction methods and different approaches to consistent term translation. Finally, we provide in-depth analyses on extracted bilingual terms and translation outputs to demonstrate why and how the proposed term translation models improve translation quality.

The remainder of this article is organized as follows. Section 2 begins with a brief introduction on statistical machine translation and topic modeling as background knowledge. Section 3 describes the process of bilingual term extraction. Section 4 elaborates the proposed three models for term translation with details of training. Section 5 introduces how we integrate the three models into SMT. Section 6 conducts experiments to validate the effectiveness of the proposed models. Section 7 presents in-depth analyses of extracted bilingual terms and translation results. Section 8 gives a brief overview of related work and highlights the differences of our work from them. Finally, we conclude and provide directions for future work in Section 9.

2. Background

Before we present our term translation models, we provide a brief introduction of statistical machine translation and topic modeling in this section. This will help build relevant background knowledge.

2.1. Statistical machine translation

SMT is one of the most popular machine translation paradigms, which relies on statistical models to capture translation equivalents between the source and target language. Most SMT systems adopt a log-linear model [5] to find the best translation $\hat{e}$ among all possible translations for a given source sentence f, which can be formulated as follows:

$$\begin{aligned}\hat{e} &= \underset{e}{\operatorname{argmax}} \left\{ \frac{\exp\left[\sum_1^M \lambda_m h_m(f,e)\right]}{\sum_{e'} \exp\left[\sum_1^M \lambda_m h_m(f,e')\right]} \right\} \\ &= \underset{e}{\operatorname{argmax}} \left\{ \exp\left[\sum_{m=1}^{M} \lambda_m h_m(f,e)\right] \right\} \\ &= \underset{e}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f,e) \right\} \end{aligned} \tag{1}$$

where $h_m(f,e)$ are feature functions defined on the source sentence f and its corresponding translation e, λ_m are weights of feature functions.

In the SMT literature, feature functions $h_m(f,e)$ are also referred to as sub-models of the log-linear model, or just models for simplicity. Normally, the log-linear model of SMT includes a language model that measures the fluency of a generated target translation, a translation model that estimates the probabilities of translation equivalents, a reordering model that captures the word order differences between the source and target language, as well as other models that incorporate knowledge useful for machine translation.

These sub-models are trained separately and independently. Trained sub-models are then combined into the log-linear model of SMT with associated weights. The weights λs are tuned via algorithms such as the Minimum Error Rate Training (MERT) [5] or Margin Infused Relaxed Algorithm (MIRA) [6]. We choose the best weights by optimizing the log-linear model towards some translation quality metrics such as BLEU [7], instead of maximizing the mutual information of the log-linear model. This is because feature weights learned by maximizing the mutual information are not necessarily optimal with respect to translation quality [5].

A full introduction of SMT is far beyond the scope of this article. We refer readers to the textbook "Statistical Machine Translation" [8] or book "Linguistically Motivated Statistical Machine Translation" [9] for more details on SMT.

2.2. Topic modeling

Generally, topic models are statistical models that automatically learn hidden topics in a collection of documents (i.e., corpus) in an unsupervised fashion. They do not require any annotations of documents. They analyze words of a corpus and infer latent distributional patterns in a text. Most topic models are extensions of Latent Dirichlet allocation (LDA) model [10], which is a generalization of the probabilistic latent semantic analysis (pLSA) [11], an early topic model.

In LDA, a document D is assumed to be a mixture of topics, or a distribution $p(\mathbf{z}|D)$ over a set of topics $\mathbf{z}$ while a topic z is defined as a distribution $p(\mathbf{w}|z)$ over a fixed vocabulary $\mathbf{w}$. The former distribution is referred to as per-document topic distribution and the latter per-topic word distribution. Both distributions are assumed to have a Dirichlet prior. Given

a corpus, the generative process of LDA is as follows: 1) draw a per-document topic distribution θ_i for the ith document from a Dirichlet distribution Dir(α); 2) for each word $w_{i,j}$ in the ith document D_i, 2.1) draw a topic assignment $z_{i,j} \sim$ Multinomial(θ_i); and 2.2) draw a word $w_{i,j} \sim \varphi_{z_{i,j}}$ where $\varphi_{z_{i,j}}$ is a per-topic word distribution drawn from a Dirichlet distribution Dir(β).

Latent variables and distributions of the LDA model, e.g., the set of topics, topic assignments of words, per-document topic distribution and per-topic word distribution, can be learned via Bayesian inference algorithms, such as variational inference [10] and collapsed Gibbs sampling [12].

Although topic models are first described and applied in the context of natural language processing, they are also adapted to find hidden patterns in other data, e.g., images. And various extensions to the LDA model have been proposed, such as polylingual topic models [13]. We refer readers to a general introduction of topic models [14] by David Blei for more details about topic modeling.

3. Bilingual term extraction

Bilingual term extraction is to extract terms from two languages with the purpose of creating or expanding a bilingual term bank, which in turn can be used to improve other tasks such as information retrieval and machine translation. In this article, we want to automatically build a bilingual term bank so that we can model term translation to improve translation quality of SMT. Particularly, our interest is to extract multi-word terms.

There are 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 measures, such as the C-value/NC-value [15,16], or other common co-occurrence measures such as the Likelihood Ratio, Dice coefficient and Pointwise Mutual Information [17,18]. The extracted monolingual terms are then paired together [19–21]. 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 [22,23].

We adopt the first strategy to build our bilingual terminology.[2] We first extract monolingual term candidates from the source and target language, and then pair them according to word alignments. The following two subsections will introduce how we extract monolingual term candidates and pair them into bilingual terms respectively.

3.1. Extracting monolingual terms

We extract monolingual terms using two methods: one with the C-value/NC-value measure and the other with the Log-Likelihood Ratio (LLR) measure. For the C-value/NC-value measure based term extraction, we implement it in the same way as described by Frantzi et al. [15]. This method combines linguistic and statistical properties of terms and correspondingly runs in two steps: linguistic and statistical step. In the linguistic step, it recognizes all linguistic units according to the three linguistic patterns (mainly noun phrases) listed as follows:

$$\begin{aligned} &Noun^{+}Noun \\ &(Adj|Noun)^{+}Noun \\ &((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)^{?})(Adj|Noun)^{*})Noun \end{aligned} \tag{2}$$

which are written as regular expressions. “NounPrep” is a combination of a noun and preposition, e.g., “language of” in the term “language of instruction”. The three linguistic patterns are used to capture linguistic structures of terms.

In the statistical step, the C-value/NC-value method measures statistical properties of multi-word term candidates that pass the linguistic filters[3] in the first step. Statistical properties of terms include their occurrence frequencies in a corpus, frequencies that nested terms (terms appearing in other longer terms) overlap with longer term candidates. The C-value is used to extract multi-word nested terms while the NC-value is used to incorporate context information into the C-value for general term extraction.

Specifically, the C-value measures the degree to which a term candidate is related to domain-specific context based on the frequency of the candidate and the frequency that the candidate is a substring of other longer candidate terms (i.e., a nested term). Given a candidate term a, the C-value of a can be formulated as follows:

$$C(a) = \log_2|a| \cdot \left(f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b \right) \tag{3}$$

where f_a is the frequency of candidate a with $|a|$ words, T_a is a set of candidate terms that contain a as a sub-part, $|T_a|$ is the number of items in T_a.

[2] Both strategies can be used to build the bilingual term bank for our models. However, since the term unithood model only requires source-side terms, the first strategy provides flexibility if we only use this model.

[3] 0.99% Chinese phrases and 1.11% English phrases pass these filters in our experiments.

Table 2
The number of terms extracted by C-value/NC-value (C/NC) and LLR from the target side of our training data.

	C/NC	LLR	Overlap
#Terms	2.46M	2.60M	0.60M

Table 3
Examples of terms extracted only by C-value/NC-value (C/NC), by LLR and by both.

Only by C/NC	special programme assistance committee government information office international campaign
Only by LLR	small island developing states privately financed infrastructure projects transnational organized crime locally recruited staff
By both	international tribunal security council central military commission international atomic energy agency

Once the C-value is calculated, it is used to further compute the NC-value that combines the C-value and a score based on context words, which is called N-value. The N-value is formulated as follows:

$$\begin{aligned} N(a) &= \sum_{b \in C_a} w(b) f(b) \\ w(b) &= t(b)/T \end{aligned} \tag{4}$$

where b is a term context word, C_a is the set of distinct term context words, $f(b)$ is the frequency of b as a context word for a, $w(b)$ is the weight of b, defined as the number of terms appearing with b (t(b)) over the total number of terms T. A term context word is an adjective, noun or verb that either precedes or follows a candidate term in a 5-word window. The NC-value is therefore computed as follows:

$$NC(a) = 0.8C(a) + 0.2N(a) \tag{5}$$

For the LLR metric based term extraction, we implement it following Daille [17], who estimates the propensity of two words to appear together as a multi-word expression. The LLR of word a and b is defined as follows:

$$\begin{aligned} LLR(a,b) &= \log L(f_{ab}, f_a, f_a/N) + \log L(f_b - f_{ab}, N - f_a, f_b/N) \\ &\quad - \log L(f_{ab}, f_a, f_{ab}/N) - \log L(f_b - f_{ab}, N - f_a, (f_b - f_{ab})/N - f_a) \end{aligned} \tag{6}$$

where f_a, f_b, f_{ab} are the frequency of the occurrence of a, b and the co-occurrence of a and b respectively, $L(x, y, z)$ is a function defined as $z^x(1-z)^{y-x}$. We also adopt LLR-based hierarchical reducing algorithm proposed by Ren et al. [21] so that we can extract terms with arbitrary lengths.

