Learning Chinese Bracketing Knowledge Based on a Bilingual Language Model

Yajuan Lü, Sheng Li, Tiejun Zhao, Muyun Yang
School of Computer Science & Engineering, Harbin Institute of Technology
Harbin, China, 150001
Email: {lyj,lish,tjzhao,ymy}@mtlab.hit.edu.cn

Abstract

This paper proposes a new method for automatic acquisition of Chinese bracketing knowledge from English-Chinese sentence-aligned bilingual corpora. Bilingual sentence pairs are first aligned in syntactic structure by combining English parse trees with a statistical bilingual language model. Chinese bracketing knowledge is then extracted automatically. The preliminary experiments show automatically learned knowledge accords well with manually annotated brackets. The proposed method is particularly useful to acquire bracketing knowledge for a less studied language that lacks tools and resources found in a second language more studied. Although this paper discusses experiments with Chinese and English, the method is also applicable to other language pairs.

Introduction

The past few years have seen a great success in automatic acquisition of monolingual parsing knowledge and grammars. The availability of large tagged and syntactically bracketed corpora, such as Penn Tree bank, makes it possible to extract syntactic structure and grammar rules automatically (Marcus 1993). Substantial improvements have been made to parse western language such as English, and many powerful models have been proposed (Brill 1993, Collins 1997). However, very limited progress has been achieved in Chinese.

Knowledge acquisition is a bottleneck for real appication of Chinese parsing. While some methods have been proposed to learn syntactic knowledge from annotated Chinese corpus, most of the methods depended on the annotated or partial annotated data(Zhou 1997, Streiter 2000). Due to the limited availbility of Chinese annotated corpus, tests of these methods are still small in scale. Although some institutions and universities currently are engaged in building Chinese tree bank, no large scale annotated corpus has been published until now because the complexity in Chinese syntatic sturcture and the difficulty in corpus annotation (Chen 1996).

This paper proposes a novel method to facilitate the Chinese tree bank construction. Based on English-Chinese bilingual corpora and better English parsing, this method obtains Chinese bracketing information automatically via a bilingual model and word alignment results. The main idea of the method is that we may acquire knowledge for a language lacking a rich collection of resources and tools from a second language that is full of them.

The rest of this paper is organized as follows : In the next section, a bilingual language model is introduced. Then, a bilingual parsing method supervised by English parsing is proposed in section 2. Based on the bilingual parsing, Chinese bracketing knowlege is extracted in section 3. The evaluation and discussion are given in section 4. We conclude with discussion of future work.

1 A bilingual language model – ITG

Wu (1997) has proposed a bilingual language model called Inversion Transduction Grammar (ITG), which can be used to parse bilingual sentence pairs simultaneously. We will give a brief description here. For details please refer to (Wu 1995, Wu 1997).

The Inversion Transduction Grammar is a bilingual context-free grammar that generates two matched output languages (referred to as L_1

and L_2). It also differs from standard context-free grammars in that the ITG allows right-hand side production in two directions: straight or inverted. The following examples are two ITG productions:

$$C \text{ -> } [A\ B]$$
$$C \text{ -> } <A\ B>$$

Each nonterminal symbol stands for a pair of matched strings. For example, the nonterminal *A* stands for the string-pair (A_1, A_2). A_1 is a sub-string in L_1, and A_2 is A_1's corresponding translation in L_2. Similarly, (B_1, B_2) denotes the string-pair generated by *B*. The operator [ ] performs the usual concatenation, so that *C* -> [*A B*] yields the string-pair (C_1, C_2), where $C_1=A_1B_1$ and $C_2=A_2B_2$. On the other hand, the operator <> performs the straight concatenation for language 1 but the reversing concatenation for language 2, so that *C* -> <*A B*> yields $C_1=A_1B_1$, but $C_2=B_2A_2$. The inverted concatenation operator permits the extra flexibility needed to accommodate many kinds of word-order variation between source and target languages (Wu 1995).

There are also lexical productions of the following form in ITG:

$$A \text{ -> } x/y$$

This means that a symbol *x* in language L_1 is translated by the symbol *y* in language L_2. *x* or *y* may be a null symbol *e*, which means there may be no counterpart string on other side of the bitext.

