Phrase Alignment Based on Combination of Multiple Strategies

Chun-Xiang Zhang
School of Software, Harbin University of Science and Technology, Harbin, 150080, China
Email: z6c6x6@yahoo.com.cn

Xue-Yao Gao
College of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China

Zhi-Mao Lu
College of Information and Communication Engineering, Harbin Engineering University, Harbin, 150080, China

Da-Song Sun
Computer Center, Harbin University of Science and Technology, Harbin, 150080, China

Yong Liu
School of Computer Science and Technology, Heilongjiang University, Harbin, 150080, 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. There is phrase boundary information in parsing trees of sentences. Linguistics knowledge in translation lexicon and semantic lexicon, and statistics results from bilingual corpus can be used to align Chinese words and English words, which will provide alignment information for extracting phrase translation pairs. In this paper, we propose a new method to extract phrase translation pairs based on aligning Chinese words and English words in bilingual corpus with multiple alignment strategies. Experimental results indicate that the extracted phrase translation pairs achieve 63.07% at accuracy, when the new method is applied.**

***Index Terms*—phrase translation pairs, natural language processing, linguistics knowledge, bilingual corpus, multiple alignment strategies**

I. INTRODUCTION

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]. Extraction 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 word-aligned bilingual corpus.

The word-aligned bilingual corpus is an important knowledge source for many tasks in natural language processing. Word alignment is an object for indicating the corresponding words between source sentences and target sentences in bilingual corpus. Word alignment results can tells us which words in target sentence are linked to words in source sentence. An English-Chinese word alignment model based on bilingual lexicon and language knowledge is given, and it is built on the theory of formal optimal partition of the bilingual sentence pairs[4]. An approach of word alignment based on multi-grain model is proposed, where a bilingual sentence pair is split into blocks in different grains, and word alignments within each corresponding blocks are extracted, which will restrict the searching space of word alignment in the relatively accurate range and reduce the mapping errors[5]. A bootstrapping frame is designed, where the bilingual corpus is aligned based on translation dictionary, and translation dictionary is expanded based on word alignment results. The process goes on until the threshold is gotten[6]. A discriminative framework for word alignment based on the linear model is proposed, where all knowledge sources are treated as feature functions that depend on source sentences, target sentences, and the alignment results between them[7]. The linear combination of features gives an overall score to each candidate alignment, from which the best alignment is selected. A discriminative word alignment method based on CRF model is proposed for aligning Mongolian-English bilingual sentence pairs, which has the ability to use a large variety of features flexibly and to combine information from various knowledge sources[8]. An unsupervised expectation maximization (EM) algorithm is proposed to align the bilingual corpus[9].

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[10]. A hierarchical phrase alignment method has been proposed, which is used to extract equivalent

doi:10.4304/jsw.7.2.332-338

phrases hierarchically from a bilingual corpus even though they belong to different language families[11]. Zhang builds a mutual information matrix to represent a bilingual sentence pair. Box-shaped region whose mutual information values are similar with each others is looked upon as a phrase translation pair[12]. In bilingual sentence pairs, English statistical dependency parser is used to determine dependency relations between English base phrases, and Japanese statistical dependency parser is utilized to determine dependency relations between Japanese base phrases. Then English base phrases and Japanese base phrases are paired with their translations[13]. 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 pre-designed heuristics. 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[14]. Dependency structures for source sentences and target sentences are obtained, which are then aligned, and structural translation correspondences are extracted from the resulting alignment[15]. Zhao proposes an algorithm for extracting phrase translation pairs, which does not need explicit word alignment results. Bilingual lexicon-based evaluation score, fertility score and center distortion score are computed for aligning source phrases and target phrases[16]. Source parsing tree and target sentence are aligned based on the word alignment results[17]. At the same time, translation literality, alignment probability, and length difference are used for select phrase translation pairs which are considered to be correct semantically. Zettlemoyer presents a technique for selecting phrase translation pairs to be included in translation tables based on their estimated quality according to a translation model[18]. 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[19].

In this paper, we propose a new method to extract phrase translation pairs from Chinese-English bilingual corpus by applying the technology of parsing analysis and the method of multiple-strategy word alignment. Experimental results show that after the new method is applied, accuracy of extracted phrase translation pairs achieves 63.07%.

The rest of this paper is organized as follows: the word alignment method based on multiple strategies including lexicon similarity, translation similarity, semantic similarity and co-occurrence is described in section II. Phrase alignment method based on multiple-strategy word alignment is proposed in section III. Experimental results are given in section IV. Conclusions of this paper are given in section V.

