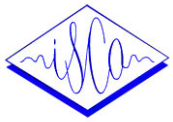

International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

Block Analysis of Bilingual Corpus for Chinese-English Statistical Machine Translation[1]

Hairong XIA, Bo XU, Taiyi HUANG
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing
{hrxia, xubo, huang}@nlpr.ia.ac.cn

Abstract

In this paper, we describe a bilingual corpus processing strategy, block analysis, from a new point of view. By this analysis strategy, we want to extract more information from bilingual corpus for future statistical machine translation. At first, we define some block types and give some statistical data from a Chinese-English bilingual corpus under this framework. Then a block-based alignment algorithm is presented, by which we can extract and align the corresponding bilingual blocks automatically. Some experimental results show that block analysis is practical and more informative than any other word-based approach.

1. Introduction

Researchers have taken notice of the value of corpus for natural language processing, so a new branch named corpus linguistics was born. As the acquisition of large corpus is much easy now, people have present a lot of processing strategies, such as Chinese sentence segmenting, part-of-speech tagging, syntax parsing, and so on. All these works are the first stages in making use of corpus.

People are also interesting in bilingual corpus, since bilingual parallel corpus can provide more information for many fields, such as lexicon compilation and machine translation. As for bilingual corpus, a key issue is to align bilingual texts at different levels. In this paper, we discuss a novel analysis approach, i.e. block analysis, and give a block-based alignment algorithm to achieve this goal.

The rest of this paper is organized as follows. In section 2, a review of corpus processing is given. After this review, we bring out our novel strategy in section 3. Then a block-based alignment algorithm is introduced. Finally, some experimental data are shown.

2. Review of conventional corpus processing

Generally, a goal of natural language processing is to parse a sentence automatically and exactly, or to understand it by some means. Because corpus consists of a large number of sentences, plentiful syntactic or semantic knowledge are contained. Corpus-based approach has been proved effective for NLP. Before we can make use of a corpus, a lot of processing work is indispensable, such as segmenting words for Chinese, tagging part-of-speech, parsing syntax and so on.

In the research of new generation of MT technology, such as statistical machine translation, parallel bilingual corpus is much more useful. We need to extract more information or train model parameters from bilingual corpus. For bilingual corpus, some researches are particular.

(1) Sentence aligning: This is the first step to make use of bilingual corpus. Many papers have been proposed for this problem. The techniques fall into two main classes: lexical and statistical. According to some papers, good results can be acquired (Brown, etc, 1993; Chen, 1993).

(2) Word aligning: Word pairs are the most direct information we can extract from bilingual corpus. For a language pair such as Chinese and English, there exist many differences in word order, phrase order, and other aspects, so it seems no satisfactory result reported now (Brown etc. 1990; Vogel 1996).

(3) Bilingual clustering: Many approaches, such as mutual information, are used widely to cluster words in NLP. However, for a bilingual application, such approaches would have many biases and generate a lot of language-special result. To solve this problem, a bilingual clustering approach was proposed by Ries (Ries, 1995).

There are also some other interesting aspects, due to limitation of space, we would pass over now. Such research works are definitely important for further processing, but they often focus on single words or sentences, so relationship between words are ignored, especially in IBM word-based alignment. We believe that bilingual analysis at structure level is more informative than at word level, since languages are always structural.

3. Block analysis of Chinese and English

Our machine translation task is to translate Chinese

[1] The research work described in this paper was supported by the National Key Fundamental Research Program (the 973 Program) of China under the grant G1998030504, the National Natural Science Foundation of China under the grant 69835030 and also the National 863 Hi-Tech Program under the grant 863-306-ZT03-02-2

spoken language to English in hotel preservation domain, and we hope this new approach can improve the performance of statistical MT. For research, we build an experimental bilingual corpus by collecting some English sentences from Microsoft SAPI technology document and translating them to Chinese. The advantage of this corpus is its diversity and complexity of sentence structure, for example, most English sentences have a clause or participle structure. Following research and examples are all based on this corpus.

3.1 Concept of Block analysis

Before introducing the concept of block analysis, we observe a sentence pair now:

English: [The interface syntax in this book] [follows] [the variable-naming convention known as Hungarian notation], [in which] [variables] [are prefixed] [with lower-case letters] [to indicate] [their data type].

