An Experiment in Hybrid Dictionary and Statistical Sentence Alignment

Nigel Collier, Kenji Ono and **Hideki Hirakawa**
Communication and Information Systems Laboratories
Research and Development Center, Toshiba Corporation
1 Komukai Toshiba-cho, Kawasaki-shi, Kanagawa 210-8582, Japan
{nigel,ono,hirakawa}@eel.rdc.toshiba.co.jp

Abstract

The task of aligning sentences in parallel corpora of two languages has been well studied using pure statistical or linguistic models. We developed a linguistic method based on lexical matching with a bilingual dictionary and two statistical methods based on sentence length ratios and sentence offset probabilities. This paper seeks to further our knowledge of the alignment task by comparing the performance of the alignment models when used separately and together, i.e. as a hybrid system. Our results show that for our English-Japanese corpus of newspaper articles, the hybrid system using lexical matching and sentence length ratios outperforms the pure methods.

1 Introduction

There have been many approaches proposed to solve the problem of aligning corresponding sentences in parallel corpora. With a few notable exceptions however, much of this work has focussed on either corpora containing European language pairs or clean-parallel corpora where there is little reformatting. In our work we have focussed on developing a method for robust matching of English-Japanese sentences, based primarily on lexical matching. The method combines statistical information from byte length ratios. We show in this paper that this hybrid model is more effective than its constituent parts used separately.

The task of sentence alignment is a critical first step in many automatic applications involving the analysis of bilingual texts such as extraction of bilingual vocabulary, extraction of translation templates, word sense disambiguation, word and phrase alignment, and extraction of parameters for statistical translation models. Many software products which aid human translators now contain sentence alignment tools as an aid to speeding up editing and terminology searching.

Various methods have been developed for sentence alignment which we can categorise as either *lexical* such as (Chen, 1993), based on a large-scale bilingual lexicon; *statistical* such as (Brown et al., 1991) (Church, 1993)(Gale and Church, 1993)(Kay and Röshcheisen, 1993), based on distributional regularities of words or byte-length ratios and possibly inducing a bilingual lexicon as a by-product, or *hybrid* such as (Utsuro et al., 1994) (Wu, 1994), based on some combination of the other two. Neither of the pure approaches is entirely satisfactory for the following reasons:

- Text volume limits the usefulness of statistical approaches. We would often like to be able to align small amounts of text, or texts from various domains which do not share the same statistical properties.
- Bilingual dictionary coverage limitations mean that we will often encounter problems establishing a correspondence in non-general domains.
- Dictionary-based approaches are founded on an assumption of lexical correspondence between language pairs. We cannot always rely on this for non-cognate language pairs, such as English and Japanese.
- Texts are often heavily reformatted in translation, so we cannot assume that the corpus will be clean, i.e. contain many one-to-one sentence mappings. In this case statistical methods which rely on structure correspondence such as byte-length ratios may not perform well.

These factors suggest that some hybrid method may give us the best combination of coverage and accuracy when we have a variety of text domains, text sizes and language pairs. In this paper we seek to fill a gap in our understanding and to show how the various components of the hybrid method influence the quality of sentence alignment for Japanese and English newspaper articles.

2 Bilingual Sentence Alignment

The task of sentence alignment is to match corresponding sentences in a text from one language to sentences in a translation of that text in another language. Of particular interest to us is the application to Asian language pairs. Previous studies such as (Fung and Wu, 1994) have commented

that methods developed for Indo-European language pairs using alphabetic characters have not addressed important issues which occur with European-Asian language pairs. For example, the language pairs are unlikely to be cognates, and they may place sentence boundaries at different points in the text. It has also been suggested by (Wu, 1994) that sentence length ratio correlations may arise partly out of historic cognate-based relationships between Indo-European languages. Methods which perform well for Indo-European language pairs have therefore been found to be less effective for non-Indo-European language pairs.

In our experiments the languages we use are English (source) and Japanese (translation). Although in our corpus (described below) we observe that, in general, sentences correspond one-to-one we must also consider multiple sentence correspondences as well as one-to-zero correspondences. These cases are summarised below.

