STATISTICAL TEXT ALLINGMENT

Shubham kumar 18030142033

Pratik shukla 18030142049

ABSTRACT:-

This paper describes an accurate and robust text alignment system for structurally different languages. Among structurally different languages such as Japanese and English, there is a limitation on the amount of word correspondences that can be statistically acquired. The proposed method makes use of two kinds of word correspondences in aligning bilingual texts. One is a bilingual dictionary of general use. The other is the word correspondences that are statistically acquired in the alignment p rocess. Our method gradually determines sentence pairs (anchors) that correspond to each other by relaxing parameters. The method, by combining two kinds of word correspondences, achieves adequate word correspondences for complete alignment. As a result, texts of various length and of various genres in structurally different languages can be aligned with high precision. Experimental results show our system outperforms conventional methods for various kinds of English texts.

INTRODUCTION:-

Corpus-based approaches based on bilingual texts are promising for various applications(i.e., lexical knowledge extraction (Kupiec, 1993; Matsumoto et al., 1993; Smadja et al., 1996; Dagan and Church, 1994; Kumano and Hirakawa, 1994; Haruno et al., 1996), machine translation (Brown and others, 1993; Sato and Nagao, 1990; Kaji et al., 1992) and information retrieval (Sato, 1992)). Most of these works assume voluminous aligned corpora. Many methods have been proposed to align bilingual corpora. One of the major approaches is based on the statistics of simple features such as sentence length in words (Brown and others, 1991) or in characters (Gale and Church, 1993). These techniques are widely used because they can be implemented in an efficient and simple way through dynamic programing. However, their main targets are rigid translations that are almost literal translations. In addition, the texts being aligned were structurally similar European languages (i.e., English-French, English-German). The simple-feature based approaches don't work in flexible translations for structurally different languages such as Japanese and English, mainly for the following two reasons. One is the difference in the character types of the two languages. Japanese has three types of characters (Hiragana, Katakana, and Kanji), each of which has different amounts of information. In contrast, English has only one type of characters. The other is the grammatical and rhetorical difference of the two languages. First, the systems of functional (closed) words are quite different from language to language. Japanese has a quite different system of closed words, which greatly influence the length of simple features. Second, due to rhetorical difference, the number of multiple match (i.e., 1-2, 1-3, 2-1 and so on) is more than that among European languages. Thus, it is impossible in general to apply the simple-feature based methods to Japanese-English translations. One alternative alignment method is the lexiconbased approach that makes use of the wordcorrespondence knowledge of the two languages. (Church, 1993) employed n-grams shared by two languages. His method is also effective for JapaneseEnglish computer manuals both containing lots of the same alphabetic technical terms. However, the method cannot be applied to general translations in structurally different languages. (Kay and Roscheisen, 1993) proposed a relaxation method to iteratively align bilingual texts using the word correspondences acquired during the alignment process. Although the method works well among European languages, the method does not work in aligning structurally different languages. In JapaneseEnglish translations, the method does not capture enough word correspondences to permit alignment. As a result, it can align only some of the two texts. This is mainly because the syntax and rhetoric are greatly differ in the two languages even in literal translations. The number of confident word correspondences of words is not enough for complete alignment. Thus, the problem cannot be addressed as long as the method relies only on statistics. Other methods in the lexicon-based approach embed lexical knowledge into stochastic models (Wu, 1994; Chen, 1993), but these methods were tested using rigid translations. To tackle the problem, we describe in this paper a text alignment system that uses both statistics and bilingual dictionaries at the same time. Bilingual dictionaries are now widely available on-line due to advances in CD-ROM technologies. For example, English-Spanish, English-

