Approach to Automatic Translation Template Acquisition Based on Unannotated Bilingual Grammar Induction

Rile HU
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
rlhu@nlpr.ia.ac.cn

Chengqing ZONG
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
cqzong@nlpr.ia.ac.cn

Bo XU
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing 100080, China
xubo@nlpr.ia.ac.cn

Abstract

In this paper, we propose a new approach which can automatically acquire translation templates from the unannotated bilingual spoken language corpora in the domain of travel information accessing. In the approach, two basic algorithms named grammar induction algorithm and dynamic programming algorithm are adopted. Our approach is an unsupervised, statistical, data-driven method which avoids the parsing procedure. Firstly, the semantic groups and the phrasal structure groups are extracted from both the source language and the target language. Then, the dynamic programming algorithm is used to align these phrasal structure groups. The aligned phrasal structure groups are treated as the translation templates. And the preliminary experimental result is also desirable.

Keywords: Unannotated bilingual corpus, Translation template acquisition, Grammar induction, Automatic alignment, Machine translation

1 Introduction

Along with the development of the corpus technology, more and more bilingual corpora are available for knowledge acquisition in machine translation and many other natural language processing tasks. Translation template is a kind of especially useful knowledge for machine translation systems. In this paper, we aim at acquiring translation templates automatically on the sentence aligned parallel English-Chinese corpus.

In some EMBT systems, the translation templates are extracted manually from the corpus. [Kitano 1993] has adopted the manual encoding of the translation rules. [Sato 1995] has also proposed an example-based system, which took manually-built matching expressions as the translation templates. However, when the corpus is large, the task of template extraction using hands becomes more and more difficult and error-prone.

Some methods to automatically acquire the translation templates also have been proposed. In [Güvenir et al. 1998], [Oz et al. 1998], and [Cicekli et al. 2001], the analogical models are adopted to learn translation templates. By grouping the similar translation examples and replacing their difference with a variable, the methods could obtain translation templates from the bilingual parallel corpus. This kind of methods needs a very large scale of bilingual parallel corpora which contain a large amount of similar sentences. Some methods based on structure alignment are also proposed by some researchers to acquire structural translation templates [Kaji et al. 1992], [Watanabe et al. 2000], [Imamura 2001]. These approaches followed a procedure what may be called “parse-parse-match” [Wu 1997]. In these approaches each language of

the parallel corpus is first parsed individually using the monolingual grammar, and then the corresponding constituents are matched using some heuristic procedures. This kind of methods needs two parsers with high performance in both the source language and the target language. [Lü et al. 2001] has proposed a method based on bilingual language model. In this method, bilingual sentence pairs are first aligned in syntactic structure by combining the language parsing with a statistical bilingual language model. The alignment results are used to extract translation templates. This method also needs a high performance parser and POS tagging systems in both sides of the source and target languages.

In this paper, we propose a statistical, data-driven approach which can acquire translation templates from unannotated bilingual corpora based on bilingual grammar induction. Remainder of this paper is organized as follows. In Section 2, our motivations and the description of our translation templates acquisition system is introduced in detail. In Section 3, a brief overview of basic algorithms including grammar induction algorithm and the dynamic programming algorithm is introduced. In Section 4, the experimental results and analysis are shown. Finally, some concluding remarks are given in Section 5.

2 Our Motivations

Now, translation templates acquisition based on structural alignment is the popular method in the area of statistical machine translation (SMT). Many researchers have done a lot of work on this research topic. Unfortunately, as we mentioned above that these methods need at least a robust parser with high performance. The parser is hard to build especially in Chinese. We focus on finding out a method based on unsupervised machine learning.

The overview of our translation templates acquisition system is shown in Figure 1:

The input of the system is sentence aligned bilingual corpus. Here, we use an English-Chinese bilingual corpus. The Chinese sentences are firstly segmented, and then, the grammar induction step is taken on both of English and the Chinese sentences. After this step, we get the phrasal structures of the both languages. Finally, the phrasal structures of the languages are aligned using the dynamic programming algorithm. These aligned phrasal structures are output of as the translation templates.

In the following, a simple example is given to explain how the translation templates are acquired from the corpora.