Chinese: [本书中的接口语法][遵从][称作匈牙利命名法的变量命名约定],[即][变量名][用小写字母][作前缀][来表示][它们的数据类型].

Clearly, those words bracketed in square brackets can be aligned tersely with a counterpart in the other sentence; in other words, we can map Chinese and English better at the level of "blocks" rather than words. If we look at a number of sentence pairs, we find a fact: given a Chinese sentence and its translation English sentence, then

1) Chinese sentence can be segmented to some phrases according to syntactic function, and the words in English sentence can also form phrases;
2) Even if the Chinese phrases may move around when aligning, their English corresponding words tend to stay together.

This fact is the base of our block analysis. Now we can give the definition of block here: a block, which may be equal to a syntactic phrase, but not always, is a sequence of words in a certain language. As for the sentence pair above, we can bracket some continuous words to a single block, and some other words to another block. Some similar techniques are proposed recently (Steven 1996), this strategy is different to them in the following two aspects: (1) A block here may not be a syntactic component; (2) The boundaries of a block are restricted by two sentences.

3.2 Block types

We define six basic block types. Although these types are summarized from our restricted corpus, we find they are also suitable for analyzing unrestricted texts. These types include:

(1) Noun block

A noun block is often equal to a noun phrase, because noun phrases often function as the subject or object of a sentence or a preposition. Recently, there are some papers proposed for extracting max-length NP or base NP (Zhou 1998), since the acquisition of NPs is very important for NLP. Here, under the framework of block analysis, we can extract corresponding NPs from bilingual corpus more easily, because two languages are more informative than a single language. The main difference of noun block between Chinese and English is the order of modifiers and headword.

In section 4, we will present a novel algorithm to acquire max-length NPs as noun blocks.

(2) Verb block

Not like syntax analysis, block analysis is only partial and shallow. Here verb block, like verb chunk (Steven, 1996), only consists of continuous verbs including all modals, auxiliaries, medial adverbs, head verb and other words such as "not", etc.

Some considerations are taken:

1) Some intransitive verbs are followed by a preposition, and they denote a sole mean, so they cannot be separated when bracketing. For example:

   Chinese: [小看][某人]

   English: [look down] [sb.]

2) Some other intransitive verbs and prepositions followed are separable. For example:

   Chinese: 由[某物]组成

   English: consist [of sth.]

3) Main verb and its auxiliaries are not always continuous, because there may have other components between them. For example:

   Chinese: 然后应用程序*就能*通过一个端口*发送*音频数据

   English: The application *can* then *send* audio data though a port.

(3) Preposition block

Preposition block consists of a preposition and a noun block. Note that not all prepositions can lead a block, such as those followed by a present participle.

1) Preposition block may function as an attribute of a noun. Under this circumstance, it should be incorporated into a noun block. For example:

   Chinese: *桌子上*的那本书

   English: the book *on the desk*

2) The block may serve as an adverbial modifier, in this instance it form an independent block. For example:

   Chinese: 对象*通过键盘*接收用户的输入

   English: the object receives user's input *though a keyboard*

3) The block may act as a complement of a verb, please refer to verb block section.

(4) Fixed block

Fixed block is flexible enough to contain word sequences with various sizes, which is similar to Example-based strategy. Idioms, custom usage, fixed combination, translation history can be viewed as a fixed block. Fixed block is able to deal with those words that cannot be aligned well at all. For example:

Chinese: 太好了

English: It is great

Recognition of fixed block depends on a machine-readable dictionary, and this dictionary is maintained as system runs.

(5) Conjunction block

This category includes all conjunctions, such as “when”, “but” and so on. These words often lead another sentence, so they are effective flags of block boundary. But we need to analysis coordinators like “and, or, etc.” in English (“以及, 或者…” in Chinese) firstly when recognizing blocks, because these words always connect two peer to peer components.

(6) Adjective block

Only those used as predicative adjectives can form an adjective block.

4. Block-based alignment

Under the framework above, an issue is how to recognize blocks and align them from a bilingual corpus. To solve this problem, we present a block-based alignment approach.

