A Simple, Straightforward and Effective Model for Joint Bilingual Terms Detection and Word Alignment in SMT

Guoping Huang[1,2], Jiajun Zhang[1], Yu Zhou[1], and Chengqing Zong[1(✉)]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{guoping.huang,jjzhang,yzhou,cqzong}@nlpr.ia.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China

Abstract. Terms extensively exist in specific domains, and term translation plays a critical role in domain-specific statistical machine translation (SMT) tasks. However, it's a challenging task to extract term translation knowledge from parallel sentences because of the error propagation in the SMT training pipeline. In this paper, we propose a simple, straightforward and effective model to mitigate the error propagation and improve the quality of term translation. The proposed model goes from initial weak monolingual detection of terms based on naturally annotated resources (e.g. Wikipedia) to a stronger bilingual joint detection of terms, and allows the word alignment to interact. The extensive experiments show that our method substantially boosts the performance of bilingual term detection by more than 8 points absolute F-score. And the term translation quality is substantially improved by more than 3.66% accuracy, as well as the sentence translation quality is significantly improved by 0.38 absolute BLEU points, compared with the strong baseline, i.e. the well tuned Moses.

1 Introduction

Terms, defined by specialists, a noun or compound word used in a specific context, deliver essential context and meaning in human languages [25], such as technical terms "header text" and "summary"[1]. Terms extensively exist in specific domains. For example, in Microsoft Translation Memory, there are 8 terms out of every 100 words, whereas named entities are nearly nonexistent. What's more, new terms are being created all the time, such as in areas of computer science and medicine. Thus, term translation plays a critical role in domain-specific statistical machine translation (SMT) tasks.

However, unlike person names or other named entities having obvious characteristics and boundary clues, it's a challenging task to extract term translation knowledge from parallel sentences in the SMT training pipeline. A typical SMT training pipeline consists of monolingual term recognition, word alignment and

[1] In this paper, we do not consider named entities (e.g., person names, location names, organization names, time and numbers) and treat named entities non-terms.

C.-Y. Lin et al. (Eds.): NLPCC-ICCPOL 2016, LNAI 10102, pp. 103–115, 2016.
DOI: 10.1007/978-3-319-50496-4_9

translation rule extraction. So, the term recognization errors will propagate into the next stages. To make matters worse, it is expensive to annotate training data, in practice, to obtain high-quality term recognizers for various specific domains.

As a result, the poor performance of term recognition further decreases the quality of word alignment and translation rule extraction. Thus, it is a challenging task to extract term translation knowledge from parallel sentences. Thus, frequent term translation errors make users hard to follow MT results in specific areas. For example, in the case of Microsoft Translation Memory, more than 10% of high-frequency terms are incorrectly translated by our baseline system, although the BLEU-score is up to 63%.

In order to mitigate the error propagation and improve the quality of term translation, we propose in this paper a simple, straightforward and effective model for jointing bilingual term detection and word alignment. The proposed model goes from the initial weak monolingual detection of terms based on naturally annotated resources, e.g., Wikipedia, to a stronger bilingual joint detection of terms, and allows the word alignment to interact. A brief overview of the proposed model is shown in Fig. 1.

In Fig. 1(a), the starting point is the weak English term recognizer, the weak Chinese term recognizer and the HMM-based word alignment model. Obviously, there are some critical errors denoted by red color (the italics words and the dotted lines).

Fortunately, based on Fig. 1(a), we have the following observations: (1) The initially recognized monolingual terms can act as anchors for further detecting terms. (2) The source terms and target terms in parallel sentences come in pair, and it provides mutual constraints for bilingual term detection. (3) The detected bilingual term pairs can further improve the performance of word alignment, in turn, word alignment can contribute to term recognition.

Based on the above observations and inspired by [2,27], the proposed model adopts the initial results as anchors, then enlarges or shrinks the boundaries of the anchors to generate new term candidates, and allows the word alignment to interact, as shown in Fig. 1(b). Finally, we get a stronger bilingual joint detection of terms and the promoted word alignment as seen in Fig. 1(c).

In the experiments, our proposed joint model has achieved remarkable results on bilingual term detection, word alignment, term translation and sentence translation. In summary, this paper makes the following contributions:

1. The proposed simple and straightforward model jointly performs bilingual term detection and word alignment for the first time.
2. The proposed joint model starts with low-quality naturally annotated monolingual resources rather than expensive human annotated data to perform initial term recognition, and allows the word alignment to interact with bilingual term detection, finally gets a stronger bilingual detection of terms.
3. The proposed model substantially boosts the performance of bilingual term detection and word alignment, and finally significantly improves the performance of term translation in the specific domain compared to a strong baseline.

