Regular Paper

Creating a Noisy Parallel Corpus from Newswire Articles Using Cross-Language Information Retrieval

Nigel Collier,[†,☆] Hideki Hirakawa[†] and Akira Kumano[†]

In this paper we present an adaptation of cross-language 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 five 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 show that alignment precision levels of over 70% at 70% recall are possible.

1. Introduction

Our goal in this paper is to show how the term vector translation model in cross-language information retrieval (CLIR) 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.

Availability of multi-lingual corpora is an important issue for the development of machine translation. The knowledge from such resources has many applications, including extraction of statistical relations for machine translation[4)], word sense disambiguation[3),5),16)], learning translation rules and templates[19)], bilingual vocabulary extraction[8),21),22)], and detecting omissions in translation[23)].

It is becoming increasingly apparent that clean-parallel corpora used in most previous studies, 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[6),11)~15),20)] 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. If we attempt sentence alignment, we will lose much information from non-corresponding sentences. Although many words and phrases correspond due to the nature of news events, we found that sentence alignment using linguistic methods could succeed for only a small fraction (4.4%) 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.

In this paper we describe the application of information retrieval (IR) techniques, based on the vector space model, for aligning parallel texts. The translation method we use is dictionary term lookup, which although simple has shown performance equal to other CLIR methods[2),7)]. The method has the advantage of being effective while being straightforward to implement and evaluate.

After presenting a summary of the CLIR task we introduce several standard models and show through simulations their effectiveness for news article alignment.

[†] Communication and Information System's Laboratory, Research and Development Centre, Toshiba Corporation

[☆] Presently with Department of Information Science, Faculty of Science, University of Tokyo

2. Task Description

A standard task in 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 by Salton[25),26)] and more recently under the title *cross-language information retrieval* is to enter the query in a different language to that of the document collection.

In Salton's early investigation the task of translating the query vector was significantly simplified. The transfer dictionary contained a controlled vocabulary with the exact translations of the terms in the query vector. Therefore the task of transfer disambiguation which results from polysemy and homography was not a significant factor. This is a major difference from the task we face in a multi-domain task, which demands a bilingual lexicon with significant coverage. This necessarily implies increasing homography as coverage increases. Salton's broad conclusion from his experiments was that the methods used for monolingual IR were almost as effective for the multilingual task. With the increased availability of bilingual lexical resources, we can remove some of the simplifying assumptions that restricted Salton's investigation and also aim at full automation of the task.

In recent CLIR studies using broader language coverage (for example, Davis, *et al.*[10)] on the Spanish TREC corpus) a significant difference has been found between the results for monolingual document retri eval and retrieval when the query has been translated from another language. This performance penalty has been linked to the degree of transfer ambiguity, that is, the number of polysemes and homonyms in the query vector.

The central issue for CLIR, as identified by Davis[9)], is whether vector matching methods can succeed given that they essentially exploit linear relations in the query and target document, relying 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 that most closely correspond to the Japanese query. The overall system is outlined in **Fig. 1** and discussed below.

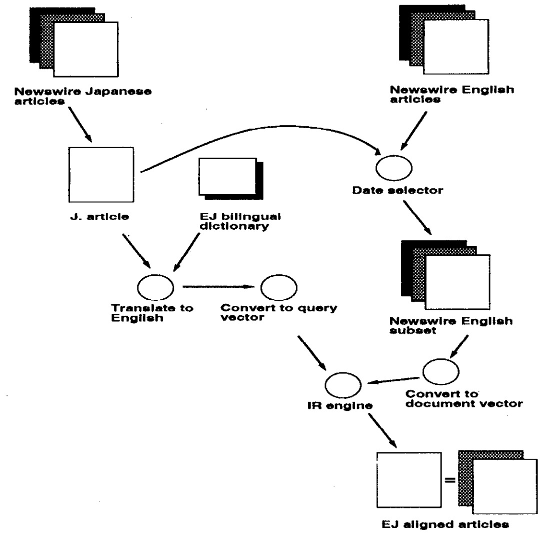


Fig. 1 System overview.

Our method of investigation differs in several ways from that of other researchers in CLIR. Previous work has focussed on analysing the relative performance of models for monolingual and multilingual IR given the addition of a query translation stage. These authors therefore used queries in the document collection language as a control and compared performance to the back-translation of hand-translated queries in a second language.

