Bilingual Chunk Alignment Based on Interactional Matching and Probabilistic Latent Semantic Indexing

Feifan Liu, Qianli Jin, Jun Zhao, Bo Xu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing P.O. Box 2728, 100080
{ffliu,qljin,jzhao,bxu}@nlpr.ia.ac.cn

Abstract. An integrated method for bilingual chunk partition and alignment, called "Interactional Matching", is proposed in this paper. Different from former works, our method tries to get as necessary information as possible from the bilingual corpora themselves, and through bilingual constraint it can automatically build one-to-one chunk-pairs associated with the chunk-pair confidence coefficients. Also, our method partitions bilingual sentences entirely into chunks with no fragments left, different from collocation extracting methods. Furthermore, with the technology of Probabilistic Latent Semantic Indexing(PLSI), this method can deal with not only compositional chunks, but also non-compositional ones. The experiments show that, for overall process (including partition and alignment), our method can obtain 85% precision with 57% recall for the written language chunk-pairs and 78% precision with 53% recall for the spoken language chunk-pairs.

Keywords: Bilingual Chunking, Alignment, Interactional Matching, PLSI

1 Introduction

The knowledge granularity in natural language processing has four levels: text, sentence, chunk, and word. In these levels, chunk has its specific priority. As a kind of knowledge granularity between "sentence" and "word", "chunk" is much more unambiguous than "word", and more recurrent than "sentence". So, chunk is quite suitable for us to use in machine translation[10], word sense disambiguation, cross-lingual information retrieval system[11] and *etc.*

The popular alignment methods for chunk-level are commonly based on parsing technology[7] or length-based algorithm. But, it has been proven that these methods have poor performance when dealing with long sentences[2]. The more serious problem is that these methods can only deal with compositional chunk, such as "natural language —— 自然/natural 语言/language", while they have no way to get non-compositional chunk-pairs, like "rain cats and dogs —— 倾盆大雨". Furthermore, some methods([1],[3],[4]) can only extract chunks partly, and discard many fragments.

In this paper, an integrated method named "Interactional Matching" for Bilingual Chunk Alignment is proposed. Our aim is to get much enough information for recognizing one-to-one chunk-pairs from untagged corpora themselves, while using as less additional manually assignment as possible. The method involves two key technologies, namely Interactional Matching and Probabilistic Latent Semantic Indexing (PLSI). The former is used to conduct cross-lingual constraint in order to integrate the whole process including partition and alignment. And the latter is used to conduct chunk matching both for compositional chunk pairs and for non-compositional ones. Our method can output one-to-one chunk-pairs, together with their confidence coefficient, or called matching probabilities.

In this paper, we use Interactional Matching algorithm to deal with English-Chinese sentence-parallel corpora. Also, it's easy to extend this method to other bilingual pairs. We test this method in two types of corpora (English-Chinese Spoken Language and Written Language). The experiments show that our method can get overall 85% precision with 57% recall for written language chunk-pairs and 78% precision with 53% recall for spoken language chunk-pairs.

2 Related Work

Most of previous work of bilingual chunking and alignment are based on complex syntax information or focus on special kinds of phrases, such as V+NP, *etc.* Generally speaking, there are four representative methods, Xtract[1], Parsing([2],[7],[13],[14]), LocalMax([3],[4]), and Crossing Constraint[5]. Their characteristics are list in table 1, and we will analyze these four methods in detail.

Table 1. Comparison between different methods

Algorithm	Xtract	Parsing	Local MAX	Crossing Constraint
Bilingual	N	Y	N	Y
Non-Compositional chunks	N	N	N	N
Having fragment	Y	N	Y	Y
Grammatical-based	N	Y	N	Y
Probabilistic-based	Y	N	Y	Y

(1) Frank Smadja (1993) proposed a probabi-listic method call Xtract, which is the first mature algorithm of extracting phrases from corpora. He defined two key measures for word sequence, cohesion degree and variance degree. Using this method, based on some central words, phrases and templates can be extracted from the single language corpora. However, he did not express how to deal with bilingual alignment.

