http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (2): 371-375, 2010
ISSN 1812-5638

Extracting Translation Equivalences Automatically Based on Tree-String

[1]Chun-Xiang Zhang, [2]Ying-Hong Liang, [3]Peng Li, [4]Zhi-Mao Lu and [5]Yong Liu
[1]School of Software, Harbin University of Science and Technology, Harbin 150080, China
[2]School of Computer Engineering, Vocational University of Suzhou City, Suzhou 215104, China
[3]College of Computer Science and Technology,
Harbin University of Science and Technology, Harbin 150080, China
[4]College of Information and Communication Engineering,
Harbin Engineering University, Harbin 150001, China
[5]School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China

Abstract: In this study, we propose a new method to align Chinese-English bilingual sentence pairs based on parsing tree of Chinese sentence and word alignment result. From tree-string alignment, translation equivalences are extracted. Experimental results indicate that the new method achieves 80.97% at precision, which is better than parse-parse-match method. The performance of tree-string alignment is better than that of parse-parse-match on extracting translation equivalences.

Key words: Translation equivalence, parsing information, word alignment result

INTRODUCTION

Acquisition of translation equivalences, 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. A bilingual sentence pair contains a source language sentence and a target language sentence with same semantic meaning. Acquired equivalences can be used in a variety of applications such as bilingual lexicography (William and Church, 1991), machine translation system (Imamura, 2002) and cross-lingual information retrieval. As a reasonable step beyond the word-to-word alignment model, several researchers propose phrase alignment models. The advantage is that word context and local reordering are implicitly taken into account in phrase alignment models (Och and Ney, 2003). 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 (DeNero and Klein, 2008). Many methods have been proposed for acquisition of translation equivalences. Parse-parse-match method is adopted firstly to extract translation equivalences (Imamura, 2002). 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 monolingual parser is needed for either language and there is always grammar disagreement between source and target languages. Dan Melamed (1997) has proposed a fast and greedy algorithm called competitive linking in order to find word-to-word equivalences. Ying and Stephan (2005) built a two-dimensional matrix to represent a bilingual sentence pair where the value of each cell corresponds to point-wise mutual information between source word and target one. Box-shaped region whose mutual information value is similar is looked upon as a translation equivalence. Bing and Stephan (2005) used translation model to calculate phrase translation probabilities. Imamura and Sumita (2002) used translation literality to evaluate literality of bilingual sentence pairs and cleans the corpus for the purpose of improving equivalences' quality. Dekai (1997) proposes a bilingual language model to parse bilingual sentence pairs simultaneously, from which translation equivalences are extracted, which avoids the impact of inaccuracy of monolingual parser. But a suitable bilingual grammar is difficult to be found in practice. Zettlemoyer and Moore (2007) presented a technique for selecting translation equivalences to be included in translation tables based on their estimated quality according to a translation model. Chiang (2007) hypothesizes

Corresponding Author: Chun-Xiang Zhang, School of Software, Harbin University of Science and Technology, P.O. Box 98, No. 52, Xuefu Road, Nangang District, Harbin 150080, Heilongjiang Province, China
Tel: 13313643466

that incorrect reordering choices would often correspond to hierarchical phrases that violate syntactic boundaries in the source language and he explores the use of a constituent feature intended to reward the application of hierarchical phrases which respect source language syntactic categories. Johnson and Martin (2007) has presented a technique for pruning the phrase table in a phrase-based SMT system using Fisher's exact test. He computes the significance value of each phrase pair and prunes the table by deleting phrase pairs with significance values smaller than a threshold.

In this study, a new method in which translation equivalences are extracted from tree-string alignment of Chinese-English bilingual sentence pairs is presented, where tree-string alignment is built based on parsing tree of Chinese sentence and word alignment result. It is called as a tree-string alignment method. The new method decreases the influence of grammar disagreement between Chinese and English in alignment process. Experimental results indicate that the performance of new method is better than that of parse-parse-match method.