1. 1:1 The sentences match one-to-one.
2. 1:n One English sentence matches to more than one Japanese sentence.
3. m:1 More than one English sentence matches ot one Japanese sentence.
4. m:n More than one English sentence matches to more than one Japanese sentence.
5. m:0 The English sentence/s have no corresponding Japanese sentence.
6. 0:n The Japanese sentence/s have no corresponding English sentence.

In the case of 1:n, m:1 and m:n correspondences, translation has involved some reformatting and the meaning correspondence is no longer solely at the sentence level. Ideally we would like smaller units of text to match because it is easier later on to establish word alignment correspondences. In the worst case of multiple correspondence, the translation is spread across multiple non-consecutive sentences.

3 Corpus

Our primarily motivation is knowledge acquisition for machine translation and consequently we are interested to acquire vocabulary and other bilingual knowledge which will be useful for users of such systems. Recently there has been a move towards Internet page translation and we consider that one interesting domain for users is international news.

The bilingual corpus we use in our experiments is made from Reuter news articles which were translated by the Gakken translation agency from English into Japanese[1] . The translations are quite literal and the contents cover international news for the period February 1995 to December 1996. We currently have over 20,000 articles (approximately 47 Mb). From this corpus we randomly chose 50 article pairs and aligned them by hand using a human bilingual checker to form a judgement set. The judgement set consists of 380 English sentences and 453 Japanese sentences. On average each English article has 8 lines and each Japanese article 9 lines.

The articles themselves form a boundary within which to align constituent sentences. The corpus is quite well behaved. We observe many 1:1 correspondences, but also a large proportion of 1:2 and 1:3 correspondences as well as reorderings. Omissions seem to be quite rare, so we didn't see many m:0 or 0:n correspondences.

An example news article is shown in Figure 1 which highlights several interesting points. Although the news article texts are clean and in machine-tractable format we still found that it was a significant challenge to reliably identify sentence boundaries. A simple illustration of this is shown by the first Japanese line J1 which usually corresponds to the first two English lines E1 and E2. This is a result of our general-purpose sentence segmentation algorithm which has difficulty separating the Japanese title from the first sentence.

Sentences usually corresponded linearly in our corpus, with few reorderings, so the major challenge was to identify multiple correspondences and zero correspondences. We can see an example of a zero correspondence as E5 has no translation in the Japanese text. A 1:n correspondence is shown by E7 aligning to both J5 and J6.

4 Alignment Models

In our investigation we examined the performance of three different matching models (lexical matching, byte-length ratios and offset probabilities). The basic models incorporate dynamic programming to find the least cost alignment path over the set of English and Japanese sentences. Cost being determined by the model's scores. The alignment space includes all possible combinations of multiple matches upto and including 3:3 alignments. The basic models are now outlined below.

4.1 Model 1: Lexical vector matching

The lexical approach is perhaps the most robust for aligning texts in cognate language pairs, or where there is a large amount of reformatting in translation. It has also been shown to be particularly successful within the vector space model in multilingual information retrieval tasks, e.g. (Collier et al., 1998a),(Collier et al., 1998b), for aligning texts in non-cognate languages at the article level.

The major limitation with lexical matching is clearly the assumption of lexical correspondence -

[1] The corpus was generously made available to us by special arrangement with Gakken

E1. Taiwan ruling party sees power struggle in China
E2. TAIPEI , Feb 9 (Reuter) - Taiwan's ruling Nationalist Party said a struggle to succeed Deng Xiaoping as China's most powerful man may have already begun.
E3. "Once Deng Xiaoping dies, a high tier power struggle among the Chinese communists is inevitable," a Nationalist Party report said.
E4. China and Taiwan have been rivals since the Nationalists lost the Chinese civil war in 1949 and fled to Taiwan.
E5. Both Beijing and Taipei sometimes portray each other in an unfavourable light.
E6. The report said that the position of Deng's chosen successor, President Jiang Zemin, may have been subtly undermined of late.
E7. It based its opinion on the fact that two heavyweight political figures have recently used the phrase the "solid central collective leadership and its core" instead of the accepted "collective leadership centred on Jiang Zemin" to describe the current leadership structure.
E8. "Such a sensitive statement should not be an unintentional mistake ...
E9. Does this mean the power struggle has gradually surfaced while Deng Xiaoping is still alive ?," said the report , distributed to journalists.
E10. "At least the information sends a warning signal that the 'core of Jiang' has encountered some subtle changes," it added .

