COMPUTATIONAL METHODS IN ENGINEERING AND SCIENCE
EPMESC X, *Aug. 21-23, 2006, Sanya, Hainan,China*

Cross-level Sentence Alignment

Anna Ho[1]*, Francisco Oliveira[1], Fai Wong[1]

[1] *Department of Computer Science, Faculty of Science and Technology of University of Macao, PO Box 3001 Macao SAR*
Email: ma36560@umac.mo

Abstract This paper describes a new model for sentence alignment system of structurally different languages such as Chinese and Portuguese. In the proposed method, we try to combine the statistical approach and lexical approach in order to achieve the efficiency and accuracy. We first complete the word level alignment by making use of the Chinese-Portuguese dictionary to get the basic translation rate between the two texts. In order to make the system more adaptive, we apply the maximum entropy model to align the named entities without perform the word segmentation for Chinese. Secondly, we apply the Singular Value Decomposition (SVD) to get the statistical information of the sentences.

Key words: alignment, Hidden Markov Model, maximum entropy model, singular value decomposition

1. INTRODUCTION

In this paper, we address the problem of cross-level sentence alignment for a structurally different bilingual corpus such as Chinese and Portuguese. There are several papers which describe the similar problem but focus on structurally similar corpus, such as: Melamed [1], Fung et. la. [2] and S. Vogel, H. Hey, and C. Tillmann [3].

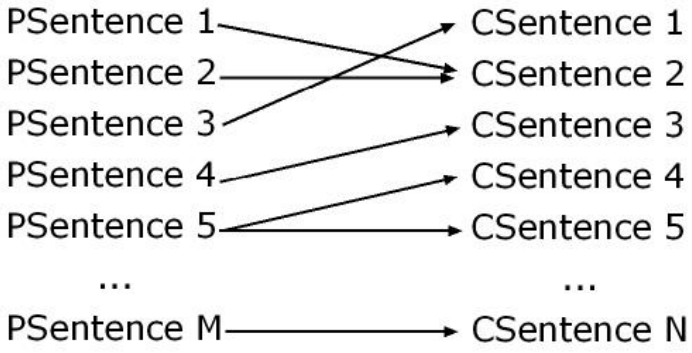

Figure 1: Possible Cross-level Sentence Alignment

In our approach, we first complete the word level alignment by making use of the Chinese-Portuguese dictionary to collect the basic translation rate between to the Chinese text and Portuguese Text. Since the dictionary only provides the common words translation and named entities are not included, the system cannot make good decision in this procedure. Therefore, we apply the maximum entropy model to align the named entities without perform the word segmentation for Chinese.

Secondly, from the word level alignment, we achieve anchor point and process the sentence level alignment. We apply SVD to get the statistical information of the sentences. For SVD model, we first set up a matrix which consists of the word level alignment statistic information. Then performs the two dimensional reconstruction of the original matrix. By analyzing the results, we can observe some relationships among the sentences and this can give an approximation of sentence alignment.

The organization of the paper is as follows: In section 2, we will have a brief review of statistical approach and lexical approach to machine translation. Section 3 presents the overall framework and algorithms of the system. We will show the experimental results in section 4. Conclusions of the paper will be discussed in section 5.

2. REVIEW

In statistical approach, Brown et al [4] and Gale and Church [5] show a method of aligning sentences based on a statistical model of character length. Brown el al makes use of the major and minor anchor points to facilitate the alignment process. On the other hand, Gale and Church's model assigns a probabilistic score to each proposed correspondence of sentence base on the difference of their lengths and the variance of this difference. S. Vogel, H. Hey, and C. Tillmann apply HMM in word alignment to an English-French corpus. This model makes the alignment probabilities depend of the alignment position of the previous world.

In lexical approach, Kay and Roscheisen [6] employ a partial alignment of the word level to introduce a maximum likelihood alignment of the sentence level. Chen [7]'s model estimates the cost of an alignment by aligning two texts *S* and *T* and divides it into sequence of sentence. Mostly, lexical approach makes use of the dictionary to assist the aligning system.

