Available online at www.sciencedirect.com

SciVerse ScienceDirect

AASRI Procedia 1 (2012) 468 – 473

AASRI Procedia

www.elsevier.com/locate/procedia

2012AASRI Conference on Computational Intelligence and Bioinformatics

A Sentence Alignment Model Based on Combined Clues and Kernel Extensional Matrix Matching Method

Wu Honglin[a], Liu Yiyang[a,]*, Liu Shaoming[a,b]

[a]*Northeastern University, Shenyang, 110004, China;*
[b]*Fuji Xerox Co., Ltd., Kanagawa, 2208668, Japan*

Abstract

A sentence alignment model based on combined clues and Kernel Extensional Matrix Matching (KEMM) method is proposed. In this model, a similarity matrix for sentence aligning is formed by the similarities of bilingual sentences calculated by the combined clues, such as lexicon, morphology, length and special symbols, etc.; then this similarity matrix is used to construct a select matrix for sentence aligning; finally, obtains the sentence alignments by KEMM. Experimental results illustrated that our model outperforms over the Gale's system on handling any types of sentence alignments, with 30% total sentence alignment error rate decreasing.

Key Words：Sentence Alignment， Combined Clues，Kernel Extensional Matrix Matching

1. Introduction

In various fields of research on natural language processing, the importance of bilingual corpus is more and more obvious. Different applications call for aligned bilingual corpus of different granularities and corresponding technology, which includes article, paragraph, sentence, phrase and word level. Sentence level aligned bilingual corpus is indispensable to example based machine translation.

*Liu Yiyang. Tel.: +86-24-83672480.
E-mail address: lyy880315@yahoo.cn.

doi:10.1016/j.aasri.2012.06.073

A sentence alignment model based on combined clues and KEMM method is proposed. In this model, a similarity matrix for sentence aligning is formed by the similarities of bilingual sentences calculated by the combined clues; then this similarity matrix is used to construct a select matrix for sentence aligning; finally, obtains the sentence alignments by KEMM.

2. Similarities of Bilingual Sentences Calculated By the Combined Clues

To the translations needed to have sentence alignment for Chinese text C and Japanese text J, we assume that the number of C is m and to be expressed as C=$CS_1CS_2\ldots CS_m$, the number of J is n and to be expressed as J=$JS_1JS_2\ldots JS_n$. The similarities of all bilingual sentences calculated by the combined clues included lexicon, morphology and length. Results by each approach are summed up after multiplied by different weights and then plus special symbols similarity to be the final similarity between Chinese and Japanese sentences. The formula to calculate similarity of Chinese sentence CS_h and Japanese sentence JS_k:

$$\begin{aligned} Sim(CS_h, JS_k) = \; & \alpha \times SimDict(CS_h, JS_k) + \beta \times SimMorph(CS_h, JS_k) + \\ & \gamma \times SimLength(CS_h, JS_k) + SValue(CS_h, JS_k) \end{aligned} \tag{1}$$

α , β and γ are the weights, and α + β + γ =0.8; 0≤SValue(CS_h,JS_k)≤0.2.

2.1. Based on Bilingual Dictionary

In this calculating method, sentences are expressed as a set of words. Chinese and Japanese sentences are expressed as $CS_h = \{c_1, c_2, \cdots, c_m\}$ and $JS_k = \{j_1, j_2, \cdots, j_n\}$ .Similarity calculatingbased on bilingual dictionary (SimDict) is:

$$SimDict(CS_h, JS_k) = \frac{|TransSetC| + |TransSetJ|}{|CS_h| + |JS_k|} \tag{2}$$

|CS_h| and |JS_k| are the number of content words in Chinese and Japanese, |TransSetC|and |TransSetC| are the number of content words that have translations.

2.2. Based on Font

Sentence is expressed as a set of characters in calculating. Chinese and Japanese sentences are expressed as $CS_h = \{cc_1, cc_2, \cdots, cc_r\}$ and $JS_k = \{jc_1, jc_2, \cdots, jc_s\}$. The calculating method based on font:

$$SimMorph(CS_h, JS_k) = \frac{|MorphSetC| + |MorphSetJ|}{|CS_h| + |JS_k|} \tag{3}$$

|CS_h| and |JS_k| are the number of characters in sentences, |MorphSetC| is the number of the same Chinese characters CS_hhave inJS_k. |MorphSetJ| is the number of the same Chinese characters JS_k have in CS_h.

2.3. Based on Sentence Length

Block=<C,J> is proposed as a translation block that have been aligned. Then alignment A=$P_1P_2P_3\ldots P_k$. P_i(1≤i≤k) is a couple of translation and P_i=<CS_i,JS_i>; CS_i is a set of Chinese sentences, it expresses from zero to more sentences in C; JS_i is a set of Japanese sentences, it expresses from zero to more sentences in J.

