Multiple Linear Regression for Extracting Phrase Translation Pairs

Chun-Xiang Zhang
School of Software, Harbin University of Science and Technology, Harbin, China
Email: zcxbysj2006@yahoo.com.cn

Ming-Yuan Ren
School of Software, Harbin University of Science and Technology, Harbin, China

Zhi-Mao Lu
College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

Ying-Hong Liang
School of Computer Engineering, Vocational University of Suzhou City, Suzhou, China

Da-Song Sun
Computer Center, Harbin University of Science and Technology, Harbin, China

Yong Liu
School of Computer Science and Technology, Heilongjiang University, Harbin, China

Abstract—**Phrase translation pairs are very useful for bilingual lexicography, machine translation system, cross-lingual information retrieval and many applications in natural language processing. Phrase translation pairs are always extracted from bilingual sentence pairs. In this paper, we extract phrase translation pairs based on word alignment results of Chinese-English bilingual sentence pairs and parsing trees of Chinese sentences, in order to decrease the influence of the grammar disagreement between Chinese and English. Discriminative features for phrase translation pairs are proposed to evaluate extracted ones in this paper, including translation literality, phrase alignment probability and phrase length difference. Multiple linear regression model combined with N-best strategy will be employed to filter phrase translation pairs, in order to improve the evaluating and filtering performance. Experimental results indicate that the filtering performance of phrase alignment probability is best in three kinds of discriminative features for evaluating Chinese-English phrase translation pairs. After multiple linear regression model combined with N-best strategy is used, its F1 achieves 86.24%.**

Index Terms—**phrase translation pairs, natural language processing, bilingual sentence pairs, parsing trees, discriminative features, multiple linear regression**

I. Introduction

Acquisition of phrase translation pairs, is a task where phrases in source language and phrases in target language, which can be translated from and to each other, are extracted from bilingual sentence pairs.

Bilingual sentence pairs have been recognized as a valuable resource for knowledge acquisition in many applications of natural language processing. A bilingual sentence pair contains a source language sentence and a target language sentence with the same semantic meaning. To make better use of them, bilingual sentence pairs are often aligned firstly. These alignments have been proven to be very useful in machine translation, word sense disambiguation, information retrieval, translation lexicon extraction, and so on. Intensive researches have been done on word level alignment and phrase level alignment. After bilingual sentence pairs are aligned, the corresponding words and corresponding phrases will be determined. These correspondences can be used directly in the acquisition of translation knowledge. Correspondence between source word and target word can be applied to acquisition of translation lexicon. The correspondence between source phrase and target phrase can be applied to acquisition of phrase translation pairs.

Phrase translation pairs are very important translation knowledge in natural language processing, which can be used in a variety of applications such as bilingual lexicography[1], machine translation system[2] and cross-lingual information retrieval[3]. Many methods have been proposed for acquisition of phrase translation pairs. John proves that finding optimal phrase alignment is NP-hard, and the problem of finding an optimal alignment can be cast as an integer linear program[4]. Parse-parse-match method is adopted firstly to extract phrase translation pairs[2]. Its main idea is that each language of bilingual corpus is parsed independently by a monolingual grammar, and then corresponding constituents are matched based on word alignment results. The disadvantage of this method is that robust

doi:10.4304/jcp.6.5.905-912

monolingual parser is needed for either language and there is always grammar disagreement between source language and target language. Melamed has proposed a fast and greedy algorithm called competitive linking in order to find word-to-word equivalences[5], which provides aligning anchors for extracting phrase translation pairs. Zhang builds a two-dimensional matrix to represent a bilingual sentence pair where the value of each cell corresponds to the point-wise mutual information between source word and target one. Box-shaped region whose mutual information values are similar with each others is looked upon as a phrase translation pair[6]. Zhang uses individually the monolingual language model to identify phrases in Chinese corpus and phrases in English corpus. Alignments are built on Chinese phrases and English phrases in order to extract phrase translation pairs which are applied to an example-based machine translation system[7]. Venugopal utilizes an improved IBM model to create knowledge sources in phrase level that effectively represent local phrasal context and global phrasal context, which can be applied to the process of phrase alignment. The method is robust to noisy alignments at both sentence level and corpus level, and can deliver phrase translation pairs in high quality that contribute to significant improvements in translation quality[8]. Philip uses a widely practised approach to get word alignments from two directions including source to target and target to source. Intersection operation and union operation can be applied to get refined word alignments with predesigned heuristics fixing the unaligned words. With this refined word alignment, target candidate phrases will be extracted for a given source phrase in the target sentence by searching the left and right projected boundaries[9]. Vogel uses translation model to calculate phrase translation probabilities[10]. Kenji uses translation literality to evaluate literality of bilingual sentence pairs and cleans the corpus in order to improve the quality of phrase translation pairs[11]. Zhao proposes an algorithm for extracting phrase translation pairs, which do not need explicit word alignment results. For each phrase translation pair, a bilingual lexicon-based evaluation score is computed to estimate the translation quality between source phrase and target phrase. A fertility score is computed to estimate how good the lengths are matched between source phrase and target phrase. A center distortion score is computed to estimate the relative position divergence between source phrase and target phrase. The method avoids the burden of testing and comparing different heuristics especially. On the other hand, the algorithm has such flexibilities that one can incorporate word alignment and heuristics in several possible stages to further improve the quality of phrase translation pairs[12]. Wu proposes a bilingual language model to parse bilingual sentence pairs simultaneously, from which phrase translation pairs are extracted, which avoids the impact of inaccuracy of monolingual parser[13]. But a suitable bilingual grammar is difficult to be found in practice. Luke presents a technique for selecting phrase translation pairs to be included in translation tables based on their estimated quality according to a translation model[14]. Vogel treats phrase alignment as a sentence splitting process which is to find the boundaries of the target phrase for a given source phrase, so that alignment lexicon probability for the overall sentence under this splitting process is optimal[15].

