自然言語処理 120−17
(1997. 7. 25)

多言語情報検索技術を用いた二か国語コーパスの自動アラインメント

Nigel Collier, 熊野 明 , 平川秀樹

(株) 東芝 研究開発センター
〒210 川崎市幸区小向東芝町 1
email: {*nigel,kmn,hirakawa*} *@eel.rdc.toshiba.co.jp*

あらまし

インターネット上の英語・日本語ニュース記事からの対訳コーパス作成に多言語情報検索技術を適用した結果について述べる。記事単位のアラインメントを行う実験において、ベクトル空間モデルに基づいたインプリメントを行い、シミュレーションを通して 6 種類の多言語検索技術の有効性を示す。各方法は、計算効率性がよく、効果を評価しやすいだけでなく、他のジャンルや言語対にも一般化することができるものである。これは、分野を限定しない対訳記事から知識抽出を行う際に重要な点である。実験によると、検索に用いる英単語の語幹化、頻度利用、語彙制限はそれぞれ精度を向上させたが、最も効果のあったのは、記事長による単純な正規化であった。

キーワード　アラインメント、コーパス、MLIR、知識獲得

Automatic Alignment of Bilingual Corpora using Multi-lingual Information Retrieval

Nigel Collier, Akira Kumano, Hideki Hirakawa

Research and Development Center
Toshiba Corporation
1 Komukai Toshiba-cho, Kawasaki-shi
Kanagawa 210, Japan
email: {*nigel,kmn,hirakawa*} *@eel.rdc.toshiba.co.jp*

Abstract

In this paper we present an adaptation of multi-lingual information retrieval for the production of an aligned bilingual corpus from noisy parallel English-Japanese newswire articles. We implement the standard vector space model and show though simulation the effectiveness of six variations for the alignment task. The methods are computationally efficient, easy to evaluate and generalizable to other genres and language pairs - an important factor if we are to use the aligned articles for knowledge acquisition in unrestricted domains. Our results indicate that while stemming, inverse document frequency and lexical filtering all improve the performance, the best overall improvement was due to simple normalization of article length.

key words　alignment, corpus, MLIR, knowledge acquisition

1 Introduction

Our goal in this paper is to show how the term vector translation model in multi-lingual information retrieval (MLIR) can be adapted to match bilingual texts for the production of aligned parallel corpora. Our investigation uses newswire texts in English and Japanese. The methods are intended to be computationally efficient and re-usable, making minimum use of external knowledge sources. The purpose of our research is to use the resulting noisy-parallel corpus for bilingual knowledge acquisition to supplement a general-purpose machine translation system. It is important therefore that the methods can be applied to unrestricted text.

It is becoming increasingly apparent that clean-parallel corpora such as the Canadian/Hong Kong Hansards are very rare, and this limits the applicability of the techniques developed for knowledge extraction from them. Recently, papers (e.g. [Fung and McKeown, 1997]) have appeared on the subject of noisy-parallel corpora where alignment does not often occur on a one-to-one sentence basis. As internet resources become more plentiful we are likely to discover many sources of noisy-parallel texts. However, the task of aligning corresponding units of text is more challenging.

In our work with Reuter bilingual English-Japanese newswire articles we have found that text units correspond poorly at the sentence level due to the heavy reformatting that occurs during translation which includes large omissions, reordering and concatenation of sentences. Consequently, the most appropriate and reliable units of alignment appear to be at the article level. The approach we present in this paper is particularly applicable when there is an absence of language independent annotations with which to establish a bilingual relation.

We have chosen to explore IR techniques for text matching which use the vector space model where the two texts are converted into word frequency vectors. After presenting a summary of the multi-lingual information retrieval task we introduce several standard models for this problem and show their effectiveness through simulations.

2 Task Description

A standard task in information retrieval (IR) is to retrieve documents from a large collection in response to a user's query. These documents are typically scored according to relevance and presented to the user in ranked order. A related task which was first investigated in [Salton, 1970] and more recently under the title *multi-lingual information retrieval* is to enter the query in a different language to the document collection.

In recent MLIR studies using broader language coverage, e.g. [Davis and Dunning, 1995] on the Spanish TREC corpus, a significant difference has been found between the results for monolingual document retrieval and retrieval when the query has been translated from another language. This performance penalty has been linked to the degree of transfer ambiguity, i.e. the number of polysemes and homonyms in the query vector.