Formally, given a Chinese sentence $c_1^J = c_1...c_j...c_J$ and its English translation $e_1^I = e_1...e_i...e_I$, block alignment can be described as follows:

$$E = (e_{11}e_{12}...e_{1l_1})(e_{21}e_{22}...e_{2l_2})...(e_{n1}e_{n2}...e_{nl_n}) = E_0E_1...E_n$$

$$C = (c_{11}c_{12}...c_{1l_1})(c_{21}c_{22}...c_{2l_2})...(c_{m1}c_{m2}...c_{ml_m}) = C_0C_1...C_m$$

Our goal is to find a series of match pairs like <*Ei*, *Cj*>. A similar structure method (Wang, 1998) allows German phrase discontinuous, but this method may add the cost of alignment. Here we take a different flexible way. Words in a Chinese block can be continuous by reordering words and adjusting the granularity of block. For example:

English: …[they] [must leave] [pause] [between words]…

Chinese: …[他们][必须][在单词中间][留出][暂停]…

To ensure the continuity of words in a block, the verb block “[must leave]” will be split to “[must]” and “[leave]”. On the other hand, the structure of noun block is severely different in Chinese and English, but for ease of implement, we recognize the head-part and modifiers separately and then to combine them into a final noun block. To solve these two problems, combining and splitting operations are needed.

Three terms are introduced here:

Sensitive words: these words in a sentence often indicate the beginning of another block. For example, a preposition can determine a boundary between it and its frontal words, generally a verb or a noun phrase. We define the words with the following part-of-speech as sensitive words in English sentence: article, numeral, verb, pronoun, preposition, conjunction and interjection. Numeral, verb, preposition, conjunction, interjection, direction-position word and auxiliary are signs for Chinese sentences.

Alignment likelihood: given an English block and a Chinese candidate, we use this criterion to evaluate the probability of the alignment. This criterion is inducted from IBM translation model. Without loss of generality, let an English block be $E = e_1...e_i...e_I$ and a Chinese candidate $C = c_1...c_j...c_J$, the alignment likelihood is defined as:

$$T(E:C) = \sum_{i=1}^{I} T_i(e_i:C) = \sum_{i=1}^{I} \underset{j=1}{\overset{J}{Max}}\, t_{ij}(e_i:c_j) \qquad (1)$$

$$t_{ij}(e_i:c_j) = \begin{cases} 0 & \text{if (e,f) not in lexicon} \\ \dfrac{1}{N_e} & \text{if (e,f) in lexicon} \end{cases}$$

where N_e is the number of Chinese words list as translations of the English word e_i. Equation 3 can be seen as the sum of the contribution of whole Chinese candidate to each English word e_i.

Relative position offset:

Given a English word sequence $...e_{j-2}e_{j-1}e_je_{j+1}e_{j+2}...e_{j+r}...$, and a Chinese sequence $...c_{i-s}..c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}...$, suppose that e_j is aligned with c_i, e_{j+1} with c_{i-2}, e_{j+2} with c_{i-1}, and e_r with c_s, determine a baseline alignment, for example (e_j, c_i), then we define average position offset simply as ratio of their geometry centers:

$$rpo = \frac{1+2+...+r}{(-1)+(-2)+...+(-s)} \qquad (2)$$

With discussion above, we can have a procedure as follows:

Input: a sentence pair *c* and *e* tagged with part-of-speech, formally like $c_1^J = c_1/t_1...c_j/t_j...c_J/t_J$ and $e_1^I = e_1/t_1...e_i/t_i...e_I/t_I$.

Output: block serials and alignment between them, $A = \{k : a_1...a_i...a_k\}$, $a_i =< c_{i1}, l_{ci}, e_{i1}, l_{ei} >$, where k is the total number of alignment matches, a_i is an alignment match pair, subscript $i1$ is the beginning position of the block, l_{ci} and l_{ei} denote the length of the Chinese block and English block, respectively.

***Block-based alignment procedure description**:*

Step 1 Preprocessing the sentence pair

Tag sensitive words, recognize all fixed blocks, and reorder Chinese words (if necessary). Then separate the sentences to some spans.