In this paper, we extract phrase translation pairs based on word alignment results of bilingual sentence pairs and parsing trees of source language sentences. Discriminative features including translation literality, phrase alignment probability, and phrase length difference are proposed for evaluating and filtering phrase translation pairs. Experimental results show that the evaluating and filtering performance of phrase alignment probability is best in three kinds of discriminative features for phrase translation pairs in open test, and its F1 achieves 85.11%. After multiple linear regression model combined with N-best strategy is used, Precision is 85.02% and Recall is 87.50%.

The rest of this paper is organized as follows: the method of acquiring phrase translation pairs based on Translation Corresponding Tree is described in Section II. Discriminative features of phrase translation pairs are proposed in Section III, and multiple linear regression model is used in order to improve the evaluating performance. Experimental results are given in Section IV. Conclusions of this paper are given in Section V.

II. Extraction Of Phrase Translation Pairs

In parse-parse-match method, source language and target language will be respectively analyzed by parsers. For Chinese and English, the alignment process will be restricted by the grammar disagreement, and lots of Chinese phrases can not be aligned to English phrases. So the number of extracted phrase translation pairs is very little, which leads that more translation knowledge will be lost. A bilingual language model which parses bilingual sentence pairs simultaneously can eliminate the influence of the grammar disagreement. After a bilingual sentence pair is parsed by a bilingual language model, phrase translation pairs will be gotten. But there is no parsing information in source part and target part. So the extent to which such phrase translation pairs are applied is very small. Wong proposes the annotation schema of translation corresponding tree (TCT) on bilingual sentence pairs, from which phrase translation pairs are extracted for constructing the example base[16]. Each TCT represents syntactic structure of source language sentence, and specifies the correspondence between source parsing tree and target string. In order to get TCT, source parsing tree and target sentence are aligned based on word alignment results. TCT can be viewed as the tree-string alignment of a bilingual sentence pair. The method can decrease the impact of the grammar disagreement between source language and target language. Phrase translation pairs can be acquired from TCT. This partly solves the problem that the alignment process is restricted by the grammar incompatibility between source language and target language. From tree-string alignment, we can extract phrase translation pairs. There are only parsing information in source parts of phrase translation pairs, and target parts do not include any parsing knowledge.

A Chinese parser tool and a Chinese–English word alignment tool are only used here. Firstly, Chinese sentence is analyzed by Chinese parser. Secondly, we use word alignment tool to align the bilingual sentence pair. At last, the tree-string alignment between Chinese and English is built according to word alignment results, from which phrase translation pairs can be extracted. The process of extracting Chinese-English phrase translation pairs from tree-string alignment is shown in Fig. 1.

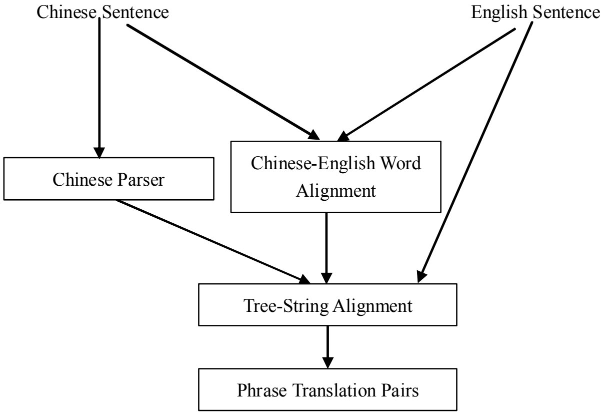

Figure 1. Extracting Chinese-English phrase translation pairs from tree-string alignment.

We use a triple sequence intervals [SNODE(n)/STREE(n)/STC(n)] encoded for each node in Chinese parsing tree to represent the corresponding relations between the structure of Chinese sentence and the substrings from both Chinese and English sentences. In tree-string alignment between Chinese sentence and English sentence, three interrelated correspondences are included. The first one is the correspondence between node n and its son nodes, encoded by the interval SNODE(n) that denotes which son node is the core node of n. The core node is very useful for acquiring translation templates and translation rules. The second one is the correspondence between the subtree and the substring of Chinese sentence, represented by the interval STREE(n) which indicates the interval of substring that is dominated by the subtree with node n as root. The last one is the correspondence between the subtree of Chinese sentence and the substring of English sentence, expressed by the interval STC(n) which indicates the interval containing the substring in English sentence corresponding to the subtree of Chinese sentence.