The central issue for MLIR, as identified by [Davis, 1996] is whether vector matching methods can succeed given that they essentially exploit linear relations in the query and target document. In other words they rely on term-for-term translations.

3 Implementation

Given a corpus of English source texts and another corpus of Japanese summary translations it was natural to consider the Japanese texts, which are typically only four or five sentences long, as IR queries. The goal of article alignment can be reformulated as an IR task by trying to find the English document(s) in the collection (corpus) of news articles which most closely corresponded to the Japanese query. The overall system is outlined in Figure 1 and discussed below.

3.1 Newswire articles

One of the richest sources of bilingual information for knowledge acquisition may be newswire sources. In particular international news stories, even though produced by different agencies, may be expected to correspond because of the common interest in the events covered.

Our English document collection consists of Reuter daily news articles taken from the inter-

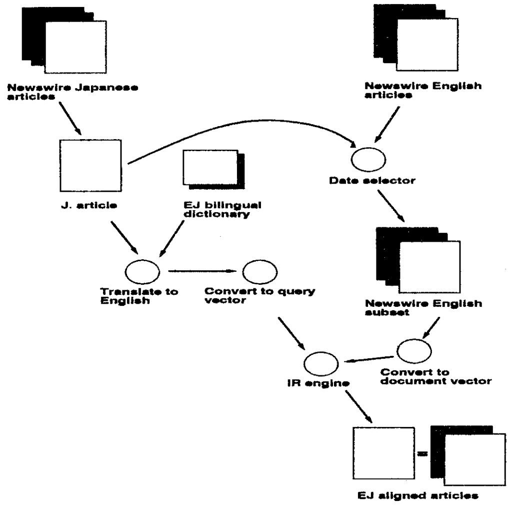

Figure 1: System Overview

net for the 7th December 1996 to the 21st of March 1997. In total we have 4689 English articles with an average of about 45 articles per day. Once pre-processing has taken place to remove hypertext and formatting characters we are left with approximately 26000 paragraphs of English text.

In contrast to the English news articles, the Japanese articles, which are also produced daily by Reuter's, are very short. The Japanese is a translated summary of an English article, but as we mentioned earlier, considerable reformatting has taken place. The 1071 Japanese articles cover the period from the 12th December 1996 to the 14th of March 1997. From this collection we selected a sample for our query set. We discuss the composition of the query and judgement set later.

3.2 Query translation

In order to match English and Japanese documents it was first necessary to translate the Japanese text into English. Query translation using term vector translation is possibly the most simple option available. Full machine translation would undoubtedly reduce the level of transfer ambiguity but it is not a very viable option for general language at this time.

The disadvantage of term vector translation using a bilingual lexicon arises from the shallow level of analysis. This leads to the incorporation of a range of polysemes and homographs which act to reduce the level of matching between the query and its corresponding English document(s). In fact we find that the greater the depth of coverage in the bilingual lexicon, the greater this problem will become. For example, terms with many translations could make the query vector too general, leading to a loss of precision in document retrieval. Furthermore, the most precisely defining terms for a news article are often proper nouns, for which the coverage in our bilingual lexicon is very low.

Another problem is the issue of how we weight the terms in the translated query vector. As we know from statistical machine translation, different terms will not have the same likelihood of translation, therefore how should we incorporate such information in the query vector? Moreover, the weighting of alternative homonyms should also be considered. In this study to assign each homonym term its full weight.

3.3 High frequency words

In most previous work the common practice is to use a stop list of function words which are then used to prevent such words entering the document index. In our simulations we noticed that some common words and even proper nouns had very high frequencies and that they were statistically irrelevant as document discriminators.

We chose not to use a hand built stop list but instead calculated word frequencies and removed the top 0.05% most frequent words from consideration. This resulted in the exclusion of 107 words which included most of the function words. Frequency thresholding also removed common nouns such as 'government', 'cabinet' and 'chief' which appeared in this top 0.05% group due to their scope of usage. Clearly these words would have greater specificity if they were used in phrasal form for example 'chief executive', 'chief secretary' or 'police chief'. The issue of phrase formation is discussed below.

3.4 Term formation

An important issue in indexing a document is term selection. We must decide for example whether to use phrases or single words, stemmed words or surface forms in the index.

There is considerable evidence that the use of phrases in the index gives superior precision compared with the use of single words (see for example the discussion in [Salton, 1989] and [Hull and Grefenstette, 1996]). However, we decided not to use phrases in the index at this stage for several reasons outlined below.