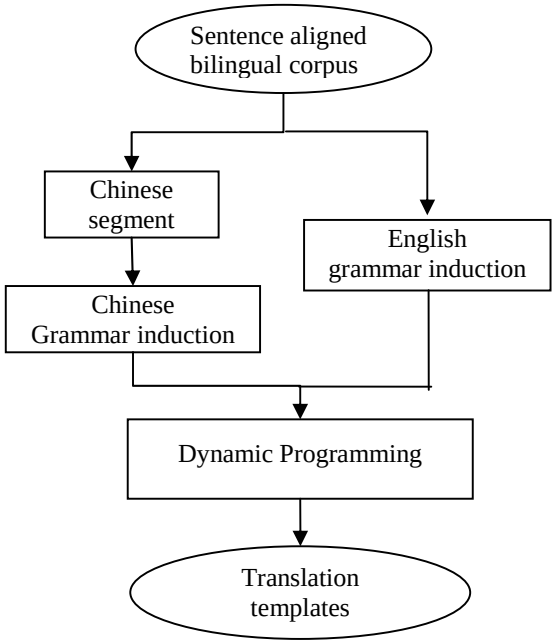

Figure 1: System Architecture

Suppose some SCi and PCi groups are gotten from the corpus, which are shown in Table 1:

Chinese part:
SCC10 → 单人间 ｜ 双人间｜ 标准间
PCC3 → 一 个
PCC8 → PCC3 SCC10
PCC12 → 我 想 预订
PCC20 → PCC12 PCC8
English part:
SCE5 → single | double | standard
PCE2 → want to
PCE4 → a SCE5 room
PCE8 →I PCE2 reserve
PCE14 → PCE8 PCE4

Table 1: Examples of the grammars acquired from the experimental corpus

With the grammars shown in Table 1, the system gets the translation template as follows:

我 想 预订 一 个 N. ⇔

I want to reserve a N^* room.

if N=单人间 then N^*=single;

if N=双人间 then N^*=double;

if N=标准间 then N^*=standard.

3 Basic Algorithms

In this section, a brief overview of basic algorithms including grammar induction algorithm and the dynamic programming algorithm is introduced.

3.1 Grammar Induction Algorithm

This method of clustering has two main steps, which are called spatially clustering and temporally clustering. In the spatially clustering step, the words which have similar contexts, including both side of left and right, are grouped together. These words generally have similar semantics. In the temporally clustering step, the words which have a high degree of co-occurrence are clustered into a group. These groups of words tend to be commonly-used phrases.

In spatially clustering, the Kullback-Leibler distance is used to describe the similarity of the distributions of the entity local contexts, which means the previous one word and the next word of the entity (1):

$$D(p_1 \| p_2) = \sum_{i=1}^{V} p_1(i) \log \frac{p_1(i)}{p_2(i)} \quad (1)$$

Where, p_1 denotes the unigram distribution of the words which appear in the local context of the entities e_1, p_2 denotes that of e_2, and V denotes the all the words which appear in the entity local contexts.

In order to acquire a symmetric distance measure, divergence is used as the measurement of the distance (2):

$$Div(p_1, p_2) = D(p_1 \| p_2) + D(p_2 \| p_1) \quad (2)$$

The distance of the two entities e_1 and e_2 is defined as it showed in (3):

$$\begin{aligned} Dist(e_1, e_2) = {} & Div(p_1^{left}, p_2^{left}) \\ & + Div(p_1^{right}, p_2^{right}) \end{aligned} \quad (3)$$

It is the sum of the divergences of the distributions of the words to the both left side and right side of the entities.

In order to make the clustering more accurate, we introduce the extended distance contexts into the measurement of the entities' distance.

We consider the words next to the context of the entities, which are called extended contexts. For the extended contexts, we calculate their Kullback-Leibler distance as the similarity of the distributions of them with Formula (1).

And we can also calculate the symmetric distance measure of these words using Formula (2).

Then, we compute the distance of the pairs of entities as the sum of the distance of the contexts and extended contexts. The distance of the entities e_1 and e_2 can be described as (4):

$$\begin{aligned} Dist^*(e_1, e_2) = {} & Div(p_1^{left}, p_2^{left}) \\ & + \tfrac{1}{2} Div_2(p_1^{left}, p_2^{left}) \\ & + Div(p_1^{right}, p_2^{right}) \\ & + \tfrac{1}{2} Div_2(p_1^{right}, p_2^{right}) \end{aligned} \quad (4)$$

Where, the expression $Div_2(p_1, p_2)$ denotes the symmetric distance of the extended contexts of the two entities e_1 and e_2.