For a bilingual sentence pair (C, E), the algorithm of tree-string alignment is shown as follows:

1. Align words between C and E by word alignment tool. Extract word links between C and E from word alignment results.

2. Parse Chinese sentence C and T is the parsing tree of C.

3. The words in C and E are assigned with their positions respectively.

4. Post-traveling parsing tree T, for every node n in T

(1)If n is a leaf node in T which is a Chinese word, SNODE(n) and STREE(n) are set to the position of this word in Chinese sentence.

(2)If n is a non-leaf node in T which is a Chinese phrase and sons of node n are m_1, m_2, …, m_k, triple sequence intervals of node n and node m_i are respectively expressed as [SNODE(n)/STREE(n)/STC(n)] and [SNODE(m_i)/STREE(m_i)/STC(m_i)].

a.According to pre-defined heuristic rules, core node m_t is selected from m_1, m_2, …, m_k, and SNODE(n) is set to the value of SNODE(m_t). Heuristic rules include v+n->v, adj+n->n, adv+v->v and so on.

b.STREE(n)=[u,v] (u=min(Left(STREE(m_1)), Left(STREE(m_2)), …, Left(STREE(m_k))), v=max(Right(STREE(m_1)), Right(STREE(m_2)), …, Right(STREE(m_k)))).

c.STC(n)=[u,v] (u=min(Left(STC(m_1)), Left(STC(m_2)), …, Left(STC(m_k))), v=max(Right(STC(m_1)), Right(STC(m_2)), …, Right(STC(m_k)))).

When the algorithm is applied to Chinese-English bilingual sentence pairs, the tree-string alignments will be gotten. From the tree-string alignments, we can extract phrase translation pairs when the TCT is post-traveled.

For example, in the case of the following bilingual sentence pair, the process of extracting phrase translation pairs is shown as follows:

Chinese-English bilingual sentence pair:

Chinese sentence: 您能找开一张 100 元的钞票吗？

English sentence: Can you break a $ 100 bill?

Word alignment results:

您$_1$ 能$_2$ 找$_3$ 开$_4$ 一$_5$ 张$_6$ 100$_7$ 元$_8$ 的$_9$ 钞票$_{10}$ 吗$_{11}$ ？$_{12}$

Can$_1$ you$_2$ break$_3$ a$_4$ \$$_5$ 100$_6$ bill$_7$?$_8$

(1:2); (2:1); (4:3); (5:4); (7:6); (10:7); (12:8);

Parsing tree of Chinese sentence:

S[您 /r VP[能 /vz VO[BVP[找 /vg 开 /vq]NP[BNT[BMP[一/m 张/q]BNT[100/m 元/q]的/usde 钞票/ng]]]] 吗/y ?/wj]

Translation corresponding tree between Chinese and English is shown in Fig. 2.

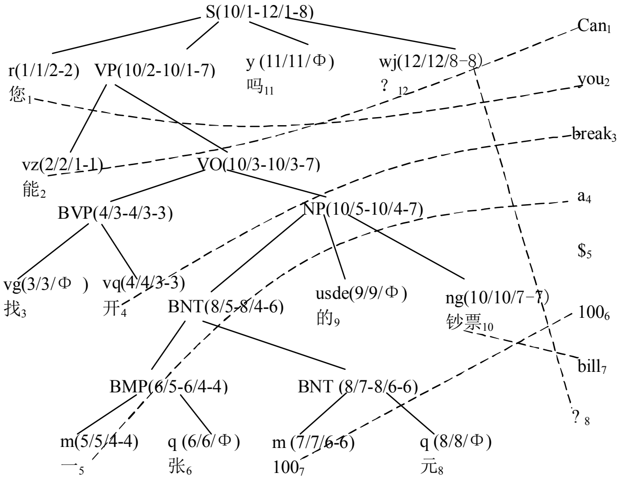

Figure 2. TCT between Chinese sentence and English sentence.

Extracted phrase translation pairs:

BMP[一/m 张/q]->a
BNT[100/m 元/q]->100
BNT[BMP[一/m 张/q] BNT[100/m 元/q]]->a $ 100
NP[BNT[BMP[一/m 张/q] BNT[100/m 元/q]的/usde 钞票/ng]]->a $ 100 bill
BVP[找/vg 开/vq]->break
VO[BVP[找/vg 开/vq] NP[BNT[BMP[一/m 张/q]BNT[100/m 元/q]]的/usde 钞票/ng]]]->break a $ 100 bill
VP[能/vz VO[BVP[找/vg 开/vq] NP[BNT[BMP[一/m 张/q]BNT[100/m 元/q] 的/usde 钞票/ng]]]]->Can you break a $ 100 bill

III. Evaluating Phrase Translation Pairs

Phrase translation pairs extracted from above include lots of noises because the whole extraction process is restricted by the accuracy of Chinese-English word alignment tool and Chinese parser tool. In order to improve the quality of phrase translation pairs, they should be evaluated and filtered. Left part of the phrase translation pair is a Chinese phrase with parsing information and the right part is only a phrase string in English. But when we evaluate and filter phrase translation pairs, phrase strings are only considered here. For example, on determining whether ‘BNT[100/m 元/q]->100’ is a correct phrase translation pair, we only consider ‘100 元->100’. In order to evaluate phrase translation pairs correctly, three kinds of discriminative features are used here, including translation literality, phrase alignment probability and phrase length difference.