Firstly we wanted to establish a base case for article alignment using single terms. Secondly, the reusability consideration led us to avoid as far as possible the use of external knowledge sources for phrase identification, although this will change as the performance characteristics of tools such as taggers become known. More important was the methodological consideration that we must be able to accurately find translations of phrases, many of which are proper nouns, in the Japanese article using a rather simple bilingual lexicon. We therefore chose to use single word terms rather than phrases.

Our bilingual dictionary which contained 65,000 Japanese words in base form with their English translations returned quite a mixed-bag in the morphological sense. In this situation it seems intuitive that we should normalise words to improve term matching. In the simulations described below we measure the effectiveness of English stemming using the Porter algorithm [Porter, 1980] as one refinement of the basic model.

The result is a balanced compromise between a purely statistical approach and the need to introduce some bilingual knowledge in order to establish a correspondence between English and Japanese articles.

4 Models

Below we present several models which calculate similarity between an English and Japanese news article with increasing sophistication.

Terminology

An index of t terms is generated from the document collection (English corpus) and the query set (Japanese articles). Each document has a description vector $D = (w_{d1}, w_{d2}, .., w_{dt})$ where w_{dk} represents the relevance of term k in document D. There are N documents in the collection, and n_k represents the number of documents in which term k appears. tf_{dk} denotes the term frequency of term k in document D. A query Q is formulated as a query description vector $Q = (w_{q1}, w_{q2}, .., w_{qt})$.

4.1 Model 1: *tf*

Our base model calculates the similarity between Q and D using a simple inner product correlation of term frequencies *tf*.

$$IP(Q, D) = \sum_{k=1}^{t} w_{qk} w_{dk} \qquad (1)$$

where

$$w_{xk} = tf_{xk} \qquad (2)$$

4.2 Model 2: *tf* with document length normalisation

Model 1 produces a score which is in theory unbounded, and in practice only bounded by the size of a document. This is unsatisfactory because longer documents will have an unfair advantage as they (a) have more terms, and (b) the terms have higher frequencies of occurrence. Model 2 uses the cosine coefficient to normalise the score by taking into account the number of terms in the query Q and document D.

$$Cos(Q, D) = \frac{\sum_{k=1}^{t} w_{qk} w_{dk}}{(\sum_{k=1}^{t} w_{qk}^2 + \sum_{k=1}^{t} w_{dk}^2)^{1/2}} \qquad (3)$$

4.3 Model 3: Lexical normalisation with English stemming

We now supplement model 2 with an English stemmer to remove the suffix variations between surface words in the English documents and the translation of the query. We decided to use the Porter algorithm [Porter, 1980], whose operational characteristics are well documented, e.g. the investigation by Hull and Grefenstette.

We found that stemming resulted in a reduction in the size of the lexicon generated from the document collection from 34574 surface words to

20438 word tokens, a reduction of approximately 40%. After stemming an average Japanese query had a cardinality of 245 English tokens and an average English document a cardinality of 156 word tokens.

4.4 Model 4: Refining weights with *idf*

Rather than simply using weights which are limited to the frequency of the term in a single document or query we would like to take account of the frequency within the document collection as a whole.

Model 4 combines the term weight in the document or query with a measure of the importance of the term in the document collection as a whole. This gives us the well-known inverse document frequency *idf*.

$$w_{xk} = tf_{xk} \times log(N/n_k) \qquad (4)$$

We note that since $log(N/n_k)$ favors rarer terms, *idf* is known to improve precision.

4.5 Model 5: Query expansion with relevance feedback

Relevance feedback aims at refining the weights in the query vector by incorporating information from previously discovered relevant documents. The query is then rerun and hopefully recall will be improved. In theory terms which are present in the retrieved documents have implicitly been identified as relevant to the query. Previous authors in IR have usually reported dramatic improvements of upto 50% as a result of relevance feedback and the early results in MLIR, e.g. [Sheridan and Ballerini, 1996], seem to share this trend.