In our work, we are using quite well-known IR models and are not so concerned with the difference between monolingual and cross-language performance. Indeed, we assume that our queries start life in a different language to the document collection, together with all the complex reformatting characteristics which we have described above. Our work also differs from previous studies in the size of the query, which is typically less than ten terms for standard document retrieval tasks; for example, Hull, *et al.*[18)] and Sheridan, *et al.*[28)] use 7 and 5 terms per query, respectively. Our queries, despite being generated from short news articles, are significantly longer, usually containing over fifty terms before they are translated. Furthermore, the size of the document search space is small in comparison with that of the standard IR task, and we require very high precision to match queries to individual documents.

国連次期事務総長にアナン氏を任命.
国連総会は17日，ガーナ出身のコフィ・アナン国連事務次長（58）$_1$ を次期国連事務総長に任命した. アナン氏は，来年1月1日から第7代事務総長に就任する. アナン氏は，事務総長への就任を宣誓した後,「アフリカ代表，また，国連職員として，皆さんの信頼に答えられるよう全力で任務を遂行する」とあいさつした. 事務総長選出については，ガリ事務総長の立候補に対し，安保理常任理事国である米国が拒否権を行使する$_2$ など難航していたが，国連安保理が13日，ガリ氏と同じアフリカ出身の候補者の中から，アナン氏を選出していた.$_3$
［国連17日ロイター］

Annan to ask U.S. Congress for U.N. arrears
By Evelyn Leopold
UNITED NATIONS, Dec 16 (Reuter) – Kofi Annan, the newly-chosen U.N. secretary-general, said on Monday he would speak to the U.S. Congress if that was what was necessary to get Washington to pay its debts. I think it will be necessary and I am prepared to do so, he said in reference to the $1.3 billion arrears owed by the United States which is bankrupting the world organisation. I hope that I'll be able to work with the administration and through them to Congress to get the United States to pay its arrears and to pay it dues because without a stable financial base, it is extremely difficult to carry on reform, he told the Newshour With Jim Lehrer in one of his first U.S. television interviews. Asked why he wanted the post, he replied, that's a good question. Someone had to do it and it turns out to be my lot. And a friend of mine described it as a job from hell. Annan, 58 the Ghanaian U.N. undersecretary-general for peacekeeping, was chosen by the 15-member Security Council on Frid ay$_3$ to succeed Secretary-General Boutros Boutros-Ghali, of Egypt, whose candidacy for a second five-year term was torpedoed by the United States. Annan will be formally appointed on Tuesday by the 185-member General Assembly$_1$. The United States said it opposed Boutros-Ghali$_2$ because he did not move quickly enough in streamlining the organisation. But Annan, who was supported by Washington, said the organisation had to decide what the purpose of reform was. Let me say that the reform is important as we have all discussed, he said...

Fig. 2 Example of sentence alignment in noisy corpora.

3.1 Newswire Articles

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

Our English document collection consists of daily news articles published by Reuters on the Internet during the period from 7th December 1996 to the 9th May 1997. In total we have 6782 English articles, with an average of about 45 articles per day. After pre-processing has taken place to remove hypertext and formatting characters, we are left with approximately 140000 paragraphs of English text.

In contrast to the English news articles, the Japanese articles, which are also produced daily by Reuters, are very short. A Japanese article is a translated summary of an English article, but as we mentioned earlier, considerable reformatting has taken place. The 1488 Japanese articles cover the same period as the English articles. From this collection we selected a sample for our query set. We discuss the composition of the query and judgement set later.

We examined several pairs of hand-aligned news articles to get a better understanding of the degree of parallelism in our corpora. We observed that in most cases the Japanese sentences are summaries of key sentences in the English article. The positioning of the key sentences varies between translation pairs as is shown in **Fig. 2**.

A further factor was that not all the key sentences were contained in a single English article. The reason appears to be that in one day several versions of an English news article appear on the newswire as the event which is described develops. The Japanese translation may draw on multiple sources including some which do not appear on the public newswire at all. These characteristics make sentence alignment very challenging, and illustrate why we chose the article as our unit of alignment. The basic task therefore is to align one Japanese news article to multiple English articles. On average the matching ratio is 1 : 4.