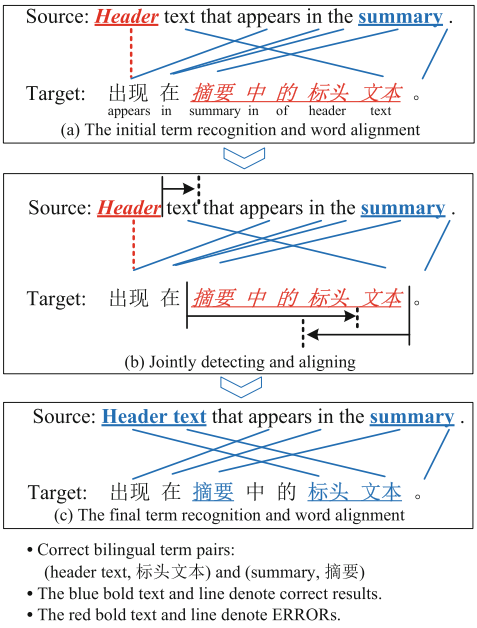

Fig. 1. A brief work flow overview of the proposed model. (Color figure online)

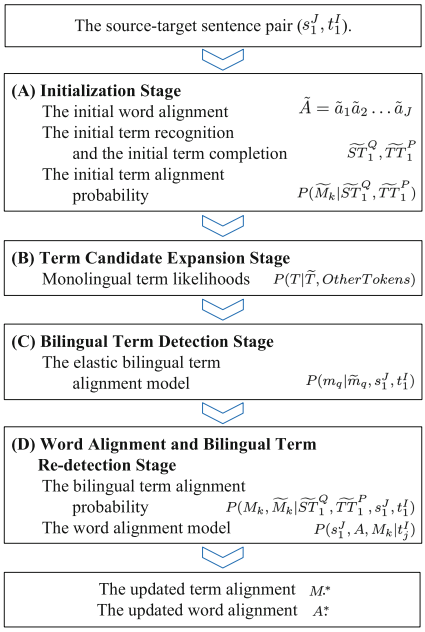

Fig. 2. The four-stage framework for joint bilingual term detection and word alignment.

2 Related Work

To automatically recognize terms, researchers have proposed many approaches, which can be divided into two types. One aims at using linguistic tools (e.g. POS tagger, phrase chunker) to filter out stop words and restrict candidate terms to noun phrases [1]. The other focuses on employing statistical measures to rank the candidate terms (n-gram sequences), such as mutual information [4], log likelihood [17], t-test [6], TF-IDF [20], C-value/NC-value [9], and many others [14,30]. More recent term recognition systems use hybrid approaches that combine both linguistic and statistical information.

However, seldom is the full range of the problem dealt with by any one method. First, most works rely on the simplifying assumption [11,15] that the majority of terms consist of multi-word, In fact, [21] claims that 85% of domain-specific terms are multi-word units, while [15] claims that only a small percentage of gene names are multi-word units. Such an assumption leads to very low recall for some domains. Second, some approaches apply frequency thresholds to reduce the algorithm's search space by filtering out low frequency term candidates. Such methods have not taken into account Zipf's law, again leading to reduced recall.

In this paper, in order to improve the recall, we adopt naturally annotated resources for term detection, such as Wikipedia, and focus on supervised machine learning approaches based recognition approaches for SMT with a wide range of domains.

Most bilingual term alignment systems first identify term candidates in the source and target languages based on predefined patterns [16], statistical measures (e.g., frequency information) [17], or supervised approaches [7], and then select translation candidates for these terms. In such pipeline approaches, the error propagation has a negative impact on the bilingual term detection and term translation.

3 The Proposed Joint Model

In this section, we first introduce the whole framework, then propose a formalized representation, and finally describe the important details.

3.1 The Framework for Jointly Detecting Bilingual Term Pairs and Aligning Words

In this paper, in order to jointly detect bilingual terms and align words, we propose a four-stage framework as shown in Fig. 2: (A) Initialization stage goes from initial weak monolingual detection of terms based on naturally annotated resources. (B) Term candidate expansion stage, expanding the associated term candidate set to remedy the errors occurred in the previous stage. (C) Bilingual term detection stage. The framework obtains a stronger bilingual joint detection of terms. (D) Word alignment and bilingual term re-detection stage. The framework allows the word alignment to interact with the bilingual term detection results. In Fig. 2, only the key points are showed.