In a couple of translation P=＜CS,JS>, to propose that l_1=length(CS) is the length of CS and l_2=length(JS) is the length of JS.In a word, l_1、l_2 are the number of characters in CS and JS. To the translation block

Block=<C,J>, sentence alignment based on length statistic is to look for the one has the largest probability in all probable alignment A. Condition below must be met:

$$A = \arg\max \Pr(A \mid Block) = \arg\max \Pr(A \mid < C, J >) \qquad (4)$$

To propose that each couple of translation is absolutely, and Pr(P|<C,J>) is not depend on translation block, it's only decided by P=<CS,JS>. It can be thought based on statistic that Pr(P|<CS,JS>) is only decided by matching pattern match(P) of P and the length of CS and JS.

Since the Bayesian theory:

$$A = \arg\max_{A} \prod_{P \in A} \Pr(< l_1, l_2 > \mid match(P)) \cdot \Pr(match(P)) \qquad (5)$$

Pr(match(P)) is the matching probability. To propose that $\delta(l_1, l_2)$ is a function satisfied Standard normal distribution, and the key of model above is the design of the evaluate function $\delta(l_1, l_2)$. It's given in Gale:

$$\delta(l_1, l_2) = \frac{l_2 - c \cdot l_1}{\sqrt{l_1 \cdot s^2}} \qquad (6)$$

$c = \sum_{P \in TC} l_2 / \sum_{P \in TC} l_1$, s^{2}is the slope of the line that the result of all points $(l_1, (l_2 - l_1)^2)$ in training corpus being linear regression analyzed. s^2 is a normalization factor to make sure $\delta(l_1, l_2)$ is a normal distribution. To the Chinese and Japanese sentence alignment, c=1.06 and S^2=6.8 are from the statistic of the corpus.

In Gale test, Pr(match(P)) in formula(6) is a penalty factor to multi-alignment. Since the sentence alignment model in this paper is a method of combining some different approaches, the final alignment similarity needs a penalty calculating. Then formula(6) becomes:

$$SimLen(CS_h, JS_k) = \arg\max_{A} \prod_{P \in A} \Pr(\delta(l_1, l_2) \mid match(P)) \qquad (7)$$

2.4. Based on Special Character

Numbers (such as 1978, 03, 24, etc.), English characters (such as China, Henry, etc.), quotation marks and brackets are used to calculate the similarity of Chinese and Japanese special characters. The maximum value of special characters' similarity (SValue) is set as 0.2. When CS$_h$ and JS$_k$ have the same special characters, SValue(CS$_h$,JS$_k$) is added 0.05, and it's added to 0.2 for max. It is a supplement for the similarity calculating of SimDict,SimMorph and SimLength.

2.5. Multi-alignment Penalty Factor

Multi-alignment penalty factor is to deal with alignment conflict. For an instance, when aligned conflicts happen and judge if the style is 1-2 or the two aligned 1-1 and 0-1, we need to calculate the similarity of these three kinds and then multiply the corresponding penalty factor. Then we can compare their similarity modified.

We stat the proportion of different alignment in Chinese and Japanese corpus including 9679 aligned sentences. Combining researches to set the penalty factor (ξ) as the table 1 shows.

Table 1 Penaltyfactor for multi-alignment

Alignment style	Penalty factor(ξ)
1-2;2-1	0.95
2-2	0.60
1-0;0-1	0.55

3. SentenceMatching Based on Kernel Extensional Matrix Method

3.1. Alignment Similarity Matrix and Alignment Selecting Matrix

An alignment similarity two-dimensional matrix (SimMatrix) is constructed by Chinese and Japanese sentences that the number of them is m and n. The element in the matrix is the similarity.

$$SimMatrix[h][k] = Sim(CS_h, JS_k) \tag{8}$$

Then we construct a sentence alignment selecting matrix (SelMatrix). Each element is the sorting information of the alignment similarity.

3.2. Sentence Matching

Sentence matching is to choose the alignment results from the sentence alignment selecting matrix.

1) Select the Alignment in “1/1” Row

Six steps are included in this process.

Step1: According to the SelMatrix to select the location where "1/1" is. And then construct a 3×3 matrix to calculate similarity and check the probability if there is some multi-alignment.

Step2: To calculate the probability of the multi-alignment.

```
  If SelMatrix[i-1][j-1] = “1/1” and
SelMatrix[i+1][j+1] = “1/1”;
          NCA = (CSi)-(JSj);
  Else If SelMatrix[i-1][j-1] or SelMatrix[i+1][j+1] is
not "1/1"
          $NCA = \arg\max_{P \in A} Sim(P)$
    MultiSim= Cal(min(SimAdd, Selmatrix[i][j]));
    Compare MutiSim with NCA;
```

If MutiSim is larger than NCA, we put the multi-alignment into CandSet. Otherwise turn to step5.

Step3:To judge the style of new member of CandSet. Processing is according to its style.