1. *Translation Literality*

A bilingual sentence pair that has many word correspondences is more literal. Translation literality is a widely used measure for weighting literality of bilingual sentence pairs[11]. It can also be used for evaluating and filtering phrase translation pairs. Specifically, translation literality for each extracted phrase translation pair can be calculated and phrase translation pairs having scores over a given threshold can be regarded as correct ones. Translation literality of phrase translation pair Ph_c->Ph_e is usually defined as formula (1).

$$L(Ph_c, Ph_e) = \frac{Link(Ph_c, Ph_e)}{Num(Ph_c) + Num(Ph_e)} \quad (1)$$

Here, Ph_c denotes Chinese phrase of the phrase translation pair and Ph_e denotes its English phrase. Link(Ph_c, Ph_e) denotes the number of word links between phrase Ph_c and phrase Ph_e. Num(X) is the number of words in phrase X.

2. *Phrase Alignment Probability*

Brown uses $P(F|E)$ to compute the alignment probability of target language string E given source language string F[17]. The alignment probability $P(F|E)$ is shown in formula (2).

$$P(F \mid E) = \frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{l} t(f_j \mid e_i) \quad (2)$$

In our approach, IBM Model-1 is applied to compute word-to-word translation probability $t(f|e)$ that word f in source language is translated given word e in target language. This probability can be reliably estimated using expectation maximization(EM) algorithm[18].

Given the training set consisting of bilingual sentence pairs: $\{(f^{(s)}, e^{(s)}), s=1, 2, \ldots, S\}$, we use formula (3) and formula (4) to train word-to-word translation probability $t(f|e)$.

$$t(f \mid e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f \mid e; f^{(s)}, e^{(s)}) \quad (3)$$

$$c(f \mid e, f^{(s)}, e^{(s)}) = \frac{t(f \mid e)}{\sum_{k=1}^{l} t(f \mid e_k)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i) \quad (4)$$

Here λ_e^{-1} is a normalization factor. $c(f \mid e, f^{(s)}, e^{(s)})$ denotes expected number of times that word e is connected with word f. We use $P(Ph_c|Ph_e)$ to calculate alignment probability between phrase Ph_c and phrase Ph_e. If $P(Ph_c|Ph_e)$ is larger, the confidence of Ph_c->Ph_e being a correct one is higher.

For phrase translation pair ‘一 张 100 元 的 钞票->a $ 100 bill’, the computing process of its $P(Ph_c|Ph_e)$ is shown in Table I. Here, Ph_c=‘一 张 100 元 的 钞票’ and Ph_e=‘a $ 100 bill’.

TABLE I.
THE PROCESS OF CALCULATING PHRASE ALIGNMENT PROBABILITY

$t(c\|e)$	a	$	100	bill	$\sum_{i=1}^{n} t(c_j \mid e_i)$
一	0.57481900	0.00000000	0.00000000	0.00066778	0.57548678
张	0.00963629	0.00000000	0.06700060	0.01455820	0.09119509
100	0.00000000	0.00000000	0.73257400	0.00000000	0.73257400
元	0.00000000	0.25799100	0.03521210	0.00000000	0.29320310
的	0.00000000	0.00000886	0.00005538	0.00000000	0.00006424
钞票	0.00000000	0.00000000	0.00000000	0.06534590	0.06534590

Phrase alignment probability $P(Ph_c|Ph_e)$ for phrase translation pair ‘一 张 100 元 的 钞票-> a $ 100 bill’ is shown in formula (5).

$$\begin{aligned} P(Ph_c \mid Ph_e) &= \frac{1}{5^6}[0.57548678 * 0.09119509 \\ &\quad * 0.73257400 * 0.29320310 \quad (5) \\ &\quad * 0.00006424 * 0.06534590] \\ &\cong 3.0285218863149944e-12 \end{aligned}$$

3. *Phrase Length Difference*

The sentence length difference is a very good indication for the alignment of bilingual sentence pairs[19]. For a given phrase translation pair, we use phrase length difference to compute the confidence that phrase in source language can be translated from and to phrase in target language. It is described in formula (6). For the language pair of Chinese and English, the phrase length can be defined in several ways. A widely used method is to segment the Chinese sentence into words and count how many words are in Chinese phrase. For English phrase, we can also count its length in words.

$$\begin{aligned} P(A \mid Ph_c, Ph_e) &= P(|Ph_c| <-> |Ph_e| \,||\, Ph_c, Ph_e) \\ &\cong P(|Ph_c| <-> |Ph_e| \,|||\, Ph_c|, |Ph_e|) \\ &\cong P(|Ph_c| - |Ph_e|) \\ &\cong P(D(|Ph_c|, |Ph_e|)) \end{aligned} \quad (6)$$

Here, $|Ph_c|$, $|Ph_e|$ denote the length of phrase Ph_c and the length of phrase Ph_e respectively. The length difference between source phrase and target phrase in a phrase translation pair can be viewed approximately as the difference between the length of source phrase and the length of target phrase.