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 simplest option available. Full machine translation would undoubtedly reduce the level of transfer ambiguity, but it would not give us the range of synonyms which we need to ensure a lexical match.

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; how, therefore, should we incorporate such information in the query vector? Moreover, the weighting of alternative homonyms should also be considered. In this study we assign each homonym term its full weight.

3.3 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[18),27)] that the use of phrases in the index gives superior precision to the use of single words. 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.

Thirdly, an issue for the cross-language IR task is the accuracy with which phrases can be translated. There is some evidence, for example in Ballesteros and Croft[2)], that incorrect phrasal analysis can lead to mistranslations which significantly reduce performance in CLIR systems. An alternative to general phrasal analysis is to translate just the proper nouns, which are known to be helpful; however, the coverage of such phrases in our bilingual lexicon is not so wide, and we decided to translate them on a term-by-term basis to maintain methodological consistency. We will comment further on this later.

Our bilingual dictionary contained 79,000 Japanese words in base form, including approximately 14,000 proper nouns relevant to international news. Lexical lookup and transfer returned quite a mixed bag in the morphological sense. It seems intuitive therefore 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[24)] 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.

Our basic approach is a simple one. We look up each term in the query vector and find its list of translated terms in the document collection language. Duplicate terms were not removed from the translation list. We incorporate word frequencies from analysis of the document collection within the matching model so that high-frequency terms will be given less weighting than low frequency terms.

The objective is then to determine for each query the relevance of each document in the collection. The relevance is determined by the number of matching terms and their distributional characteristics both in the document itself and also in the rest of the collection.

Figure 3 shows an example translation of a single Japanese sentence used as part of a query. Firstly, we notice that our bilingual lexicon contains an uneven coverage of proper nouns. For example, "パキスタン" receives its full translation in English as "Islamic Republi c of Pakistan", whereas "タリバン" (*Taliban*) is not found and is left untranslated. We believe that this is an accurate reflection of the proper noun coverage in many publicly available bilingual lexicons that have been added to ad hoc.

In the example we also see the generation of various synonyms such as "market", "mart", and "marketplace" from "市場". The lexicon also produces a range of inflected forms, such

また，パキスタンにある別の通信社によれば，カブール中心街に位置する市場で，反タリバン派が爆弾を仕掛けたことから，3 人が死亡，37 人が負傷している．

again some day later for the second time also next moreover and or nor Pakistan Islamic Republic of Pakistan Islamic Rep. of Pakistan in on be exist happen have possess keep occur occur situate locate site find show illustrate with having another other different distinct separate new other than apart from aside from besides except except for particularly in particular especially independently separately additionally apart news agency Kabul center middle middle core centre main leading major central principle focusing on emphasizing laying stress on attaching importance to centering on centering around revolving around consisting mainly of town city street locate situate position situation location place site rank station stand lie market mart marketplace タリバン bundle bomb devise set in stall start set injured hurt injury wound cut bruise wounded injure get

Fig. 3 Example of lexical transfer.

as “injure”, “injured”, and “injury”. Unfortunately the inflectional coverage of word forms in the lexicon is not so consistent, so we perform word normalisation with a stemming algorithm.

The translation of numerals is an interesting issue, but has not been dealt with yet in our system. For example, if we read “1 万 5000” in Japa nese, this could be translated into English as either “15,000”, “15000” or “fifteen thousand”. Clearly some sort of numeric normalisation routine is required to make use of these language-independent clues. Surprisingly though, we have found that many of the numerals in corresponding articles are different. Often exact numerals appear in one article and a rounded figure in another for monetary and population units. A second factor seems to be that a series of English news articles is produced as the news event develops with statistics being refined. Dates on the other hand are constant except where they refer to the date of publication of the article and its translation.

4. Models

Below we present several models that calculate the 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 $D = (w_{d1}, w_{d2}, \ldots, w_{dt})$, where w_{dk} represents the weight of term k in document D. The set of documents in the collection is N, 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 $Q = (w_{q1}, w_{q2}, \ldots, w_{qt})$.

Model 1: *tf*

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

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

where

$$w_{xk} = tf_{xk}. \tag{2}$$

Model 2: *tf* with document length normalisation

Model 1 produces a score that in theory is unbounded, and in practice is bounded only by the size of a document. This is unsatisfactory because longer documents will have an unfair advantage since (a) they contain 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.