J1. 台湾国民党報告書、中国共産党指導部の権力争い表面化を指摘 ［台北 9日 ロイター］ 台湾の与党、国民党は 、中国の最高実力者、トウ小平氏の後継者争いが、すでに始まっている可能性が高い、と指摘した。
J2. 同党は、記者らに配布した報告書のなかで、「トウ小平氏の死後、中国共産党の指導部で、権力争いが起きるのは確 実だろう」と述べた。
J3. 中国と台湾は、国民党が、１９４９年に中国共産党との内戦に敗れて、台湾に移って以来、敵対関係にある。
J4. 報告書によると、トウ氏の後継者に選ばれている江沢民国家主席の立場が、最近、微妙に弱まった可能性が高い。
J5. 同党の見解は、中国政府の実力者２人が、最近の発言のなかで、現在の指導部の構造を、"堅実な中央集団指導制と その核心 " という言葉を使ったことに基づいているという。
J6. これまでは、" 江主席を核とする中央集団指導制 " という言葉が使われてきた。
J7. 報告書では、「このような微妙な発言で、無意識に誤ったとは考えられない。
J8. これは、トウ小平氏がまだ生存しているのに、徐々に権力争いが表面化してきたことを示しているのではないか。
J9. 少なくとも、" 江主席を核とする " という部分に、微妙な変化があったことは確かだ」と述べている。

Figure 1: Example English-Japanese news article pair

which is particularly weak for English and Asian language pairs where structural and semantic differences mean that transfer often occurs at a level above the lexicon. This is a motivation for incorporating statistics into the alignment process, but in the initial stage we wanted to treat pure lexical matching as our baseline performance.

We translated each Japanese sentence into English using dictionary term lookup. Each Japanese content word was assigned a list of possible English translations and these were used to match against the normalised English words in the English sentences. For an English text segment E and the English term list produced from a Japanese text segment J, which we considered to be a possible unit of correspondence, we calculated similarity using Dice's coefficient score shown in Equation 1. This rather simple measure captures frequency, but not positional information. The weights of words are their frequencies inside a sentence.

$$Dice(E,J) = \frac{2f_{EJ}}{f_E + f_J} \quad (1)$$

where f_{EJ} is the number of lexical items which match in E and J, f_E is the number of lexical items in E and f_J is the number of lexical items in J. The translation lists for each Japanese word are used disjunctively, so if one word in the list matches then we do not consider the other terms in the list. In this way we maintain term independence.

Our transfer dictionary contained some 79,000 English words in full form together with the list of translations in Japanese. Of these English words some 14,000 were proper nouns which were directly relevant to the vocabulary typically found in international news stories. Additionally we perform lexical normalisation before calculating the matching score and remove function words with a stop list.

4.2 Model 2: Byte-length ratios

For Asian language pairs we cannot rely entirely on dictionary term matching. Moreover, algorithms which rely on matching cognates cannot be applied easily to English and some Asian language. We were motivated by statistical alignment models such as (Gale and Church, 1991) to investigate whether byte-length probabilities could improve or replace the lexical matching based method. The underlying assumption is that characters in an English sentence are responsible for generating some fraction of each character in the corresponding Japanese sentence.

We derived a probability density function by making the assumption that English and Japanese sentence length ratios are normally distributed. The parameters required for the model are the mean, μ and variance, σ, which we calculated from a training set of 450 hand-aligned sentences. These are then entered into Equation 2 to find the probability of any two sentences (or combinations of sentences for multiple alignments) being in an alignment relation given that they have a length ratio of x.

$$F(x) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right] e^{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2} \qquad (2)$$

The byte length ratios were calculated as the length of the Japanese text segment divided by the length of the English text segment. So in this way we can incorporate multiple sentence correspondences into our model. Byte lengths for English sentences are calculated according to the number of non-white space characters, with a weighting of 1 for each valid character including punctuation. For the Japanese text we counted 2 for each non-white space character. White spaces were treated as having length 0. The ratios for the training set are shown as a histogram in Figure 2 and seem to support the assumption of a normal distribution.