The C-value/NC-value extraction method obtain terms strictly satisfying the linguistic rules defined in the equation (2). In contrast, the LLR method extracts terms without using any linguistic constraints. Table 2 shows the number of terms extracted from the target side of our training data by these two methods. As shown in the table, LLR extracted more terms than C-value/NC-value does as it does not need to satisfy linguistic constraints. Terms extracted by both LLR and C-value/NC-value account for around 25% of terms extracted by either of the two methods.

We also give examples of terms that are extracted only by C-value/NC-value, only by LLR or by both in Table 3. From the table, we can observe that the C-value/NC-value method is not able to find some term patterns that do not satisfy the linguistic rules in the equation (2). For example, both “privately financed infrastructure projects” and “transnational organized crime” cannot be extracted by C-value/NC-value. However, these terms can be detected by LLR as words in them often appear together.

As shown in Tables 2 and 3, C-value/NC-value and LLR are complementary to each other. We therefore combine the two sets of term candidates that are separately extracted by these two methods to construct our monolingual term bank.

3.2. Pairing monolingual terms

After we extract two sets of monolingual terms from the source and target side with the methods described above, we pair monolingual terms into bilingual terms based on word alignments. In particular, for each extracted source term t_f, we find all target phrases A_{t_f} that are aligned to t_f. And for each of these target phrases $t_e \in A_{t_f}$,

Table 4
Term examples showing the changes after we use the heuristic pairing rules. $i:j$ denotes the word alignment between a source word i and a target word j.

Eg. 1	before	guo2ji4 shi4wu4 | handling international affairs | 1:1 1:2 2:3
	after	guo2ji4 shi4wu4 | international affairs | 1:1 2:2
Eg. 2	before	xin1wen2 zhu3bo1 | news anchor who | 1:1 2:2 2:3
	after	xin1wen2 zhu3bo1 | news anchor | 1:1 2:2
Eg. 3	before	lv4se4 ping2zhang4 | "green screen" | 1:1 1:2 2:3 2:4
	after	lv4se4 ping2zhang4 | green screen | 1:1 2:2

- If t_e is also a term on the target side, we store the source and target term as a bilingual term pair $\langle t_f, t_e \rangle$.
- If not, we use the following heuristic rules.
  - If the probability[4] that the leftmost word of t_e is aligned to a word on the source side is less than a threshold τ, we change t_e to $t_e^{\dashv}$ by removing the leftmost word from t_e (see Eg. 1 in Table 4). If $t_e^{\dashv}$ is a term, we store $\langle t_f, t_e^{\dashv} \rangle$ in our bilingual term bank.
  - Similarly, if the alignment probability of the rightmost word of t_e is less than τ, we remove the rightmost word of t_e to obtain ${}^{\vdash}t_e$ (see Eg. 2 in Table 4). If ${}^{\vdash}t_e$ is a term, we get a bilingual term pair $\langle t_f, {}^{\vdash}t_e \rangle$.
  - If the alignment probabilities of the leftmost and rightmost word are both less than τ, we remove them to obtain ${}^{\vdash}t_e^{\dashv}$ (see Eg. 3 in Table 4), which will be paired with t_f if it is a term.

These heuristic rules are used to reduce the impact of word alignment errors on bilingual term extraction. Some examples are provided in Table 4 to show how we use these rules. We empirically set the threshold $\tau = 0.2$ according to our preliminary experiment results.

4. Models

In this section, we elaborate the three models proposed for term translation. They are the (A) term translation disambiguation model, (B) term translation consistency model and (C) term unithood model respectively.

4.1. Model A: term translation disambiguation

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.

There are a great variety of different topic models that can infer topic distributions of documents. As we mentioned in Section 2.2, most of them use Latent Dirichlet Allocation (LDA) [10] as their foundation. Without loss of generality, we exploit the LDA topic model for inferring topic distributions of documents.

For each term pair in the bilingual term bank created from training data as described in the last section, we calculate the source-to-target term translation probabilities 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.

The probability $p(t_e|t_f, z=k)$ is computed via maximum likelihood estimation using counts from training data. For each bilingual term pair $\langle t_f, t_e \rangle$, the source part of which occurs in a document D with a topic distribution $p(z|D)$ estimated via the LDA model, we collect an instance $(t_f, t_e, p(z|D), c)$, where c is the fraction count of the instance as described by Chiang [25]. In this way, we can obtain a set of instances $\mathcal{I} = \{(t_f, t_e, p(z|D), c)\}$ with different per-document topic distributions for each bilingual term pair. Using these instances, we calculate the probability $p(t_e|t_f, z=k)$ as follows:

$$p(t_e|t_f, z=k) = \frac{\sum_{i \in \mathcal{I}, i.t_e = t_e, i.t_f = t_f} i.c \times p(z=k|D)}{\sum_{i \in \mathcal{I}, i.t_f = t_f} i.c \times p(z=k|D)} \tag{7}$$

Table 5 displays examples of bilingual terms with their topic-conditioned translation probabilities. For each bilingual term, we only show the topic t where the bilingual term has the highest topic-conditioned translation probability $p(t_e|t_f, z=t)$, ignoring translation probabilities under other topics. We can clearly see that the same source term "fang2yu4 xi4tong3" is translated differently under different topics. For example, in the health domain, the term is tightly related to immune mechanisms. But in the military/politics domain, "defense system" is a widely-used translation for this term. This is also true for translations of the source term "zhan4lue4 si1xiang3" that occurs in different domains with slight meaning shifts.

[4] The probability is computed as $(p(e|f) + p(f|e))/2$, where $p(e|f)$ and $p(f|e)$ are lexical translation probabilities estimated using relative counts [24].

Table 5
Examples of bilingual terms with topic-conditioned translation probabilities. TD: topic descriptions that are manually generated according to top-15 topical words.

Source	Target	$p(t_e\|t_f, z=k)$	TD
fang2yu4 xi4tong3	defence mechanisms	0.00097	Health
	defense program	0.011	Science
	defense system	0.87	Politics
	prevention system	0.033	News
zhan4lue4 si1xiang3	strategic concept	0.15	Crime
	strategic doctrines	0.34	Military
	strategic principle	0.038	Reform
	strategic thoughts	0.040	Reform
	strategic vision	0.73	Administration

We associate each extracted bilingual term pair in the bilingual term bank with its corresponding topic-conditioned translation probabilities estimated in the equation (7). When translating sentences of document D', we first get the topic distribution of D' via LDA. Given a sentence which contains T terms $\{t_f^i\}_1^T$ in D', our term translation disambiguation model *TermDis* can be denoted as

$$TermDis = \prod_i^T P_d(t_e|t_f^i, D') \tag{8}$$

where the conditional source-to-target term translation probability $P_d(t_e|t_f^i, D')$ given the document D' is formulated as follows:

$$P_d(t_e|t_f^i, D') = \sum_{k=1}^{K} p(t_e|t_f^i, z=k) \times p(z=k|D') \tag{9}$$

Whenever a source term t_f^i is translated into t_e, we check whether the pair of t_f^i and its translation t_e can be found in our bilingual term bank. If can be found, we calculate the conditional translation probability from t_f^i to t_e given the document D' according to the equation (9).

Through the term translation disambiguation model, 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. Model B: term translation consistency

The term translation disambiguation model enables the decoder to 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. A straightforward way to address this issue is to simply count how many times a term translation is reused in recently translated sentences. Those term translations frequently reused will be encouraged. However, such term translations selected by the method are not necessarily correct or appropriate translations under current topics. We therefore propose a topic-based term consistency model. Instead of simply counting the number of times of a term translation being used in previously translated sentences, we enable the decoder to detect terms that are highly consistently translated under given topics in training data with the proposed model, and encourage the decoder to translate these terms.

The essential component of our term translation consistency model is the translation consistency strength of a source term estimated under the per-document topic distribution. We describe how to calculate it before introducing the whole model.

For 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 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:

$$\begin{aligned} cons(t_f, k) &= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\left(\frac{q_m^n \times p(k|D_m)}{Q_k}\right)^2 \\ Q_k &= \sum_{m=1}^{M}\sum_{n=1}^{N_m} q_m^n \times p(k|D_m) \end{aligned} \tag{10}$$

Table 6
Values of $cons(t_f, 30)$ for the source term "fang2yu4 xi4tong3" under the topic 30. DID is the document ID.

DID	t_e	$(\frac{q_D^n \times p(k\|D)}{Q_k})^2$	$\sum_{n=1}^{N_D}(\frac{q_D^n \times p(k\|D)}{Q_k})^2$
...	...	...	...
10797	defense system	3.96e-05	4.21e-05
	defensive system	2.48e-06	
50648	defence system	2.54e-07	5.08e-07
	defense programmes	2.54e-07	
66379	defense system	3.80e-06	4.23e-06
	missile system	4.23e-07	
...	...	...	...
$\sum_{m=1}^{M}\sum_{n=1}^{N_D}(\frac{q_D^n \times p(k\|D)}{Q_k})^2$	0.00407		

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 D_m, q_m^n is the frequency of the nth translation of t_f ($\in Set(t_e)$) in document D_m, $p(k|D_m)$ is the conditional probability of document D_m over topic k, and Q_k is the normalization factor. All translations of t_f are from $Set(t_e)$.