French, English-German, English-Japanese, Japanese-French, Japanese-Chinese and other dictionaries are now commercially available. It is reasonable to make use of these dictionaries in bilingual text alignment. The pros and cons of statistics and online dictionaries are discussed below. They show that statistics and on-line dictionaries are complementary in terms of bilingual text alignment. Statistics Merit Statistics is robust in the sense that it can extract context-dependent usage of words and that it works well even if word segmentation 1 is not correct. Statistics Demerit The amount of word correspondences acquired by statistics is not enough for complete alignment. Dictionaries Merit They can contain the information about words that appear only once in the corpus. Dictionaries Demerit They cannot capture context-dependent keywords in the corpus and are weak against incorrect word segmentation. Entries in the dictionaries differ from author to author and are not always the same as those in the corpus. Our system iteratively aligns sentences by using statistical and on-line dictionary word correspondences. The characteristics of the system are as follows. • The system performs well and is robust for various lengths (especially short) and various genres of texts. • The system is very economical because it assumes only online-dictionaries of general use and doesn't require the labor-intensive construction of domain-specific dictionaries. • The system is extendable by registering statistically acquired word correspondences into user dictionaries.

SYSTEM OVERVIEW:-

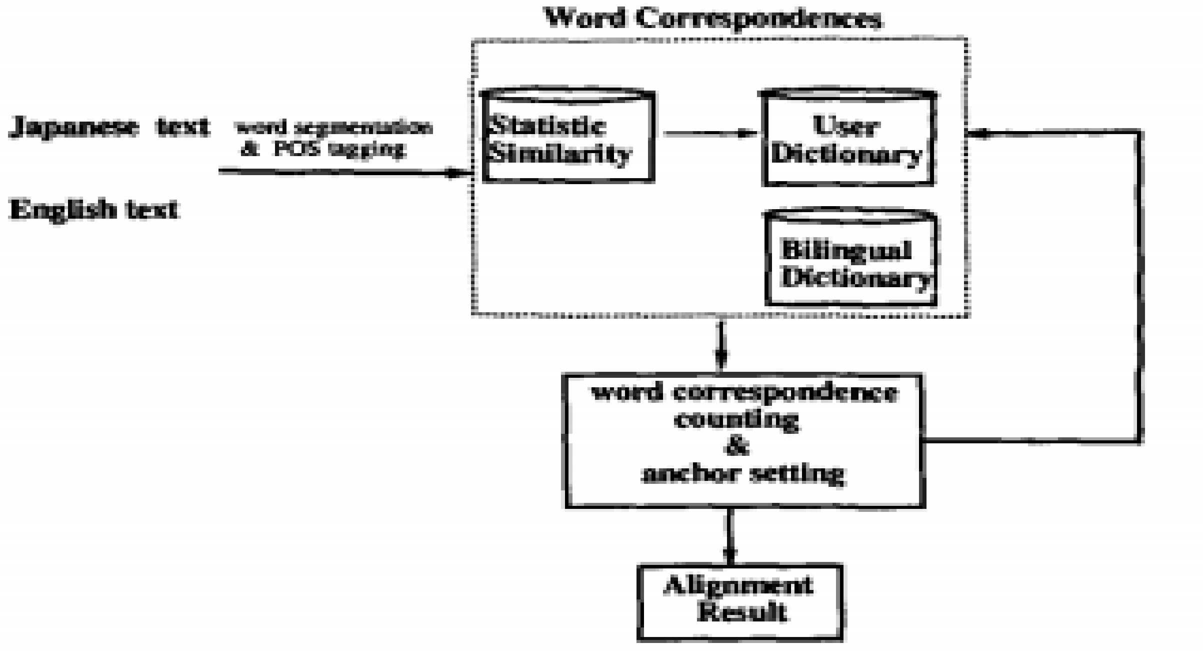


Figure 1: Overview of the Alignment System

Figure 1 overviews our alignment system. The input to the system is a pair of Japanese and English texts, one the translation of the other. First, sentence boundaries are found in both texts using finite state transducers. The texts are then partof-speech (POS) tagged and separated into original form words z. Original forms of English words are determined by 80 rules using the POS information. From the word sequences, we extract only nouns, adjectives, adverbs verbs and unknown words (only in Japanese) because Japanese and English closed words are different and impede text alignment. These pre-processing operation can be easily implemented with regular expressions.