II. Word Alignment With The Combination of Multiple Strategies

There are lots of linguistic knowledge in translation lexicon and semantic lexicon. Statistics information can be gotten from bilingual corpus. Linguistic knowledge and statistics information are very useful for align Chinese words and English words in bilingual sentence pairs. Lü combines lexicon similarity, translation similarity, semantic similarity and co-occurrence to measure the alignment degree between Chinese words and English words in bilingual corpus, and they are described as follows[20].

Bilingual lexicon is the most direct source of linguistic knowledge for evaluating the similarity between source-target word pairs. Based on bilingual lexicon, lexicon similarity SimL(c, e) can be computed to measure the similarity between Chinese word c and English word e. SimL(c, e) is 1 when English word e is a translation of Chinese word c, or Chinese word c is a translation of English word e. Otherwise SimL(c, e) is 0.

Although English word e does not appear in lexicon translations of any Chinese word, there are common parts between c and lexicon translations of e. Translation similarity SimT(c, e) can be utilized to measure the similarity between Chinese word c and English word e. SimT(c, e) is computed as formula (1) describes.

$$SimT(c,e) = \max_{d \in DT(e)} \frac{2*|d \cap c|}{|d|+|c|}. \quad (1)$$

Here, DT(e) is the set containing all Chinese translations of English word e in bilingual lexicon, and d∩c denotes common Chinese words which d and c all contains. |X| is the number of words in X.

Tonyi cilin is a Chinese semantic lexicon. Tonyi Cilin and English-Chinese translation lexicon are combined to compute semantic similarity between c and e. Semantic similarity SimS(c, e) is computed in formula (2).

$$SimS(c,e) = \max_{d \in DT(e)} \max_{\substack{S_m \in Classof(d) \\ S_n \in Classof(c)}} 1/SDist(S_m, S_n). \quad (2)$$

Here, DT(e) is the set containing all Chinese translations of English word e in bilingual lexicon, and *Classof*(X) is the set containing all sense code of Chinese word X in Tonyi Cilin. *SDist*(S_m, S_n) is the distance between S_m and S_n.

Asso(c, e) measures the similarity between Chinese word c and English word e in terms of co-occurrence, and it can be calculated by X^2[1]. Asso(c, e) can be computed in a large number of bilingual sentence pairs.

SimL(c, e) is applied to align Chinese words and English words in a bilingual sentence pair firstly. Secondly, SimT(c, e) is used to align those Chinese words and English words which are not aligned. Thirdly, SimS(c, e) is utilized to align those Chinese words and English words which are not aligned. Lastly, Asso(c, e) is applied to align those Chinese words and English words which are not aligned.

III. Extracting Phrase Translation Pairs

In parsing tree with phrase category, nodes are organized hierachically. Nodes describe the syntactic structures of phrases, which are categorized into base

phrases and complex phrases. Complex phrases are made up of base phrases nestedly. There are phrase boundaries in parsing tree. When Chinese sentence is analyzed and English sentence is analyzed, we can get Chinese parsing tree and English parsing tree. There is the correspondence between Chinese words and English words in a bilingual sentence pair. The correspondences between word pairs can provide the information for aligning syntactic nodes in Chinese parsing tree with syntactic nodes in English parsing tree. Then syntactic nodes in two parsing trees can be matched. Syntactic nodes are corresponded with syntactic phrases. When Chinese parsing tree is traced, Chinese syntactic phrases are obtained and corresponding English syntactic phrases are also gotten according to the matching results of two parsing trees. When Chinese syntactic phrase is used as source part and the corresponding English syntactic phrase is utilized as target part, a phrase translation pair will be obtained.

For a bilingual sentence pair (C, E), the steps of aligning Chinese parsing tree and English parsing tree is shown as follows:

(1)Chinese words in C and English words in E are aligned with the multiple-strategy word alignment method mentioned in section II. Extract word links between C and E from word alignment results.

(2)Parse Chinese sentence C and *CT* is the parsing tree of C. Parse English sentence E and *ET* is the parsing tree of E.

(3)Trace parsing tree *CT* top-down, and get its all syntactic nodes CT_1, CT_2, …, CT_m. Trace parsing tree *ET* top-down, and get its all syntactic nodes ET_1, ET_2, …, ET_m.