D($|Ph_c|$, $|Ph_e|$) denotes the difference between the length of phrase Ph_c and the length of phrase Ph_e, which is assumed to be a normal distribution[19]. It is computed according to formula (7).

$$D(|Ph_c|, |Ph_e|) = \frac{|Ph_e| - c|Ph_c|}{\sqrt{(|Ph_c|+1)\sigma^2}} \sim N(0,1) \quad (7)$$

Phrase length difference $P(D(|Ph_c|, |Ph_e|))$ between phrase Ph_c and phrase Ph_e is defined in formula (8).

$$P(D(|Ph_c, Ph_e|)) = P(\frac{|Ph_e| - c|Ph_c|}{\sqrt{(|Ph_c|+1)\sigma^2}}) \quad (8)$$

Here c is a constant indicating the mean length ratio which is the expected number of words in English phrase Ph_e per word in Chinese phrase Ph_c. σ^2 is the variance of c. For training set of phrase translation pairs $\{Ph^i_c\text{-}>Ph^i_e|\mathrm{i}=1, 2, \ldots, n\}$, c is computed according to formula (9). σ^2 is computed as formula (10) describes.

$$c = \frac{\sum_{i=1}^{n} Num_{word}(Ph_e^i)}{\sum_{i=1}^{n} Num_{word}(Ph_c^i)} \quad (9)$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{Num_{word}(Ph_e^i)}{Num_{word}(Ph_c^i)} - c)^2 \quad (10)$$

For the test set of phrase translation pairs $\{Ph^i_c\text{-}>Ph^i_e|i=1, 2, \ldots, m\}$, translation literality $L(Ph_c, Ph_e)$, phrase alignment probability $P(Ph_c|Ph_e)$ and phrase length difference $P(D(|Ph_c|, |Ph_e|))$ are used to score every phrase translation pair respectively. Then phrase translation pairs in test set are ranked in descending order according to their evaluation scores. N-best strategy is applied to select the front N phrase translation pairs whose evaluation scores are highest, and they will be labeled as positive instances. Others will be labeled as negative ones.

A linear combination model based on multiple discriminative features is used to evaluate phrase translation pairs in order to improve the filtering performance. Here, discriminative features including translation literality $L(Ph_c, Ph_e)$, phrase alignment probability $P(Ph_c|Ph_e)$, and phrase length difference $P(D(|Ph_c|, |Ph_e|))$ are used in the linear combination model. For the given phrase translation pair $Ph_c\text{-}>Ph_e$, its evaluation score $y(Ph_c\text{-}>Ph_e)$ is calculated as formula (11) describes. If the value of $y(Ph_c\text{-}>Ph_e)$ is larger, the confidence of $Ph_c\text{-}>Ph_e$ being a correct phrase translation pair is higher.

$$\begin{aligned} y(Ph_c - > Ph_e) &= w_1 * L(Ph_c, Ph_e) + w_2 * P(Ph_c \mid Ph_e) \\ &+ w_3 * P(D(|Ph_c|, |Ph_e|)) \end{aligned} \quad (11)$$

Actually the problem can be viewed as multiple linear regression model[20]. The purpose of multiple linear regression is to find a hyperplane which can reflect the real distribution of training data. The values of parameters w_1, w_2, and w_3 will differ from domains. When the model is applied to the filtering task in a new domain, we can automatically train parameters to determine their values.

On the training set $\{Ph^i_c\text{-}>Ph^i_e|i=1, 2, \ldots, n\}$, translation literality $L(Ph_c, Ph_e)$, phrase alignment probability $P(Ph_c|Ph_e)$ and phrase length difference $P(D(|Ph_c|, |Ph_e|))$ are used to score every phrase translation pair respectively. The evaluation matrix $A = [L(Ph_c, Ph_e), P(Ph_c \mid Ph_e), L]_{n*3}$ for training set of phrase translation pairs will be gotten. H is the manually-annotated results for training set of phrase translation pairs and it is a n-dimensional column vector. Multiple linear regression method will find the optimized weight vector W^*, which makes the automatically-labeled results consistent with manually-annotated ones as much as possible. The W^* is computed as formula (12) describes.

$$W^* = \min_W \left\| AW^T - H \right\| \quad (12)$$

Here, parameter $W=(w_1, w_2, w_3)$. The solving process of parameter W in multiple linear regression model is shown in formula (13) [20].

$$W^* = ((A^T A)^{-1} A^T H)^T \quad (13)$$

At the same time, we can get the evaluation matrix $B = [L(Ph_c, Ph_e), P(Ph_c \mid Ph_e), L]_{m*3}$ for test set of phrase translation pairs $\{Ph^i_c\text{-}>Ph^i_e|i=1, 2, \ldots, m\}$. With the optimized weight vector W^*, multiple linear regression model can be used to score phrase translation pairs in test set. The evaluation scores of phrase translation pairs in test set are computed according to formula (14). H^T is automatically-labeled results for test set of phrase translation pairs.

$$H^T = B \cdot W^* \quad (14)$$

Based on H^T, the front N phrase translation pairs whose evaluation scores are highest will be selected as positive instances, and others will be annotated as negative ones.