ITG based parsing matches constituents for an input sentence-pair. For example, Figure 1 shows an ITG parsing tree for an English-Chinese sentence-pair. The inverted production is indicated by a horizontal line in the parsing tree. The English text is read in the usual depth-first left to right order, but for the Chinese text, a horizontal line means the right sub-tree is traversed before the left. The generated parsing results are:

(1) [[[Mr. Wu]$_{BNP}$ [[plays basketball]$_{VP}$ [on Sunday]$_{PP}$]$_{VP}$]$_S$.]$_S$
(2) [[[吴 先生] [星期天 [打 篮球]]] 。]

We can also represent the common structure of the two sentences more clearly and compactly with the aid of <> notation:

(3) [[<Mr./先生 Wu/吴>$_{BNP}$ < [plays/打 basketball/篮球]$_{VP}$ [on/e Sunday/星期天]$_{PP}$ >$_{VP}$]$_S$ ./。]$_S$

where the horizontal line from Figure 1 corresponds to the <> level of bracketing.

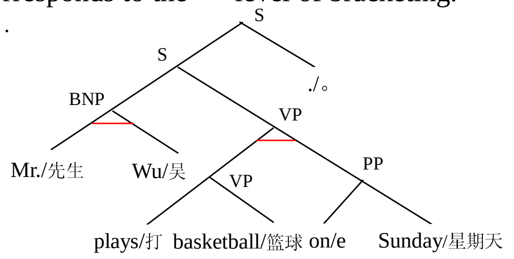


Figure 1 Inversion transduction Grammar parsing

Any ITG can be converted to a normal form, where all productions are either lexical productions or binary-fanout nonterminal productions(Wu 1997). If probability is associated with each production, the ITG is called the Stochastic Inversion Transduction Grammar (SITG).

2 English parsing supervised bilingual bracketing

Because of the difficulty in finding a suitable bilingual syntactic grammar for Chinese and English, a practical ITG is the generic Bracketing Inversion Transduction Grammar (BTG)(Wu 1995). BTG is a simplified ITG that has only one nonterminal and does not use any syntactic grammar. A Statistical BTG (SBTG) grammar is as follows:

$$A \xrightarrow{a} [AA]; \quad A \xrightarrow{a} <AA>; \quad A \xrightarrow{b_{ij}} u_i / v_j\ ;$$
$$A \xrightarrow{b_{ie}} u_i / e; \quad A \xrightarrow{b_{ej}} e / v_j$$

SBTG employs only one nonterminal symbol *A* that can be used recursively. Here, "*a*" denotes the probability of syntactic rules. However, since those constituent categories are not differentiated in BTG, it has no practical effect here and can be set to an arbitrary constant. The remaining productions are all lexical. b_{ij} is the translation probability that source word u_i translates into target word v_j. b_{ij} can be obtained using a statistical word-translation model (Melamed 2000) or word alignment(Lü 2001a). The last two productions denote that the word in one language has no counterpart on other side of the bitext. A small constant can be chosen for the probabilities b_{ie} and b_{ej}.

In BTG, no language specific syntactic

grammar is used. The maximum-likelihood parser selects the parse tree that best satisfies the combined lexical translation preferences, as expressed by the b_{ij} probabilities. Because the expressiveness characteristics of ITG naturally constrain the space of possible matching in a highly appropriate fashion, BTG achieves encouraging results for bilingual bracketing using a word-translation lexicon alone (Wu 1997).

Since no syntactic knowledge is used in SBTG, output grammaticality can not be guaranteed. In particular, if the corresponding constituents appear in the same order in both languages, both straight and inverted, then lexical matching does not provide the discriminative leverage needed to identify the sub-constituent boundaries. For example, consider an English-Chinese sentence pair:

(4) English: That old teacher is our adviser.
Chinese: 那个老教师是我们的顾问。

Using SBTG, the bilingual bracketing result is :

(5) [[[[[[The/那个 old/老] teacher/教师] is/是] our/我们的] adviser/顾问] ./。]

The result is not consistent with the expected syntactic structure. In this case, grammatical information about one or both of the languages can be very helpful. For example, if we know the English parsing result shown in (6), then the bilingual bracketing can be determined easily; the result should be (7).

(6) [[That old teacher]$_{BNP}$ [is [our adviser]$_{BNP}$]$_{VP}$.]$_{S}$
(7) [[That/那个 old/老 teacher/教师] [is/是 [our/我们的 adviser/顾问]] ./。]

From the example, we can see that if one language parser is available, the induced bilingual bracketing result would be more accurate. English parsing methods have been well studied and many powerful models have been proposed. It will be helpful to make use of English parsing results. In the following, we will propose a method of bilingual bracketing supervised by English parsing.