Though statistical approach gives a high performance, lexical approach provides a lot of confirmation details of alignments. Therefore, Li et. la. [8] combines both statistical (length-based) and lexical methods in sentence alignment. Some lexical cues or parameters are included in order to maximum the result of a probability function.

In the above reviews of several papers, some does not support cross-level alignment but focus on structurally different corpus, while others support cross-level alignment but focus on structurally similar corpus. The following sections illustrate our approach to solve the cross-level alignment which focuses on structurally different corpus.

3. FRAMEWORK OVERVIEW

In this section, we discuss the aligning model of our system. Without relying on syntactic knowledge from either the Portuguese side or the Chinese side, we find there are several valuable features that can be used for the cross-level alignment. Considering the advantages of SVD which can help analyzing the relationship between sentences and integrating different kinds of methods, the system framework can be basically divided into three procedures: statistical information retrieval, learning named-entities and sentence relation analysis.

3.1 CONSTRUCTING THE PRELIMINARY MATRIX

In the first step, we need to estimate the word correspondence between the sentences in the two texts. To calculate this, we set up an *M x N preliminary matrix*, where *M* represents the number of sentences appear in the Portuguese text and *N* represents the number of sentences appear in the Chinese text. With the assistance of the built-in Portuguese-Chinese dictionary, cell entries are the number of times that every Portuguese word in each Portuguese sentence that can be found an exactly translated Chinese word from each Chinese sentence.

$$cell_i = \sum_{a=0}^{m} trans(p_a, C) \quad (1)$$

We calculate the preliminary cell entries from (1), where *m* is the length of a Portuguese sentence, *N is* the total number of the Chinese sentences, *p* is the Portuguese word and *P* is a Portuguese sentence which contains a set of *p*, *C* is the Chinese sentence. Since we do not rely on the syntactic knowledge of the Chinese side, we do not perform segmentation in Chinese sentence. Therefore, we segment the Portuguese sentence by recognizing the space, and perform the translation matching process for each Portuguese word with each Chinese sentence. After scanning all the sentences, a matrix will be formed as show in *Table 1*.

Table 1: The preliminary matrix formed by applying Equation 1

	C_1	C_2	C_3	C_4	C_5	C_6	C_n
P_1	6	2	2	6	2	2	trans(P_1,C_n)
P_2	5	4	4	4	4	4	trans(P_2,C_n)
P_3	1	0	0	1	0	0	trans(P_3,C_n)
P_4	1	1	1	1	1	1	trans(P_4,C_n)
P_5	2	1	1	2	1	1	trans(P_5,C_n)
P_6	1	1	1	1	1	1	trans(P_6,C_n)
P_7	1	0	0	1	0	0	trans(P_7,C_n)
P_8	0	0	0	0	0	0	trans(P_8,C_n)
P_9	11	7	7	12	7	7	trans(P_9,C_n)
P_{10}	3	3	3	3	3	3	trans(P_{10},C_n)
P_n	trans(P_n,C_1)	trans(P_n,C_2)	trans(P_n,C_3)	trans(P_n,C_4)	trans(P_n,C_5)	trans(P_n,C_6)	trans(P_n,C_n)

We call this preliminary matrix because the statistic information retrieved just only base on the Portuguese-Chinese dictionary and named-entities are not considered which lost many valuable information that can be provided to the matrix. So, in the next section, we describe how the system learns the named-entities and make contribution to the matrix.

3.2 ALIGNING NAMED-ENTITIES

As we cannot rely on the dictionary information anymore and do not want to perform segmentation in Chinese sentence to increase the efficiency, we defined two functions which find the co-occurrences and distortion of the Portuguese and Chinese named-entities in the whole corpus. Feng, et al [9] presents a method of aligning Chinese and English named-entities by applying maximum entropy model in four different features: translation score, transliteration score, co-occurrence score and distortion score. In our case, we choose to ignore the transliteration score because it is time consuming for converting the Chinese named-entities to PinYin string.

