

2001 Johns Hopkins University Summer
Workshop
Automatic Summarization of Multiple
(Multilingual) Documents
Section: Sentence Alignment

DANYU LIU

email: liudy@cis.uab.edu

January 22, 2002

Abstract: This report presents a project to align and match English-Chinese bilingual news documents came from *Hong Kong News Bilingual Corpus*. The work involves the alignment of bilingual documents at the sentence levels. These news documents have both their own characteristics and some properties that previously alignment work had reported in the literature. To align the news documents we apply the length-based statistical approach. To get alignment pairs, we employ *Dynamic Programming* to align at the paragraph level, then align at the sentence level. The precision and recall of the alignment are quite good for translation documents. Finally, to determine the various parameters used in aligning and matching, we utilize a statistical approach to obtain their optimized values.

Key words:

sentence alignment, dynamic programming, Hong Kong News Corpus

The **Automatic Summarization of Multiple (Multilingual) Documents Project** at *2001 JHU summer workshop* is to study the integration of cross-lingual information retrieval and subsequent multi-document summarization. As one part of the whole project, we investigate some existing alignment algorithms and produce a special-purpose experimental alignment system for Hong Kong News Corpus.

Sentence alignment is the problem of making explicit the relations that exist between the sentences within two documents that are known to be mutual translations. Generally, automatic sentence alignment methods face two kinds of difficulties. First, there are discrepancies between the source document and its translation such as differences in layout, omissions, etc. Sentence alignment programs must be designed to manage these cases. Second, sometimes alignment is a nontrivial matter: some pairs of sentences are even difficult for humans to make decisions. There are many situations where alignments are preferable. One benefit of this application is that it can obtain translation lexicon for particular domains or genres. We also find another benefit that it can provide a new way to get corresponding summaries for bilingual documents if you already get the monolingual summaries.

Many methods of automatic alignment have already been reported in the literature. Most existing approaches fall into one of three categories: namely, the lexical approach, the length-based statistical approach, and the combination of them. The lexical approach as in the works of Kay and Roscheisen(1991, 1993) and of Hwang and Nagao(1994), finds relationships between lexical contents of the bilingual documents in order to get alignment pairs. The statistical approach as presented in the methods by Brown et al.(1991) and by Gale and Church(1991, 1993), on the other hand, uses statistical correlation between sentence lengths of the bilingual documents as the basis of matching. Finally, other methods such as those demonstrated by Simard et al.(1992), Wu(1994), Tan & Nagao(1995) and Langlais(1997) combine both approaches to align documents.

Our alignment work is based on a statistical model of character lengths that is borrowed from Gale and Church (1991). Gale algorithm used a statistical modelization of translations that only considered the length of the sentence segments and was depended on a dynamic programming scheme to find the best sentence alignment pairs. The statistical approach to alignment of Gale algorithm can be briefly summarized as follows:

Gale used an estimate of $-\log Prob(match|\delta)$ as the distance measure to compare two individual elements, where δ depends on l_1 and l_2 , the lengths of the two portions of text under consideration. This distance measure is based on the assumption that each character in one language, L_1 , gives rise to a random number of characters in the other language, L_2 . Given the assumption that these random variables are independent and identically distributed with a normal distribution. The model is then specified by the mean, c , and variance, s^2 , of this distribution. c is the expected number of characters in L_2 per character in L_1 , and s^2 is the variance of the number of characters in L_2 per character in L_1 . δ can be defined to be $(l_2 - l_1c)/\sqrt{l_1s^2}$ so that it has a normal distribution with mean zero

and variance one. Using Bayes's Rule, $Prob(match|\delta)$ can be computed as a constant times $Prob(\delta|match)Prob(match)$. The conditional probability $Prob(\delta|match)$ can be estimated by

$$Prob(\delta|match) = 2(1 - Prob(|\delta|)) \quad (1)$$

where $Prob(|\delta|)$ is the probability that a random variable, z , with a standardized normal distribution, has magnitude at least as large as $|\delta|$. That is, where

$$Prob(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz \quad (2)$$

The program computes δ directly from the lengths of the two portions of text, l_1 and l_2 , and the two parameters, c and s^2 . That is, $\delta = (l_2 - l_1c)/\sqrt{l_1s^2}$. Then, $Prob(\delta)$ is computed by integrating a standard normal distribution.

Gale algorithm has been tested on *Canadian Hansard Bilingual Corpus* where documents are kept in full translation in both English and French. Our project use a Chinese-English bilingual corpus, *Hong Kong News Corpus* developed by LDC and Hong Kong Government. Although it was said that length-based methods were language independent, parameter adjustment is still needed if applying the Gale algorithm to Chinese-English bilingual corpus. In Chinese, document consists of an unsegmented character stream without marked word boundaries, it would not be possible to count the number of words in a sentence without first parsing it. Therefore our methods should first segment Chinese characters, then separate Chinese document by sentence units. On the other hand, we only need to segment sentences rather than characters in English documents for alignment experiments. However, before the extended Gale algorithm can be applied to Chinese, it is important to define the length of Chinese characters, because generally Chinese document contain Chinese characters and punctuation and English characters. Our method is to count each English character or punctuation as length one and each Chinese character or punctuation as length two. Gale algorithm gave two parameters, mean $c = 1.06$ as the expected number of French Characters per English character and $s^2 = 5.6$ as the variance of the number of characters in French per character in English. We used a statistical approach to determine the appropriate values for mean c , variance s^2 and other parameters used in algorithm. Using the training corpus randomly selected from original corpus, first we manually align all these Chinese and English sentence pairs then compute the length rate between Chinese and English characters. Finally, we get the mean $c = 0.6322$ as the expected number of Chinese Characters per English character and $s^2 = 0.175$ as the variance of the number of characters in Chinese per character in English.