(4)For Chinese syntactic node CT_i, formula (3) is utilized to determine its English syntactic node corresponded with CT_i. Here, *Link*(*X*, *Y*) denotes the number of word links between Chinese phrase corresponded with syntactic node *X* in *CT* and English phrase corresponded with syntactic node *Y* in *ET*. *Number*(*X*) is the number of words in phrase corresponded with syntactic node *X*.

$$ET_j = \max_{u \in \{ET_1, ET_2, \ldots, ET_n\}} \frac{Link(CT_i, u)}{Num(CT_i) + Num(u)}. \quad (3)$$

(5)Call the algorithm of solving boundary conflicts to process translation boundary conflicts between syntactic nodes with the same height in Chinese parsing tree *CT*.

There may be translation boundary conflicts between different syntactic nodes with the same height in Chinese parsing tree. It is shown in Figure 1.

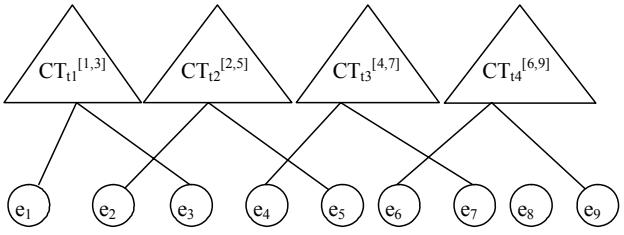

Figure 1. Boundary conflicts between syntactic nodes with the same height in Chinese parsing tree.

Here, $CT_{t1}^{[1,3]}$, $CT_{t2}^{[2,5]}$, $CT_{t3}^{[4,7]}$ and $CT_{t4}^{[6,9]}$ are all syntactic nodes whose height are t in Chinese parsing tree *CT*. In syntactic node $CT_{ti}^{[m,n]}$, t is its height in Chinese parsing tree and *i* is its label. At the same time, [*m*, *n*] means that English translation of syntactic node $CT_{ti}^{[m,n]}$ is the string from the *m*th word to the *n*th word in English sentence. [*m*, *n*] can be called as the translation boundary of this syntactic node. Translation extent *SP* of syntactic node $CT_{ti}^{[m,n]}$ is computed in formula (4).

$$SP(CT_{ti}^{[m,n]}) = n - m. \quad (4)$$

There are translation correspondences in bilingual sentence pairs. For example, equivalent word pairs, phrase translation pairs, and bilingual sentence pairs. So, there should not be translation boundary conflicts between syntactic nodes with the same height in Chinese parsing tree. In order to acquire more translation knowledge, we should be sure that the number of phrase translation pairs extracted from every layer of Chinese parsing tree is the largest and the extent to which these phrase translation pairs cover English sentence is the biggest.

Here, translation association degree δ is used to describe the translation boundary conflicts between syntactic nodes with the same height. $ST_{t1}^{[j_1,k_1]}, ST_{t2}^{[j_2,k_2]}, \ldots, ST_{tm}^{[j_m,k_m]}$ are all non-leaf syntactic nodes whose height are t in Chinese parsing tree. The association degree δ of $ST_{tu}^{[j_u,k_u]}$ is the number of syntactic nodes whose height is also t, and there are translation boundary conflicts between $ST_{tu}^{[j_u,k_u]}$ and these syntactic nodes. The value of δ is computed as formula (5) describes.

$$\begin{aligned} &\delta = \sum_{n=1}^{m} \mathrm{sgn}(j_n \in [j_u, k_u] + k_n \in [j_u, k_u]) - 1 \\ &x \in [j_u, k_u] = 1 \quad iff\,(x > j_u)\,\&\&\,(x <= k_u) \\ &x \in [j_u, k_u] = 0 \quad iff\,(x < j_u)\,||\,(x > k_u) \\ &\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x <= 0 \end{cases} \end{aligned} \quad . \quad (5)$$

Translation boundary conflicts are solved top-down in Chinese parsing tree. The algorithm of solving boundary conflicts is shown as follows:

1. Initialization, t=0, IsExist=0, A=Φ, B=Φ.
2. While (t<the height of *CT*)
   ①All non-leaf syntactic nodes whose heights are t in *CT*, are collected into the set A.
   ②*SP* and δ of nodes in set A are computed respectively according to formula (4) and formula (5).
   ③If there is node whose δ is larger than 0, then IsExist is set to 1. Otherwise, IsExist is set to 0.
   ④While(IsExist)

a. Nodes in set A are sorted in descending order based on the value of translation association degree δ.
b. Node m whose translation extent *SP* is smallest, is selected from all nodes whose translation association degree δ are the largest in set A.
c. A=A-{m}.
d. δ values of nodes in set A are computed.
e. If there is node whose δ is larger than 0, then IsExist is set to 1. Otherwise, IsExist is set to 0.