FRAMEWORK OF EXTRACTING TRANSLATION EQUIVALENCES

In parse-parse-match method, source language and target language will be respectively analyzed by parser. For Chinese and English, the alignment process will be restricted by grammar disagreement and lots of Chinese phrases can not be aligned to English phrases. So the number of extracted translation equivalences is very little, which leads more translation knowledge will be lost. A bilingual language model which parses bilingual sentence pairs simultaneously can eliminate the influence of grammar disagreement. After a bilingual sentence pair is parsed by a bilingual language model, translation equivalences will be gotten. But there is no parsing information in source and target part of translation equivalences. So, the extent to which such translation equivalences are applied is very small. In order to avoid the restriction of grammar disagreement and utilize more parsing knowledge, we will use parser to analyze Chinese sentence and its parsing tree is gotten. Then Chinese parsing tree and English sentence are aligned based on word alignment result. When the new method is used, we can get the tree-string alignment of a Chinese-English bilingual sentence pair. From tree-string alignment, we can extract translation equivalences. But there are only parsing information in source part of translation equivalences and target part does not include any parsing knowledge. The process of extracting Chinese-English translation equivalences from tree-string alignment is shown in Fig. 1.

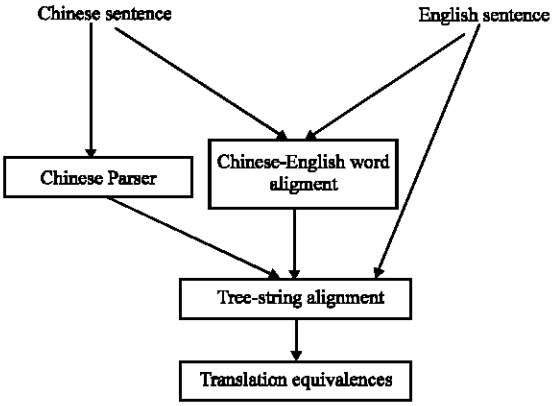

Fig. 1: Extracting Chinese-English translation equivalences from tree-string alignment

Only a Chinese parser and a Chinese–English word alignment tool are used here. Firstly, Chinese sentence is analyzed by Chinese parser. Secondly, we use word alignment tool to align Chinese sentence and English sentence. At last, the tree-string alignment between Chinese and English is built according to word alignment results, from which translation equivalences can be extracted.

EXTRACTING TRANSLATION EQUIVALENCES 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 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 Chinese-English sentence pair (C, E), the algorithm of building tree-string alignment is as follows:

- Align words between C and E by word alignment tool. Extract word links between C and E from word alignment result
- Parse Chinese sentence C and T is parsing tree of C
- The words in C and E are assigned with their positions, respectively. For example, 您$_1$ 能$_2$ 找$_3$ 开$_4$ 一$_5$ 张$_6$ 100$_7$ 元$_8$ 的$_9$ 钞票$_{10}$ 吗$_{11}$?$_{12}$ for Chinese sentence, as well as for English sentence
- Post-traveling parsing tree T, for every node n in T
  - 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
  - If n is a non-leaf node in T which is a Chinese phrase and node n has sons $m_1, m_2, \ldots, m_k$, triple sequence intervals of node n is expressed as $\{[SNODE(m_i)/STREE(m_i)/STC(m_i)] \mid I=1, 2,\ldots, k\}$

According to pre-defined heuristic rules, core node m_t is selected from $m_1, m_2, \ldots, m_k$, SNODE(n)=SNODE(m_t). Heuristic rules includes v+n->v, adj+n->n, adv+v->v and so on.

$$STREE(n) = [u,v]\ (u = \min(Left(STREE(m_1)), Left(STREE(m_2)),\ldots, Left(STREE(m_k))),\ v = \max(Right(STREE(m_1)), Right(STREE(m_2)),\ldots, Right(STREE(m_k))))$$

$$STC(n) = [u,v]\ (u = \min(Left(STC(m_1)), Left(STC(m_2)), \ldots, Left(STC(m_k))), v=\max(Right(STC(m_1)), Right(STC(m_2)), \ldots, Right(STC(m_k))))$$

When the algorithm is applied to Chinese-English bilingual sentence pairs, the tree-string alignment is gotten. From the tree-string alignment, we extract translation equivalences when Chinese parsing tree is post-traveled.