(2) Syntax-parsing-based method is popular for structure alignment. Dekai Wu (1997) put forward inversion transduction grammar (ITG) for bilingual parsing. The stochastic ITG brings bilingual constraints to bear upon problematic corpus analysis tasks such as segmentation, bracketing, phrasal alignment and paring. This method got very good results for short and regular sentences. However, it is difficult to write a broad bilingual grammar to deal with long sentences in written lan-guage and irregular syntax in spoken language. Sun Le (2000) and Watanabe (2000) proposed methods based on parsing, bilingual lexicon and heuristic information, whereas these two methods can't treat with non-compositional chunks and can't also partition sentences completely.

(3) LocalMax was developed by Silva (1999), based on the Xtract method. He adopted two new association measures called Symmetric Conditional Probability and Mutual Expectation for extraction multiword lexical units. This idea produced a better result than Xtract, but still many fragments were discarded, and it can only deal with Monolingual corpora.

(4) Wei Wang and Ming Zhou (2001) proposed an integrated algorithm to do structure alignment, in which parsing and alignment are conducted together, and got good precision. However, this method needs quite a lot of prior knowledge, such as tree bank and word alignment information, *etc.* These knowledge, especially word alignment information, is very difficult to get.

In summary, the above methods have difficulties in complete chunking and alignment. In order to break the limitation, we propose an integrated algorithm of partition and alignment for the untagged bilingual corpora. Our method has the following characteristics:

a) Completely partition sentences into chunks (not extract collocations from corpora);

b) Independent on the syntax information, and grammatical rules;

c) Use bilingual constraint and integrate partition process and alignment process;

d) Can deal with compositional chunks as well as non-compositional chunks.

3 Interactional Matching Algorithm for Automatic Chunking and Alignment

3.1 Overview

We proposed a new integrated method for chunk partition and alignment, named "Interactional Matching". The purpose is to get one-to-one chunk-pairs from untagged bilingual corpora.

Input: Bilingual sentence-aligned corpora

Output: Chunk-pairs with confident coeffi-cients (compositional and non-compositional)

Technology: Bilingual constraint and PLSI

3.2 Interactional Matching Algorithm

Considering that our purpose is to find out one-to-one chunk-pairs from bilingual sentence pairs without fragments, we use the most general formula to represent the integrated algorithm.

Let i denote the number of one-to-one chunk-pairs partitioned from a sentence-pair; Let j denote the jth partition mode of English Sentence when partitioned into $i$chunks; Let k denote the kth partition mode of Chinese Sentence when partitioned into $i$chunks; $(i^*, j^*, k^*)$is the optimized result of partition and alignment.Then

$$(i^*, j^*, k^*) = \arg\max_{i,j,k}\{\alpha \times [K_e(i,j) + K_c(i,k)] + (1-\alpha) \times Align(i,j,k)\} \quad (1)$$

where $K_e(i,j)$ denotes the probability of partitioning the English sentence into i chunks, with the jth partition mode. $K_c(i,k)$ denotes the probability of partitioning the Chinese sentence into i chunks, with the kth partition mode. $Align(i,j,k)$ is the probability of aligning i one-to-one chunk pairs, based on the jth partition mode of English sentence and kth partition mode of Chinese sentence. And α is the coefficient, which can be adjusted based on the characteristics of the corpora.

For the N-words-long sentence, if it is partitioned into i chunks, there will be $[(N-1)!]/[(i-1)!(N-i)!]$ partition modes. So, for a sentence pair (N-words-long English sentence and M-words-long Chinese sentence), if both sentences are partitioned into i chunks, the total number of partition modes will be formalized as follows: $(N-1)!(M-1)!/\{[(i-1)!]^2(N-i)!(M-i)!\}$.

The meaning of formula (1) is to conduct any possible partition modes and to compare the integrated probabilities of all these situations to meet the optimization of partition and alignment. The coefficient $\alpha$for written language is usually bigger than the one for spoken language, because the written language has more regular in syntax.