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

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[24], whose operational characteristics are well documented, for example, in 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 37554 surface words to 25816 word tokens, a reduction of approximately 31%. After stemming an average Japanese query had a cardinality of 211 English tokens and an average English document a cardinality of 174 word tokens.

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.

Model 5: Query expansion with local relevance feedback

Local relevance feedback[1] 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 top retrieved documents have implicitly been identified as relevant to the query. Previous authors in IR have reported dramatic improvements of up to 50% as a result of relevance feedback, and the early results in CLIR, for example, those given by Sheridan and Ballerini[28], seem to share this trend.

We use a variation of the Rocchio method to refine and expand the query description vector Q. We find all relevant documents in the collection which match the query Q better than a threshold, and sum the weights of this subcollection. In our tests we used a threshold of 0.10, which we found to be almost optimal. The weights from the relevant documents and those from non-relevant documents are mixed with the original query weights as follows:

$$Q'_j = \alpha Q_j + \beta \frac{1}{|R|} \sum_{D_i \in R} D_i - \lambda \frac{1}{|N-R|} \sum_{D_i \in N-R} D_i, \qquad (5)$$

where R is the set of relevant documents found and Q_j is the query to be refined. Small modifications to term weights are known to be most effective in Rocchio relevance feedback (see, for example, Salton[27]) and following this we found that, in our simulations, values of $\alpha = 1.0$, $\beta = 0.2$, and $\lambda = 0.2$ worked quite well.

5. Evaluation

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 100 Japanese queries with 454 relevant English documents. Some 24 Japanese queries had no corresponding English document at all. These irrelevant queries can be thought of as "distractors" and the large percentage is a particular feature of this alignment task, emphasizing the necessity for the matching method to have fine precision as well as good recall.

This set was then given to a bilingual checker, who was asked to asign each aligned article pair to one of the following categories:

(1) The two articles are translations of each other.

(2) The two articles are strongly contextually related.

(3) No match.

We then removed type-3 correspondences so that the judgement set contained pairs of articles which at least shared the same context, that is, referred to the same news event. Given that the distinction between type-1 and type-2 correspondences is not always clear because of partial translation correspondences, we felt that it was better to include type-2 correspondences, to capture as much bilingual information as possible.

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

6. Results and Discussion

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

$$recall = \frac{\text{no. of relevant items retrieved}}{\text{no. of relevant items in collection}} \qquad (6)$$

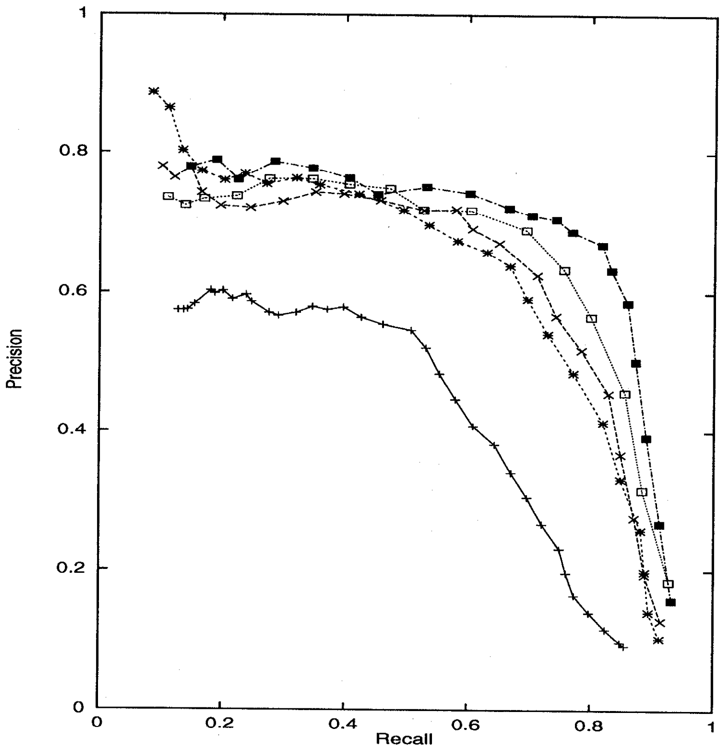


Fig. 4 Recall and precision for English-Japanese article alignment: 100 Japanese queries over 6782 English documents, with a mean search range of 135 documents per query. +: Model 1, ×: Model 2, ∗: Model 3, open square: Model 4, close square: Model 5.