(A) Initialization Stage

The first stage includes the following steps: initial word alignment, initial term recognition, initial term completion and initial term alignment. Let $s_1^J = s_1 s_2 \dots s_J$ denote the source sentence, and $t_1^I = t_1 t_2 \dots t_I$ denote the target sentence, where J and I are the numbers of words in source sentence and target sentence, respectively.

Initial Word Alignment and Initial Term Recognition: Given the source-target sentence pair (s_1^J, t_j^I), we can get the initial word alignment $\widetilde{A} = \tilde{a}_1 \tilde{a}_2 \dots \tilde{a}_J$, the initial recognized source terms $\widetilde{ST}_1^Q$, and the initial recognized target terms $\widetilde{TT}_1^P$, where Q and P are the numbers of initially recognized terms of the source and the target sentence, respectively. In word alignment, $\tilde{a}_j = \{i|a(j) = i\}$, and the expression $a(j) = i$ denotes that the target word t_i is connected to the source word s_j.

For this work, the word alignment refers to the HMM-based word alignment model by default. The term recognition tool is based on the Stanford

Classifier [19], which is trained by naturally annotated Wikipedia monolingual sentences, e.g., hyperlinks, boldfaces and quotes. And a beam search style decoding algorithm is employed to convert the classification results to appropriate term recognition results. As a result, we can get initial weak monolingual term detectors.

Initial Term Completion: In order to prevent the incorrect term alignment caused by the initial term recognition errors, $\widetilde{ST}_1^Q$ and $\widetilde{TT}_1^P$ will be fixed by the following operation: if none of aligned target words of the source term $\widetilde{ST}_q$ is recognized as the term, then the one, which is most likely to be a term, of them will be added into $\widetilde{TT}_1^P$; the same operation will be applied to the target terms.

Initial Term Alignment: We construct the initial term alignment set $\widetilde{M} = \widetilde{M}_1^{(P^Q)}$ by generating a Cartesian product of the source term set $\widetilde{ST}_1^Q$ and the target term set $\widetilde{TT}_1^P$. We rank each candidate $\widetilde{M}_k$ of the initial term alignment set in descending order with the score calculated by the Viterbi algorithm [8] using the pre-trained term alignment model. The k-th initial term alignment is denoted by $\widetilde{M}_k = \widetilde{m}_1\widetilde{m}_2 \ldots \widetilde{m}_Q$, where $\widetilde{m}_q = (\widetilde{ST}_q, \widetilde{TT}_p)$.

In the first stage, the initial term alignment is based on the pre-trained term alignment model, which is implemented according to the HMM-based word alignment model. And the training data is the bilingual term dictionary consisting of Wikipedia titles and the domain-specific term database.

Example: For the example in Fig. 1, the input of the first stage is the following:

(1) The English (source) tagged sentence "<Header> text that appears in the <summary> ." and the Chinese (target) tagged sentence "出现 在 <摘要 中 的 标头 文本> 。".
(2) The initial English terms ([header], [summary]) and the initial Chinese terms ([摘要 中 的 标头 文本]).
(3) The initial word alignment "NULL{3} 出现{1,4} 在{5,6} 摘要{7} 中{} 的{} 标头{} 文本{2} 。{8}".

And the output is the following result:

(1) The fixed initial English terms ([header], [summary]) and the fixed initial Chinese term ([出现], [摘要中的标头文本]).
(2) The initial term-alignment set ({[header]::[出现], [summary]::[摘要 中 的 标头 文本]}; {[header]::[摘要 中 的 标头 文本], [summary]::[出现]}; {[header]::[摘要 中 的 标头 文本], [summary]::[摘要 中 的 标头 文本]}; {[header]::[出现], [summary]::[出现]}).

(B) Term Candidate Expansion Stage

In order to mitigate the error occurred in the previous stage, we generate another two term candidate sets $ST_1^{Q'}$ and $TT_1^{P'}$ sets by allowing the initial term to enlarge/shrink its boundaries up to four words on each side. Each time, when the one of the boundaries is enlarging/shrinking, the another one should be fixed. And finally we get a series of term candidates. The limitation "four words" is an empirical value. In addition, the regenerated terms in this stage are not allowed to overlap different initial terms, but they can share the same base initial term.

Example: For the example in Fig. 1, the input of the second stage is the initial term-alignment set, and the output is the following result:

(1) The regenerated English term set ([header] → {[header text], [header text that], [header text that appears], [header text that appears in]}; [summary] → {[summary], [the summary], [in the summary], [appears in the summary], [that appears in the summary]}).
(2) The regenerated Chinese term set([出现] → {[出现 在]}, [摘要 中 的 标头 文本] → {[在 摘要 中 的 标头 文本], [摘要 中 的 标头 文本 。], [摘要 中 的 标头], [摘要 中 的], [摘要], [中 的 标头 文本], [的 标头 文本], [标头 文本], [文本]}).