Step 2 Phrasing and pre-aligning

1) For English, use a series of finite-state transducers to recognize basic blocks that would be block candidates.

For Chinese, finite-state transducers are also employed to generate some simple blocks.

2) For each certain English block, try to find its counterpart by computing their translation likelihood

$t(e:c)$ with Chinese block candidates. We choose the candidate with highest score and then record an alignment pair. Here block style is an important heuristic information.

Step 3 Combining and splitting

Scan all alignment pairs, if condition 1 is fulfilled, do combining operation; or if condition 2 is fulfilled, do splitting operation:

Condition 1: syntactic restriction exists between two adjoining blocks, and the average position offset is not more than a threshold.

Condition 2: two blocks (in Chinese) correspond to part words of a block in English separately, and vice versa.

To implement this algorithm, we need an electric dictionary which list detailed lexical information.

5. Experimental result and discussion

According to the discussion above, we tagged about 617 sentence pairs manually, some statistical data are showed in table 1.

Table 1: Statistics of the corpus

	Chinese	English
Total words	9118	9440
Vocabulary	1367	1237
Total blocks	3764	3604
Average number of block per sentence	6	6

As we can see, total blocks of Chinese and English are almost equal, that is, some differences at word level have been hidden. In addition, we align the sentence pairs manually (see Table 2), we find that most of sentences can be tidily matched. In these pairs, blocks can be aligned one to one according to syntax function or meaning. There are also some blocks that cannot find a counterpart, because block analysis cannot eliminate the differences between English and Chinese completely.

Table 2: Tagging result

	Chinese	English
Aligned blocks	3252	3252
Unaligned blocks	512 (13.6%)	352(9.8%)

As a contrast, we implement the algorithm mentioned above, and test it on the same texts, the effect is listed in table 4. The criterion *ac* denotes the accuracy of recognized Chinese blocks, and *ae* for English blocks. Another criterion *pa* describes the accuracy of alignment pairs found by the algorithm. The value *ac* is less than *ae* because English structure is more distinct than Chinese, and the finite-state transducers run better. Precision of alignment is satisfactory, which means that most of blocks acquired from corpus can be aligned accurately.

Table 4: The effect of alignment algorithm

pc	79.8%
pe	89.4%
pa	86%

6. Conclusion

In this paper, we presented a block analysis strategy for bilingual corpus processing. The approach is based on the fact the two languages are structural and can be mapped better at structure level. Basic ideas are to phrase words with finite-state transducers and align blocks according to alignment likelihood. Through these two operations, bilingual corpus can be aligned much more tidily. Future work involves how to improve the alignment performance and to mine useful knowledge from aligned blocks for statistical machine translation.

7. Reference

1. P. F. Brown, Cocke, J., V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "A Statistical Approach to Machine Translation", in *Computational Linguistic*, Vol. 16, No. 2, pp. 79-85, 1990.
2. S. Vogel, H. Ney, and C. Tillmann, "HMM-Based Word Alignment in Statistical Translation", in *Proc. Of the International Conference on Computational Linguistics*, pp. 836-841, Copenhagen, Denmark, August 1996.
3. Ries, Klaus, Finn Dag Bu ϕ , and Ye-Yi Wang. "Improved Language Modelling by Unsupervised Acquisition of Structure". In *ICASSP '95. IEEE*.
4. Ye-Yi Wang and Alex Waibel. "Modeling with Structure in Statistical Machine Translation". In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistic*, 1998.
5. Ker, S. J. and Chang, J. S., "A Class-Based Approach to Word Alignment", Computational Liguistic, 27(2): pp. 313-343, 1997.
6. Brown. P. F., Lai. J.C. and Mercer. R. L., "Aligning sentences in Parallel Corpora", in *Proc. of the 29th Annual Meeting of the ACL*, 1993.
7. Chen. S. F., "Aligning Sentence in Bilingual Corpora Using Lexical Information", in *Proc. of the 31th Annual Meeting of the ACL*, 1993.
8. Steven Abney, "Partial Parsing via Finite-State Cascades", in *Proc of the ESSLLI'96 Robust parsing Workshop*.
9. Qiang Zhou. "Automatic recognition of Chinese max-length NP", Chinese Journal of Software, 1998.