The resulting normal curve with $\sigma = 0.33$ and $\mu = 0.76$ is given in Figure 3, and this can then be used to provide a probability score for any English and Japanese sentence being aligned in the Reuters' corpus.

Clearly it is not enough simply to assume that our sentence pair lengths follow the normal distribution. We tested this assumption using a standard test, by plotting the ordered ratio scores against the values calculated for the normal curve in Figure 3. If the

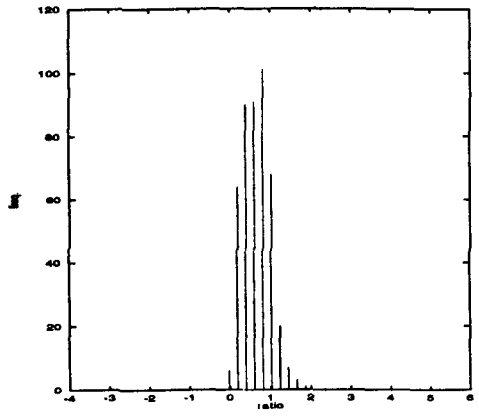

Figure 2: Sentence length ratios in training set

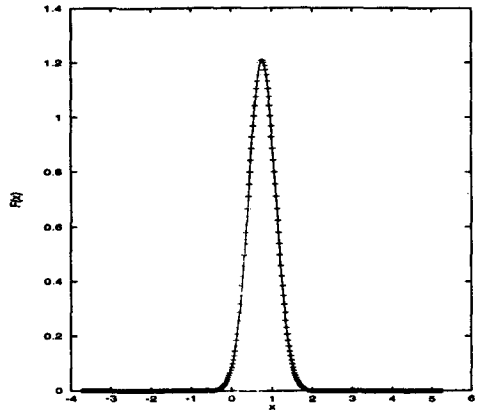

Figure 3: Sentence length ratio normal curve

distribution is indeed normal then we would expect the plot in Figure 4 to yield a straight line. We can see that this is the case for most, although not all, of the observed scores.

Although the curve in Figure 4 shows that our training set deviated from the normal distribution at