The term consistency strength $cons(t_f, k)$ measures how consistently a term is translated in a domain. It is computed according to two factors: 1) the number of translation variations of a term t_f in a domain and 2) topic-related frequencies of these variations $q_m^n \times p(k|D_m)$. If a term is always translated into only one unique translation in a domain, the consistency strength will be the highest. If there are multiple translation variations for a term in a domain and most translations are concentrated on one translation variation, the consistency strength will remain high. If there are multiple variations and the majority of translations do not concentrate, the consistency strength will be low.

Table 6 shows an example of a source term "fang2yu4 xi4tong3" with its various translations in different documents. We calculate the translation consistency values of the term in each document under the topic 30 (shown in the last column of Table 6). Summing up all these values in all documents, we can obtain the term translation consistency strength of "fang2yu4 xi4tong3" under the topic 30 (as shown in the last row of Table 6).

We adapt Itagaki et al. [26]'s translation consistency index for terms to our topic-based translation consistency measure in the equation (10). The significant difference between our term translation consistency measure and their consistency index is that we take the topic distributions of documents into account when we calculate the term translation consistency strength. Table 7 shows the translation consistency strength values of term "fang2yu4 xi4tong3" under different topics. We display the first 5 topics with the highest values of $cons(t_f, k)$ and the last 5 topics with lowest values of $cons(t_f, k)$ in the table. From Table 7, we can observe that

- In the topics that are most closely related to the term "fang2yu4 xi4tong3" (e.g., military and war), the term is translated flexibly. This may be because the term has several different target translations, all of which are acceptable in these domains.
- In a general domain like news, "fang2yu4 xi4tong3" is consistently translated into a target term that is widely accepted by most people. Yet another reason for the high consistency strength value in the news domain is because newswire services normally use translation memories or handbooks to encourage consistency.[5]
- In the topics that are not quite related to the term "fang2yu4 xi4tong3", such as religion, the term is not consistently translated either.

These suggest that the consistency of term translation is topic-sensitive. The same terms may be translated in a different consistency pattern in different topics. Therefore we calculate the translation consistency strength of a source term t_f based on topic distributions.

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 values calculated in the equation (10). When translating sentences of document D, we first get the topic distribution of D. Given a sentence which contains T terms $\{t_f^i\}_1^T$ in D, our term translation consistency model *TermCons* can be denoted as follows:

$$TermCons = \prod_{i}^{T} \exp(S_c(t_f^i|D)) \tag{11}$$

where the strength of term translation consistency for t_f^i given the document D is formulated as follows:

$$S_c(t_f^i|D) = \log(\sum_{k=1}^{K} cons(t_f^i, k) \times p(k|D)) \tag{12}$$

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

Table 7
Values of $cons(t_f, k)$ for the source term "fang2yu4 xi4tong3". Descriptions for topics are manually generated according to their top-15 topical words.

Topic ID k	Topic description	$cons(t_f, k)$
109	News	0.0682
26	Aerospace	0.0271
34	Party building	0.0241
36	Finance	0.0228
22	Ecology	0.0222
...	...	...
50	Energy	0.00394
53	Military	0.00393
8	War	0.00389
48	Politics	0.00389
90	Religion	0.00387

We translate a document in a sentence-by-sentence manner. During decoding of a sentence, whenever a hypothesis 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 term translation consistency for t_f^i given the document D according to the equation (12). If the topic-dependent consistency strength is very high, this indicates that: 1) the number of translation variations of the term is small or 2) the majority of translations concentrate under a topic distribution similar to that of D in training data. Our model will encourage the decoder to translate this term consistently using translations from the bilingual term bank. If the strength is low, this suggests that the term has many scattered translation variations under the current topic distribution. Our model will either choose to translate a larger term with a high consistency strength, which subsumes the term as a nested term, or just let other models decide how to translate this term. All these topic-sensitive term translation consistency patterns (e.g., patterns shown in Table 7) are learned by the term translation consistency model from training data and used during decoding.

4.3. Model C: term unithood

The term translation disambiguation model and consistency model concern the term translation accuracy with domain information. We further propose a term unithood model to guarantee the integrality of term translation. Xiong et al. [27] propose a syntax-driven bracketing model for phrase-based translation, which predicts whether a target translation of a phrase is contiguous with rich syntactic constraints. It is also desirable for multi-word terms to be contiguous units after translation. We therefore adapt Xiong et al. [27]'s bracketing approach to term translation and build a classifier to measure the probability that a source term should be translated into a contiguous unit.

For all source parts of the extracted bilingual terms, 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 true unithood instance, otherwise a false unithood instance. With these instances, we train a maximum entropy (MaxEnt) binary classifier to predict the unithood ($u \in \{true, false\}$) probability of a given source term t_f within particular contexts $c(t_f)$. The binary classifier is formulated as follows:

$$P_u(u|c(t_f)) = \frac{\exp(\sum_j \theta_j h_j(u, c(t_f)))}{\sum_{u'} \exp(\sum_j \theta_j h_j(u', c(t_f)))} \tag{13}$$

where $h_j \in \{0, 1\}$ is a binary feature function and θ_j is the weight of h_j.

The feature h_j takes the following binary form:

$$h_j(u, c(t_f)) = \begin{cases} 1, & \text{if } u = \mu \text{ and } c(t_f).\psi = \nu \\ 0, & \text{else} \end{cases} \tag{14}$$

where $\mu \in \{\text{true}, \text{false}\}$, ψ represents a contextual element for the source term t_f and ν denotes the value of the element. We use the following contextual elements of a source term to define our 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. Taking the third contextual element as an example, we can define a feature as follows:

$$h_j(u, c(t_f)) = \begin{cases} 1, & \text{if } u = \text{true and } c(t_f).lastword = \text{"sai4"} \\ 0, & \text{else} \end{cases}$$

Given a source sentence which contains T terms $\{t_f^i\}_1^T$, our term unithood model *TermUnit* can be denoted as follows:

$$TermUnit = \prod_i P_u(u|c(t_f^i)) \tag{15}$$

Whenever a hypothesis translates a source term t_f^i into a contiguous unit on the target side, we calculate the unithood probability of t_f^i according to the equation (13). For those source terms that are not translated into a whole unit, we do not calculate this probability.

5. Integration of the three models into SMT

Our Models can be integrated into any SMT formalisms that use the log-linear model for feature combination as described in Section 2.1. This is because we only need to integrate the three models as new features into the log-linear model. The integration itself will neither change values of other features nor methods to calculate these values.

Without loss of generality, we choose hierarchical phrase-based SMT [3], one of state-of-the-art SMT formalisms, to show how we integrate the three models into SMT and to validate their effectiveness on term translation. The integration can be easily adapted to other SMT formalisms. Before we describe the integration, we give a short introduction of hierarchical phrase-based SMT, especially translation rules and the log-linear model of it.

The rules used in hierarchial phrase-based SMT are synchronous context-free grammar rules, which can be represented as follows:

$$X \rightarrow \langle \alpha, \beta, \sim \rangle \tag{16}$$

where X is a nonterminal, α and β are strings that contain terminals (words) and nonterminals on the source and target side respectively, $\sim$ denotes the one-to-one alignment between nonterminals in α and nonterminals in β. These rules can be automatically extracted from word-aligned bilingual training data. If rules only contain terminals, we refer to them as phrase rules since they are the same as bilingual phrase pairs used in phrase-based SMT [24]. Our extracted bilingual term pairs described in Section 3 are exactly phrase rules. In addition to rules that are extracted from bilingual training data, the grammar of hierarchical phrase-based SMT also includes two special glue rules that concatenate nonterminal Xs in a monotonic fashion.

The log-linear model of hierarchical phrase-based SMT can be formulated as follows:

$$w(\mathcal{D}) = \exp\left(\sum_{r \in D} \log(t(r)) + \lambda_{lm} \log P_{lm}(e) + \lambda_{wp}|e| + \lambda_{rp} I\right) \tag{17}$$

where $\mathcal{D}$ is a derivation, $w(\mathcal{D})$ is the score of D, $t(r)$ is the translation probability of rule r, $P_{lm}(e)$ is the language model, $|e|$ is the number of words in the target translation e and I is the number of rules in $\mathcal{D}$. The derivation $\mathcal{D}$ is defined as a set of triples (r, i, j), each of which denotes an application of a rule that spans words from i to j on the source side.

Each of the tree term translation models is treated as a new feature when we integrate them into hierarchical phrase-based SMT. With these three features, the log-linear model shown in the equation (17) is reformulated as follows:

$$\begin{aligned} w(\mathcal{D}) &= \exp(\sum_{r \in D} \log(t(r)) + \lambda_{lm} \log P_{lm}(e) + \lambda_{td} \log(\mathit{TermDis}) \\ &\quad + \lambda_{tc} \log(\mathit{TermCons}) + \lambda_{tu} \log(\mathit{TermUnit}) + \lambda_{wp}|e| + \lambda_{rp} I) \end{aligned} \tag{18}$$

where *TermDis*, *TermCons*, *TermUnit* are the term translation disambiguation, consistency and unithood model calculated according the equation (8), (11), (15) respectively, $\lambda_{td}, \lambda_{tc}, \lambda_{tu}$ are their weights that are tuned with other feature weights via the Minimum Error Rate Training [5]. The integrated three models contribute to term translation selection in proportion to their weights. They collectively enable the decoder to select translation hypotheses that translate terms appropriately and consistently as a whole unit.