Here, English parsing supervised BTG means using an English parser's bracketing information as a boundary restriction in the BTG language model. But this does not necessitate parsing Chinese completely according to the same parsing boundary of English. If the English parsing structure is totally fixed, it is possible that the structure is not linguistically valid for Chinese under the formalism of Inversion Transduction Grammar. To illustrate this, see the example shown in Figure 2.

If you want to lose weight, you had better **eat less bread** .
如果 你 想 减轻 体重，最好 **少 吃 面包** 。

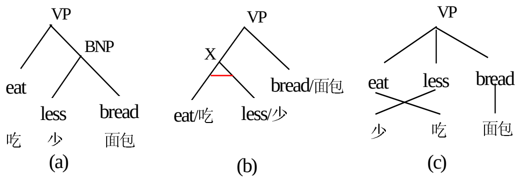


Figure 2 A example of mismatch subtree

The sub-tree for blacked underlined part of English and corresponding Chinese are shown in Figure 2(a). We can see that the Chinese constituents do not match the English counterparts in the English structure. In this case, our solution is that: the whole English constituent of "VP" is aligned with the whole Chinese correspondence; i.e., "eat less bread" is matched with "少吃面包" shown in Figure 2(b). At the same time, we give the inner structure matching according to ITG regardless of the English parsing constraint. An "X" tag is introduced to indicate that the sub-bilingual-parsing-tree is not consistent with the given English sub-tree. Our result can also be understood as a flattened bilingual parsing tree as shown in Figure 2(c). This means that when the bilingual constituents couldn't match in the small syntactic structure, we will match them in a larger structure.

The main idea is that the given English parser is only used as a boundary constraint for bilingual parsing. When the constraint is incompatible with the bilingual model ITG, we use ITG as the default result. This process enables parsing to go on regardless of some failures in matching.

We heuristically define a constraint function $F_e(s, t)$ to denote the English boundary constraint, where s is the beginning position and t is the end. There are three cases of structure matching: *violate match, exact match* and *inside match. Violate match* means the bilingual parsing conflicts with the given English bracketing boundary. For example, given the following English bracketing result (8), (1,2), (1,3), (2,3), (2,4) etc. are *Violate match*es. We assign a minimum $F_e(s, t)$ (0.0001 at present) to prevent the structure match from being chosen when an alternative match is available. *Exact match* means the match falls exactly on the English parsing boundary, and we assign a high $F_e(s, t)$ value (10 at present) to emphasize it. (1,6), (2,5), (3,5) are examples. (3,4), (4,5) are examples of *inside match,* and the value 1 is assigned to these $F_e(s, t)$ functions.

(8) [She/1 [is/2 [a/3 lovely/4 girl/5]] ./6]

Let the input English and Chinese sentences be $e_1,...e_T$ and $c_1,...c_V$. As an abbreviation we write $e_{s...t}$ for the sequence of words $e_{s+1}, e_{s+2,} ...e_t$, and similarly write $c_{u...v}$. The local optimization function $\delta(s,t,u,v) =$ $\max P[e_{s..t} / c_{u..v}]$ denotes the maximum probability of sub-parsing-tree of node *q* and that both the sub-string $e_{s...t}$ and $c_{u...v}$ derive from node *q*. Thus, the best parser has the probability $\delta(0,T,0,V)$. $\delta(s,t,u,v)$ is calculated as the maximum probability combination of all possible sub-tree combinations(Wu 1995). To insert English parsing constraints in bilingual parsing, we integrate the constraint function $F_e(s, t)$ into the local optimization function. Computation of the local optimization function is then modified as given below:

$$\delta(s,t,u,v) = \max[\delta^{[]}(s,t,u,v), \delta^{<>}(s,t,u,v)],$$
$$\delta^{[]}(s,t,u,v) = \max_{\substack{s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \ne 0}} F_e(s,t)\delta(s,S,u,U)\delta(S,t,U,v),$$
$$\delta^{<>}(s,t,u,v) = \max_{\substack{s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \ne 0}} F_e(s,t)\delta(s,S,U,v)\delta(S,t,u,U).$$