There are many ways to fulfill the functions$K_e(i,j)$, $K_c(i,k)$ and $Align(i,j,k)$. Here probabilistic-based method is adopted to design them as the following two parts, (A) and (B).

(A) We use the same model to design $K_e(i,j)$and $K_c(i,k)$, and the formula can be written as follows:

$$K(i,j) = \Pr(i|language, length) \times \log[\theta + CohesionDegreeCluster(i,j)] \quad (2)$$

where $\Pr(i|language, length)$ is the prior distribution probability of number of chunks (NOC) related to the language type, and sentence length, which can be estimated by manually-chunked sentences. And the second logarithm item represents the weight of cohesion degree clustering, related to the number of chunks i and the partition mode j, where:

$$CohesionDegreeCluster(i,j) = [\sum_{q=1}^{N-i} Reserve(q)/(N-i) - \sum_{p=1}^{i-1} Cut(p)/(i-1)]$$
$$\div[\sum_{p_1=1}^{i-2}\sum_{p_2=p_1+1}^{i-1}(Cut(p_1) - Cut(p_2))^2 + \sum_{q_1=1}^{N-i-1}\sum_{q_2=q_1+1}^{N-i}(Reserve(q_1) - Reserve(q_2))^2] \quad (3)$$

Here N is the length of the sentence, p is an index for the partition points in mode j, q is an index for the non-partition points in mode j, $Cut(p)$ is cohesion degree of two words adjacent to the pth partition point in this sentence, $Reserve(p)$ is cohesion degree of two words adjacent to the pth non-partition point in this sentence. Also θ in formula (2) is a coefficient between 0 and 1. The cohesion degree of two words (w_1,w_2) is simply defined as:

$$D(w_1, w_2) = (1 - \beta) \times MI(w_1, w_2) + \beta \times Times(w_1, w_2) \tag{4}$$

where MI is mutual information between two words. $Times(w_1, w_2)$ is just the co-occurrence times of word sequence (w_1, w_2), which is used to compensate irregular word sequence, such as some spoken expression. And β is a coefficient between 0 and 1, which is usually bigger for the spoken language than the one for the written language.

The value $K(i,j)$indicates the possibility of partitioning the sentence into i chunks, with the jth partition mode. The formula (2) has two parts, first of which is the prior probability, and the second represents the weight of cohesion degree clustering, which indicates the characteristic of the sentence. Given the number of chunks i and partition mode j, all the values of cohesion degrees in a sentence are separated into two classes, $Cut(p)$ and $Reserve(q)$. $Cut(p)$denotes the cohesion degree of pth partitioned point, and $Reserve(q)$ denotes the cohesion degree of qth reserved point. In order to give a effective weight to this classification result, we developed a self-clustering-based measurement algorithm, which is represented as formula (3). The numerator of formula (3) is the interactional distance between two classes of cohesion degrees, and the denominator is the internal distance of the two classes. So, $CohesionDegreeCluster(i, j)$ will be equal to a big value if the result of cohesion degree classification is reasonable; or it will be very small when the classification is unreasonable. And logarithm operator and θ are just used to adjust the dynamic range.

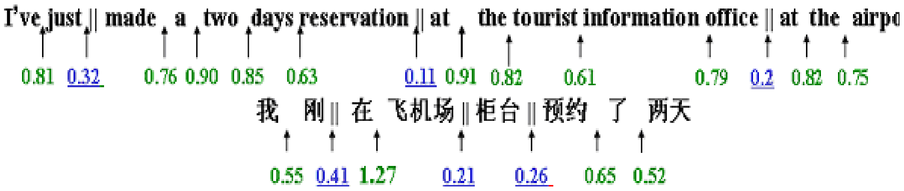
I've just || made a two days reservation || at the tourist information office || at the airpo
0.81 0.32 0.76 0.90 0.85 0.63 0.11 0.91 0.82 0.61 0.79 0.2 0.82 0.75
我 刚 || 在 飞机场 || 柜台 || 预约 了 两天
0.55 0.41 1.27 0.21 0.26 0.65 0.52