(C) Bilingual Term Detection Stage

The third stage is to jointly perform monolingual term detection and bilingual term alignment. We conduct a beam search process to select the top K updated term alignment set $M = M_1^K$ based on the initial term alignment set $\widetilde{M}$, the re-generated source terms $ST_1^{Q'}$ and the re-generated target terms $TT_1^{P'}$. The searching process will keep removing those overlapping terms from the candidate list. The k-th updated term alignment is denoted as $M_k = m_1 m_2 \dots m_Q$ where $m_q = (TT_p, ST_q)$. We can get the probability of each updated term alignment $P(M_k|ST_1^{Q'}, TT_1^{P'})$ for each k. As a result, the proposed framework obtains a stronger bilingual term detection.

Example: For the example in Fig. 1, the input of the third stage includes the regenerated English term set and the regenerated Chinese term set, and the output is the following result:

The updated-term-alignment set ({[header text]::[标头文本], [summary]::[摘要中]}; {[header text]::[的标头文本], [summary]::[摘要]}; ⋯Total 132 (11 × 12) term pairs will be generated).

(D) Word Alignment and Bilingual Term Re-detection Stage

In the last stage, the framework allows the word alignment to interact with the bilingual term detection results through jointly executing bilingual term re-detection and word alignment via a generative model. The joint word alignment tool in this stage is the extension for the initial word alignment tool in the first stage. As a result, we can get the final word alignment $A^* = a_1^* a_2^* \dots a_J^*$ and the final term alignment $M^* = m_1^* m_2^* \dots m_Q^*$ using the generative word alignment model based on the constraint of the updated term alignment M.

Example: For the example in Fig. 1, the input of the last stage is the updated-term-alignment set, and the output is the following result:

(1) The re-ranked updated-term-pair set({[header text]::[标头 文本], [summary]::[摘要]}; {[header text]::[的 标头 文本], [summary]::[摘要]}; ⋯).
(2) The top 1 word alignment "NULL{6} 出现{4} 在{5} 摘要{7} 中{3} 的{} 标头{1} 文本{2} 。{8}".
(3) The top 1 term alignment in updated-term-pair set({[header text]::[标头 文本], [summary]::[摘要]}).

3.2 The Joint Model

We put all the four stages together, and the proposed joint model can be formulated as:

$$(A^*, M^*) = \underset{(M_k, A)}{\operatorname{argmax}} \left[\max_{\widetilde{M}_k} P(M_k, \widetilde{M}_k | \widetilde{ST}_1^Q, \widetilde{TT}_1^P, s_1^J, t_1^I) \times P(s_1^J, A, M_k | t_1^I) \right] \quad (1)$$

where $P(M_k, \widetilde{M}_k|\widetilde{ST}_1^Q, \widetilde{TT}_1^P, s_1^J, t_1^I)$ refers to the bilingual term alignment probability, and $P(s_1^J, A, M_k|t_j^I)$ refers to the the word alignment model based on the constraint of the updated term alignment M_k.

The following steps are executed jointly with respect to $\widetilde{ST}_1^Q$, $\widetilde{TT}_1^P$, s_1^J and t_1^I: monolingual term recognition, bilingual term alignment and word alignment. And there is no independence assumption among those term pairs including in the associated term-pair sequence.

Next, we will introduce the important derivation details. The derivation looks like a somewhat complicated framework, but it's not so hard to comprehend and implemented.

3.3 Derivation Details

In Eq. (1), the bilingual term alignment probability, in the fourth stage as shown in Fig. 2, is computationally infeasible and will be simplified and derived as follows:

$$P(M_k, \widetilde{M}_k|\widetilde{ST}_1^Q, \widetilde{TT}_1^P, s_1^J, t_1^I) \approx P(\widetilde{M}_k|\widetilde{ST}_1^Q, \widetilde{TT}_1^P) \times \prod_{m_q \in M_k} \prod_{\widetilde{m}_q \in \widetilde{M}_k} P(m_q|\widetilde{m}_q, s_1^J, t_1^I) \tag{2}$$

It implies that monolingual term recognition and bilingual term alignment are executed jointly. In Eq. 2, $P(\widetilde{M}_k|\widetilde{ST}_1^Q, \widetilde{TT}_1^P)$ denotes the initial term alignment probability in the first stage, and $P(m_q|\widetilde{m}_q, s_1^J, t_1^I)$ denotes the elastic bilingual term alignment model in the third stage.