Initialization is as follows :

$$\delta_{t-1,t,v-1,v} = b(e_t / c_v), \qquad 1 \le t \le T, \ \ 1 \le v \le \mathrm{V}$$
$$\delta_{t-1,t,v,v} = b(e_t / e), \qquad 1 \le t \le T, \ \ 1 \le v \le \mathrm{V}$$
$$\delta_{t,t,v-1,v} = b(e / c_v), \qquad 1 \le t \le T, \ \ 1 \le v \le \mathrm{V}$$

where, T ,V is the length of English and Chinese sentence respectively. $b(e_t / c_v)$ is the probability of translating English word e_t into Chinese word c_v . A minimal probability can be assigned to empty word alignment $b(e_t / e)$ and $b(e / c_v)$.

The optimal bilingual parsing tree for a given sentence-pair can be computed using dynamic programming (DP) algorithm(Wu 1997). Using the standard SBTG local optimization fuction, the obtained bilingual parsing result for the given sentence-pair(4) is shown as example (5); when using the above modified local optimization function, the parsing result is that shown as example (7). Comparing the two results, we can see that by intergrating English parsing constraints into BTG, the bilingual parsing becomes more grammatical. Our experiments showed that this English parsing supervised BTG would improve the accuracy of bilingual bracketing by nearly 20% (Lü 2001b).

The obtained bilingual parsing tree is in the normal form of ITG, that is each node in the tree is either a lexical node or a binary-fanout nonterminal node. We can combine the subtree to restore the fanout flexibility using the production characters [[*AA*]*A*]=[*A*[*AA*]]=[*AAA*] and <<*AA*>*A*>= <*A*<*AA*>>=<*AAA*>. The combining operation could not cross the given English parisng boundary.

3 Chinese bracketing knowledge extraction

Table 1 shows some bilingual bracketing examples obtained using the above method. To understand easily, we give the tree form of the first example in Figure 3(a). The leaf node is the aligned words of the two languages and their POS tag categories. These POS tags are generated from an English and a Chinese POS tagger respectively. The English POS tag and phrase tag set are the same as those of the Penn Tree Bank (Marcus 1993) and the Chinse POS tag set please refer to the web site: http://mtlab.hit.edu.cn. The nonterminal node are labeled using English sub-tree tags.

Based on the bilingual parsing result, it is easy to extract the Chinese bracketing structure according to the Inversion Transduction Grammar. For the normal node, the Chinese text is traversed in depth-first left to right order, but for an inverted node (indicated by a horizontal line in the parsing tree or indicated by a <> notation in bracketing expression), the right sub-tree is traversed before the left. Thus, the Chinese parsing tree corresponding to Figure 3(a) is shown in Figure 3(b). The nonterminal labels are derived from the English sub-tree. The extracted Chinese bracketing results from Table1