We use *local feedback* to refine the query description vector Q by finding the most relevant document D in the collection, $D^m = (w_{m1}, w_{m2}, .., w_{mt})$. If $Cos(Q, D^m)$ exceeds some threshold and D^m is the best matching document then we mix Q and D^m, i.e. we refine the query. The assumption is that D^m contains more information about the document set we are looking for that Q by itself.

$$w_{qk} = \alpha w_{qk} + \beta w_{mk} log(\frac{p_k(1-q_k)}{q_k(1-p_k)}) \qquad (5)$$

where

$$p_k = rd_k/R \qquad (6)$$

and

$$q_k = \frac{n_k - rd_k}{N - R} \qquad (7)$$

where rd_k denotes the number of relevant documents in which term k appears and R is the number of relevant documents found. In our simulations we set α to 0.5 and β to 0.5.

4.6 Model 6: Lexical filtering

In the final model we remove the influence of terms in the query description vector Q which are not present in the document collection through a process of *lexical filtering*. This prevents unknown terms in Q from reducing the matching score.

5 Evaluation

In order to automatically evaluate fractional recall and precision it was necessary to construct a representative set of Japanese articles with their correct English article alignments. We call this a judgement set.

The judgement set consists of 60 Japanese queries with 227 relevant English documents. Some 17 Japanese queries (approximately 30%) had no corresponding English document at all. This large percentage of irrelevant queries can be thought of as 'distractors' and is a particular feature of this alignment task, emphasizing the necessity for the matching method to have fine precision as well as good recall.

Following inspection of matching articles we used the heuristic that the search space for each Japanese query was three days of English articles, which was on average 135 articles. This is small by the standards of conventional IR tasks, but given the large number of distractor queries and the low level of analysis in translating the query from Japanese to English, the task is challenging.

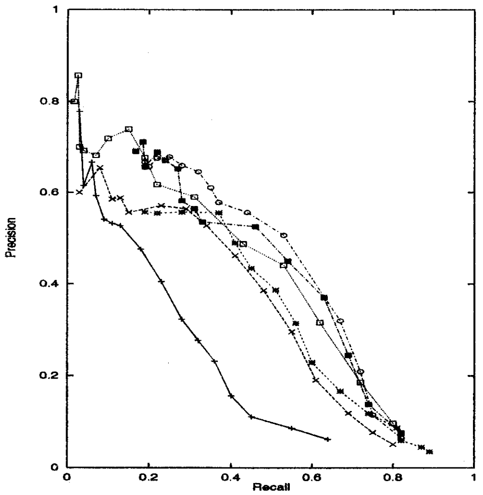

Figure 2: Recall and precision for English-Japanese article alignment: 60 Japanese queries over 4689 English documents, with a mean search range of 135 documents per query. +: model 1, ×: model 2, *: model 3, open square: model 4, close square: model 5, *o*: model 6.

6 Results and Discussion

We define recall and precision in the usual way as follows:

$$recall = \frac{\text{no. of relevant items retrieved}}{\text{no. of relevant items in collection}} \quad (8)$$

$$precision = \frac{\text{no. of relevant items retrieved}}{\text{no. of items retrieved}} \quad (9)$$

Model	Mean precision
1	0.20
2	0.45
3	0.47
4	0.52
5	0.53
6	0.57

Table 1: Mean precision for all methods. The figures are calculated over the 0.2 to 0.6 recall range from curve interpolation on the simulation results.

Figure 2 shows the recall-precision curves for the six methods. The area of most interest to us for our application of knowledge acquisition is in the 0.2 to 0.6 recall range. Mean precision levels are summarized for this range in Table 1.

Model 2 improves greatly over model 1. Our results reflect the belief in IR (e.g. [Singhal *et al.*, 1996]) that cosine normalisation benefits the retrieval of shorter documents. A detailed inspection of the results showed that this was generally true, but a significant fraction of very short English documents of 7 sentences or less in length could not be recalled. Almost 10% of the English articles in our sample were of very short length, of which only 4% were found with model 2. We also see in Figure 2 that length normalisation is of less benefit as recall increases past 0.6. This shows that model 2 cannot compensate for low term overlap.

In line with results for monolingual IR (e.g. [Hull and Grefenstette, 1995]) we see that stemming in model 3 improves recall by upto 5%. The overall effect on mean precision in the Table 1 though is only +2%. This is surprising given the wide morphological variation we found in the bilingual lexicon which was then introduced in lexical transfer. One explanation is that stemming improves matching of cognate terms, but also increases the generality of terms so that we have improved recall at the expense of reduced precision. For higher precision levels we see that stemming actually appears to be counterproductive, showing that well-matching queries are becoming more general.