IV. Experiment

81204 Chinese-English bilingual sentence pairs from traveling field are collected to acquire phrase translation pairs. The extraction method of phrase translation pairs described in Section II is used here. 286790 phrase translation pairs are obtained. Here word alignment tool and Chinese parser tool[21] are developed by MOE-MS Key Laboratory of Natural Language Processing and Speech in Harbin Institute of Technology. Their performances are shown in Table II.

TABLE II.
WORD ALIGNMENT TOOL AND CHINESE PARSER

	Precision	Recall
Word alignment tool	86%	89%
Chinese parser	78%	79%

We randomly select 6041 phrase translation pairs from these 286790 ones. Two human annotators are asked to manually annotate these 6041 phrase translation pairs. 4440 phrase translation pairs are annotated as positive instances and 1601 phrase translation pairs are annotated as negative ones. We divide these 6041 phrase translation pairs into two parts. One is training set of phrase translation pairs and the other is test set of phrase translation pairs. They are described in Table III.

TABLE III.
TRAINING DATA AND TEST DATA

	Positive	Negative
Training data	3697	1338
Test data	743	263

c and σ^2 are parameters of phrase length difference $P(D(|Ph_c|, |Ph_e|))$. We employ positive instances in training set of phrase translation pairs to estimate parameter c according to formula (9). Parameter σ^2 is estimated by formula (10) based on positive instances in training set of phrase translation pairs. The results are shown in Table IV.

TABLE IV.
ESTIMATION OF PARAMETERS C AND σ^2

Parameter	c	σ^2
Value	0.8320	0.1859

We design Precision, Recall and F1 to measure the performance of filtering phrase translation pairs. S is the set of phrase translation pairs which are labeled automatically as positive instances, and T is the set consisting of phrase translation pairs which are annotated manually as positive ones. Precision, Recall and F1 are shown respectively in formula (15), formula (16) and formula (17).

$$\mathrm{Pr}\,ecision = \frac{|S \cap T|}{|S|} * 100\% \quad (15)$$

$$\mathrm{Re}\,call = \frac{|S \cap T|}{|T|} * 100\% \quad (16)$$

$$F1 = \frac{2 * \mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} * 100\% \quad (17)$$

We sort phrase translation pairs of training set in descending order according to their evaluation scores and employ N-best strategy to label phrase translation pairs. We set N=500, 1000, 1500, …, 5000 respectively, and label phrase translation pairs according to evaluation scores under different N-best strategy, in which the front N phrase translation pairs whose evaluation scores are highest will be labeled as positive instances. Then automatically-labeled results are evaluated according to manually-annotated results.

When Precision is used as measure to evaluate filtering performance of different discriminative features, evaluation results are shown in Fig. 3.

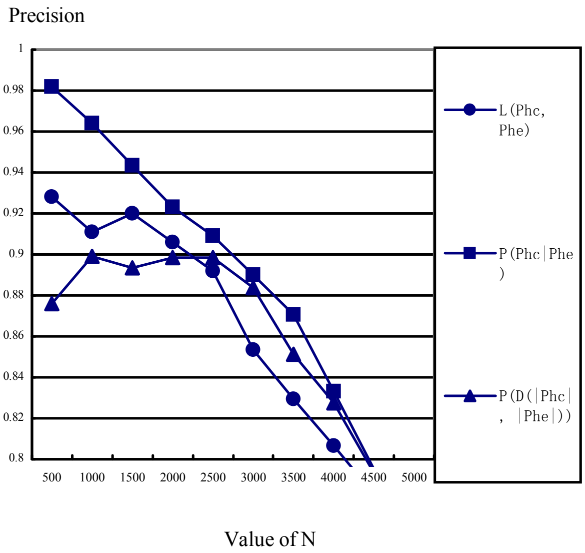

Figure 3. Precision of discriminative features under different N-best strategy.

We also use Recall as measure to evaluate filtering performance of different discriminative features. The evaluation results are shown in Fig. 4.