⑤B=B+A, t=t+1, A=Φ.

Chinese parsing tree is traced top-down. Then Chinese syntactic node and the corresponding English syntactic node are extracted. Syntactic node is corresponded with syntactic phrase. Chinese syntactic phrase is used as source part, and English syntactic phrase is used as target part. Then phrase translation pairs are obtained.

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: 我们想要张靠窗户的桌子。

English sentence: We want to have a table near the window.

Word alignment results:

我们/1 想要/2 张/3 靠/4 窗户/5 的/6 桌子/7 。/8

We/1 want to/2 have/3 a/4 table/5 near/6 the/7 window/8 ./9

(1:1); (2:2); (4:6); (5:8); (7:5); (8:9);

Parsing tree of Chinese sentence:

S[我们/r/1 VO[想要/vg/2 NP[张/q/3 NP[VO[靠/vg/4 窗户/ng/5]的/usde/6 桌子/ng/7]]]。/wj/8]

Parsing tree of English sentence:

S[We/PRP/1 VP[want to/VBP/2 VP[have/VB/3 NP[BNP[a/ART/4 table/NN/5] PP[near/IN/6 BNP[the/ART/7 window/NN/8]]]]] ./FSP/9]

The correspondence between Chinese parsing tree and English parsing tree is shown in Figure 2.

Extracted phrase translation pairs:

VO[想要/vg NP[张/q NP[VO[靠/vg 窗户/ng] 的/usde 桌子 /ng]]]->VP[want to/VBP VP[have/VB NP[BNP[a/ART table/NN] PP[near/IN BNP[the/ART window/NN]]]]]

NP[张/q/3 NP[VO[靠/vg/4 窗户/ng/5]的/usde/6 桌子 /ng/7]]->NP[BNP[a/ART table/NN] PP[near/IN BNP[the/ART window/NN]]]

NP[VO[靠 /vg/4 窗 户 /ng/5] 的 /usde/6 桌 子 /ng/7]->NP[BNP[a/ART table/NN] PP[near/IN BNP[the/ART window/NN]]]

VO[靠 /vg/4 窗 户 /ng/5]->PP[near/IN BNP[the/ART window/NN]]

This kind of phrase translation pairs is very useful in machine translation, bilingual lexicography, and translation ordering model. This is because that there are correspondences of morphology, part of speech, semantics and syntax between its source part and target part. For example, we can get the template of translation ordering 'VO+的+ng->BNP+PP'.

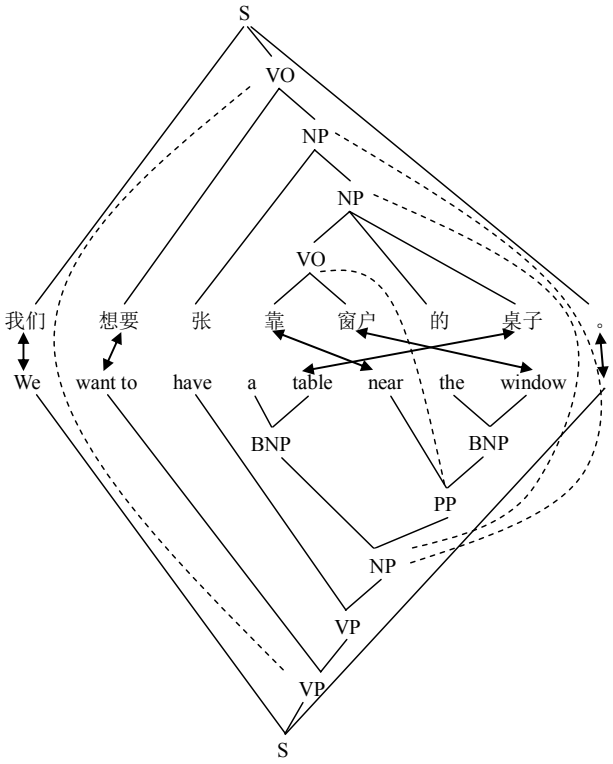

Figure 2. Correspondence between Chinese parsing tree and English parsing tree.

IV. EXPERIMENT

81204 Chinese-English bilingual sentence pairs from traveling field are collected to acquire phrase translation pairs. The performance of bilingual corpus is described in Table I.