Translation score can be obtained in the stage of constructing preliminary matrix easily. For the co-occurrences function $P_{co}(ne_c \mid ne_p)$, if both named-entities co-occur very often, the probability that they align to each other is very large. When the system forms the preliminary matrix, it also forms a co-occurrence matrix which holds the co-occurrence probability of the Portuguese named-entities and Chinese named-entities.

$$P_{co}(ne_c \mid ne_p) = \frac{count(ne_c, ne_p)}{\sum count(*, ne_p)} \quad (2)$$

The co-occurrence probability is calculated by (2), where *count(ne_c,ne_p)* = the number of times ne_c and ne_p appear together, *count(*,ne_p)* = the number of times that ne_p appears.

The second function is the distortion function $P_{dist}(ne_c, ne_p)$. We choose to include distortion because we notice that the difference of the named-entities position in both languages can determine their relation. The bigger the difference, the less probability they can be translated to each other. For example, if the Portuguese named-entity starts at position *i*, and the length of the whole sentence is *m*, the relative position of the Portuguese named-entity is defined as (3):

$$pos_p = \frac{i}{m} \tag{3}$$

Similarly, for the relative position of the candidate Chinese named-entity, we defined it as pos_c. So we have $0 \le pos_p, pos_c \le 1$ and defined the distortion function with (4):

$$P_{dist}(ne_c, ne_p) = 1 - |(pos_p - pos_c)| \tag{4}$$

If there are multiple identical candidate Chinese named-entities at different positions, we choose the one with the highest distortion score.

With the above three featured scores, we apply the maximum entropy model to align the named-entity. Firstly, we perform the selection of the named-entity candidates. For those Portuguese words that are in capital letters or the first letter is capital and cannot find an exact translation when we constructing the preliminary matrix, we considered it as the named-entity candidates. For those candidates, we try to release the rules find the partial translation, in this case we can find several Chinese named-entity candidates. Now, we have the seed data finally we can train the maximum entropy model. We use the published package YASMET to process the training. *Table 2* shows the procedure of the training.

Table 2: Procedure of the training

1. Set the coefficients λ_i as uniform distribution;
2. Calculate all the features scores to get the N-best list of the Chinese NE candidates;
3. Candidates with their values over a given threshold are considered to be correct and put into the re-ranking training set;
4. Retrain the parameters λ_i with YASMET;
5. Repeat from Step 2 until λ_i converge, and take the current ranking as the final result.

3.3 SENTENCE REALTIONSHIP ANALYSIS

After the final matrix has constructed, we can predict the sentence alignment by looking at the scores in each cell. Since the scores in each row represents how strong that a Portuguese sentence corresponds to the Chinese sentence, the highest the score in cell(i, j), the most probably that P_i can be aligned to C_j. However, this may lead to wrong prediction when there are two same highest scores in a row. To prevent this case, we apply SVD to the final matrix. *Figure 2* shows the concept of SVD.

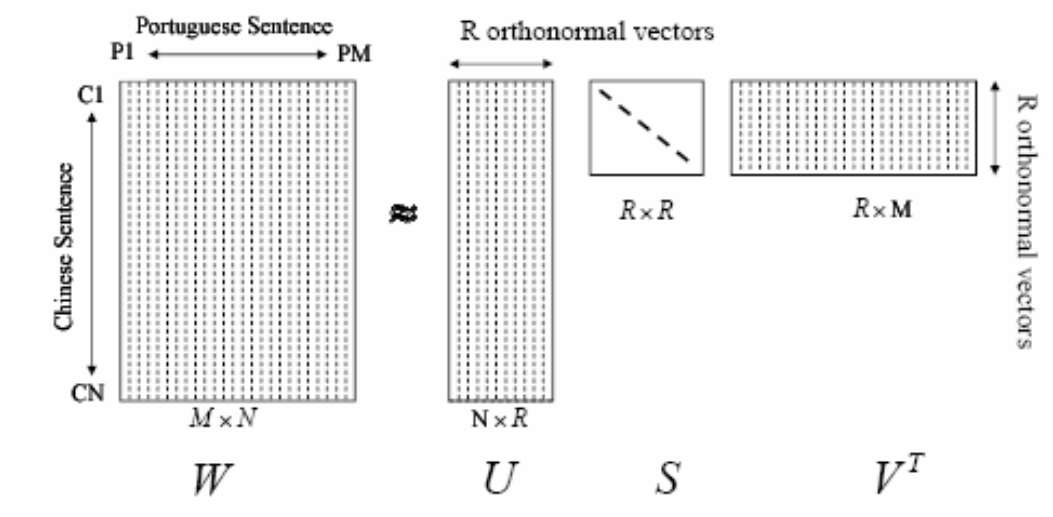

Figure 2: Process of Singular Value Decomposition

Base on our final matrix, we decompose it into a product of three matrixes by SVD as show in *Formula 5.*