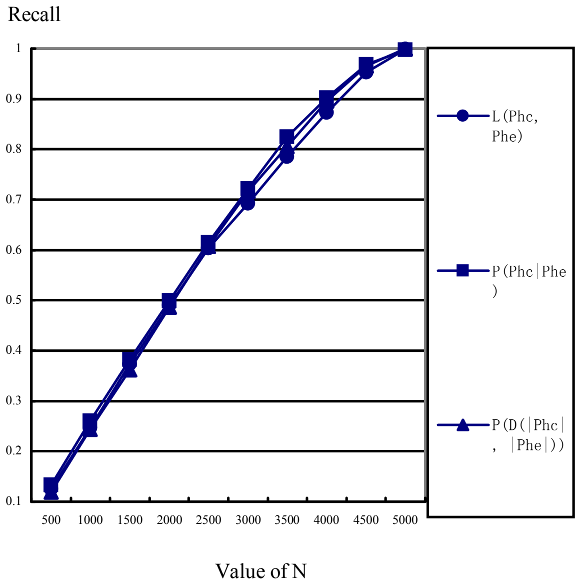

Figure 4. Recall of discriminative features under different N-best strategy.

At the same time, when F1 is used as measure to evaluate filtering performance of different discriminative features, evaluation results are shown in Fig. 5.

From Fig. 3, Fig. 4 and Fig. 5, we can see that phrase alignment probability $P(Ph_c|Ph_e)$ does better than other discriminative features on filtering performance. This is because that when $P(Ph_c|Ph_e)$ is computed, discriminative feature for phrase translation pairs can be divided into discriminative features for evaluating multiple Chinese-English word pairs. The computation process is very precise and reasonable. It can reflect the degree that phrase Ph_c and phrase Ph_e can be translated from and to each other.

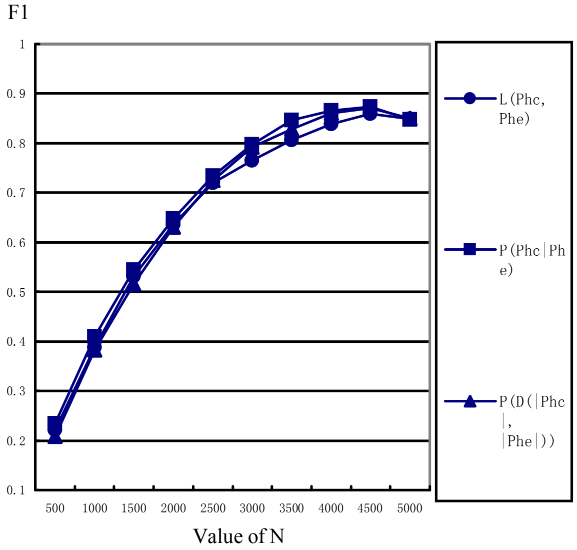

Figure 5. F1 of discriminative features under different N-best strategy.

Discriminative features including $L(Ph_c, Ph_e)$, $P(Ph_c|Ph_e)$ and $P(D(|Ph_c|, |Ph_e|))$ are applied respectively to score phrase translation pairs in test set. Then phrase translation pairs are sorted in descending order based on their evaluation scores. In training set, 73.43 percent of phrase translation pairs are positive instances. So, when phrase translation pairs in test set are sorted in descending order and N-best strategy is used, the value of N should be set to 739. This is because that the number of phrase translation pairs in test set is 1006(739:1006≈73.43%). The front 739 phrase translation pairs whose evaluation scores are highest will be labeled as positive instances, and others are labeled as negative ones. At the same time, Precision, Recall, and F1 are used as measures to evaluate the filtering performance in open test. The results are shown in Table V.

From Table V, we can find that when phrase alignment probability $P(Ph_c|Ph_e)$ is used to evaluate phrase translation pairs in test set and N-best (N=739) strategy is used to select the front N phrase translation pairs which have highest scores, its filtering performance is best in three kinds of discriminative features. Precision is 84.71%, Recall is 85.52%, and F1 achieves 85.11%. This is because that $P(Ph_c|Ph_e)$ can evaluate phrase translation pairs better. After multiple linear regression model combined with N-best (N=739) strategy is used, F1 is 86.24%. The filtering performance is improved further in open test.

TABLE V.
ANALYSIS OF FILTERING PERFORMANCE (N=739)

	Precision	Recall	F1
$L(Ph_c,Ph_e)$*+N-Best*	81.87%	82.65%	82.26%
$P(Ph_c|Ph_e)$*+N-Best*	**84.71%**	**85.52%**	**85.11%**
$P(D(|Ph_c|, |Ph_e|))$*+N-Best*	84.17%	84.97%	84.57%
Multiple linear regression Model+N-Best	***85.02%***	***87.50%***	***86.24%***

V. CONCLUSIONS

In this paper, phrase translation pairs are extracted based on word alignment results of bilingual sentence pairs and parsing trees of source language sentences. Discriminative features including translation literality, phrase alignment probability, and phrase length difference are proposed for evaluating and filtering phrase translation pairs. Experimental results show that the evaluating and filtering performance of phrase alignment probability is best in three kinds of discriminative features in open test. Multiple linear regression model and N-best strategy are applied to improve the filtering performance further. In the future, we will apply acquired phrase translation pairs to bilingual lexicography and cross-lingual information retrieval in order to test their performance.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China under Grant Nos. 60903082, 60603092, Science and Technology Research Funds of Education Department in Heilongjiang Province under Grant Nos. 11541045, Top-Notch Talent Funds of Harbin University of Science and Technology, School Foundation of Harbin University of Science and Technology under Grant Nos. 2008XQJZ017, and Jiangsu Province Support Software Engineering R&D Center for Modern Information Technology Application in Enterprise under Grant Nos. SX200907. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

REFERENCES

[1] W. A. Gale, and K. W. Church, "Identifying word correspondences in parallel texts," *Proceedings of the 4th DARPA Workshop on Speech and Natural Language*, pp. 152-157, 1991.