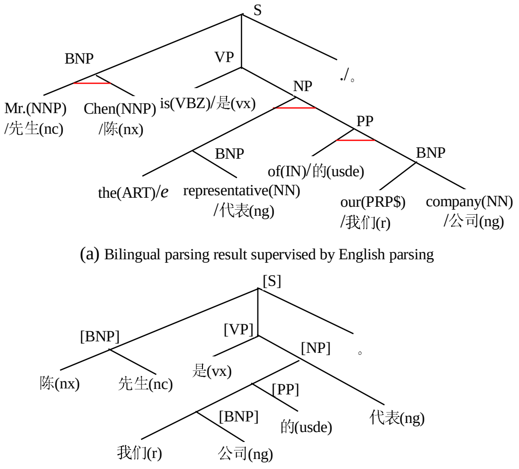
Table 1 Bilingual bracketing examples

1. [<Mr.(NNP)/先生(nc) Chen(NNP)/陈(nx) >$_{BNP}$ [is (VBZ) /是(vx) < [the(ART)/e representative(NN)/代表(ng)]$_{BNP}$ <of (IN) /的(usde) [our (PRP$)/我们(r) company(NN)/公司(ng)]$_{BNP}$ >$_{PP}$ >$_{NP}$]$_{VP}$.(.)/。(wj)]$_{S}$
2. [Spring(NN)/春天(t) [is(VBZ)/是(vx) <[the(ART)/e first(JJ)/第一(m) e/个(q) season(NN)/季节(ng)]$_{BNP}$ <in(IN)/里(f) [a(ART)/一(m) year(NN)/年(q)]$_{BNP}$ >$_{PP}$ >$_{X}$]$_{VP}$.(.)/。(wj)]$_{S}$
3. [[The(ART)/e window(NN)/窗子(ng)]$_{BNP}$ [is/e/VBZ <[e/更(d) narrower(JJR)/狭窄(a)] [than(IN)/比(p) [the(ART)/e door(NN)/门(ng)]$_{BNP}$]$_{PP}$ >$_{ADJP}$]$_{VP}$.(.)/。(wj)]$_{S}$
4. [<[The(ART)/e policeman(NN)/警察(ng)]$_{BNP}$ [who(WP)/e [reported(VBD)/报告(vg) [the(ART)/这(r) e/一(m) accident(NN)/事故(ng)]$_{BNP}$]$_{VP}$ e/的(usde)]$_{SBAR}$ >$_{NP}$ [thinks(VBZ)/认为(vg) [it(PRP)/那(r) [was(VBD)/是(vx) [Tom(NNP)/汤姆(ny) 's(PRP$)/的(usde) fault(NN)/错(ng)]$_{BNP}$]$_{VP}$]$_{S}$]$_{VP}$.(.)/。(wj)]$_{S}$
5. [[The(ART)/e Beijing(NNP)/北京(nd) zoo(NN)/动物园(ng)]$_{BNP}$ [is(VBZ)/是(vx) <[the(ART)/e largest(JJS)/最大(a) e/的(usde) zoo(NN)/动物园(ng)]$_{BNP}$ [I(PRP)/我(r) [e/所(ussu) have(VBP)/e ever(RB)/e visited(VBN)/参观(vg) e/过(ut) e/的(usde)]$_{VBP}$]$_{S}$ >$_{NP}$]$_{VP}$.(.)/。(wj)]$_{S}$

Table 2 The extracted Chinese bracketing results corresponding to Table 1

1. [[陈/nx 先生/nc]$_{BNP}$ [是/vx [[[我们/r 公司/ng]$_{BNP}$ 的/usde]$_{PP}$ 代表/ng]$_{NP}$]$_{VP}$ 。/wj]$_{S}$
2. [春天/t [是/vx [[一/m 年/q]$_{BNP}$ 里/f]$_{PP}$ [第一/m 个/q 季节/ng]$_{BNP}$]$_{VP}$ 。/wj]$_{S}$
3. [窗子/ng [[比/p 门/ng]$_{PP}$ 更/d 狭窄/a]$_{VP}$ 。/wj]$_{S}$
4. [[[[报告/vg [这/r 一/m 事故/ng]$_{BNP}$]$_{VP}$ 的/usde]$_{SBAR}$ 警察/nc]$_{NP}$ [认为/vg [那/r [是/vx [汤姆/ny 的/usde 错/ng]$_{BNP}$]$_{VP}$]$_{S}$]$_{VP}$ 。/wj]$_{S}$
5. [[北京/nd 动物园/ng]$_{BNP}$ [是/vx [[我/r [所/ussu 参观/vg 过/ut 的/usde]$_{VBP}$]$_{S}$ [最大/a 的/usde 动物园/ng]$_{BNP}$]$_{NP}$]$_{VP}$ 。/wj]$_{S}$

are listed in Table 2.



(a) Bilingual parsing result supervised by English parsing

(b) The Chinese parsing result extracted from (a)

Figure 3 Extract Chinese Bracketing structure from Bilingual Parsing

It can be seen from Table 2 that the automatic acquired bracketing results reflect the Chinese structure well though some English phrase tags are not suitable to label the corresponding Chinese phrase directly. For example, in Table 2, the English tags "PP (preposition phrase)" in sentence 1 and "SBAR(clause)" in sentence 4 are incorrectly tag the corresponding Chinese structure. We don't care about the phrase tags here. Our main concern is the bracketing boundary of the syntactic structure. The bracketing boundary knowledge has been proved to be valuable for Chinese grammar induction (Zhou 1997). The advantage of our method is that the bracketing knowledge is acquired from bilingual corpus automatically. It reduces the manual labour for corpus tagging, which are time-consuming and error-prone.

4 Evaluation and discussion

To evaluate the quality of the acquired Chinese bracketing boundaries, we compared them with the parsing annotation based on an existed Chinese syntax annotation scheme. Detail of the Chinese syntax annotation scheme and a annotated corpus can be download from the website http://mtlab.hit. edu.cn.