ALGORITHMS:-

STATISTICS USED:- In this section, we describe the statistics used to decide word correspondences. From many similarity metrics applicable to the task, we choose mutual information and t-score because the relaxation of parameters can be controlled in a sophisticated manner. Mutual information represents the similarity on the occurrence distribution and t-score represents the confidence of the similarity. These two parameters permit more effective relaxation than the single parameter used in conventional methods(Kay and Roscheisen, 1993). Our basic data structure is the alignable sentence matrix (ASM) and the anchor matrix (AM). ASM represents possible sentence correspondences and consists of ones and zeros. A one in ASM indicates the intersection of the column and row constitutes a possible sentence correspondence. On the contrary, AM is introduced to represent how a sentence pair is supported by word correspondences. The i-j Element of AM indicates how many times the corresponding words appear in the i-j sentence pair. As alignment proceeds, the number of ones in ASM reduces, while the elements of AM increase. Let pi be a sentence set comprising the ith Japanese sentence and its possible English correspondences as depicted in Figure 3. For example, P2 is the set comprising Jsentence2, Esentence2 and Esentencej, which means Jsentence2 has the possibility of aligning with Esentence2 or Esentencej. The pis can be directly derived from ASM.

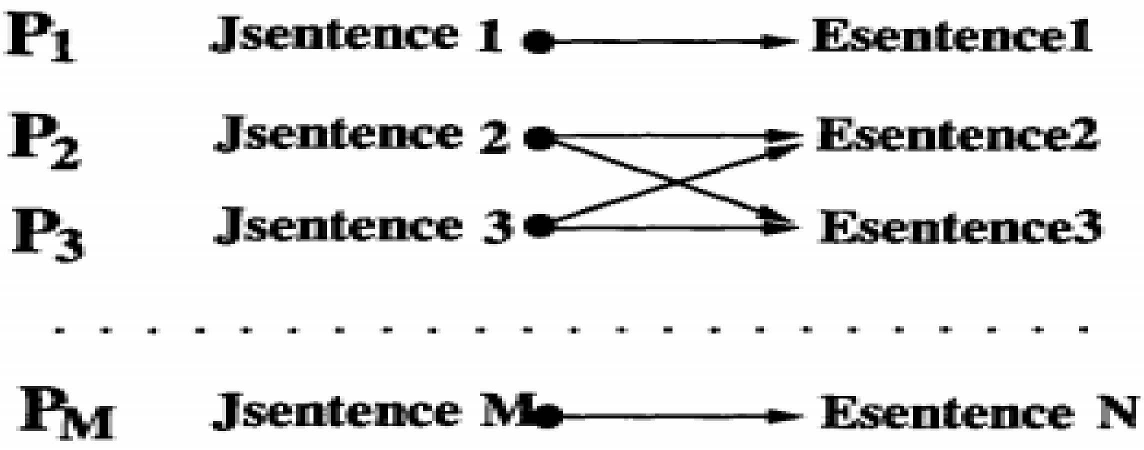


Figure 3: Possible Sentence Correspondences

BASIC ALGORITHMS:- Our basic algorithm is an iterative adjustment of the Anchor Matrix (AM) using the Alignable Sentence Matrix (ASM). Given an ASM, mutual information and t-score are computed for all word pairs in possible sentence correspondences. A word combination exceeding a predefined threshold is judged as a word correspondence. In order to find new anchors, we combine these statistical word correspondences with the word correspondences in a bilingual dictionary. Each element of AM, which represents a sentence pair, is updated by adding the number of word correspondences in the sentence pair. A sentence pair containing more than a predefined number of corresponding words is determined to be a new anchor. The detailed algorithm is as follows.