[2] K. Imamura, "Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT," *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 74-84, 2002.

[3] D. W. Oard, and B. J. Dorr, *A survey of multilingual text retrieval*, Technical Report, Institute for Advanced Computer Studies, University of Maryland, 1996.

[4] D. N. John and K. Dan, “The complexity of phrase alignment problems,” *Proceedings of ACL-08: HLT*, pp. 25-28, 2008.

[5] I. D. Melamed, “A word-to-word model of translational equivalence,” *Proceedings of Conference of the Association for Computational Linguistics*, pp. 490-497, 1997.

[6] Y. Zhang, S. Vogel and A. Waibel, “Integrated phrase segmentation and alignment model for statistical machine translation,” *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, 2003.

[7] Y. Zhang, R. D. Brown, and R. E. Frederking, “Adapting an example-based translation system to Chinese,” *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1-4, 2001.

[8] A. Venugopal, S. Vogel, and A. Waibel, “Effective phrase translation extraction from alignment models,” *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp.319-326, 2003.

[9] K. Philip and K. Kevin, “Feature-rich statistical translation of noun phrases,” *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, 2003.

[10] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel, *The CMU statistical machine translation system*, Language Technologies Institute, Carnegie Mellon University, 2003.

[11] K. Imamura and E. Sumita, “Bilingual corpus cleaning focusing on translation literality,” *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 1713-1716, 2002.

[12] B. Zhao and S. Vogel, “A generalized alignment-free phrase extraction,” *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 141-144, 2005.

[13] D. K. Wu, “Stochastic inversion transduction grammars and bilingual parsing of parallel corpora,” *Computational Linguistics*, vol. 23, no. 3, pp. 377-404, 1997.

[14] L. Zettlemoyer and R. Moore, “Selective phrase pair extraction for improved statistical machine translation,” *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 209-212, 2007.

[15] S. Vogel, S. Hewavitharana, and M. Kolss, “The ISL statistical translation system for spoken language translation,” *Proceedings of the International Workshop on Spoken Language Translation*, pp. 65-72, 2004.

[16] F. Wong, D. C. Hu, Y. H. Mao and M. C. Dong, “A flexible example annotation schema: translation corresponding tree representation,” *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1079-1085, 2004.

[17] P. F. Brown, “The mathmatics of statistical machine translation: parameter estimation,” *Computational Linguistics*, vol. 19, no. 2, pp. 263-311, 1993.

[18] W. B. Cavnar, and J. M. Trenkle, “N-gram-based text categorization,” *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175, 1994.

[19] K. W. Church, “Char_align: a program for aligning parallel texts at the character level,” *Proceedings of ACL-93*, 1993.

[20] H. Trevor, T. Robert, and F. Jerome, *The elements of statistical learning: data mining, inference and prediction*, Springer Publisher, 2001.

[21] H. L. Cao, T. J. Zhao, M. Y. Yang, and S. Li, “Parsing Chinese with head-driven model,” *Proceedings of International Conference on Machine Learning and Cybernetics*, pp. 2618-2622, 2004.

Chun-Xiang Zhang is Ph.D. and graduates from MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, in Harbin Institute of Technology. He is also an associate professor in Harbin University of Science and Technology. His research interests are natural language processing, machine translation and machine learning. He has authored and coauthored more than twenty journal and conference papers in these areas.

Ming-Yuan Ren is Ph.D. candidate in School of Astronautics, in Harbin Institute of Technology. He is also a lecturer in Harbin University of Science and Technology. His research interests are natural language processing, machine translation and machine learning. He has authored and coauthored more than ten journal and conference papers in these areas.

Zhi-Mao Lu is Ph.D. and graduates from MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, in Harbin Institute of Technology. He is also a professor and Ph.D. supervisor in Harbin Engineering University. His research interests are natural language processing, machine translation and machine learning. He has authored and coauthored more than thirty journal and conference papers in these areas.

Ying-Hong Liang is Ph.D. and graduates from MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, in Harbin Institute of Technology. She is also an associate professor in School of Computer Engineering, Vocational University of Suzhou City. Her research interests are natural language processing, machine translation and machine learning. She has authored and coauthored more than twenty journal and conference papers in these areas.

Da-Song Sun is an associate professor in Harbin University of Science and Technology. His research interests are natural language processing, and machine learning.

Yong Liu is Ph.D. and graduates from School of Computer Science and Technology, in Harbin Institute of Technology. He is also a lecturer in Heilongjiang University. His main research interests include data mining and graph data management.