In the next subsections, we will introduce how to compute the important submodels embedded in the four stages as shown in Fig. 2.

(1) The Initial Term Alignment Probability

The initial term alignment probability, in the first stage, is based on the maximum entropy model [3]. In this paper, we design a set of feature functions $h_f(\widetilde{M}_k, \widetilde{ST}_1^Q, \widetilde{TT}_1^P)$, where $f = 1, 2, \ldots, F$. Let λ_f be the weight corresponding to the feature function. We adopt GIS algorithm [5] to train the weight λ_f. According to [22], we have the following initial term alignment model:

$$P(\widetilde{M}_k|\widetilde{ST}_1^Q, \widetilde{TT}_1^P) = \frac{\exp\left[\sum_{f=1}^F \lambda_f h_f(\widetilde{M}_k, \widetilde{ST}_1^Q, \widetilde{TT}_1^P)\right]}{\sum_{\widetilde{M}_k'} \exp\left[\sum_{f=1}^F \lambda_f h_f(\widetilde{M}_k', \widetilde{ST}_q, \widetilde{TT}_p)\right]} \tag{3}$$

In order to calculate the initial term alignment model, we employ the following three feature functions in this paper: phrase translation probability (denoted as h_1), lexical translation probability (h_2) and co-occurrence feature (h_3).

The phrase translation probability h_1 is calculated by the pre-trained term word alignment model as follows:

$$h_1(\widetilde{M}_k, \widetilde{ST}_1^Q, \widetilde{TT}_1^P) = \log P(\widetilde{ST}_1^Q|\widetilde{TT}_1^P, \widetilde{M}_k) + \log P(\widetilde{TT}_1^P|\widetilde{ST}_1^Q, \widetilde{M}_k) \tag{4}$$

The lexical translation probability h_2 is calculated by the pre-trained term word alignment:

$$h_2(\widetilde{M}_k, \widetilde{ST}_1^Q, \widetilde{TT}_1^P) = \log lex(\widetilde{ST}_q^Q | \widetilde{TT}_1^P, \widetilde{M}_k) + \log lex(\widetilde{TT}_1^P | \widetilde{ST}_1^Q, \widetilde{M}_k) \quad (5)$$

The co-occurrence feature h_3 is calculated based the current parallel corpus:

$$h_3(\widetilde{M}_k, \widetilde{ST}_1^Q, \widetilde{TT}_1^P) = \log \prod_{q=1}^{Q} \left(\frac{count(\widetilde{ST}_q, \widetilde{TT}_{\widetilde{m}(q)})}{count(*, \widetilde{TT}_{\widetilde{m}(q)})} + \frac{count(\widetilde{TT}_{\widetilde{m}(q)}, \widetilde{ST}_q)}{count(*, \widetilde{ST}_q)} \right) \quad (6)$$

(2) The Monolingual Term Likelihoods

This is the key step of the third stage as well as the whole joint model. Given the initial term $\widetilde{T} = \widetilde{T}_1^{\widetilde{H}} = \widetilde{w}_1 \widetilde{w}_2 \dots \widetilde{w}_{\widetilde{H}}$, where $\widetilde{w}_i$ refers to the i-th word, and $\widetilde{H}$ is the number of words. Then, the re-generated term T can be formulated as $T = T_1^H = w_1 w_2 \dots w_H = \widetilde{w}_{-d_L} \dots \widetilde{w}_{-1} \widetilde{w}_1 \widetilde{w}_2 \dots \widetilde{w}_{\widetilde{H}} \widetilde{w}_{+1} \dots \widetilde{w}_{+d_R}$, where d_L refers to the left distance, namely numbers of words enlarged ($d_L \geq 1$) or shrunk ($d_L \leq -1$) from the left boundary; similarly, d_R refers to the right distance. In fact, $\tilde{t}_1$ and $\tilde{t}_{\widetilde{H}}$ are the anchor points that we can enlarge or shrink the initial recognized term. Then, the monolingual term likelihoods can be derived as:

$$\begin{aligned} P(T|\widetilde{T}, OtherTokens) \approx P(T)^{\beta_1} \times (1 - P(\widetilde{w}_{-d_L} \dots \widetilde{w}_{-1}))^{\beta_2} \times \\ (1 - P(\widetilde{w}_{+1} \dots \widetilde{w}_{+d_R}))^{\beta_3} \times P(\widetilde{T})^{\beta_4} \end{aligned} \quad (7)$$

where $P(*)$ refers to the probability that $*$ is a term given by the initial monolingual term recognition model; $1 - P(*)$ refers to the probability that the enlarged/shrunk part $*$ is not a term; β refers to the corresponding weight (the optional value is 0.25).