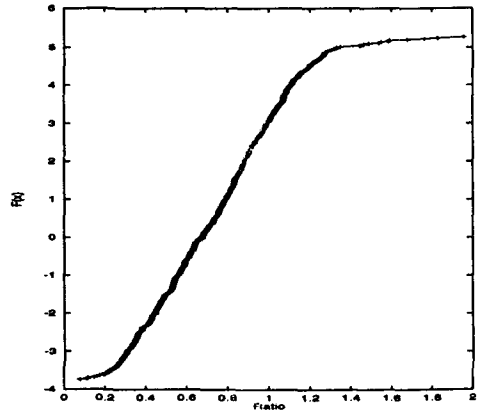

Figure 4: Sentence length ratio normal check curve

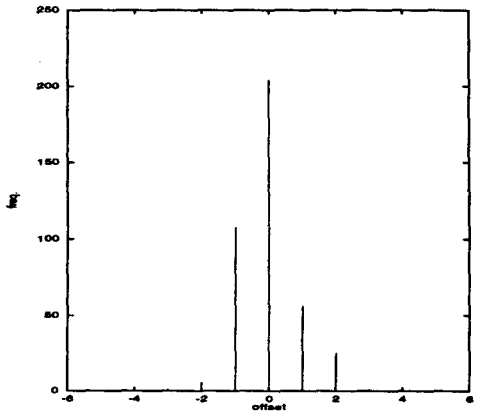

Figure 5: Sentence offsets in training set

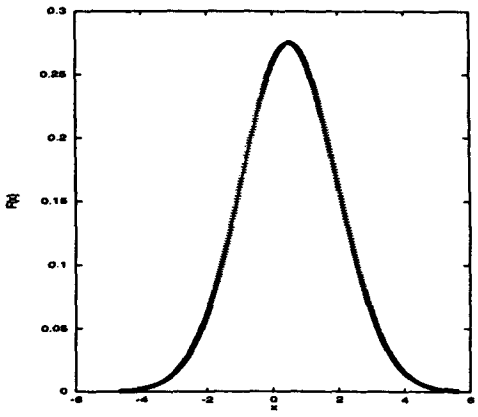

Figure 6: Sentence offsets normal curve

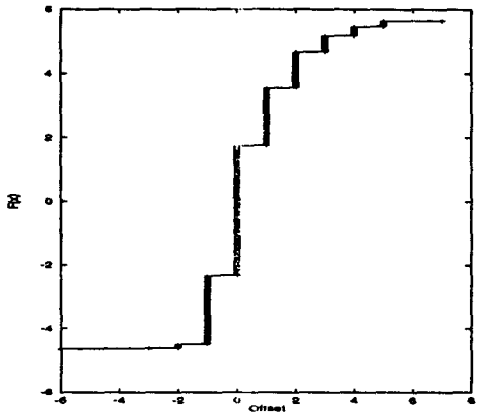

Figure 7: Sentence offsets normal check curve

the extremes we nevertheless proceeded to continue with our simulations using this model considering that the deviations occured at the extreme ends of the distribution where relatively few samples were found. The weakness of this assumption however does add extra evidence to doubts which have been raised, e.g. (Wu, 1994), about whether the byte-length model by itself can perform well.

4.3 Model 3: Offset ratios

We calculated the offsets in the sentence indexes for English and Japanese sentences in an alignment relation in the hand-aligned training set. An offset difference was calculated as the Japanese sentence index minus the English sentence index within a bilingual news article pair. The values are shown as a histogram in Figure 5.

As with the byte-length ratio model, we started from an assumption that sentence correspondence offsets were normally distributed. We then calculated the mean and variance for our sample set shown in Figure 5 and used this to form a normal probability density function (where $\sigma = 0.50$ and $\mu = 1.45$) shown in Figure 6.

The test for normality of the distribution is the same as for byte-length ratios and is given in Figure 7. We can see that the assumption of normality is particularly weak for the offset distribution, but we are motivated to see whether such a noisy probability model can improve alignment results.

5 Experiments

In this section we present the results of using different combinations of the three basic methods. We combined the basic methods to make hybrid models simply by taking the product of the scores for the models given above. Although this is simplistic we felt that in the first stage of our investigation it was better to give equal weight to each method.

The seven methods we tested are coded as follows: *DICE*: sentence alignment using bilingual dictionary and Dice's coefficient scores; *LEN*: sentence alignment using sentence length ratios; *OFFSET*: sentence alignment using offset probabilities.

We performed sentence alignment on our test set of 380 English sentences and 453 Japanese sentences. The results are shown as recall and precision which we define in the usual way as follows:

$$recall = \frac{\#\text{correctly matched sentences retrieved}}{\#\text{matched sentences in the test collection}} \quad (3)$$

$$precision = \frac{\#\text{correctly matched sentences retrieved}}{\#\text{ matched sentences retrieved}} \quad (4)$$

The results are shown in Table 1. We see that the baseline method using lexical matching with a bilingual lexicon, *DICE*, performs better than either of the two statistical methods *LEN* or *OFFSET* used separately. Offset probabilities in particular performed poorly showing that we cannot expect the correctly matching sentence to appear constantly in

the same highest probability position.

Method	Rec. (%)	Pr. (%)
DICE (baseline)	84	85
LEN	82	83
OFFSET	50	57
LEN+OFFSET	70	70
DICE+LEN	89	87
DICE+OFFSET	80	80
DICE+LEN+OFFSET	88	85

Table 1: Sentence alignment results as recall and precision.

Considering the hybrid methods, we see significantly that *DICE+LEN* provides a clearly better result for both recall and precision to either *DICE* or *LEN* used separately. On inspection we found that *DICE* by itself could not distinguish clearly between many candidate sentences. This occured for two reasons.