The three models only apply to translation rules that contain a source term. On the one hand, as the values of the three models calculated on these rules are either probabilities (the term translation disambiguation and unithood model) or between 0 and 1 (the term translation consistency model), the integrated values of the three models will be minus (after taking the logarithm). Generally, the more rules are applied, the lower the overall score $w(\mathcal{D})$ calculated in the equation (18) will be. On the other hand, the best translation is generated from the best derivation with the highest score according to the equation (18). This indicates that the integration of the three models has a bias towards translation hypotheses generated by fewer rules against those generated by more translation rules. We don't want this. Therefore we add a counting feature to balance this as follows:

$$\begin{aligned} w(\mathcal{D}) &= \exp(\sum_{r \in D} \log(t(r)) + \lambda_{lm} \log P_{lm}(e) + \lambda_{td} \log(\mathit{TermDis}) \\ &\quad + \lambda_{tc} \log(\mathit{TermCons}) + \lambda_{tu} \log(\mathit{TermUnit}) + \lambda_{wp}|e| + \lambda_{rp} I + \lambda_{tw} T) \end{aligned} \tag{19}$$

where T is the number of source terms translated by the derivation $\mathcal{D}$. This counting feature will reward translation hypotheses that are generated with more translation rules.

We can also integrate the three models into SMT one-by-one with the term counting feature. For example, we can only integrate the term translation disambiguation model into SMT as follows:

Table 8
BLEU-4 scores (%) of the combined model (Combined-Model) on the development test set MT06 and the final test set MT08. */+: significantly better than the baseline/CountFeat ($p < 0.01$).

Models	MT06	MT08	Avg
Moses-Chart	32.03	23.91	27.97
Baseline	32.41	24.14	28.28
CountFeat	32.77	24.29	28.53
Combined-Model	**33.68*+**	**25.06*+**	**29.37**

$$w(\mathcal{D}) = \exp(\sum_{r \in D} \log(t(r)) + \lambda_{lm} \log P_{lm}(e) + \lambda_{td} \log(TermDis) + \lambda_{wp}|e| + \lambda_{rp} I + \lambda_{tw} T) \quad (20)$$

6. Experiments

In this section, we evaluated the effectiveness of the proposed term translation models and their combination and variants. We also conducted experiments to investigate the impact of various factors, such as the number of topics and the size of extracted bilingual term bank, on our models. Please refer to Appendix A for details of our experiment setup.

6.1. Overall performance

We incorporated the proposed three models simultaneously into the decoder so as to investigate whether they can collectively obtain significant improvements over the baseline system, a state-of-the-art hierarchical phrase-based system that is built following Hiero [25]. In addition to the baseline, we also compared against an open source system Moses-Chart that is a re-implementation of hierarchical phrase-based SMT system in the Moses framework [28] and a method ("CountFeat") proposed by Ren et al. [21] who use a binary feature to indicate whether a bilingual phrase contains a term. The reason that we compared against the "CountFeat" method is because it is one of the few approaches that consider term translations in the context of SMT. The binary feature in this method is 1 if a target language phrase contains a term otherwise 0.

We set the number of topics $k = 150$ and used the combination of bilingual terms extracted by the LLR method and C-value/NC-value method with all bilingual training data. For more details on the impact of these parameters, please refer to Section 6.2.

Results are reported in Table 8. Our re-implemented hierarchical phrase-based system (baseline) is more competitive than the Moses re-implementation. We also find that the combination of the three models (Combined-Model) achieves higher BLEU scores than the baseline, Moses-Chart and the CountFeat method. The final gain of Combined-Model over the baseline is 1.27 BLEU points and 0.92 points on MT06 and MT08 respectively. It is also significantly better than the CountFeat method by 0.77 BLEU points on the MT08 test set.

6.2. Impacts of various factors

We carried out a series of experiments to further investigate the impacts of various factors on our models. Particularly, we would like to answer the following three questions by this group of experiments.

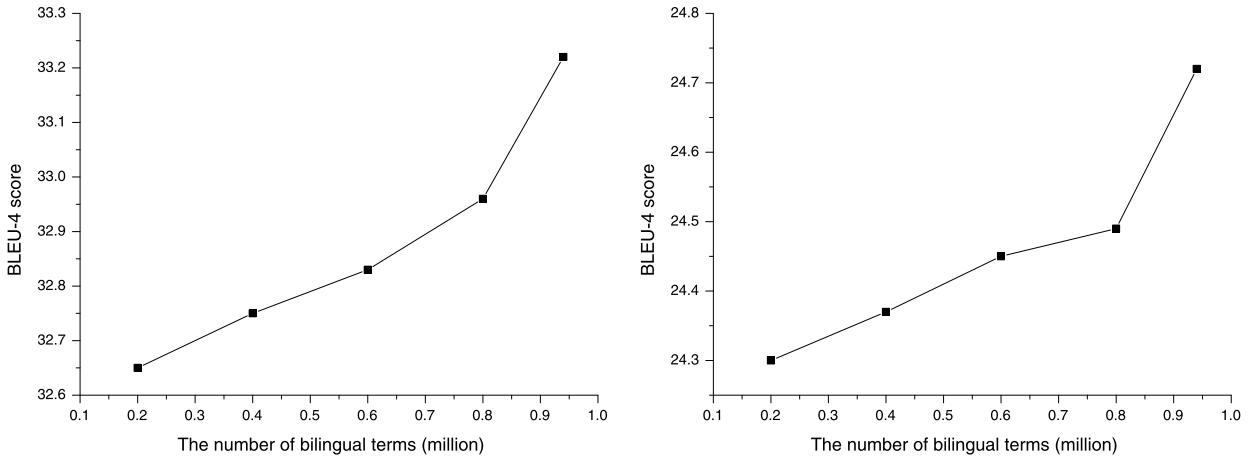
- What's the impact of the topic number k on the term translation disambiguation and consistency models that explore document-level topic information?
- What's the impact of the size of extracted bilingual term bank on the term translation disambiguation and consistency model?
- Which term extraction method, LLR, C-value/NC-value or their combination, is the best method for our term translation models?

6.2.1. Topic number k

As we integrate document-level topic information into the term translation disambiguation and consistence model, we want to study the impact of the number of topics on these two models. We therefore carried out experiments on the development test set with the number of topics k varying from 50 to 200. Table 9 shows the results. From the table, we clearly find that the performance of both models in terms of BLEU goes up when we increase k from 50 to 150 and down when k continuously rises to 200. This trend can be observed again on the final test set. This suggests that the two models obtain the best performance when $k = 150$.

Table 9
BLEU-4 scores (%) of the term translation disambiguation model (Dis-Model) and the term translation consistency model (Cons-Model) on the development test set MT06 and the final test set MT08 with the number of topics $K \in \{50, 100, 150, 200\}$.

Model	Topic number	MT06	MT08	Avg
Dis-Model	$K = 50$	32.97	24.55	28.76
	$K = 100$	33.02	24.69	28.86
	$K = 150$	**33.22**	**24.72**	**28.97**
	$K = 200$	33.12	24.65	28.89
Cons-Model	$K = 50$	32.98	24.67	28.83
	$K = 100$	33.07	24.77	28.92
	$K = 150$	**33.20**	**24.86**	**29.03**
	$K = 200$	33.10	24.62	28.86



(a) MT06 (b) MT08

Fig. 1. The impact of the size of bilingual term bank on the term translation disambiguation model on the test sets MT06 and MT08. The number of term pairs in the extracted bilingual term bank varies from 0.2 million to 0.94 million which is the number of all bilingual terms extracted from the whole bilingual training data.

In order to investigate why this happens, let us suppose that we use the most probable topic of a document as the topic assignment for the document.[6] We find that terms occur in 50 different topics if we set the topic number k to 50, 86 topics if $k = 100$, 110 topics if $k = 150$ and 116 topics if $k = 200$ in our training data. The data sparseness problem is becoming serious when $k = 200$ as terms do not occur in many topics (about 42%). However, the data sparseness problem is under control when $k = 150$ since only 26.7% topics do not have terms. This is further alleviated as we use per-document topic distributions instead of the most probable topics on all documents. The value of $k = 150$ is the best tradeoff between the benefit from using topic information and the cost of data sparseness caused by a larger topic number. Therefore we set the number of topics k to 150 for all experiments hereafter.

6.2.2. Bilingual term bank size

We built several bilingual term banks with different numbers of bilingual term pairs by varying the size of bilingual data, from which we extracted bilingual term pairs according to the method described in Section 3. Specifically, we built 5 different bilingual term banks $\{TB_i\}_1^5$, where $TB_i \subset TB_{i+1}$. We then evaluated the term translation disambiguation model with these 5 bilingual term banks on the test sets MT06 and MT08. The BLEU scores are plotted in Fig. 1. From the figure, we can obviously see that the BLEU score of the term translation model gradually rises from 32.65 to 33.22 on MT06 and 24.30 to 24.72 on MT08 when we feed more bilingual terms to the model. On average, we can obtain an improvement of 0.1 BLEU points for each increase of 0.2 million bilingual terms on the development test set MT06. The curves do not level off even if we extract all bilingual terms from our bilingual training data, this indicates that the upward trend does not stop and that we can potentially achieve further improvement if we have more bilingual training data.