$$\mathit{precision} = \frac{\text{no. of relevant items retrieved}}{\text{no. of items retrieved}} \tag{7}$$

Figure 4 shows the recall-precision curves for the five methods. Results are calculated by increasing the threshold of the matching score. The area of most interest to us for our application of knowledge acquisition is in the 0.2 to 0.8 recall range. Mean precision levels for this range are summarized in **Table 1**.

Table 1 Mean precision for all methods. The figures are calculated over the 0.2 to 0.8 recall range from curve interpolation on the simulation results.

Model	Mean precision
1	0.47
2	0.69
3	0.68
4	0.72
5	0.75

Model 2 improves greatly over Model 1. Our results reflect the belief in IR (for example, in Singhal, *et al.*[29]) 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 with a length of 7 sentences or less could not be recalled. Almost 10% of the English articles in our sample were of very short length, and only 4% of these were found with Model 2. We also see in Fig. 4 that length normalisation becomes of slightly less benefit as recall increases past 0.8. This shows that even though Model 2 can compensate for differing absolute term overlap it cannot compensate as well for low term overlap in matching documents which is found at high recall and low precision.

In line with results for monolingual IR, for example Hull, *et al.*[17], we see that stemming in Model 3 improves recall by up to 5%, but only at recall levels below 40%. The effect of stemming at higher recall levels is to *reduce* precision quite substantially. The overall effect on mean precision in Table 1, however, is −1%. This is surprising given that no morphological analysis has taken place in lexical transfer. One explanation is that stemming improves matching of

Table 2 Example of erroneous article matches caused by an *ill-defined* query.

Article title	Ill-defined terms
アフガニスタンの首都カブールで，空爆・爆発相次ぐ．	
S. African police arrest two whites after explosions	explosion, bomb, blast
Algeria says rebels killed 18 people overnight	rebel, kill, force
Explosion reported in Belfast centre-police	explosion, centre
Explosions shatter peace at Peru siege site	explosion, injury, rebel, force

cognate terms, but also increases the generality of terms, so that we have improved precision where a query and a document are already well matched at the expense of reduced precision where the correspondence is not so good.

Incorporating *idf* in model 4 led to another significant improvement in the mean precision of +4%. This confirms that term relevance should be considered in CLIR. 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 show an overall improvement in the mean precision of +3% in the key recall range, but also much more significant improvements at higher levels of recall. This is natural, as relevance feedback is designed to find those documents which are not matched strongly with the original query. At the lexical level, relevance feedback should reduce the weighting of irrelevant homonyms and increase that of relevant homonyms.

Generally we see that precision peaks at about 75%. When we analysed the erroneously recalled article pairs at this level of precision, we found that many of the English articles were somehow related to the Japanese article, but could not be classified as having a "strong" contextual relationship in our scheme, which requires not only that the articles refer to the same news event, but also that they refer to the same "aspect" of that event. This illustrates the somewhat fuzzy boundary between levels of correspondence.

Furthermore, one or two queries were responsible for a large amount of the error (fallout) at high precision levels as shown by **Fig. 5**. Although such *ill-defined* queries have proper nouns to anchor the Japanese to the English articles, other terms acquired disproportionate importance due to their low frequency in the corpus as a whole. An example of an *ill-defined* query title is shown in **Table 2**, where the Japanese article tells us about a bomb explosion in Kabul, the Afghan capital, killing a number of people. The disproportionate key terms are shown against the incorrectly matching English document titles.