The most similar pairs of entities are grouped into a semantic group which is labeled as SCi. In another word, the pairs of entities which have the minimum of distance shown in (4) are clustered together.

After the spatially clustering, the words that have been grouped are substituted as the group labels in the corpus. Then the temporally clustering is taken into the clustering process.

In temporally clustering, the cohesion degree (CD) is used to describe the degree of the co-occurrence of the two entities e_1 and e_2. This is the metric for the clustering process. The cohesion degree is defined in (5):

$$\begin{aligned} CD(e_1, e_2) = {} & \alpha \times \frac{\sum_{i=1}^{V} Ngram(i)}{n} \\ & + \beta \times MI(e_1, e_2) \end{aligned} \quad (5)$$

where,

$$Ngram(i) = \theta \times p(e_2 \mid e_1) + (1-\theta)\left[p(e_2 \mid e_0, e_1) + p(e_3 \mid e_1, e_2)\right] \quad (6)$$

and V denotes all the probable word sequences (e_0,e_1,e_2,e_3) in the corpus, n denotes the number of these sequences. The parameters α β θ in (6) are estimated based on experience. The different corpus may have different parameters.

And the MI in (5) is the Mutual Information shown in (7):

$$MI(e_1, e_2) = P(e_1, e_2) \log \frac{P(e_2 \mid e_1)}{P(e_2)} \quad (7)$$

The entities which have the highest cohesion degree are clustered into the phrasal groups labeled as PCi. These entity pairs are substituted with their PC labels. And then the process turns to another iteration of spatial clustering.

The grammar induction approach described in this section can capture the semantic and phrasal structures from the unannotated corpus. After the clustering algorithm, the semantic groups and the phrasal structure groups can be extracted from the corpus.

3.2 Dynamic Programming Algorithm

Let the source language sentence be $S: ws_1, ws_2, \cdots, ws_{NS}$, the target language sentence be $T: wt_1, wt_2, \cdots, wt_{NT}$. The sequence of the aligned phrases of the source and target languages is $Q = \{< bs_k, bt_k >, k \in [0, K]\}$, K is the number of these aligned phrasal groups. f(n) denotes the estimated minimum cost of the path which starts from the source node *s*, passes through the node *n*, and finally reaches the target node *t*. f(n) contains two parts named as g(n) and h(n). g(n) denotes the estimated minimum cost of the path from source node *s* to the node *n*. And h(n) denotes the estimated minimum cost of the path from node *n* to the target node *t*.

In the procedure of the dynamic programming algorithm, we define the cost of the path as it shown in (8):

$$g(k) = \sum_k -\log p(bt_k \mid bs_k) \quad (8)$$

So, the goal of the searching is to find out the aligned phrasal structure which makes the g(k) minimum. It is shown in (9):

$$\min[\sum_{k=1}^{K} -\log p(bt_k \mid bs_k)]$$
$$S = bs_1 \cdots bs_k ; T = ts_1 \cdots ts_k \quad (9)$$

Based on the Bayes' theorem, we can get (10):

$$p(bt_k \mid bs_k) = \frac{p(bt_k) \cdot p(bs_k \mid bt_k)}{p(bs_k)} \quad (10)$$

Where, $p(bs_k)$ and $p(bt_k)$ can be calculated from the Ngram models of the source language and the target language. In this paper, we use Bigram models shown in (11):

$$p(bs_k) = \prod_{j=1}^{m_k} p(ws_j \mid ws_{j-1}); \quad p(bt_k) = \prod_{i=1}^{l_k} p(wt_i \mid wt_{i-1}) \quad (11)$$

Where, m and l denotes the length of the phrasal groups of the source language and the target language.