[6] Note that we actually use per-document topic distributions $p(z|D)$ over all topics rather than the most probable topics when we estimate term frequencies of occurrence in all our models.

Table 10
BLEU-4 scores (%) of the three models on the development and final test set with bilingual terms extracted using the C-value/NC-value (NC), LLR method and their combination (NC + LLR) described in Section 3.

Model	Term extraction	MT06	MT08	Avg
Dis-Model	NC	33.01	24.53	28.77
	LLR	33.05	24.65	28.85
	NC + LLR	**33.22**	**24.72**	**28.97**
Cons-Model	NC	33.09	24.71	28.90
	LLR	33.12	24.71	28.92
	NC + LLR	**33.20**	**24.86**	**29.03**
Unit-Model	NC	33.12	24.72	28.92
	LLR	33.05	24.84	28.94
	NC + LLR	**33.29**	**24.93**	**29.11**

Table 11
BLEU-4 scores (%) of the term translation disambiguation model (Dis-Model) on the development test set MT06 and the final test set MT08. "Baseline" is the traditional hierarchical phrase-based system. "CountFeat" is the method that adds a counting feature to reward translation hypotheses containing bilingual terms [21]. */+: significantly better than the baseline/CountFeat ($p < 0.01$).

Models	MT06	MT08	Avg
Baseline	32.41	24.14	28.28
CountFeat	32.77	24.29	28.53
Dis-Model	**33.22*+**	**24.72***	**28.97**

6.2.3. LLR vs. C-value/NC-value in bilingual term extraction

We further carried out experiments to empirically compare different term extraction methods (i.e., LLR vs. C-value/NC-value) on the proposed three term translation models. We report the experiment results in Table 10. From the table, we can observe that

- First, for all three models, the LLR method is marginally better than the C-value/NC-value term extraction method. This may be because that the LLR method extracted more bilingual term pairs than the C-value/NC-value method (1.01M vs. 0.90M).
- Second, if we combine the bilingual terms extracted by the C-value/NC-value and LLR method together, we achieve the best performance for all three models. The combination of the C-value/NC-value and LLR outperforms the C-value/NC-value by up to 0.2 BLEU points.

6.3. Effect of the proposed three models

In this section, we thoroughly validated the effectiveness of the proposed term translation disambiguation model, consistency model and unithood model respectively. In particular,

- We evaluated the term translation disambiguation, consistency and unithood models against the baseline systems.
- We derived new variations of the term translation consistency and unithood models with or without topic information to study the impact of topic information on these two models.
- We compared our term translation consistency model against an alternative consistency method that rewards hypotheses with repeated terms from previously translated sentences.

We used the same parameter setting as described in Section 6.1 in all experiments of this section.

6.3.1. Term translation disambiguation model

We carried out experiments to investigate the effect of the term translation disambiguation model (Dis-Model) and report the results in Table 11. According to the table, our Dis-Model gains higher performance in terms of BLEU than both the baseline system and the "CountFeat" method. The "CountFeat" method rewards translation hypotheses that contain bilingual terms. It does not explore any domain information. In contrast, our Dis-Model incorporates document-level topic information to conduct translation disambiguation. Particularly, the term translation disambiguation model outperforms the "CountFeat" method by up to 0.45 BLEU points. It is also significantly better than the baseline by 0.81 and 0.58 BLEU points on MT06 and MT08, respectively. The final gain over the baseline is 0.69 BLEU points on average.

Table 12
BLEU-4 scores (%) of the term translation consistency model (Cons-Model) on the development test set MT06 and the final test set MT08. "Cons-Count" is a model that rewards a translation hypothesis whenever a previously used target term repeats. "Cons-No-Topic" is a simplified version of Cons-Model, which does not use any topic information. */+: significantly better than the baseline/CountFeat ($p < 0.01$).

Models	MT06	MT08	Avg
Baseline	32.41	24.14	28.28
CountFeat	32.77	24.29	28.53
Cons-Model	**33.20***	**24.86*+**	**29.03**
Cons-Count	32.96*	24.53	28.75
Cons-No-Topic	32.99*	24.56	28.78

6.3.2. Term translation consistency model

We conducted experiments to study whether the term translation consistency model (Cons-Model) is able to improve the performance in terms of BLEU. Results are shown in Table 12. Cons-Model gains significant improvements of 0.79 and 0.72 BLEU points over the baseline system on MT06 and MT08, respectively. It also outperforms the "CountFeat" method by 0.5 BLEU points on average.

We also compared our term translation consistency model against another two models.

- "Cons-Count": This model rewards a translation hypothesis whenever a target term in the hypothesis has already occurred in recently translated sentences. We maintain a counter for each sentence and store all target terms in previously translated sentences in the same document. If a previously used target term repeats itself in the current translation hypothesis, we accumulate the counter.
- "Cons-No-Topic": This model is a simplified version of our term translation consistency model, which does not use any topic information. In this model, the term translation consistency strength $cons(t_f)$ is calculated as follows:

$$\begin{aligned} cons(t_f) &= \sum_{m=1}^{M}\sum_{n=1}^{N_m}\left(\frac{q_m^n}{Q}\right)^2 \\ Q &= \sum_{m=1}^{M}\sum_{n=1}^{N_m} q_m^n \end{aligned} \tag{21}$$

where M, N_m, q_m^n are the same as defined in the equation (10).

Although these two models are better than the baseline and the "CountFeat" method as shown in Table 12, they are worse than our term translation consistency model. Our model outperforms the "Cons-Count" model by up to 0.33 BLEU points. This suggests that the strategy of capturing information of how terms are consistently translated in the training data is better than the strategy of counting the times of a term being consistently translated in a test set on the fly. Our model is also better than the "Cons-No-Topic" model by 0.25 BLEU points on average. This indicates that topic information is useful in modeling term translation consistency.

6.3.3. Term unithood model

We compared the term unithood model against the baseline. Additionally, we also compared the term unithood model against its two variations: "Unit-All" and "Unit-Topic". The Unit-All model predicts the unithood property for any source phrases of length up to 6 words (not limited to terms). In order to train this model, we extracted all source phrases that are translated as a whole unit as true unithood instances, otherwise false unithood instances. The second variation Unit-Topic model incorporates document topics as features into the MaxEnt classifier to predict whether a source term is translated as a whole unit into the target language. We set the topic number $k = 150$.

Experiment results are shown in Table 13. From the table, we can observe that

- The term unithood model achieves an improvement of 0.83 BLEU points over the baseline on average. It also outperforms the "CountFeat" method. This suggests that our term unithood model is useful for term translation.
- The "Unit-All" model also outperforms the baseline. However it is worse than the term unithood model. This might suggest that terms are more likely to be translated as a whole unit than other source phrases (ordinary phrases). The errors in predicting the unithood property of ordinary phrases may be more serious than those for terms. These errors, in turn, jeopardise translation quality.
- The "Unit-Topic" model is marginally worse than the term unithood model. This seems to suggest that document-level topic information is not helpful for predicting the unithood property of terms.

Table 13
BLEU-4 scores (%) of the term unithood model (Unit-Model) on the development test set MT06 and the final test set MT08. "Unit-All" is a model that predicts the unithood property for all source phrases of length up to 6 words. "Unit-All" is a model that incorporates document-level topic information into the unithood prediction. */+: significantly better than the baseline/CountFeat ($p < 0.01$).

Models	MT06	MT08	Avg
Baseline	32.41	24.14	28.28
CountFeat	32.77	24.29	28.53
Unit-Model	**33.29*+**	**24.93*+**	**29.11**
Unit-All	33.03*	24.43	28.73
Unit-Topic	33.11*	24.73*	28.92

Table 14
Statistics of bilingual terms extracted from the training data and sentence pairs containing bilingual terms.

Type	Statistics
Bilingual terms	1.81M
Source terms	1.08M
Sentences with bilingual terms	2.72M

Table 15
Examples of bilingual terms extracted from the training data. "#Doc" means the total number of documents in which the corresponding source term occurs and "#Mdoc" 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 23 different translations of the source term "fang2yu4 xi4tong3".

Source	Target	#Doc	#Mdoc
fang2yu4 xi4tong3	defence mechanisms		
fang2yu4 xi4tong3	defence programmes		
fang2yu4 xi4tong3	defence systems		
fang2yu4 xi4tong3	defense plan	470	56
fang2yu4 xi4tong3	defense programmes		
fang2yu4 xi4tong3	prevention systems		
...	...		
zhan4lve4 dao3dan4	strategic missile	7	0
fang2yu4 xi4tong3	defense system		

7. Analysis

In this section, we will provide more details of our three term translation models from two distinct perspectives: extracted bilingual terms and translations generated by the baseline and the system enhanced with our models. In the first perspective, we will study the distributions of bilingual terms in sentences and documents and evaluate the quality of extracted bilingual terms. In the second perspective, we will take a deeper look at the differences that our models make on target translations. The analysis from these two angles will help us gain some insights into why we need to propose the three models and how the proposed models improve term translation in SMT.