The test set consisted of 3,000 English-Chinese bilingual sentence-pairs that come from the machine translation evaluation corpus(Duan 1996). The average length is 9.1 words for English sentences and 12.6 Chinese characters for Chinese sentences. The test sentence pairs were first aligned at the word level based on statistics and lexicon with a accuracy of nearly 90%(Lü 2001a). The English and Chinese sentences were parsed based on the Penn Tree bank tag set and the Chinese syntax annotation scheme respectively. Both the English and the Chinese parsing results were manually corrected. The corrected Chinese parsing results are used as the standard test set.

We acquired Chinese bracketing results using the proposed method. The previous defined *exact match* , *violate match,* and *inside match* are used to evaluate the accordance between acquired bracketing result and the standard parsing result. Here, *exact match* means the acquired structure are the same as the standard structure; *violate match* means the acquired structure conflict with the standard structure. Otherwise, the acquired structure is called a *inside match.* In example (9), A is the standard bracketing result, B is the acquired bracketing result and C demonstrates the classification of the acquired structures. The structure of whole sentence are not participate in evaluation. *Exact match rate(EMR), violate match rate(VMR),* and *inside match rate(IMR)* denote the ratio of three types of bracketing numbers in all bracketing numbers respectively.

(9) A. [白色 、 红色 和 蓝色]$_{\text{NP}}$ 是 [[[很多 女孩]$_{\text{BNP}}$ [所/
(white red and blue are many girls)
喜欢]$_{\text{VSUO}}$]$_{\text{SS}}$ 的 [[三 种]$_{\text{BMP}}$ 颜色]$_{\text{BNP}}$]$_{\text{NP}}$ 。
(like three colors)
(In English : White , red and blue are the three colors which many girls like .)

B. [[白色 、 红色 和 蓝色]$_{\text{BNP}}$ [是 [[[[很多 女孩 所]$_{\text{BNP}}$ 喜欢]$_{\text{S}}$ 的]$_{\text{SBAR}}$ [三 种 颜色]$_{\text{BNP}}$]$_{\text{NP}}$]$_{\text{VP}}$ 。]$_{\text{S}}$

C. a) *exact match*：[白色、红色和蓝色]；[三种颜色]；[很多女孩所喜欢]; [很多女孩所喜欢的三种颜色]

b) *violate match*：[很多女孩所]；

c) *inside match*：[很多女孩所喜欢的]；[是很多女孩所喜欢的三种颜色]；

Table 3 gives the evaluation result. The evaluation results for acquired Chinese structure corresponding to six main English phrases (BNP, NP, VP, ADJP, ADVP and PP) are also given in detail.

From the results we can see that only a fraction of the learned structures are *violate match*(14.03%), most of them are *exact match* (55.46%). In addition, there are also many *inside match.* These *inside match*es occured due to the difference standard in phrase merging between Penn Tree bank and the standard Chinese annotation scheme. The English phrase structure are labeled with more details. While for Chinese, the main phrase in the level of sentence are not merged futher. For example, the verb and object in sentence level are not combined. That is why most of the verb phrases(VP) are inside match (53.28%). The bracketing boundary of inside match can be either right or wrong. We checked the correctness of inside match manually and got a average accuray of 79.37%. Then the accuracy of all acquired structure bracketing is 79.68% (EMR＋IMR × Accuracy of IM).

The *violate match*es acquired in bilingual parsing are mainly due to the empty word alignments. Such as in the special strucures " 把 ..." and " 被 ..." in Chinese. The word " 把 "and" 被 " has no counterpart word in English.They are usually merged with the neighboring noun word as shown in example (10)，thus lead to a violate match. It is neccessary to build special patterns to handle these structures. Word alignment errors also produce *violate match*es in bilingual bracketing.