Fig. 1. Cohesion degrees under one partition mode with 4 chunks per sentence

Figure1 shows the example of chunking with above algorithm, which shows cohesion degrees between every two adjacent words in a sentence pair. Places marked with "||", are the partitioned points, which belong to the class $Cut(p)$. And other places are the reserved points, which belong to the class $Reserve(q)$. Use the formula (2) and (3), we can obtain the corresponding values such as

$K_e(i = 4, j) = 0.32$ and $K_c(i = 4, k) = 0.29$, when the prior probabilities $Pr(4|Eng, 15) = 0.42$ $\mathrm{Pr}(4|Chn, 8) = 0.30$, and coefficient θ=0.5.

(B) Different alignment has different matching measurement for one partition mode. Since every partitioned chunk can be represented as a word vector, we can get the matching value of one match mode by computing the similarity between vectors which can be realized using word similarity matrix $P(W, W)$. So we can choose one match mode of one-to-one chunk-pair by maximizing the matching degree which can be defined as follows:

$$
\begin{aligned}
Align(i, j, k) &= \max_{m} \sum_{p=1}^{i} ScoreV(i, j, k, p, m) \\
&= \max_{m} \sum_{p=1}^{i} [(wVectorE(i, j, p, m) \cdot P(W, W)) \cdot (wVectorC(i, k, p, m) \cdot P(W, W))^T]
\end{aligned}
\qquad (5)
$$

where $m$denotes the mth match mode. $wVectorE(i, j, p, m)$ is the word vector, built from the pth chunk in mth match mode of English sentence partition result with i chunks at the jth partition mode. $wVectorC(i, k, p, m)$ is the word vector, built from the pth chunk in mth match mode of Chinese sentence partition result with i chunks at the kth partition mode. Note that here “pth chunk” denotes the sequence number of chunks in one match mode, not direct sequence number in original sentence. $P(W, W)$ is the word similarity matrix, with is built by probabilistic latent semantic indexing (PLSI). And $ScoreV(i, j, k, p, m)$ is the corresponding alignment probability between $wVectorE(i, j, p, m)$ and $wVectorC(i, k, p, m)$.

For the situation in Figure1, the following table can be easily acquired.

Table 2. Aligning probabilities of matching mode

ScoreV (10^{-3})	我刚	在飞机场	柜台	预约两天
I’ve just	**0.65**	0.16	0.07	0.21
made a two days reservation	0.13	0.19	0.09	**0.78**
at the tourist information office	0.12	0.34	**0.15**	0.20
at the airport	0.07	**0.75**	0.05	0.13

After comparing all the match modes, we select the maximal one as the value of $Align(i, j, k)$which represents one optional alignment for the whole process. That is: $Align(i = 4, j, k) = (0.65 + 0.78 + 0.15 + 0.75) \times 10^{-3}$.

We use PLSI[6] to estimate the word similarity matrix, because the traditional statistical method cannot associate the meanings between different words, while LSI can. For instance, “computer” and “software” appear together several times in the corpus, and the same as “software” and “hardware”. Because of no co-occurrence between “computer” and “hardware”, in the traditional method, the correlation degree between them is zero or a fixed small value. It’s quite unreasonable. However, LSI can deal with this meaning clustering, and give a reasonable correlation degree between “computer” and “hardware”. That’s why LSI can deal with non-combinatorial chunks.

4 Experiment Results

First, we use roughly 3000 manually partitioned sentences to estimate the prior probabilities $\Pr(i|language, length)$ as Figure2 shows.