$$W \approx \hat{W} = USV^T \quad (5)$$

In here, *W* represents our final matrix, $\hat{W}$ is the optimal approximation with the best rank-*R* of *W*. SVD provides an analysis of the data which descries the relationship among the sentences. For example, there are two same highest score which tells Portuguese sentence 1 is able to be aligned with Chinese sentence 2 and Chinese sentence 10, without SVD, the system is not able to make a correct decision on which Chinese sentence the Portuguese sentence 1 should be aligned to. However, SVD has a capability to calculate the similarity of data, it can provide information of which Portuguese sentence should align to which Chinese sentence.

4. EXPERIMENTAL RESULTS

Our experiments based on the corpus based machine translation. We evaluate our algorithm by using cultural information related bilingual corpus. This kind of corpus contains many named-entities such as name of countries, and name of persons. According to our algorithm we break down the paragraph into sentences according to punctuation marks. *Table 3* shows the improvement of accuracy from Step 1 to Step 3.

Table 3: Results of sentence alignment

	Step 1	**Step 2**	**Step 3**
Total (in sentence)	412	412	412
Correct	280	363	404
Incorrect	132	49	8
Percentage	68%	88%	96%

The accuracy has a great improvement from step 1 to step 2 because the corpus we choose contains many named-entities which causes this result. There is also 8% improvement from step 2 to step 3. The improvement is not as great as from step 1 to step 2 because the information among the sentences are not resembled to each other. There is great improvement from step 2 to step 3 mostly when the length and the meaning of sentences are similar. In this case, statistic information retrieved from step 2 is not enough to guarantee

5. CONCLUSIONS

In this paper, we have presented a mixed-features alignment model in parallel corpus. From the experiment results, we observe that the choice of corpus affect the accuracy in different stages. As our system can learn named-entity by itself, it can handle the corpus with a lot of named-entity which cannot be found from the dictionary. Moreover, our system can handle the cross-level sentence alignment by analyzing the relationship information among the sentences in the parallel texts. However, as the matrix will grow enormous when corpus becomes larger, the efficiency will drop accordingly. Therefore, there are still rooms for improvement in tuning the algorithm to be more efficiency.

REFERENCES

1. Melamed, I. D.. Bitext Maps and Alignment Via Pattern Recognition. Computational Linguistics, 1999, pp. 107 – 130.
2. FUNG, P. K. C.. Kvec: A New Approach for Aligning Parallel Texts. Proceedings of COLING 94, Kyoto, Japan, 1994, pp. 1064 – 1102.
3. S. Vogel, H. Ney, and C. Tillmann. HMM-Based Word Alignment in Statistical Translation. Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996, pp, 836 – 841.
4. Brown, P. F., J. C. Lai and R. L. Mercer. Aligning Sentences in Parallel Corpora, 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA., 1991, pp. 169-176.
5. Gale, William A. & Kenneth W. Church. A program for aligning sentences in bilingual corpora. Computational Linguistics, 1993, vol. 19, pp. 75 – 102.
6. Kay, M., & M. Roscheisen. Text-Translation Alignment. Computational Linguistics, 1993, pp. 121 – 142.
7. Chen S. F.. Aligning Sentences in Bilingual Corpora Using Lexical Information[C]. Proceedings of 31st Annual Meeting of the Association for Computational Linguistics. Columbus, OH: ACL, 1993, pp. 9 – 16.
8. Yiping Li, Chiman Pun, Fei Wu. Portuguese-Chinese Machine Translation in Macao. Proceedings of Machine Translation, SUMMIT VIT'99, Singapore, 1999, pp. 236 – 243.
9. Feng, D., Lv, Y. and Zhou, M. A New Approach for English-Chinese Named Entity Alignment. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 372 – 379.