CONSTRUCTING INITIAL ASM:- This step constructs the initial ASM. If the texts contain M and N sentences respectively, the ASM is an M x N matrix. First, we decide a set of anchors using article boundaries, section boundaries and so on. In the most general case, initial anchors are the first and last sentences of both texts as depicted in Figure 2.

Next, possible sentence correspondences are generated. Intuitively, true correspondences are close to the diagonal linking the two anchors. We construct the initial ASM using such a function that pairs sentences near the middle of the two anchors with as many as O(~/~) (L is the number of sentences existing between two anchors) sentences in the other text because the maximum deviation can be stochastically modeled as O(~rL) (Kay and Roscheisen, 1993). The initial ASM has little effect on the alignment performance so long as it contains all correct sentence correspondences.

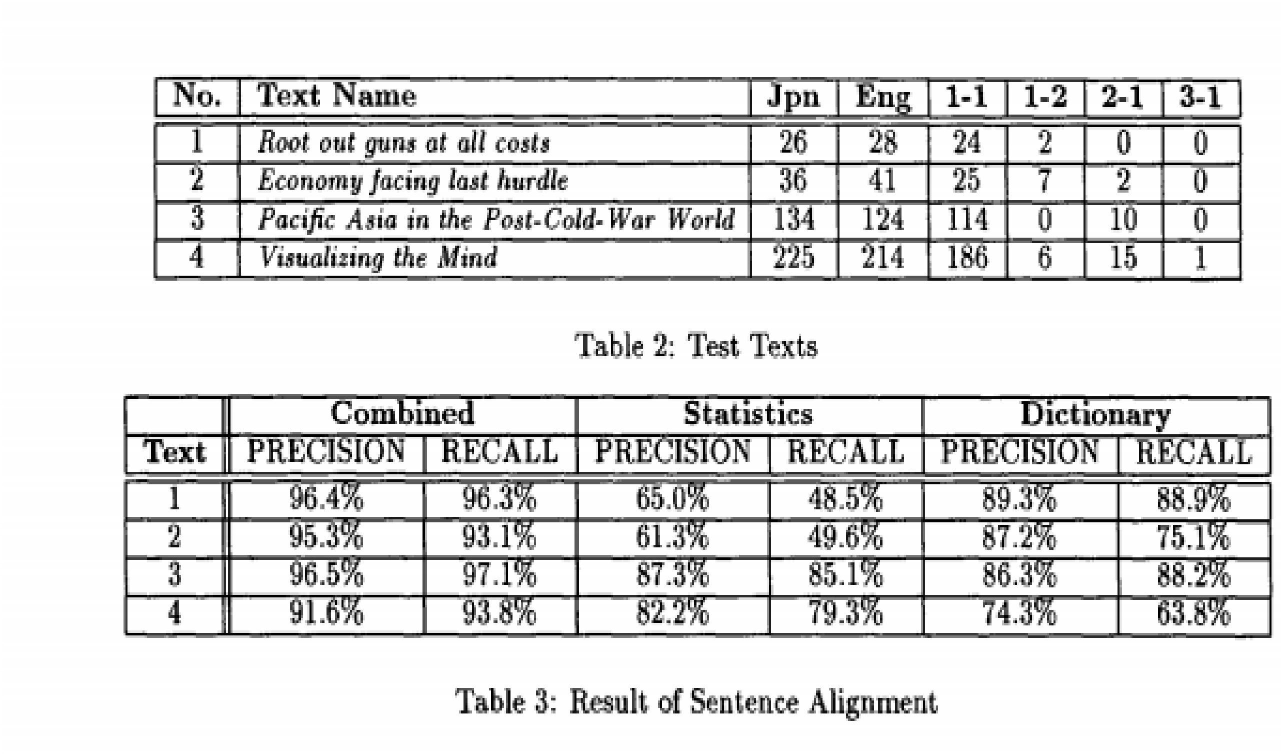
<u>CONSTRUCTING AM:-</u> d a bilingual dictionary. Let thigh, tlow, Ihigh and Izow be two thresholds for t-score and two thresholds for mutual information, respectively. Let ANC be the minimal number of corresponding words for a sentence pair to be judged as an anchor. First, mutual information and t-score are computed for all word pairs appearing in a possible sentence correspondence in ASM. We use hereafter the word correspondences whose mutual information exceeds Itow and whose t-score exceeds ttow. For all possible sentence correspondences Jsentencei and Esentencej (any pair in ASM), the following operations are applied in order. 1. If the following three conditions hold, add 3 to the i-j element of AM. (1) Jsentencei and Esentencej contain a bilingual dictionary word correspondence (wjpn and w,ng). (2) w~na does not occur in any other English sentence that is a possible translation of Jsentencei. (3) Jsentencei and Esentencej do not cross any sentence pair that has more than ANC word correspondences. 2. If the following three conditions hold, add 3 to the i-j element of AM. (1) Jsentencei and Esentencej contain a stochastic word correspondences. 3. If the following three conditions hold, add 1 to the i-j element of AM. (1) Jsentencei and Esentencej contain a stochastic word correspondence (wjp~ and we~g) that has mutual correspondence (wjpn and w~na) that has mutual information Ihig h and whose t-score exceeds thigh. (2) w~g does not occur in any other English sentence that is a possible translation of Jsentencei. (3) Jsentencei and Esentencej do not cross any sentence pair that has more than ANC word