(10) [[把/p 他的/r 画/ng] [委托/vg [给/vg [我们/r 照顾/vg]]]。/wj]

The Chinese bracketing accuracy obtained using our method is comparable to that of the example-based Chinese parser(77%) (Streiter 2000), but it is lower than that of the PCFG Chinese parser(84%)(Yao1998). However, unlike these two parser, our method needn't any

Table 3 Evaluation on acquired Chinese bracketing results

Type	Bracket number	EMR	IMR	VMR	Accuracy of IM	Accuracy
all	10119	55.46%	30.51%	14.03%	79.37%	79.68%
BNP	2533	71.69%	18.95%	9.36%	49.58%	81.09%
NP	675	65.63%	20.30%	14.07%	70.80%	80.00%
VP	1676	28.46%	53.28%	18.26%	92.72%	77.86%
ADJP	192	60.94%	26.04%	13.02%	86.00%	83.33%
ADVP	120	45.83%	38.33%	15.83%	76.09%	74.99%
PP	1198	52.92%	27.46%	19.62%	82.67%	75.62%

Chinese annotated training corpus, which is difficult to accumulate. Another advantage of our method is that the Chinese bracketing result is derived based on English parsing and parallel corpus, which make it particularly benefit for research on the corresponding relationship between Chinese and English phrase. In (Lü 2001b), we used bilingual bracketing result for automatic translation templates acquisition, which turns out to be very useful for structure transfer in machine translation. In addition, the acquired bracketing corpus can be applied to many Chinese NLP tasks. It can be used as the foundation for further Chinese treebank annotation, which will save human labour in a great deal. It can also be used to improve the efficiency and accuracy in Chinese grammar induction (Zhou 1997). Grammar rules can also be extracted from the bracketing corpus. For example, we can obtain the following BNP rules from the acquired bracketing results in Table 2:

BNP->nx+nc;   BNP->r+ng;   BNP->m+q;
BNP->m+q+ng; BNP->r+m+ng; BNP->ny+usde+ng;
BNP->nd+ng;   BNP->a+usde+ng;

Conclusion

In this paper, we have presented a method to learn Chinese syntactic structure from English parsing based on a bilingual language model. The method creates structure bracketing Chinese corpora automatically by taking full advantage of English parsing and bilingual corpora. The created corpora are very useful for further Chinese corpus annotation and parsing knowledge acquisition. Primary experiment proved the feasibility and validity of the method. Although this paper is related to Chinese and English, the method is also applicable to other language pairs. Obviously, if the concerned languages come from same language family, such as English and French, the method would be more effective.

Acknowledgements

This research was funded by High Technology Research and Development Program of China (2001AA114101). We also would like to thank the Institute of Computational Linguistics at Peking University for providing bilingual corpora for test.

References

Dekai Wu (1995) *An algorithm for simultaneously bracketing parallel texts by aligning words.* Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics, pp. 244-251.

Dekai Wu (1997). *Stochastic inversion transduction grammars and bilingual parsing of parallel corpora*. Computational Linguistics, 23(3), pp. 377-403

E. Brill (1993) *Transformation-based error driven parsing.* Proceedings of International Workshop on Parsing Technologies.

Huiming Duan and Shiwen Yu (1996). *Report for machine translation evaluation*. Computer World, 1996.3:183 (in Chinese)

I. Dan Melamed (2000). *Models of Translational Equivalence among words.* Computational Linguistics 26(2), pp. 221-249

Keh-Jiann Chen (1996). *A model for robust Chinese parser*. Computational Linguistics and Chinese Language Processing. 1(1), pp.183-204

Marcus M. P., Marcinkiewicz M. A. and Santorini B (1993). *Building a large annotated corpus of English: the Penn Treebank*. Computational Linguistics, 19(2), pp. 313-330

Michael Collins (1997).*Three generative, lexicalised models for statistical parsing.* Proceedings of the 35th Annual Meeting of the ACL, Madrid

Qiang Zhou and Changning Huang. A Chinese syntactic parser based on bracket matching principle. Communication of COLIPS, 1997, 7(2), pp.53-59

Streiter O. and Chen K.J. (2000). *Experiments in example-based parsing.* In Dialogue 2000, International Seminar in Computational Linguistics and Applications, Tarusa, Russia.

Yajuan Lü, Tiejun Zhao, Sheng Li and Muyun Yang. (2001a) *English-Chinese word alignment based on statistic and lexicon*. Proceedings of 6th Joint Symposium of Computational Linguistics, TaiYuan, China, pp. 108-115. (in Chinese)

Yajuan Lü, Ming Zhou, Sheng Li, Changning Huang, Tiejun Zhao (2001b). *Automatic translation template acquisition based on bilingual structure alignment*. International Journal of Computational Linguistics and Chinese Language Processing. 6(1), pp. 1-26.

Yuan Yao and Kim Teng Lua. (1998). *A Probabilistic Context-Free Grammar Parser for Chinese*. Computer Processing of Oriental Language,11(4), pp.393-407