Incorporating *idf* in model 4 led to another significant improvement in mean precision of +5%. This confirms that term relevance should be considered in MLIR. The *idf* method does not however capture all the relevance information, and we would like to incorporate term frequency *within* documents in future implementations.

The results from model 5 are quite mixed. Relevance feedback as we have defined it should enhance the chance of finding a correct document by reducing the weighting of irrelevant homonyms and increasing that of relevant homonyms. This works well when the query is much more general than the document, but actually reduces precision when the query and the document were a good match before relevance feedback. The overall effect in the key recall range is zero on mean precision. Further investigation is needed to find better methods of query refinement.

Model 6 provides another improvement in recall-precision. Clearly lexical filtering is beneficial as it removes terms from the query vector about which we have no knowledge in the document collection. However, it is unknown whether this improvement in performance will be sustained as the document collection, and therefore the lexicon, grows and increases coverage.

7 Future Research

We have presented preliminary results for the application of MLIR to article alignment for knowledge acquisition. Clearly an extension in the scope of the simulations is needed to test how generalizable our conclusions are. We therefore plan on repeating the simulations with larger document and query sets for six and twelve months. An increase in the size of the judgement set is also needed to guarantee accuracy.

It is important to remember that our results have *not* shown general article alignment using MLIR techniques. We have shown how the methods perform on a particular genre (international news stories) and a particular language pair (English and Japanese). In order to gain a better understanding of the processes which influence article alignment we need to extend the simulations to cover other document collections.

The approach presented here is deliberately very general and we have not used linguistic analysis in lexical transfer. Improvements in performance could be made if the user had access to more sophisticated analysis algorithms to remove or reduce the level of transfer ambiguity, e.g. to reduce the number of homonyms generated for each Japanese term.

8 Conclusion

In this paper we have shown the application of MLIR to bilingual corpus alignment where we cannot rely on language-independent alignment clues. We have shown through simulations that automatic alignment of noisy parallel English-Japanese texts is practical using only the most basic linguistic resources.

Given the small search space for news article matching we would expect a very high level of precision. Clearly though the results show that the task of MLIR is not so simple. Lexical transfer of news articles makes the problem challenging, especially as we have very few proper nouns in our bilingual lexicon. Another factor is our use of contextually related article pairs in the judgement set. Our results may show that this category needs a tighter definition as very loosely related articles fail to match.

Among the major influencing factors are the degree of polysemy in the bilingual lexicon. As our lexical coverage for matching increases, our algorithm must improve in sophistication to improve precision. One such method would be to improve the analysis of the Japanese article and to reduce ambiguity. Another method would be to identify phrases for indexing the news articles.

The size of the news articles has largely been compensated for with length normalisation, but very short articles of less than 7 sentences are still difficult to match. This is a problem because a large proportion of articles in our collection are very short. Stemming, inverse document frequency and lexical filtering all improved recall-precision marginally.

The methods we have used are all easily generalizable to other text collections and languages. This will be needed if we are to acquire a broad range of bilingual knowledge in our next task which is knowledge acquisition.

Acknowledgements

We would like to thank Miwako Shimazu for all her help, especially in evaluating the query judgement set used in our simulations and Satoshi Kinoshita for useful discussions. We gratefully acknowledge the kind permission of Reuters for the use of their newswire articles in our research.

References

[Davis and Dunning, 1995] M. Davis and T. Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference (TREC-4)*, November 1995.

[Davis, 1996] M. Davis. New experiments in cross-language text retrieval at NMSU's computing research lab. In *Fifth Text Retrieval Conference (TREC-5)*, 1996.

[Fung and McKeown, 1997] P. Fung and K. McKeown. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation - to appear*, 1997.

[Hull and Grefenstette, 1995] D. Hull and G. Grefenstette. A detailed analysis of English stemming algorithms. Technical report, Rank Xerox technical report MLTT-023, 6 chemin de Maupertuis, 38240 Meylan, France, 1995.

[Hull and Grefenstette, 1996] D. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 49–57, 18–22 August 1996.

[Porter, 1980] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[Salton, 1970] G. Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Sciences*, 21:187–194, 1970.

[Salton, 1989] G. Salton. *Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1989.

[Sheridan and Ballerini, 1996] P. Sheridan and J. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 58-65, 18–22 August 1996.

[Singhal *et al.*, 1996] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pages 21–29, 18–22 August 1996.