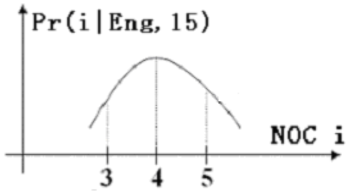

Fig. 2. Distribution of Number of Chunks(NOC) to sentence length and language

Then we take two types of corpora as our testing data, which has been included in www.Chinese-LDC.org:

A) English-Chinese Spoken Language:

Travel Domain: 20000 Sentences Pairs (ECSL)

B) English-Chinese Written Language:

General Domain: 12000 Sentences Pairs (ECWL)

We try to get enough information from the testing data themselves for chunking and alignment without any additional annotated training corpora, so we need to preprocess the testing corpora by replacing some named entities with unified tag, which can be implemented automatically and easily by virtue of existing tools.

4.1 Overall Accuracy

Considering the whole process (partition and alignment), we can get the overall result. In this experiment, parameters in formula (1) (2) and (4) have been set and listed below:

Corpora	α	β	θ
ECSL	0.50	0.05	0.50
ECWL	0.60	0.045	0.50

A general concept of the precision-recall measure is used here to testify our algorithm, and the results are presented in table 3. Note that there is no manual interference in the whole process.

Table 3. Overall results of the experiments

Corpora	No	Overall Accuracy		
		Precision (%)	**Recall (%)**	**Standard F Mea-sure**
EC**SL**	1	90	32	0.47
	2	**78**	**53**	**0.63**
	3	60	57	0.58
EC**WL**	1	90	40	0.55
	2	**85**	**57**	**0.68**
	3	60	66	0.63

The column "No" indicates three threshold of confidence coefficient for correct chunk-pairs. Obviously, the performance on written language is much better by reason of its regular expressions.

And when adjusting the parameter α, we get the following Figure3, which shows the best F measure of overall result related to the α.This result proves that the written language needs a bigger αthan spoken language.

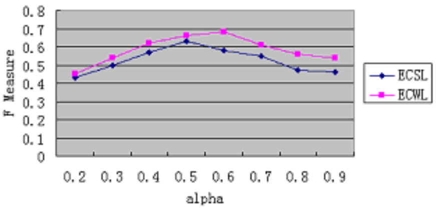

Fig. 3. The best F-measure related to α

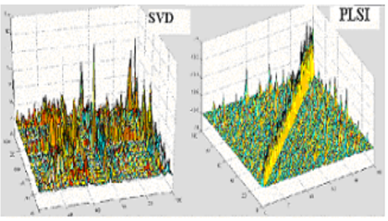

Fig. 4. $P(W, W)$ sample of SVD and PLSI

It is also proved by other experiments that the coefficient$\beta$is around the value of 0.05 and $\theta$is around the value of 0.5, which are not sensitive to the overall result.

4.2 Partition Accuracy

In this section, the first part of formula (1) is used to do the partition experiment only. That is: $(i^*, j^*, k^*) = \arg\max_{i,j,k} \alpha \times [K_e(i, j) + K_c(i, k)]$

Table4 shows that bilingual chunking does much better than monolingual chunking because of the bilingual constraint implemented by our Interactional Matching algorithm. Also we can see English sentences have better performance, because they are less flexible in expressions.

4.3 Alignment Accuracy

Manually correcting the results of partition, we get 100% precision partition data, which was used to do alignment experiments (see the formula (5)).

Table 4. Comparison between monolingual and bilingual partition accuracy

Corpora	**Language**	**Partition Precision (%)**	
		Monolingual	**Bilingual**
ECSL	English	75	77
	Chinese	66	77
ECWL	English	81	86
	Chinese	74	86

Table 5. Alignment results using DMM and PLSI

Corpora	**DMM (%)**		**PLSI (%)**	
	Precision	**Recall**	**Precision**	**Recall**
ECSL	82	75	91	80
ECWL	90	83	94	87

It is found from table5 that PLSI outperforms the traditional Dictionary Matching Method (DMM), which just measures the matching similarity of chunks via a bilingual lexicon. It can be to some extent ascribed to the fact that meaning clustering has been realized in PLSI while DMM is only limited to morphological analysis. PSLI can therefore cope with most of non-compositional chunk pairs,such as "here and there —— 各处", but DMM can't.