7.1. Analysis on extracted bilingual terms

We provide some statistics of bilingual terms extracted from the training data in this section. First, we show the total number of bilingual terms extracted from the training data and the number of sentences that contain bilingual terms in the training data in Table 14. We can see that the majority of sentences contain bilingual terms ($2.72\text{M}/4.28\text{M} \approx 65.07\%$). On average, a source term has about 1.70 different translations ($1.81\text{M}/1.08\text{M} \approx 1.68$). These statistics indicate that terms are frequently used in real-world data and that a source term can be translated into different target terms.

Second, we also present some bilingual term examples extracted from the training data in Table 15. 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 "fang2yu4 xi4tong3" has 23 different translations in total. They are distributed in 470 documents in the training data. In 414 documents, "fang2yu4 xi4tong3" has only one single translation. However, in the other 56 documents it has different translations. This indicates that "fang2yu4 xi4tong3" is not consistently translated in these 56 documents. Different from this, the source term "zhan4lve4 dao3dan4 fang2yu4 xi4tong3" 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. This percentage sharply increases to 23.49% in the development and test set. The examples and statistics shown here suggest 1) that source terms have domain-specific translations and 2) that terms are not necessarily translated

Table 16
The statistics of bilingual terms in terms of their unithood property. "Certainly True Unithood Ratio" is the percentage of source terms which are always translated into contiguous target strings as a whole unit among all extracted source terms. "True Unithood Instances Ratio" is the percentage of true unithood instances in all instances (true and false unithood instances) which are used to train the term unithood prediction model.

Type	Statistics
Certainly True Unithood Ratio	63.87%
True Unithood Instances Ratio	52.22%

Table 17
The percentages that the source/target/both parts of randomly selected bilingual terms are manually judged as real terms.

Type	Percentage
Source	64.6%
Target	72.3%
Both	64.0%

Table 18
Examples of extracted terms where both/one/neither sides of examples are manually judged as real terms.

Side	Examples
Both	dian4zi3 tong1xun4 she4bei4 ||| telecommunications equipment zhu4ce4 kuai4ji4shi1 ||| chartered certified accountants da4xue2 jiao4yu4 zheng4ce4 bai2pi2shu1 ||| university education policy white paper zhong1yang1 chu4li3 xi4tong1 ||| mainframe systems
Source or target	lian2he2guo2 xin1wen2 zhong1xin1 yong4 ||| united nations information centre zhu4yi4 guo2ji4 an1quan2 ||| international security jian4li4 shu4zi4 wei4xing1 tu2xiang4 ||| digital satellite imagery gong1zhong4 yi4jian4 ting1zheng4hui4 ||| large public hearing
Neither	guo2nei4 zi1yuan2 xi1shao3 ||| scarce domestic resources guo4du4 re4zhong1yu2 ||| excessive enthusiasm jian4li4 huo3ban4 ||| build partnerships jian4kang1 zhuang4kuang4 bu2duan4 e4hua4 ||| deteriorating health

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.

Third, we also give the statistics of the unithood information of extracted bilingual terms in Table 16. We can see that the percentage of source terms that are always translated into contiguous target strings is 63.87% among all extracted source terms. It indicates that 36.13% source terms are not translated into target strings as a whole unit. This is the reason why we propose a unithood model to predict whether a source term remains contiguous or not after translation.

Finally, we conducted a human evaluation on the quality of extracted bilingual terms. We randomly selected 1000 bilingual terms from our bilingual term bank. We showed the selected bilingual terms to a Chinese native speaker who is also fluent in English. We asked her to judge whether the source part of a given term pair is a term, whether the target part of a given term pair is a term and whether two parts are both terms and translations of each other. Based on her judgment, we calculated the percentage that items are judged as real terms among all given items. Table 17 shows the manual evaluation results. 64% of extracted bilingual terms are real bilingual terms according to the human evaluation. Table 18 display some examples of extracted terms, where both/one/neither sides of examples are manually judged as real terms.

In the majority of cases where only one side (source or target) is judged as a real term, boundary words (leftmost/rightmost) on the false term side is incorrectly aligned to the real term on the other side. If we remove these incorrectly aligned boundary words, both sides are real terms. For example, word "zhu4yi4" in "zhu4yi4 guo2ji4 an1quan2 ||| international security" in Table 18 is wrongly aligned to "international security". If it is removed, the remaining phrase will be a term. This is especially true for the source side, which explains why the term accuracy of the target side is much higher than that of the source side as shown in Table 17 (72.3% vs. 64.6%). We can enhance our monolingual term pairing procedure described in Section 3.2 by removing these incorrectly aligned boundary words. We believe that a higher percentage of real terms on both sides will further improve our models. We leave this to our future work.

Table 19
Percentages (%) of final translations where the corresponding models (i.e., the combined model or the three single models) are activated on the development test set MT06 and the final test set MT08. Improvements of BLEU scores (ΔBLEU) are also provided.

Models	MT06		MT08	
	Perc.	ΔBLEU	Perc.	ΔBLEU
Dis-Model	52.64	1.09	50.33	0.90
Cons-Model	49.10	1.34	45.54	0.50
Unit-Model	55.47	1.24	50.70	1.21
Combined-Model	62.32	1.80	56.48	1.24

Table 20
Consistency index values of term translation in MT06 and MT08.

Models	MT06	MT08
Baseline	1953.9	1358.3
Cons-Model	2131.9	1501.8

7.2. Analysis on generated target translations

In this section, we investigate to what extent the proposed models affect the translations on the test sets from a high level. We also provide translation examples to visualize how our models improve translation quality.

In Table 19, we show the percentages of final translations where the combined model and the three single models are activated on the development test set MT06 and the final test set MT08. We can see that final translations of sentences affected by any of the proposed models account for a high proportion (45 % ∼ 55 %) on both MT06 and MT08. This indicates that the three proposed models indeed significantly affect the translations of the two sets. For the combined model, if any of the three models are activated on a given final translation hypothesis, we consider the combined model is activated on this translation. Based on this counting strategy, we find that final translations affected by the combined model account for 62.32% and 56.48% on MT06 and MT08 respectively. Table 19 also presents the improvements of BLEU scores obtained by our models on sentences where our models are activated. We can achieve an improvement of up to 1.80 BLEU points over the baseline on these sentences.

In order to investigate how source terms are consistently translated in the two test sets, we calculate the term translation consistency index, similar to Itagaki et al. [26]. Table 20 report the consistency index values of the baseline and our term translation consistency model on the two sets MT06 and MT08. We can clearly observe that the consistency index values of Cons-Model are higher than those of the baseline system. This strongly suggest that terms are more consistently translated if we integrate the proposed term translation consistency model into the decoder.

Table 21 displays 4 translation examples which visualize how our models affect translation hypotheses. In the first example, the baseline system incorrectly translates the source phrase “zhu3liu2 min2yi4” into “mainstream opinion” while the Dis-model produces the correct translation “mainstream public opinion” with topic information. Table 22 shows the translation probabilities of different target translations for the source term “zhu3liu2 min2yi4” calculated by our term translation disambiguation model given the topic $z = 32$. Obviously our Dis-Model favors the target translation “mainstream public opinion” against the translation “mainstream opinion” selected by the baseline system (0.56 vs. 0.24).

In the second example, the source term “huan2jing4 zhi2fa3 ren2yuan2” has only one translation “environmental law enforcement personnel” in our training data. Therefore the consistency strength of the term is the highest 1. We encourage the decoder to translate this term with the translation from our bilingual term bank. In contrast, the nested term “zhi2fa3 ren2yuan2” has more than 25 different translations in our training data. It has a very low consistency strength 0.0236804 under the current topic. If we choose to translate this nested term, our translations for the whole source term “huan2jing4 zhi2fa3 ren2yuan2” will be not consistent across sentences.

In the third translation example, the source term “qu1ru3 li4shi3” is translated by the baseline system into two discontinuous strings “history” and “humiliation”, separated by the translation of “ge1rang4”. However, our term unithood model successfully translates this source term as a whole unit into “humiliating history”, which is the same as the reference translation.

The three translation examples discussed above show how our term translation disambiguation, consistency and unithood model improve translation quality. In the final example, our model recognizes two phrases as terms. The first phrase “jie2guo3 jie1xiao3” is actually not a term. Due to this noisy term in our bilingual term bank, our model wrongly translates this phrase into “results announced”. The other phrase “ban1jiang3 yi2shi4” is a correct term and our model correctly translate it. This suggests that our models, to some extent, are sensitive to noises of extracted terms.

Table 21
Translation examples showing the differences (in bold) between the baseline and the Dis-Model, Cons-Model, Unit-Model and Combined-Model.