(3) The Elastic Bilingual Term Alignment Model

The elastic bilingual term alignment model, in the third stage, can be further decomposed:

$$P(m_q|\widetilde{m}_q, s_1^J, t_1^I) = \sum_{L_k} P(L_k|ST_q, TT_p) \times P^{'}(m_q|\widetilde{m}_q, s_1^J, t_1^I) \quad (8)$$

where L_k denotes internal component alignment, $P^{'}(m_q|\widetilde{m}_q, s_1^J, t_1^I)$ denotes the elastic bilingual term model, and the word alignment probability $P(L_k|ST_q, TT_p)$ is determined by the pre-trained term alignment model. The elastic bilingual term model can be derived based on the monolingual term likelihoods as follows:

$$P^{'}(m_q|\widetilde{m}_q, s_1^J, t_1^I) \approx P(ST_q|\widetilde{ST}_q, OtherTokens) \times P(TT_p|\widetilde{TT}_p, OtherTokens) \quad (9)$$

(4) The Word Alignment Model

The word aligned model, in the last stage, is calculated according to the HMM word alignment model [26]:

$$P(s_1^J, A, M_k|t_j^I) = \prod_{j=1}^{J} p(a_j, M_k|a_{j-1}, I) \times P(s_j|t_{a_j}) \quad (10)$$

where $P(s_j|t_{a_j})$ denotes the word translation probability.

Let $p(a_j|a_{(j-1)}, I)$ be the HMM alignment probability according to [26], and $conflict(j, M_k)$ be the indicator which indicates whether the current word alignment a_j has a conflict with the term alignment M_k, then:

$$p(a_j, M_k|a_{(j-1)}, I) = \begin{cases} 0 & if\ conflict(j, M_k) = true \\ p(a_j|a_{(j-1)}, I) & if\ conflict(j, M_k) = false \end{cases} \quad (11)$$

At last, about the computational cost of our implementation, the time tends to increase 3–4 times more than the baseline HMM-based word alignment, and the memory requirement rises at nearly 2–3 times.

4 Experiments

We conduct the experiments to test the performance of our four-stage joint model in improving the performance of bilingual term detection and word alignment. In addition, we will check how much improvement the proposed model can achieve on the final SMT result. The performance of recognition and alignment is evaluated by precision (P), recall (R) and F-score (F); the quality of term translation and sentence translation is evaluated by precision (P) and BLEU, respectively.

Table 1. The performance of term recognition.

	P/%	R/%	F/%
En-Baseline	62.94	65.61	64.25
Ch-Baseline	57.21	66.67	61.58
En-Joint-C-Stage	67.35	71.47	69.34
Ch-Joint-C-Stage	65.13	74.86	69.65
En-Joint-D-Stage	71.20**	76.84**	**73.91****
Ch-Joint-D-Stage	67.89**	75.03**	**71.28****

Table 2. The performance of bilingual term alignment

	P/%	R/%	F/%
Baseline	49.38	56.41	52.66
Joint-C-Stage	53.47	59.44	56.29
Joint-D-Stage	58.29**	63.78**	**60.91****

Table 3. The performance of word alignment

	P/%	R/%	F/%
GIZA++	69.28	75.83	72.41
Baseline-1	67.06	73.18	69.99
Baseline-2	64.47	70.62	67.41
Joint-C-Stage	69.45	76.49	72.80
Joint-D-Stage	71.19**	78.51**	**74.67****

Table 4. The performance of translation

	Term/P/%	Sent/BLEU/%
Moses	87.30	63.58
Baseline-1	86.53	63.09
Baseline-2	78.43	62.68
Joint-C-Stage	87.73	63.54
Joint-D-Stage	**91.04****	**63.96****

"**" means the scores are significantly better than the corresponding previous line with $p < 0.01$.

4.1 Experimental Setup

All the experiments are conducted on our in-house developed SMT toolkit including a typical phrase-based decoder [28] and a series of tools, including term recognition, term alignment, word alignment and phrase table extraction.

We test our method on English-to-Chinese translation in the field of software localization. The training data (1,199,589 sentences) and annotated test data (1,100 sentences) are taken from Microsoft Translation Memory, which is a domain-specific dataset. And additional data employed by this paper includes Wikipedia terms (1,133,913) and Microsoft Terminology Collection (24,094 terms). The gold standard of term recognition and word alignment are human annotated. What's more, all data have been submitted for public. The statistical significance test is performed by the re-sampling approach [12].