ADJUSTING ASM:- This step adjusts ASM using the AM constructed by the above operations. The sentence pairs that have at least ANC word correspondences are determined to be new anchors. By using the new set of anchors, a new ASM is constructed using the same method as used for initial ASM construction. Our algorithm implements a kind of relaxation by gradually reducing flow, Izow and ANC, which enables us to find confident sentence correspondences first. As a result, our method is more robust than dynamic programing techniques against the shortage of word-correspondence knowledge.

EXPERIMENTAL RESULTS:- In this section, we report the result of experiments on aligning sentences in bilingual texts and on statistically acquired word correspondences. The texts for the experiment varied in length and genres as summarized in Table 2. Texts 1 and 2 are editorials taken from 'Yomiuri Shinbun' and its English version 'Daily Yomiuri'. This data was distributed electrically via a WWW server 4. The first two texts clarify the systems's performance on shorter texts. Text 3 is an essay on economics taken from a quarterly publication of The

International House of Japan. Text 4 is a scientific survey on brain science taken from 'Scientific American' and its Japanese version 'Nikkei Science '5. Jpn and Eng in Table2 represent the number of sentences in the Japanese and English texts respectively. The remaining table entries show categories of matches by manual alignment and indicate the difficulty of the task. Our evaluation focuses on much smaller texts than those used in other study(Brown and others, 1993; Gale and Church, 1993; Wu, 1994; Fung, 1995; Kay and Roscheisen, 1993) because our main targets are well-separated articles. However, our method will work on larger and noisy sets too, by using word anchors rather than using sentence boundaries as segment boundaries. In such a case, the method constructing initial ASM needs to be modified. We briefly report here the computation time of our method. Let us consider Text 4 as an example. After 15 seconds for full preprocessing, the first iteration took 25 seconds with tto~ = 1.55 and Izow = 1.8. The rest of the algorithm took 20 seconds in all. This experiment was performed on a SPARC Station 20 Model tIS21. From the result, we may safely say that our method can be applied to voluminous corpora.