1. As a result of the limited domain in which news articles report, there was a strong lexical overlap between candidate sentences in a news article.
2. Secondly, where the lexical overlap was poor between the English sentence and the Japanese translation, this leads to low *DICE* scores.

The second reason can be attributed to low coverage in the bilingual lexicon with the domain of the news articles. If we had set a minimum threshold limit for overlap frequency then we would have ruled out many correct matches which were found. In both cases *LEN* provides a decisive clue and enables us to find the correct result more reliably. Furthermore, we found that *LEN* was particularly effective at identifying multi-sentence correspondences compared to *DICE*, possibly because some sentences are very small and provide weak evidence for lexical matching, whereas when they are combined with neighbours they provide significant evidence for the *LEN* model.

Using all methods together however in *DICE+LEN+OFFSET* seems less promising and we believe that the offset probabilities are not a reliable model. Possibly this is due to lack of data in the training stage when we calculated σ and μ, or the data set may not in fact be normally distributed as indicated by Figure 7.

Finally, we noticed that a consistent factor in the English and Japanese text pairs was that the first two lines of the English were always matched to the first line of the Japanese. This was because the English text separated the title and first line, whereas our sentence segmenter could not do this for the Japanese. This factor was consistent for all the 50 article pairs in our test collection and may have led to a small deterioration in the results, so the figures we present are the minimum of what we can expect when sentence segmentation is performed correctly.

6 Conclusion

The assumption that a partial alignment at the word level from lexical correspondences can clearly indicate full sentence alignment is flawed when the texts contain many sentences with similar vocabulary. This is the case with the news stories used in our experiments and even technical vocabulary and proper nouns are not adequate to clearly discriminate between alternative alignment choices because the vocabulary range inside the news article is not large. Moreover, the basic assumption of the lexical approach, that the coverage of the bilingual dictionary is adequate, cannot be relied on if we require robustness. This has shown the need for some hybrid model.

For our corpus of newspaper articles, the hybrid model has been shown to clearly improve sentence alignment results compared with the pure models used separately. In the future we would like to make extensions to the lexical model by incorporating term weighting methods from information retrieval such as inverse document frequency which may help to identify more important terms for matching. In order to test the generalisability of our method we also want to extend our investigation to parallel corpora in other domains.

Acknowledgements

We would like to thank Reuters and Gakken for allowing us to use the corpus of news stories in our work. We are grateful to Miwako Shimazu for hand aligning the judgement set used in the experiments and to Akira Kumano and Satoshi Kinoshita for useful discussions. Finally we would also like express our appreciation to the anonymous reviewers for their helpful comments.

References

P. Brown, J. Lai, and R. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, USA.*

S. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. *31st Annual Meeting of the Association of Computational Linguistics, Ohio, USA*, 22–26 June.

K. Church. 1993. Char_align: a program for aligning parallel texts at the character level. In *31st Annual Meeting of the Association for Computational Linguistics, Ohio, USA*, pages 1–8, 22-26 June.

N. Collier, H. Hirakawa, and A. Kumano. 1998a. Creating a noisy parallel corpus from newswire articles using multi-lingual information retrieval. *Trans. of Information Processing Society of Japan (to appear).*

N. Collier, H. Hirakawa, and A. Kumano. 1998b. Machine translation vs. dictionary term translation - a comparison for English-Japanese news article alignment. In *Proceedings of COLING-ACL'98, University of Montreal, Canada,* 10th August.

P. Fung and D. Wu. 1994. Statistical augmentation of a Chinese machine readable dictionary. In *Second Annual Workshop on Very Large Corpora,* pages 69–85, August.

W. Gale and K. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics (ACL-91), Berkeley, California,* pages 177–184.

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

M. Kay and M. Röshcheisen. 1993. Text-translation alignment. *Computational Linguistics,* 19:121–142.

T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and N. Nagao. 1994. Bilingual text matching using bilingual dictionary and statistics. In *COLING-94, 15th International Conference, Kyoto, Japan,* volume 2, August 5–9.

D. Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA,* pages 80–87, June 27–30.