TABLE I.
PERFORMANCE OF CHINESE-ENGLISH BILINGUAL CORPUS

Chinese Corpus	
Number of Chinese sentences	81204
Number of Chinese words	18309
Average length of Chinese sentence (Bytes)	23.3
Average length of Chinese sentence (Words)	8.2
English Corpus	
Number of English sentences	81204
Number of English words	19362
Average length of English sentence (Bytes)	37.6
Average length of English sentence (Words)	8.0

Chinese parser tool is used to analyze Chinese sentence in bilingual corpus, and its parsing tree is gotten. English parser tool is applied to analyze English sentence in bilingual corpus, and its parsing tree is obtained. Here, Chinese parser tool[21] and English parser tool 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.
ENGLISH PARSER TOOL AND CHINESE PARSER TOOL

	Precision	Recall
English parser tool	77%	80%
Chinese parser tool	78%	79%

Multiple strategies of word alignment are combined to align Chinese words and English words in bilingual sentence pairs, as described in section II. At the same time, the phrase alignment method proposed in section III is used to match syntactic nodes between Chinese parsing tree and English parsing tree based on the results of word alignment. Then Chinese parsing trees are traced top-down, Chinese syntactic nodes are gotten. The corresponding English syntactic nodes are obtained according to the matching results. Chinese syntactic phrase is used as source part and English syntactic phrase is utilized as target part. Then phrase translation pairs are obtained. There are 30 categories of syntax labels. Phrase translation pairs are categorized according to syntactic labels. They are shown in Table III.

TABLE III.
NUMBER OF EXTRACTED PHRASE TRANSLATION PAIRS

	Number		Number
#BAP	10389	#BDP	219
#BMP	14177	#BNP	38111
#BNS	1180	#BNT	5744
#BVP	13923	#AP	4210
#ASIDE	131	#CO	224
#DP	9	#INP	617
#MP	1290	#NDE	762
#NP	28810	#NS	203
#NT	1774	#VP	56883
#VV	1333	#PFP	5820
#PP	11704	#SS	16404
#VBA	2370	#VBEI	506
#VC	6583	#VJ	3401
#VO	64923	#VOO	2785
#VSUO	107	#XP	0

In order to compare the performance of the proposed method in this paper, we collect 1000 bilingual sentence pairs as testing data from these 81204 Chinese-English ones. Two groups of experiments have been conducted on testing data. Chinese parser tool is used to analyze Chinese sentence and English parser tool is used to analyze English sentence in these 1000 bilingual sentence pairs. In Experiment 1, lexicon similarity SimL(c, e) is applied to align Chinese-English bilingual sentence pairs in testing data firstly, as described in section II. According to the lexicon-based word alignment results, English translations of Chinese syntactic phrases are obtained, and phrase translation pairs are gotten. In Experiment 2, multiple strategies of word alignment including lexicon similarity SimL(c, e), translation similarity SimT(c, e), semantic similarity SimS(c, e) and co-occurrence Asso(c, e) are applied to align Chinese words with English words step by step in bilingual sentence pairs of testing data firstly. Based on the results of multiple-strategy word alignment, English translations of Chinese syntactic phrases are obtained, and phrase translation pairs are gotten. Because the same Chinese parser has been applied to analyze Chinese sentences of bilingual sentence pairs in testing data, the same 4127 Chinese phrases have been acquired in Experiment 1 and Experiment 2. Because the different methods of phrase alignment have been used, we may extract different English translations for the same Chinese phrase in 2 groups of experiments. Then two human annotators will manually annotate these phrase translation pairs. If English phrase can interpret semantically Chinese phrase in a phrase translation pair, it is annotated as a positive instance. Otherwise it is viewed as a negative instance. We design accuracy to measure the quality of extracted phrase translation pairs. The computation of accuracy is shown in formula (6).

$$accuracy = \frac{N_P}{N_P + N_N} * 100\% . \quad (6)$$

Here, N_P is the number of positive instances and N_N is the number of negative instances. The accuracy of phrase translation pairs extracted in two groups of experiments is shown in Table IV.