Step4: The new multi-alignment processing can be considered as 10 categories, and one of them can be made formal to :

```
If $(\exists k:(1 \le k \le m) \wedge (k \ne i):(CS_k)-(JS_{j-1}JS_j) \in CA)$
          If $Sim(NCA) > Sim((CS_k)-(JS_{j-1}JS_j))$
            $CA = CA / \{(CS_k)-(JS_{j-1}JS_j)\}$
            $CA = CA \cup \{NCA\}$
            return true;
          return false;
```

Step5: Add the ensured alignment in the aligned queue and record the aligning status of sentences.

Step6: To judge whether all "1/1" locations are dealt with. If not, turn to step1.

2) Select the Alignment not in “1/1” Row

This processing is to look for the locations without "1/1" and look for the minimum sum of similarity in its row. And then to judge whether cross-aligned will be happened in this location.

And the processing in column is similar with the processing above.

3) Spatial Alignment Processing

Step1: To look for unaligned row.

Step2: To select the point "1/1" in the previous row and back row. And then to calculate the similarity of the unaligned row with previous row and back row.

Step3: To compare the modified multi-alignment similarity with the point.

Step4: Accoring to the comparing results to deal with them further.

Step5: To store the results and record the alignment status. And to look for if there is any unaligned row.

4. Expriments and Results Analysis

The test set used in sentence alignment expriment is randomly selected from the Chinese and Japanese chapter-aligned corpus and they are artifically marked standard answers of stence alignment.The test set includes 558 alignments, the style distribution shown as table 2.

To verify the effectiveness of the model put forward in this paper, we achieve the sentence alignment system with the method given in this paper. We test the error rate of the system we construct on the test set and compare the results with Gale system which is famous in sentence alignment field.

Table 2 Alignment types distribution of Testing set

Style	Distribution	Distribution Rate(%)
1-0;0-1	6	1
1-1	505	90.5
1-2;2-1	47	8.5

Table 3 shows the distributionof all kinds of alignment style in our test set and corpus. We can see that the distribution of our test set is approaching the distribution of our corpus. So our test set is representative, it can represent the test result of our corpus.

Table 3 Comparison of alignment types distribution

Style	Test Set(%)	Bilingual Corpus(%)	Gale(%)
1-0;0-1	1	0.2	1
1-1	90.5	93.8	89
1-2;2-1	8.5	5.8	9

We achieve the align method in Gale as beseline system. But the test results of the system have differences from reference (it's worse than the results in reference). So the data is not faith to Gale system. We use the better experimental results from Gale shown in reference. As shown in table 3, the distribution of our test set is approaching the test set in Gale. So this comparison is faithful.

Test set in Gale includes some rare alignment and they don't appeared in our test set. Since the experiental results in Gale show that the error rate of processing the styles is 100%, the difference cannot make effects. So the comparison is sloped to Baseline.

In summary, we try our best to ensure the faith and accurate in experiment. Table 4 shows the experiental results of sentence alignment.

From the table 4, we can know that our system has a lower error rate. We analysis the test results and find that the reson for lower error rate is to use the KEMM to avoid error spreading phenomenon.

On all kinds of alignment styles our system always has a lower error rate than Gale system. The error rate of our system is 2.3% and Gale is 3.5%. The exprimental results and analysis above fully prove the effectiveness of our calculating model.

Table 4 Experiment result: error rate

Align Style	Test Set Align Num	Our Error Num	Our Error rate (%)	Gale Error rate (%)
1-0;0-1	6	2	33.3	100
1-1	505	7	1.4	2
1-2;2-1	47	4	8.5	9
Sum	558	13	2.3	3.5

5. Conclusion

A sentence alignment model based on combined clues and Kernel Extensional Matrix Matchingmethod is proposed in this paper.

In this model, a similarity matrix for sentence aligning is formed by the similarities of bilingual sentences calculated by the combined clues, such as lexicon, morphology, length and special symbols, etc.; then this similarity matrix is used to construct a select matrix for sentence aligning; finally, obtains the sentence alignments by KEMM.

Experimental results illustrated that our model outperforms over the Gale's system on handling any types of sentence alignments, with 30% total sentence alignment error rate decreasing. These can prove the effectiveness of our model proposed in this paper.

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities of Northeastern University(N100304006).

References

[1] Gale W, Church K. A program for aligning sentences in bilingual corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley 1991;177-184.

[2] Elithorn A., Banerji R.Nagao. M.A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. Artifical and Human Inteligence. New York: Elsevier Science Publishers Corporation; 1984; 173-180.

[3] Brown, R.D. Automated Generalization of Translation Examples. Proceedings of the Eighteenth International Conference on Computational Linguistics 2000;125-131.

[4] Halil Altay Guvenir, Ilyas Cicekli. Learning Translation Templates from Examples. Information Systems 1998; 23(6):353--363.

[5] Arnold D., Balkan L., Humphreys R. Lee, Meijer S., Sadler L.. Machine Translation. 1994.