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

Chinese-English bilingual sentence pair:

- Chinese: 您 能 找 开 一 张 100 元 的 钞票 吗 ?
- English: Can you break a $ 100 bill?

Word alignment result: 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 result 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]

Tree-string alignment: The tree-string alignment of the bilingual sentence pair is shown in Fig. 2.

Extracted translation equivalences:

- 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

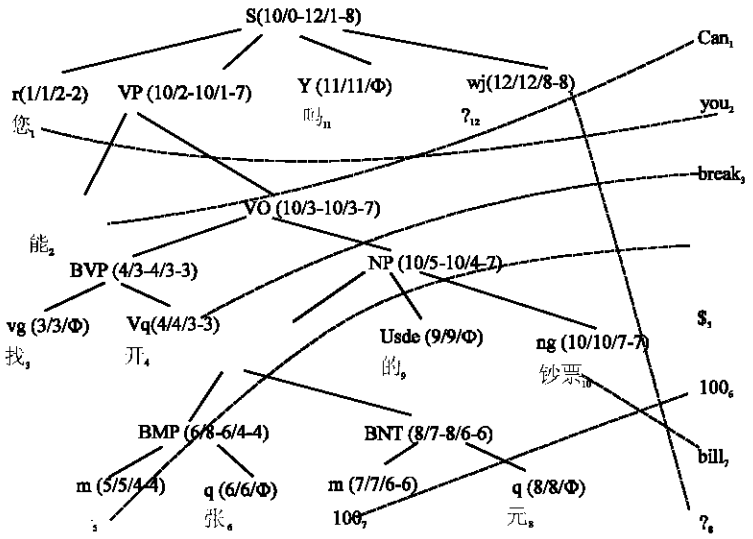

Fig. 2: Tree-string alignment between Chinese-English

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

EXPERIMENT

1000 Chinese-English bilingual sentence pairs are used for experiments, which are collected from traveling, news and traffic areas. Two groups of experiments are designed to evaluate the performance of the proposed method. Note that Exp1 applies the parse-parse-match method to align parsing tree of Chinese sentence and parsing tree of English sentence. Then translation equivalences are extracted from tree-tree alignment. Exp 2 employs the proposed method to align parsing tree of Chinese sentence and English sentence. Then translation equivalences are extracted from tree-string alignment.

Here word alignment tool, English parser and Chinese parser are used for experiments, which are developed by MOE-MS Key Laboratory of Natural Language Processing and Speech in Harbin Institute of Technology. Their performances are shown in Table 1.

For a Chinese-English bilingual sentence pair, Chinese parsing trees are the same in these two groups of experiments because the same Chinese parser is used. The English parts of extracted translation equivalences in Exp1 and Exp2 may be different. So, 3031 translation equivalences are, respectively extracted in Exp 1 and 2.

$Node(Ph_c)$->Ph_e is a translation equivalence, where Ph_c is Chinese phrase and Ph_e is English phrase. $Node(Ph_c)$ is Chinese parsing phrase. If Ph_e can interpret Ph_c semantically, $Node(Ph_c)$->Ph_e is annotated as a positive instance. Otherwise, it is viewed as a negative instance. We design accuracy to measure the performance of extracting translation equivalences. S is the set of positive translation equivalences and T is the set consisting of all extracted ones. It is shown in Eq. 1.

$$\text{Accuracy} = \frac{|S|}{|T|} \times 100\% \quad (1)$$

Four human annotators are asked to manually annotate these translation equivalences. They are divided into two teams. The first team manually annotate these 3031 translation equivalences extracted in Exp 1 and the second team are asked to manually annotate those 3031 translation equivalences extracted in Exp 2. Then crossing validation is conducted. We use accuracy as measure to evaluate respectively the performance of extracting translation equivalences in Exp 1 and 2. The evaluation results are shown in Table 2.