SENTENCE ALLINGMENT:- Table 3 shows the performance on sentence alignments for the texts in Table 2. Combined, Statistics and Dictionary represent the methods using both statistics and dictionary, only statistics and only dictionary, respectively. Both Combined and Dictionary use a CD-ROM version of a JapaneseEnglish dictionary containing 40 thousands entries. Statistics repeats the iteration by using statistical corresponding words only. This is identical to Kay's method (Kay and Roscheisen, 1993) except for the statistics used. Dictionary performs the iteration of the algorithm by using corresponding words of the bilingual dictionary. This delineates the coverage of the dictionary. The parameter setting used for each method was the optimum as determined by empirical tests. In Table 3, PRECISION delineates how many of the aligned pairs are correct and RECALL delineates how many of the manual alignments we included in systems output. Unlike conventional sentencechunk based evaluations, our result is measured on the sentence-sentence basis. Let us consider a 3-1 matching. Although conventional evaluations can make only one error from the chunk, three errors may arise by our evaluation. Note that our evaluation is more strict than the conventional one, especially for difficult texts, because they contain more complex matches. For Text 1 and Text 2, both the combined method and the dictionary method perform much better than the statistical method. This is obviously because statistics cannot capture wordcorrespondences in the case of short texts. Text 3 is easy to align in terms of both the complexity of the alignment and the vocabularies used. All methodsperformed well on this text. For Text 4, Combined and Statistics perform much better than Dictionary. The reason for this is that Text 4 concerns brain science and the bilingual dictionaries of general use did not contain domain specific keywords. On the other hand, the combined and statistical methods well capture the keywords as described in the next section. Note here that Combined performs better than Statistics in the case of longer texts, too. There is clearly a limitation in the amount of word correspondences that can be captured by statistics. In summary, the performance of Combined is better than eitherStatistics or Dictionary for all texts, regardless of text length and the domain.



No.	Text Name	Jpn	Eng	1-1	1-2	2-1	3-1
1	*Root out guns at all costs*	26	28	24	2	0	0
2	*Economy facing last hurdle*	36	41	25	7	2	0
3	*Pacific Asia in the Post-Cold-War World*	134	124	114	0	10	0
4	*Visualizing the Mind*	225	214	186	6	15	1

Table 2: Test Texts

	Combined		Statistics		Dictionary	
Text	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
1	96.4%	96.3%	65.0%	48.5%	89.3%	88.9%
2	95.3%	93.1%	61.3%	49.6%	87.2%	75.1%
3	96.5%	97.1%	87.3%	85.1%	86.3%	88.2%
4	91.6%	93.8%	82.2%	79.3%	74.3%	63.8%

Table 3: Result of Sentence Alignment

CONCLUSION:- We have described a text alignment method for structurally different languages. Our iterative method uses two kinds of word correspondences at the same time: word correspondences acquired by statistics and those of a bilingual dictionary. By combining these two types of word correspondences, the method covers both domain specific keywords not included in the dictionary and the infrequent words not detected by statistics. As a result, our method outperforms conventional methods for texts of different lengths and different domains.

REFERENCES:-

S F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In the 31st Annual Meeting of ACL, pages 9-16. K W Church. 1993.

Char_align: A program for aligning parallel texts at the character level. In the 31st Annual Meeting of ACL, pages 1-8

. Ido Dagan and Ken Church. 1994. Termight: identifying and translating technical terminology. In Proc. Fourth Conference on Apolied Natural Language Processing, pages 34-40.

Pascale Fung and K W Church. 1994. K-vec: A new approach for aligning parallel texts. In Proc. 15th COLING, pages 1096-1102.

Pascale Fung. 1995. A pattern matching method for finding noun and proper nouns translations from noisy parallel corpora. In Proc. 33rd ACL, pages 236-243

. W A Gale and K W Church. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75-102, March

. Masahiko Haruno, Satoru Ikehara, and Takefumi Yamazaki. 1996

. Learning Bilingual Collocations by Word-Level Sorting,. In Proc. 16th COLING. Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. 1992.

Learning translation templates from bilingaul text. In Proc. 14th COLING, pages 672-678. Martin Kay and Martin Roscheisen. 1993

. Texttranslation alignment. Computational Linguistics, 19(1):121-142, March. Akira Kumano and Hideki Hirakawa.

1994. Building an MT dictionary from parallel texts based on linguisitic and statistical information. In Proc. 15th COLING, pages 76-81.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In the 31st Annual Meeting of A CL, pages 17-22.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer juman. In Proc.