Eg. 1	Source	shi2ji4shang4 ne0, ta1men1 zhe4ge4 liang3 da4 bao4zhang1 dui4 ri4ben3 de0 **zhu3liu3 min2yi4** ne0 ying1gai1 hai2shi4 you3 zhe1 fei1chang2 da4 de0 zhe4ge4 ying3xiang3li4 de0
	Base	actually, they are the two major newspapers for the Japanese **mainstream opinion** we should still have this huge influence
	Dis-Model	actually, the two major newspapers on the **mainstream public opinion** in Japan, is still the huge influence
	Reference	actually, these two large newspapers of theirs should have a very large influence on Japanese **mainstream public opinion**
Eg. 2	Source	hei1long2jiang1 jiang1 pai4chu1 **huan2jing4 zhi2fa3 ren2yuan2** du1cu4 zhi3dao3 ge4 di4 kai1zhan3 xiao3 liu2yu4 huan2jing4 zong1he2 zheng3zhi4
	Base	heilongjiang will send **environmental law enforcement** supervision, guiding the development of an integrated watershed environmental control
	Cons-Model	heilongjiang **environmental law enforcement personnel** will be fielded to supervise and guide the small river valleys comprehensive environmental management
	Reference	heilongjiang will dispatch **environmental law enforcement personnel** to supervise and guide the comprehensive environmental treatment of small watersheds in various places
Eg. 3	Source	shi4 dui4 ge1rang4 xiang1gang3dao3 **qu1ru3 li4shi3** de0 gao4bie2
	Base	and was bid farewell to the **history** of the ceded **humiliation**
	Unit-Model	was ceded to bid farewell to **humiliating history**
	Reference	bidding farewell to the **humiliating history** of the cession of Hong Kong Island
Eg. 4	Source	ping2xuan3 **jie2guo3 jie1xiao3** hou4, jiang1 yu2 2007 nian2 chu1 zai4 ao4men2 te4qu1 ju3xing2 **ban1jiang3 yi2shi4**.
	Base	selection **result is announced**, to be held in the Macao Special Administrative Region (SAR) at the beginning of 2007 after **awarding ceremony**
	Combined-Model	**results announced** after the selection, will **award ceremony** held in the Macao Special Administrative Region in early 2007
	Reference	following **announcement of the evaluation** results, an **award ceremony** will be held in the Macao SAR in early 2007

Table 22
Topic-conditioned translation probabilities of different target translations for the source term "zhu3liu2 min2yi4" calculated according to the equation (7).

Target t_e	$p(t_e \mid t_f, z = 32)$
Mainstream opinion	0.240309
Mainstream public	0.0662834
Mainstream public opinion	0.555217
Mainstream public opinions	0.0804339
Popular opinion	0.0577565

8. Related work

In this section, we introduce related work and highlight the differences between our work and previous studies. The exploration of statistical term translation in SMT is quite limited. To the best of our knowledge, our work is the first attempt to systematical investigation of term translation in the context of statistical machine translation.

Bilingual Terminology for Machine Translation: Itagaki and Aikawa [29] employ bilingual term bank as a dictionary for machine-aided translation. Ren et al. [21] propose a binary feature (0/1) to indicate whether a bilingual phrase contains a bilingual term pair. Arcan et al. [30] extract and integrate bilingual terminology into SMT in a CAT environment. Weller et al. [31] mine bilingual terminology from comparable corpora to enhance SMT in domain adaptation. These studies either focus on a specific single issue of term translation or use extracted bilingual terminology as an additional resource to enhance machine translation. They do not systematically investigate term translation in the context of SMT as we do. Furthermore, document-level information is not used to assist term translation in their work.

Finding Term Translations: A number of approaches have been proposed to find term translations or extract bilingual terminology from parallel/comparable corpora [32,33,23,34,35]. To name a few, Fung and Mckeown [32] propose a method to extract terminology translations from non-parallel corpora. Lefever et al. [23] introduce a language-independent method for bilingual terminology extraction from a word-aligned parallel corpus. We also extract bilingual terminology from our parallel training data via a strategy that pairs source and target term candidates based on word alignments. We will explore more different methods for bilingual terminology extraction in our future work.

Translation Consistency and Term Unithood: A variety of methods have been proposed to encourage translation consistency, ranging from cache-based models [36,37] and post-editing methods [38] to soft constraints as additional features of the log-liner model [39]. Our consistency model is most related to the method by Itagaki et al. [26] who propose a statistical method to evaluate translation consistency for terms. Partially inspired by them, we introduce a topic-based term translation consistency metric. The differences between our term translation consistency model and their consistency index are twofold. First, we introduce per-document topic distributions into our model to calculate topic-dependent term translation consis-

tency strengths for terms as shown in the equation (10). Second, we integrate the proposed term translation consistency model into an actual SMT system, which has not been done by Itagaki et al. [26].

Term unithood is first described by Kageura and Umino [2]. Lefever et al. [23] use the mutual expectation measure to estimate term unithood. Our term unithood model is different from theirs in that we use a classifier to predict the probability of term unithood after translation. Our model is related to Xiong et al.[27]'s syntax-driven bracketing model for phrase-based translation, which predicts whether a phrase is translated as a whole unit with rich syntactic constraints. The difference is that we construct the model with automatically created bilingual terms and do not depend on any syntactic knowledge.

Topic Modeling for SMT: As we approach term translation disambiguation and consistency via topic modeling, our models are related to previous work that explores topic models for machine translation [40–44]. Among them, our topic-based term translation disambiguation model is most related to the work of Xiao et al. [42], who propose a topic similarity and sensitivity model for translation rule selection in hierarchical phrase-based SMT. Although we also use topic information to help disambiguate term translation, our term translation disambiguation model is significantly different from Xiao et al. [42]'s topic similarity model in that they estimate the rule-topic distributions $p(z|r)$ while we estimate term translation probabilities $p(t_f|t_e, z)$ conditioned on topics. Furthermore, we focus on the three translation issues (disambiguation, consistency and unithood) of terms that are special phrases with linguistic and statistical properties.

Document-Level Machine Translation: Finally our work is also related to previous work [45–50] on document-level machine translation in that we use document-level information for term translation. The significant difference between our work and these studies is that term translation has not been investigated in these document-level machine translation models.

9. Conclusions and future work

We have studied the three issues of term translation in the context of SMT and proposed three different term translation models to address these issues. The term translation disambiguation model enables the decoder to favor the most suitable domain-specific translations with document-level information for source terms. The term translation consistency model encourages the decoder to translate source terms with a high topic-dependent translation consistency strength into consistent target terms. Finally, the term unithood model rewards hypotheses that translate 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 NIST Chinese–English translation with large-scale training data. Experiment results show that

- The term translation disambiguation model is able to obtain a substantial improvement of 0.58 BLEU points over the baseline on the test set.
- The term translation consistency model outperforms the baseline by 0.72 BLEU points. We also observe 1) that topic information improves the model as term translation consistency is topic-sensitive and 2) that modeling how terms are translated consistently in training data is better than counting the number of times that they are translated consistently in a test set.
- The term unithood model is also better than the baseline by 0.79 BLEU points on the test set. Additionally, we find that document-level topic information cannot improve the model.
- The combination of the three models can obtain further improvement, which is 0.92 BLEU points over the baseline on the final test set.

Our experiments also disclose 1) that the more bilingual terms we extract, the better translation quality will be, 2) and that the LLR method is marginally better than the C-value/NC-value method in the bilingual term extraction and the combination of the two methods achieves the best performance. Our in-depth analyses further validate that the proposed three term translation models are indeed able to improve term translation.

As shown in our analysis with translation examples, noises of extracted bilingual terms will guide our models to wrong translations. Therefore, we want to improve the procedure of bilingual term extraction so that we can further improve the performance of our method in the future. Additionally, we also plan to extend our models for the purpose of multilingual text analysis as well as multilingual terminology and ontology construction for specific domains as terms are able to convey concepts of a text or a domain.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61403269), Natural Science Foundation of Jiangsu Province (No. BK20140355) and National Social Science Foundation of China (No. 14CTQ032). Fandong Meng's work was supported by National Key Technology R&D Program of China (No. 2012BAH39B03) and CAS Action Plan for the

[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.

Development of Western China (No. KGZD-EW-501). Qun Liu's work was partially supported by Science Foundation Ireland (No. 07/CE/I1142) as part of the CNGL at Dublin City University.

Appendix A. Experiment setup

Our training data consist of 4.28M sentence pairs extracted from LDC[8] data with document boundaries explicitly provided. The bilingual training data contain 67,752 documents, 124.8M Chinese words and 140.3M English words. We used the ICTCLAS segmenter [51] for Chinese word segmentation. We chose NIST MT05 as the development set for MERT tuning, NIST MT06 as the development test set, and NIST MT08 as the final test set. The numbers of documents/sentences in the NIST MT05, MT06 and MT08 are 100/1082, 79/1664 and 109/1357 respectively. There are 4 different human-generated reference translations for each source sentence in these dev/test sets.

The word alignments were obtained by running GIZA++[9] [52] on the corpora in both directions and using the "grow-diag-final-and" balance strategy [24]. We adopted SRI Language Modeling Toolkit [53] to train a 4-gram language model with modified Kneser–Ney smoothing on the Xinhua portion of the English Gigaword corpus (306 million words).

We used the Stanford natural language processing toolkit[10] to perform part-of-speech tagging. The tagger detects nouns, adjectives and prepositions for the linguistic filter in the C-value/NC-value based monolingual term extraction (see Section 3). Empirically, we set the maximum length of a term to 6 words.[11] 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. Additionally, we set the C-value/NC-value score threshold to 0 and LLR score threshold to 10 based on our preliminary experiments on the training corpora.

For the topic model, we used the open source LDA topic modeling tool GibbsLDA++[12] with the default setting for training and inference. We performed 100 iterations of the L-BFGS algorithm implemented in the maximum entropy classifier toolkit[13] with both Gaussian prior and event cutoff set to 1 to train the term unithood prediction model (Section 4.3).