Table 1: The performance of word alignment tool and English parser and Chinese parser

Tool and parser	Precision (%)	Recall (%)
Word alignment tool	86	89
Chinese parser	78	79
English parser	77	80

Table 2: Accuracy of translation equivalences in Exp 1 and 2

Experiments	Accuracy (%)
1 (Parse-parse-match)	76.37
2 (Tree-string alignment)	80.97

From Table 2, we can find that the performance of tree-string alignment is better than that of parse-parse-match. The accuracy of translation equivalences extracted in Exp 2 will achieve 80.97%. This is because that the influence of grammar disagreement between Chinese and English will be decreased greatly in tree-string alignment method and more translation equivalences will be extracted from bilingual sentence pairs. When we check the results in these two experiments, we find that for the same Chinese parsing phrase, it has corresponding English translation in Exp 2. But it has no English translation in Exp 1. More translation knowledge will be acquired in Exp 2.

CONCLUSION

In this study, a new method called as tree-string alignment to extract translation equivalences from Chinese-English bilingual sentence pairs is presented. Tree-string alignment of bilingual sentence pairs is built based on parsing tree of Chinese sentence and word alignment result. The new method decreases the influence of grammar disagreement between Chinese and English in alignment process. More translation equivalences will be acquired and Chinese part of translation equivalences has parsing information. Experimental results indicate that the performance of the new method is better than that of parse-parse-match.

ACKNOWLEDGMENTS

This study is supported by National Natural Science Foundation of China under Grant Nos. 60903082, Science and Technology Research Funds of Education Department in Helongjiang 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 and D Center for Modern Information Technology Application in Enterprise(SX200907).

REFERENCES

Bing, Z. and V. Stephan, 2005. A generalized alignment-free phrase extraction. Proceedings of the ACL Workshop on Building and Using Parallel Texts, June 29-30, Association for Computational Linguistics, USA., pp: 141-144.

Chiang, D., 2007. Hierarchical phrase-based translation. Comput. Linguistics, 33: 201-228.

Dan Melamed, I., 1997. A word-to-word model of translational equivalence. Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, Madrid, Spain, July 7-12, Association for Computational Linguistics Morristown, NJ, USA., pp: 490-497.

De Nero, J. and D. Klein, 2008. The complexity of phrase alignment problems. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Columbus, Ohio, June 16-17, Association for Computational Linguistics, Morristown, NJ, USA., pp: 25-28.

Dekai, W., 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput. Linguistics, 23: 377-404.

Imamura, K., 2002. 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, March 13-17, Keihanna, Japan, pp: 74-84.

Imamura, K. and E. Sumita, 2002. Bilingual corpus cleaning focusing on translation literality. Proceeding of the 7th International Conference on Spoken Language Processing, (ICSLP'02), Kyotop, Japan, pp: 1713-1716.

Johnson, H. and J. Martin, 2007. Improving translation quality by discarding most of the phrasetable. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Canada, June 2007, Association for Computational Linguistics, USA., pp: 967-975.

Och, F.J. and H. Ney, 2003. A systematic comparison of various statistical alignment models. Comput. Linguistics, 29: 19-51.

William, A.G. and K.W. Church, 1991. Identifying word correspondences in parallel texts. Proceedings of the Workshop on Speech and Natural Language Pacific Grove, California, Feb. 19-22, Association for Computational Linguistics Morristown, NJ, USA., pp: 152-157.

Ying, Z. and V. Stephan, 2005. Competitive grouping in integrated phrase segmentation and alignment model. Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, June 2005, Association for Computational Linguistics, USA., pp: 159-162.

Zettlemoyer, L. and R. Moore, 2007. Selective phrase pair extraction for improved statistical machine translation. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York, April 22-27, Association for Computational Linguistics, Morristown, NJ, USA., pp: 209-212.