TABLE IV.
THE ACCURACY OF PHRASE TRANSLATION PAIRS IN TWO EXPERIMENTS

	accuracy(%)
Experiment 1 SimL(c, e)	39.11%
Experiment 2 SimL(c, e)+SimT(c, e)+SimS(c, e)+Asso(c, e)	63.07%

From Table IV, we can find that when multiple strategies of word alignment including SimL(c, e), SimT(c, e), SimS(c, e) and Asso(c, e) are applied to align Chinese words and English words step by step in bilingual sentence pairs, and Chinese syntactic node is matched with English syntactic node based on the results of word alignment, the accuracy of phrase translation pairs extracted in Experiment 2 is 63.07%. But accuracy of phrase translation pairs extracted in Experiment 1 is only 39.11%. This is because that translation lexicon can only cover a little language phenomenon. We can find that there are no corresponding English phrases for lots of Chinese phrases in Experiment 1. After SimT(c, e), SimS(c, e) and Asso(c, e) are applied to the alignment process step by step, more Chinese words and English words will be aligned. So, we can find corresponding English translations for more Chinese phrases in testing data.

V. CONCLUSION

In this paper, we propose a new method to extract phrase translation pairs based on aligning Chinese words and English words with multiple strategies. Multiple strategies of word alignment including lexicon similarity, translation similarity, semantic similarity and co-occurrence are applied to align Chinese words and English words in bilingual sentence pairs. Chinese parser tool and English parser tool are used analyze respectively Chinese sentences and English sentences in bilingual corpus. An algorithm of matching Chinese syntactic nodes and English syntactic nodes are proposed to extract phrase translation pairs. Comparative experiments have been conducted. Experimental results show that when the new method is applied to phrase alignment process, the accuracy of extracted phrase translation pairs is best in two experiments.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China under Grant Nos. 60903082, 60975042, Chun-Hui Cooperated Project of the Ministry of Education of China under Grant Nos. S2009-1-15002, and Science and Technology Research Funds of Education Department in Heilongjiang Province under Grant Nos. 11541045. 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, UMIACS-TR-96-19, University of Maryland, 1996.

[4] Y. Xu, H. F. Wang, and X. Q. Lü, "Research of English-Chinese alignment at word granularity on parallel corpora," *Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science*, pp. 223–228, 2008.

[5] Y. Q. He, Y. Zhou and C. Q. Zong, "Word alignment based on multi-grain model," *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing*, pp. 269–272, 2008.

[6] D. Q. Zhu, and B. B. Chang, "Bootstrapping word alignment by automatically generated bilingual dictionary," *Proceedings of the 2008 International Conference on Natural Language Processing and Knowledge Engineering*, 2008.

[7] Y. Liu, Q. Liu and S. X. Lin, "Discriminative word alignment by linear modeling," *Computational Linguistics*, vol. 36, pp. 303–40, 2010.

[8] G. H. Zhang, "A discriminative model for Mongolian-English word alignment," *Proceedings of the 2010 Chinese Conference on Pattern Recognition*, pp. 316–320, 2010.

[9] S. Ananthakrishnan, R. Prasad, and P. Natarajan, "An unsupervised boosting technique for refining word alignment," *Proceedings of the 2010 IEEE Spoken Language Technology Workshop*, pp. 177–182, 2010.

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

[11] K. Imamura, "Hierarchical phrase alignment harmonized with parsing," *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pp. 377–384, 2001.

[12] 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.

[13] K. Yamamoto, Y. Matsumoto, "Acquisition of phrase-level bilingual correspondence using dependency structure," *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 933–939, 2000.

[14] 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.

[15] H. Watanabe, S. Kurohashi, and E. Aramaki, "Finding structural correspondences from bilingual parsed corpus for corpus-based translation," *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 906–912, 2000.

[16] 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.

[17] C. X. Zhang, S. Li, and T. J. Zhao, "Discriminative models for automatic acquisition of translation equivalences," *International Journal of Control, Automation, and Systems*, vol. 5, pp. 99–103, 2007.

[18] 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.

[19] 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.

[20] Y. J. Lu, S. Li , and T. J. Zhao, "Automatic extraction of translational equivalence based on bilingual corpora", *High Technology Letters*, vol. 13, pp. 19–24, 2003.

[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.

Xue-Yao Gao is Ph.D. and graduates from College of Computer Science and Technology, in Harbin University of Science and Technology. She is also a lecturer in Harbin University of Science and Technology. Her research interests are natural language processing and machine learning. She 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.

Da-Song Sun is an associate professor in Harbin University of Science and Technology. His research interests are natural language processing, and machine learning. He has authored and coauthored more than ten journal and conference papers in these areas.

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. He has authored and coauthored more than ten journal and conference papers in these areas.