Simultaneously, we observe the news corpus and obtain some characteristics from it. Each news document consists of several paragraphs which are separated by delimiters. Each paragraph can be divided into sentences using the punctuation, such as “!?” in Chinese and “!?” in English. Obviously, the punctuation “.” doesn’t definitely mean the end of English documents. So our sentence segmenter for English document has been intergrated several methods to deal with almost all those cases. For instance, we check whether the characters before and after “.”(“,”) are both digits to decide whether “.”(“,”) is a part of a number. Also the abbreviations with a “.” such as “U.S.” or “Mr.” can be picked out by a special dictionary look-up. After we get sentence segment of Chinese and English documents, it is time for us to begin the alignment phrase. Our alignment work includes two levels. The first is to align paragraphs for each pair of English and Chinese files, and the second is to align sentences for each pair of aligned English and Chinese paragraphs. This hierarchical alignment method reduces the recursion depth of dynamic programming and makes the algorithm faster than directly aligning at the sentence level for the entire documents.

For simplicity, our work only considers six alignment classes, which include 0 – 1, 1 – 0, 1 – 1, 2 – 1, 1 – 2 and 2 – 2. For instance, 1 – 2 represents one Chinese sentence is aligned with two corresponding English sentences. Table 1 shows the probabilities which are used by Gale algorithm.

Table 1: Prob(match)

Category	Frequency	Prob(match)
1-1	1167	0.89
1-0 or 0-1	13	0.0099
2-1 or 1-2	117	0.089
2-2	15	0.011
	1312	1.00

We test the alignment program on the whole *Hong Kong News Corpus* and Table 2 shows the length-based alignment statistical results from the output of the program.

Table 2: Alignment statistical results

	Number of Pairs	Percentage
0 \Rightarrow 1	357	0.13%
1 \Rightarrow 0	751	0.28%
1 \Rightarrow 1	215,296	81.65%
1 \Rightarrow 2	17,412	6.60%
2 \Rightarrow 1	29,056	11.04%
2 \Rightarrow 2	801	0.30%

Total sentence number: 312,544

To date we have finished aligning all documents in the *Hong Kong News Corpus* and the alignment program produced 312,544 confident pairs of English and Chinese sentence alignments. Among those alignment, 81.65% sentence pairs belong to class 1 – 1, while only 0.30% to class 2 – 2. We have checked 20 randomly chosen document pairs which contain 263 sentence pairs and found that the precision and recall was 95.5% and 95.5% respectively. Without loss of generality, we can assume that the alignment matching quality is very good, so we can depend on the alignment results to do some related research.

Future Work

This alignment method only generate a special-purpose experimental system. We have focused on the news documents collected from *Hong Kong News Corpus*, and the system can only work with these news documents according to their corresponding *XML* format. Besides, the corpora are not large enough. Therefore a likely ensuing work is to collect more parallel news from possibly other resources, and modularize the programs so that they can easily deal with other news formats. Another possible extension to this system may integrate some lexical approach and refine the alignment results.

Reference

1. Brown, Peter F., Lai, Jennifer C. and Mercer, Robert L.: 1991, ‘Aligning sentences in parallel corpora’, in *Proceedings of 29th Annual Meetings of the ACL*, pp.169-176.
2. Gale, William A. and Church, Kenneth.W.: 1991, ‘A program for aligning sentences in bilingual corpora’ *Proceedings of 29th Annual Meeting of the ACL*, pp.177-184.
3. Gale, William A. and Church, Kenneth. W.: 1993, ‘A program for aligning sentences in bilingual corpora’ *Computational Linguistics*, vol.19, no.1, pp.75-102.

4. Hwang, Dosam and Nagao, Makoto: 1994, 'Aligning of Japanese and Korean texts by analogy', 94-NL-99, vol. 94, no.9, pp.87-94.
5. Kay, Martin and Roscheisen, Martin: 1993, 'Text-translation alignment,' *Computational Linguistics*, vol.19, no.1, pp.121-142.
6. Langlais, Philippe, 'A System to Align Complex Bilingual Corpora', in TMH-QPSR 4/1997, KTH, Stockholm, Sweden, 1997.
7. Simard, Michel; Foster, George. F. and Isabelle, Pierre: 1992, 'Using Cognates to Align Sentences in Bilingual Corpora', in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pp.67-81, Montreal, Canada.
8. Tan, Chew Lim and Nagao, Makoto: 1995, 'Automatic Alignment of Japanese-Chinese Bilingual Texts', *IEICE Transactions On Information and Systems* Vol.E78-D, No.1.
9. Wu, Dekai: 1994, 'Aligning Parallel English Chinese Text Statistically with Lexical Criteria,' ACL-94