$p(bs_k \mid bt_k)$ can be calculated by translation model. Here, we use IBM Model one (12):

$$p(bs_k \mid bt_k) = p(l_k \mid m_k) \cdot \prod_{j=1}^{m_k} \sum_{i=1}^{l_k} p(ws_j \mid wt_i) \quad (12)$$

In Equation (12), $p(ws_j \mid wt_i)$ is the direct translation probability from the source language to the target language, which is trained by the EM algorithm. And $p(l_k \mid m_k)$ is the length probability, which is estimated approximately by Poisson Distribution.

According to the Equation (10) (11) (12) and (8), we get the estimated cost of the path from the beginning node *s* to the current node *k* (13):

$$g(k) = \sum_k \{-\sum_{i=1}^{l_k} \log p(wt_i \mid wt_{i-1}) - \log p(l_k \mid m_k) - \sum_{j=1}^{m_k} \log \sum_{i=1}^{l_k} p(ws_j \mid wt_i) + \sum_{j=1}^{m_k} \log p(ws_j \mid ws_{j-1})]\} \quad (13)$$

The minimum cost of the path from current node *k* to the target node *t* is defined as (14):

$$h(k) = -\log p(bt_{rest} \mid bs_{rest}) = -\log \frac{\prod_{i=k+1}^{NT} p(wt_i \mid wt_{i-1}) \times \prod_{j=k+1}^{NS} \sum_{i=k+1}^{NT} p(ws_j \mid wt_i)}{\prod_{j=k+1}^{NS} p(ws_j \mid ws_{j-1})} \quad (14)$$

Where, NS and NT denote the length of the source language sentence and the target language sentence.

So, the estimated minimum cost of the path which starts from the source node *s*, passes through the node *n*, and finally reaches the target node *t* can be defined as (15):

$$f(k) = g(k) + h(k) = -[\sum_{k} \log p(bt_k \mid bs_k) + \log p(bt_{rest} \mid bs_{rest})] \quad (15)$$

The description of the dynamic programming algorithm is shown in Table2:

Step 1: Initialization. OPEN := (*s*), g(*s*) := 0;

Step 2: IF OPEN = ( ) THEN EXIT (FAIL);

Step 3: *n* := FIRST(OPEN);

Step 4: IF reach the end of the sentence, THEN EXIT(SUCCESS);

Step 5: REMOVE (*n*, OPEN), ADD(*n*,CLOSED);

Step 6: EXPAND $(n) \rightarrow \{m_l\}$ Search the next possible phrasal group in the source language, and search all the possible corresponding target language phrasal groups. Then calculate the cost ;

$$f(n, m_l) = g(n, m_l) + h(m_l)$$

Step 7:ADD$(m_l, OPEN)$,and mark the pointer from m_l to n ;

Step 8: If there are more than one paths can reach one public node, only keep the path which has the minimum cost, and delete others. Then rank the nodes in OPEN form the minimum to the maximum by their cost f(n).

Step 9: GOTO Step 2.

Table 2: The description of the dynamic programming algorithm

Using this dynamic programming algorithm, we align the phrasal structures of the source language and the target language, which are acquired from the grammar induction algorithm.

4 Experiment and Discussion

4.1 Experimental Corpus

The corpus used in our experiment is collected in the domain of travel information accessing, which consists of 2,950 bilingual parallel utterances. The Chinese vocabulary size of the corpus is 989. The English vocabulary size of the corpus is 1074.

4.2 Experimental results

We define the accuracy of the results as:

$$Acc = \frac{Nr}{N} \times 100\% \quad (16)$$

Where, *Nr* denotes the number of correct translation templates manually judged, and *N* denotes the total number of the translation templates acquired by the algorithm. The parameters α β θ are set as 0.5,1 and 0.7.

The test result is shown in Table 3:

Nr	*N*	*Acc(%)*
352	476	73.95

Table 3: the Experimental result

4.3 Analysis of the Experimental Results

There are two kinds of errors in the experimental results. The first is the errors occurring in the step of grammar induction. This is because that this algorithm does not use the information contained in the corpus adequately. So, some irrelevant entities are clustered into one group. The second is the errors occurring in the step of dynamic programming. This is due to the approximate parameter estimation and the special translation of some idioms.

Anyway, we are planning to use a synonym dictionary to reduce the first kind of errors and introduce dictionary information and do some pre-processing work to reduce the second kind of errors.