We used the case-insensitive 4-gram NIST BLEU[14] as our evaluation metric, which measures modified precisions of n-grams against multiple reference translations. As terms occur frequently in text (more than 65% sentences in our corpus contains terms according to our statistics), changes in term translations can be captured by BLEU. In order to alleviate the impact of the instability of MERT, 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. [54].

References

[1] M. Vasconcellos, B. Avey, C. Gdaniec, L. Gerber, M. León, T. Mitamura, Terminology and machine translation, in: Handbook of Terminology Management, vol. 2, 2001, pp. 697–723.
[2] K. Kageura, B. Umino, Methods of automatic term recognition: a review, Terminology 3 (2) (1996) 259–289.
[3] D. Chiang, A hierarchical phrase-based model for statistical machine translation, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005, pp. 263–270.
[4] F. Meng, D. Xiong, W. Jiang, Q. Liu, Modeling term translation for document-informed machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Doha, Qatar, Association for Computational Linguistics, 2014, pp. 546–556.
[5] F.J. Och, Minimum error rate training in statistical machine translation, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, 2003, pp. 160–167.
[6] D. Chiang, Y. Marton, P. Resnik, Online large-margin training of syntactic and structural translation features, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics, 2008, pp. 224–233.
[7] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318.
[8] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2009.
[9] D. Xiong, M. Zhang, Linguistically Motivated Statistical Machine Translation, Springer-Verlag, 2015.
[10] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[11] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99, ACM, New York, NY, USA, 1999, pp. 50–57.
[12] T.L. Griffiths, M. Steyvers, Finding scientific topics, in: Proceedings of the National Academy of Sciences, 2004, pp. 5228–5235.
[13] D. Mimno, H.M. Wallach, J. Naradowsky, D.A. Smith, A. McCallum, Polylingual topic models, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 2, EMNLP'09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 880–889.
[14] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84, http://dx.doi.org/10.1145/2133806.2133826.

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

[9] GIZA++ is an open source tool that runs the Expectation–Maximization (EM) algorithm to align words in sentence aligned bilingual corpora.

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

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

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

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

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

[15] K.T. Frantzi, S. Ananiadou, J. Tsujii, The C-value/NC-value method of automatic recognition for multi-word terms, in: Research and Advanced Technology for Digital Libraries, Springer, 1998, pp. 585–604.
[16] T. Vu, A.T. Aw, M. Zhang, Term extraction through unithood and termhood unification, in: Proceedings of the Third International Joint Conference on Natural Language Processing, 2008, pp. 631–636.
[17] B. Daille, Study and implementation of combined techniques for automatic extraction of terminology, in: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, vol. 1, 1996, pp. 49–66.
[18] S.S. Piao, G. Sun, P. Rayson, Q. Yuan, Automatic extraction of Chinese multiword expressions with a statistical tool, in: Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context held in conjunction with the 11th EACL, Trento, Italy, 2006, pp. 17–24.
[19] H. Hjelm, Identifying cross language term equivalents using statistical machine translation and distributional association measures, in: Proceedings of NODALIDA, Citeseer, 2007, pp. 97–104.
[20] X. Fan, N. Shimizu, H. Nakagawa, Automatic extraction of bilingual terms from a Chinese–Japanese parallel corpus, in: Proceedings of the 3rd International Universal Communication Symposium, ACM, 2009, pp. 41–45.
[21] Z. Ren, Y. Lü, J. Cao, Q. Liu, Y. Huang, Improving statistical machine translation using domain bilingual multiword expressions, in: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, 2009, pp. 47–54.
[22] M. Merkel, J. Foo, Terminology extraction and term ranking for standardizing term banks, in: Proceedings of 16th Nordic Conference of Computational Linguistics, Nodalida, University of Tartu, 2007, pp. 349–354.
[23] E. Lefever, L. Macken, V. Hoste, Language-independent bilingual terminology extraction from a multilingual parallel corpus, in: Proceedings of the 12th Conference of the European Chapter of the ACL, EACL 2009, Athens, Greece, Association for Computational Linguistics, 2009, pp. 496–504.
[24] P. Koehn, F.J. Och, D. Marcu, 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, vol. 1, 2003, pp. 48–54.
[25] D. Chiang, Hierarchical phrase-based translation, Comput. Linguist. 33 (2) (2007) 201–228.
[26] M. Itagaki, T. Aikawa, X. He, Automatic validation of terminology translation consistency with statistical method, in: Proceedings of MT Summit XI, 2007, pp. 269–274.
[27] D. Xiong, M. Zhang, A. Aw, H. Li, 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, 2009, pp. 315–323.
[28] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: open source toolkit for statistical machine translation, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, Association for Computational Linguistics, 2007, pp. 177–180.
[29] M. Itagaki, T. Aikawa, Post-MT term swapper: supplementing a statistical machine translation system with a user dictionary, in: Proceedings of the International Conference on Language Resources and Evaluation, LREC, 26 May–1 June 2008, Marrakech, Morocco, 2008, pp. 1584–1588.
[30] M. Arcan, M. Turchi, S. Tonelli, P. Buitelaar, Enhancing statistical machine translation with bilingual terminology in a CAT environment, in: Proceedings of the Eleventh Biennial Conference of the Association for Machine Translation in the Americas, AMTA 2014, Vancouver, BC, 2014, pp. 54–68.
[31] M. Weller, A. Fraser, U. Heid, Combining bilingual terminology mining and morphological modeling for domain adaptation in SMT, in: Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation, EAMT, Dubrovnik, Croatia, 2014, pp. 11–18.
[32] P. Fung, K. Mckeown, Finding terminology translations from non-parallel corpora, in: Proceedings of the Fifth Workshop on Very Large Corpora, 1997, pp. 192–202.
[33] H. Déjean, É. Gaussier, F. Sadat, An approach based on multilingual thesauri and model combination for bilingual lexicon extraction, in: Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002, 2002, pp. 1–7.
[34] M. Erdmann, K. Nakayama, T. Hara, S. Nishio, Improving the extraction of bilingual terminology from Wikipedia, ACM Trans. Multimed. Comput. Commun. Appl. 5 (4) (2009) 1–17, http://dx.doi.org/10.1145/1596990.1596995.
[35] L. Macken, E. Lefever, V. Hoste, TExSIS: bilingual terminology extraction from parallel corpora using chunk-based alignment, Terminology 19 (1) (2013) 1–30.
[36] J. Tiedemann, 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, Uppsala, Sweden, Association for Computational Linguistics, 2010, pp. 8–15.
[37] Z. Gong, M. Zhang, G. Zhou, Cache-based document-level statistical machine translation, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, Association for Computational Linguistics, 2011, pp. 909–919.
[38] T. Xiao, J. Zhu, S. Yao, H. Zhang, Document-level consistency verification in machine translation, in: Proceedings of the 2011 MT Summit XIII, Xiamen, China, 2011, pp. 131–138.
[39] F. Ture, D.W. Oard, P. Resnik, Encouraging consistent translation choices, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada, Association for Computational Linguistics, 2012, pp. 417–426.
[40] B. Zhao, E.P. Xing, Bitam: bilingual topic admixture models for word alignment, in: Proceedings of the COLING/ACL Main Conference Poster Sessions, 2006, pp. 969–976.
[41] J. Su, H. Wu, H. Wang, Y. Chen, X. Shi, H. Dong, Q. Liu, 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, vol. 1, 2012, pp. 459–468.
[42] X. Xiao, D. Xiong, M. Zhang, Q. Liu, S. Lin, A topic similarity model for hierarchical phrase-based translation, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, 2012, pp. 750–758.
[43] E. Hasler, P. Blunsom, P. Koehn, B. Haddow, 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, Association for Computational Linguistics, 2014, pp. 328–337.
[44] Y. Hu, K. Zhai, V. Eidelman, J. Boyd-Graber, Polylingual tree-based topic models for translation domain adaptation, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, Baltimore, Maryland, Association for Computational Linguistics, 2014, pp. 1166–1176.
[45] C. Hardmeier, J. Nivre, J. Tiedemann, 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, 2012, pp. 1179–1190.
[46] B. Wong, C. Kit, 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, 2012, pp. 1060–1068.
[47] D. Xiong, G. Ben, M. Zhang, Y. Lü, Q. Liu, Modeling lexical cohesion for document-level machine translation, in: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, 2013, pp. 2183–2189.
[48] D. Xiong, Y. Ding, M. Zhang, C.L. Tan, Lexical chain based cohesion models for document-level statistical machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1563–1573.
[49] D. Xiong, M. Zhang, 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 2013, pp. 977–983.

[50] D. Xiong, M. Zhang, X. Wang, Topic-based coherence modeling for statistical machine translation, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (3) (2015) 483–493.
[51] H.-P. Zhang, H.-K. Yu, D.-Y. Xiong, Q. Liu, HHMM-based Chinese lexical analyzer ICTCLAS, in: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, vol. 17, SIGHAN'03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 184–187.
[52] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Linguist. 29 (1) (2003) 19–51.
[53] A. Stolcke, et al., SRILM – an extensible language modeling toolkit, in: Proceedings of the International Conference on Spoken Language Processing, vol. 2, 2002, pp. 901–904.
[54] J.H. Clark, C. Dyer, A. Lavie, N.A. Smith, 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, vol. 2, 2011, pp. 176–181.