4.2 Results and Analysis

(1) The Term Recognition Tests

First, we compare the performances of term recognition in the different joint stages with the baseline system, e.g., the pipeline approach. The corresponding systems are denoted as "En-baseline", "Ch-Baseline", "En-Joint-C-Stage", "Ch-Joint-C-Stage", "En-Joint-D-Stage" and "Ch-Joint-D-Stage", respectively. "*-Baseline" refers to that term recognition and bilingual term alignment are executed individually. "*-C-Stage" means that only term recognition and term alignment are executed jointly. "*-D-Stage" refers the proposed four-stage framework. We report all the term recognition results in Table 1.

In contrast to the pipeline approach, the figures in Table 1 show that the initially detected terms can act as quite useful anchors for further detection, and the performance of monolingual term recognition has been increased by at least 9.66 points absolute F-score through the proposed four-stage framework. According to the bold figures in Table 1, we can draw a conclusion that word alignment can substantially increase the performance of monolingual term recognition.

(2) The Bilingual Term Alignment Tests

Second, we compare the performances of bilingual term alignment in different stages. We report all the bilingual term alignment results in Table 2. The bold figures in Table 2 indicate that the performance of bilingual term alignment has been increased by 8.25 points absolute F-score, with the feedback of word alignment and the constraint of source terms and target terms being pairing off.

(3) The Word alignment Tests

Third, we evaluate the performance of proposed joint model on word alignment. Both GIZA++ [23] and the HMM-based approach "Baseline-1" take no account of terms. Then, the term pipeline approach is implemented as our "Baseline-2". The term pipeline approach means that the following steps will be accomplished sequentially without feedback: term recognition, bilingual term alignment and

word alignment. "Joint-C-Stage" means that word alignment is executed individually in the fourth stage. And "*-D-Stage" refers the proposed four-stage framework. In this paper, we adopted the balanced F-measure [10,18] as our evaluation metric for word alignment. All results are reported in Table 3.

In Table 3, "Baseline-1" is the pure HMM-based word alignment, while GIZA++ enables IBM model 1–5, HMM and other alignment improvements. Thus, the word alignment result of "Baseline-1" is worse than that of GIZA++. And the pipeline approach ("Baseline-2") cannot improve the performance of word alignment, because the performance of monolingual term recognition is too weak for the scarcity of specialized annotated data. The bold figures in Table 3 show that our proposed joint model has increased the performance of word alignment by 4.68 and 2.26 points absolute F-score, compared to the HMM-based method and GIZA++, respectively.

(4) The SMT Translation Tests

Finally, we test whether the proposed joint model can further improve the performance of term and sentence translation. The Moses (GIZA++) and the HMM-based approach "Baseline-1" take no account of terms. Then, the term pipeline approach is implemented as our "Baseline-2". The word alignment was conducted bidirectionally and then symmetrized for extracting phrases as Moses [13] does. All the MT systems are trained by the same training set and tuned by the development set (1,100 sentences) using ZMERT [29] with the objective to optimize BLEU [24]. The test set includes 1,100 sentences with 1,208 bilingual term pairs altogether. In order to highlight the performance of term translation, we count the number of terms that is translated exactly correctly, and the term translation results are denoted as "Term/P" (exact match). The sentence translation results are labeled "Sent/BLEU". We report all the translation results in Table 4.

In Table 4, GIZA++ makes the SMT result of "Baseline-1" are worse than Moses. However, with the help of the proposed joint model, the term translation quality is significantly improved by more than 3.66% accuracy. Non-term words are also strongly improved by the joint model, because the accuracy rating of term words alignment has been much improved and fewer non-term words are aligned incorrectly to term words. In sentence translation, the bold figures in Table 4 demonstrate that it improves the translation quality by 0.38 absolute BLEU points, compared with the strong baseline system, i.e., well tuned Moses. Considering one term on average in a single sentence in the test set, the BLEU scores are very promising actually, and our goals on term translation have been achieved.

For the example in Fig. 1, with the aid of the joint model, the SMT system acquired more reliable term translation knowledge from training sentences, such as "header text ||| 标头文本". For the source sentences "header text is not included", the result of the baseline systems is "不包含头部文本, head text is not included". Fortunately, we can achieve the correct term translation result "不包含标头文本" from the system "Joint-D-Stage".