5 Conclusion

In this paper, we present an approach that

automatically acquires the translation templates from the unannotated bilingual parallel corpus. Our grammar induction algorithm extracts the grammar of semantic and phrasal structures of both source language and target language from the corpus. Based on the grammars, the phrasal structures are aligned by dynamic programming algorithm. The aligned structures are treated as the translation templates. The results of the preliminary experiment show that our approach is desirable though we are facing many difficult problems including improvement of the quality of the grammar induction and alignment.

Anyway in the next step, we will use a synonym dictionary in the process of clustering algorithm and introduce more linguistic information into our approach.

Acknowledgments

The research work described in this paper is supported by the National Natural Science Foundation of China under the grant No.60175012 and 60121302, and also supported by the High-Tech Program (the 863 Program) under the grant 2002AA117010, the outstanding overseas Chinese Scholars Fund of Chinese Academy of Sciences under grant No. 2003-1-1, and also PRA project under grant No. SI02-05.

References

Chin-Chung Wong, Helen Meng and Kai-Chung Siu. 2001. Learning Strategies In A Grammar Induction Framework. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pp 153-157.

Dekai Wu, 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. In Computational Linguistics, vol.23, No.3, pp. 377-403.

F. Och. 2002. Statistical Machine Translation: From Single Word Models to Alignment Templates. Ph.D. thesis, RWTH Aachen, Germany.

H. Altay Güvenir, and Ilyas Cicekli, 1998. Learning Translation Templates from Examples, Information Systems, Vol. 23, No. 6, pp. 353-363.

H. Kaji, Y. kida, and Y. Morimoto. 1992. Learning Translation Templates from Bilingual Texts. In Proceedings of the 14th International Conference on Computational Linguistics, pp 672-678.

H. Kitano. 1993. A Comprehensive and Practical Model of Memory-based Machine Translation. In 13. IJCAI. Chambery, France.

Helen M. Meng and Kai-Chung Siu. 2002. Semi-Automatic Acquisition of Domain-Specific Semantic Structures, IEEE Transactions on Knowledge and Data Engineering, vol 14, n 1, January/February, pp 172-180.

H. Watanabe, S. Kurohashi, and E. Aramaki. 2000. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In Proceedings of the 18th International Conference on Computational Linguistics, pp 906-912.

Ilyas Cicekli and Halil Altay Guvenir, 2001. Learning translation Templates from Bilingual Translation Exmples. In Applied Intelligence, Vol. 15, No. 1 pp. 57-76.

K. Goodman and H. Nirenburg. 1992. KBMT-89: A Case Study in Knowledge Based Machine Translation. Morgan Kafmann.

K. Imamura. 2001. Hierarchical Phrase Alignment Harmonized with Parsing. In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pp 377-384.

Rile Hu, Chengqing Zong and Bo Xu, 2003. Semiautomatic Acquisition of Translation Templates from Monolingual Unannotated Corpora. In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering , October,Beijng, pp 163-167.

Satoshi Sato. 1991. Example-Based Translation Approach, In Proc. of ATR Workshop, July.

Satoshi Sato. 1995. MBT2: a method for combining fragments of examples in example-based translation. Artificial Intelligence, 75: 31-50.

Wei Cheng, 2003. Research on the Statistical Approach of Chinese-English Spoken-Language Translation in a Limited-Domain. Ph.D. thesis. National Laboratory of Pattern Recognition. Institute of Automation, Chinese Academic of Sciences, China.

Yajuan Lü, Ming Zhou, Sheng Li, Changning Huang and Tiejun Zhao. 2001. Automatic Translation Template Acquisition Based on Bilingual Structure Alignment. Computational Linguistics and Chinese Language Processing. Vol.6, No.1, February, pp. 83-108

Yeyi Wang, 1998. Grammar Inference and Statistical Machine Translation. Ph.D. thesis. Carnegie Mellon University, USA.

Zeynep Oz, and Ilyas Cicekli, 1998. Ordering Translation Templates by Assigning Confidence Factors, in: Proceedings of AMTA'98-Conference of the Association for Machine Translation in the Americas, Lecture Notes in Computer Science 1529, Springer Verlag, October, Langhorne, PA, USA, pp:51-61.

Zeynep Orhan, 1998. Confidence Factor Assignment to Translation Templates, M.S. Thesis, Bilkent University, September.