4.4 LSI Analysis

The accuracy of matrix $P(W, W)$ decides the precision of chunk alignment. Jin (2003) made effective improvements to the standard EM algorithm for PLSI[15], and calculated P matrix using 1000 bilingual texts. 100×100 matrix sample is shown in figure4. Note that in the Matrix P, there are no difference between English words and Chinese words. SVD and PLSI results are showed in Figure4.The diagonal of matrix $P(W, W)$ represents the self-correlation values, which certainly should higher than cross-correlation values. So we can see that PLSI is better than SVD, not only on correlation, but also on the equilibrium of the matrix.

5 Conclusion

This paper presents an integrated method for bilingual chunk partition (not extraction) and alignment, called "Interactional Matching". Different from existing approaches, this method can automatically find out the one-to-one chunk-pairs through mining enough information from untagged bilingual corpora, without relying heavily on corpora which have been assigned with POS tags and syntax information. Furthermore, with the technology of PLSI, this algorithm is able to deal with both compositional and non-compositional chunks.

Experiment shows that this method gets good performance for partition and alignment on both written language and spoken language. And also this chunking technique can be applied to many tasks of natural language processing, such as machine translation and cross-lingual information retrieval.

6 Acknowledgments

This work was supported by the Natural Sciences Foundation under grant No. 60272041, 60372016, 60121302.

References

1. Frank Smadja: Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1), (1993) 143-177
2. Qiang Zhou: Automatically Bracket and Tag Chinese Phrase. *Journal of Chinese Information Processing*, 11(1), (1997) 1-10
3. Boxing Chen and Limin Du: Alignment of Single Source Words and Target Multi-word Units from Parallet Corpus. In: 1^{st} *Students' Workshop on Computational Linguistics Proceedings*, August 20-23,(2002) 318-127
4. Silva J.F., Dias G., Guillor S. and Lopes J.G.P.: Using Localmaxs Algorithm for Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: 9^{th} *Portuguese Conference in Artificial Intelligence*, Lecture Notes, Spring-Verlag, Universidade de Evora, Evora Portugal (1999)
5. Wei Wang, Ming Zhou, Jinxia Huang and Changning Huang: Structure Alignment Using Bilingual Chunking. In: *Proceedings of COLING 2002*, 24 August – 1 September, Taipei (2002)
6. Thomas Hofmann: Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, Berkeley, Cali-fornia, August, (1999) 50–57
7. Dekai Wu: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3), (1997) 377-400
8. David Blei, Andrew Y. Ng and Michael Jordan: Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (2003) 993-1022
9. Gene Golub, Knut Solna, and Paul Van Dooren: Computing the SVD of a General Matrix Product/Quotient, *SIAM Journal on Matrix Analysis and Applications*, 22(1), (2000) 1-19
10. Wei Cheng, Jun Zhao, Bo Xu and Feifan Liu: Bilingual Chunking for Chinese-English Spoken-language Translation. *Journal of Chinese Information Processing*, 17(2), (2003) 21-27
11. Jun Zhao: The Framework of Cross-lingual Information Retrieval. *Chinese-Japanese Natural Language Processing Proseminar (2nd)* (2002)
12. Cong Li and Hang Li: Word Translation Disambiguation Using Bilingual Bootstrapping. In *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, July, (2002)
13. Watanabe H., Kurohashi S., Aramaki E.: Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. *COLING 2000* (2000)
14. Le, S., Youbing, J., Lin, D. and Yufang, S.: Word Alignment of English-Chinese Bilingual Corpus Based on Chunks, *Proc. 2000 EMNLP and VLC* (2000) 110-116
15. Qianli Jin, Zhao, J., Xu, B.: Weakly-Supervised Probabilistic Latent Semantic Analysis and its Applications in Multilingual Information Retrieval. In: *Proceedings of* 7^{th} *Joint Symposium on Computational Linguistics*, 9-11 August, (2003)