In summary, we can draw the conclusion that the proposed four-stage joint model significantly improves the performance of monolingual term recognition, bilingual term alignment and word alignment, and further significantly improves the performance of SMT in term translation and sentence translation.

5 Conclusion

In this paper, we have presented a simple, straightforward and effective joint model for bilingual term detection and word alignment. The proposed model starts with weak monolingual term detection based on naturally annotated monolingual resources, then jointly performs bilingual term detection and word alignment, finally substantially boosts bilingual term detection and word alignment, and significantly improves the quality of term translation and sentence translation. The experimental results are promising.

Acknowledgments. The research work has been funded by the Natural Science Foundation of China under Grant No. 61403379.

References

1. Ananiadou, S.: A methodology for automatic term recognition. In: Proceedings of COLING 1994 (1994)
2. Chen, Y., Zong, C., Su, K.Y.: A joint model to identify and align bilingual named entities. Comput. Linguist. **39**(2), 1–64 (2012)
3. Chieu, H.L., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: Proceedings of the 19th International Conference on Computational Linguistics (2002)
4. Daille, B.: Study and implementation of combined techniques for automatic extraction of terminology. Balanc. Act: Comb. Symb. Stat. Approaches Lang. **1**, 49–66 (2002)
5. Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Stat. **43**(5), 1470–1480 (1972)
6. Fahmi, B.I., Bouma, G., Plas, L.V.D.: Improving statistical method using known terms for automatic term extraction. In: Computational Linguistics in the Netherlands-CLIN 2007 (2007)
7. Fan, X., Shimizu, N., Nakagawa, H.: Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In: International Universal Communication Symposium 2009 (2009)
8. Forney, G.D.: The viterbi algorithm. Proc. IEEE **61**(3), 268–278 (1973)
9. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. Int. J. Digit. Libr. **3**(2), 115–130 (2000)
10. Fraser, A., Marcu, D.: Measuring word alignment quality for statistical machine translation. Fraser Alexander Daniel Marcu **33**(3), 293–303 (2007)
11. Kageura, K., Umino, B.: Methods of automatic term recognition: a review. Terminology **3**(2), 259–289 (1996)
12. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of the EMNLP 2004 (2004)

13. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R.: Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL 2007 (2007)
14. Kostoff, R.N., Block, J.A., Solka, J.L., Briggs, M.B., Rushenberg, R.L., Stump, J.A., Johnson, D., Lyons, T.J., Wyatt, J.R.: Literature-related discovery. Ann. Rev. Inf. Sci. Technol. **43**(1), 171 (2009)
15. Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. J. Biomed. Inform. **37**(6), 512–526 (2004)
16. Kupiec, J.: An algorithm for finding noun phrase correspondences in bilingual corpora. In: Proceedings of ACL 1993 (1993)
17. Lefever, E., Macken, L., Hoste, V.: Language-independent bilingual terminology extraction from a multilingual parallel corpus. In: Proceedings of EACL 2009 (2009)
18. Liu, Y., Liu, Q., Lin, S.: Discriminative word alignment by linear modeling. Comput. Linguist. **36**(3), 303–339 (2010)
19. Manning, C., Dan, K.: Optimization, maxent models, and conditional estimation without magic. In: Proceedings of the NAACL 2003 (2003)
20. Medelyan, O., Witten, I.H.: Thesaurus based automatic keyphrase indexing. In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (2006)
21. Nakagawa, H., Mori, T.: Nested collocation and noun for term extraction. In: Proceedings of the First Workshop on Comutational Terminology (COMPUTERM 1998) (1998)
22. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of ACL 2002 (2002)
23. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**(1), 19–51 (2003)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL 2002 (2002)
25. Sager, J.C., Dungworth, D., McDonald, P.F.: English Special Languages: Principles and Practice in Science and Technology. John Benjamins Publishing Company, Amsterdam (1980)
26. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: Proceedings of the 16th Conference on Computational Linguistics, vol. 2, pp. 836–841 (1996)
27. Wang, M., Che, W., Manning, C.D.: Joint word alignment and bilingual named entity recognition using dual decomposition. In: Proceedings of ACL 2013 (2013)
28. Xiong, D., Liu, Q., Lin, S.: Maximum entropy based phrase reordering model for statistical machine translation. In: proceedings of COLING-ACL 2006 (2006)
29. Zaidan, O.F.: Z-MERT: a fully configurable open source tool for minimum error rate training of machine translation systems. Prague Bull. Math. Linguist. **91**, 79–88 (2009)
30. Zhang, Z., Iria, J., Brewster, C.: A comparative evaluation of term recognition algorithms. In: LREC 2008 (2008)