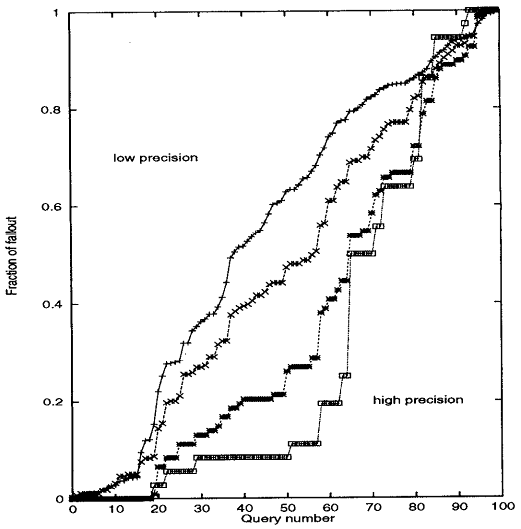


Fig. 5 Percentage contribution to fallout (of Model 4) by query at increasing precision levels.

7. Future Research

Clearly an extension in the scope of the simulations is needed to test how generalizable our conclusions are. We have shown how well the methods perform for 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.

Despite these limitations, the results are important in that they are indicative of how we can expect CLIR systems to perform for English and Japanese. The news article alignment task can be thought of as a half-way point to full CLIR, because the queries are quite short in comparison to full news articles, but longer than the single sentences common in some TREC tasks.

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, for example with a machine translation system, to remove or reduce the level of transfer ambiguity, that is, to reduce the number of homonyms generated for each Japanese term. This raises the interesting question of how the trade-off between reduced ambiguity in machine translation will compensate for the synonym choices available in dictionary term lookup. A recent study by Collier, *et al.*[7] explores this cross-method comparison.

The experiments given in this paper assume a term-to-term translation model which ignores the effects of using phrases in the index. Despite the generally held assumption that index phrases are better than terms, a recent study by Ballesteros, *et al.*[2] has found that the worsening effects resulting from poor analysis of general phrases can outweigh the benefits of proper noun identification for translation in CLIR. This aspect of CLIR deserves further investigation, and we would like to incorporate appropriate experiments into our work in the future.

The presence of appropriate proper nouns in the bilingual lexicon is likely to have a significant influence on performance. We envisage that our system will form part of a knowledge acquisition cycle in which bilingual knowledge is extracted from the aligned texts and used to improve the bilingual lexicon for future news article matching. Previous studies as Refs.[6],[13],[15],[20], etc. have shown the feasibility of bilingual lexical knowledge acquisition for noisy-parallel texts.

8. Conclusion

In this paper we have shown the application of CLIR 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 CLIR 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, since 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 by 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 (at lower recall ranges), inverse document frequency, and relevance feedback 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.

Acknowledgments We would like to thank Miwako Shimazu for her help in evaluating the query judgement set used in our simulations and to Satoshi Kinoshita and Kenji Ono for their useful comments on this paper. We gratefully acknowledge the kind permission of Reuters for the use of their newswire articles in our research.

References

1) Attar, R. and Fraenkal, A.S.: Local Feedback in Full-Text Retrieval Systems, *J. ACM*, Vol.24, pp.397–417 (1977).

2) Ballesteros, L. and Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, *Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, pp.84–91 (1997).

3) Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R.: Word Sense Disambiguation Using Statistical Methods, *Annual Meeting of the Association for Computational Linguistics*, pp.264–270 (1991).

4) Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, No.2,

pp.263–299 (1993).

5) Collier, N.: *English-Japanese Lexical Transfer Using a Hopfield Neural Network*, Ph.D. Thesis, Department of Language Engineering, UMIST, Manchester (1996).

6) Collier, N., Kumano, A. and Hirakawa, H.: Acquisition of English-Japanese Proper Nouns from Noisy-Parallel Newswire Articles Using Katakana Matching, *Natural Language Pacific Rim Symposium* (*NLPRS-97*), Phuket (1997).

7) Collier, N., Hirakawa, H. and Kumano, A.: Machine Translation vs. Dictionary Term Translation – A Comparison for English-Japanese News Article Alignment, *Proc. COLING-ACL'98*, University of Montreal (1998).

8) Dagan, I., Church, K. and Gale, W.: Robust Bilingual Word Alignment for Machine Aided Translation, *Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pp.1–8 (1993).

9) Davis, M.: New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab, *Fifth Text Retrieval Conference* (*TREC-5*) (1996).

10) Davis, M. and Dunning, T.: A TREC Evaluation of Query Translation Methods for Multilingual Text Retrieval, *Fourth Text Retrieval Conference* (*TREC-4*) (1995).

11) Davis, M., Dunning, T. and Ogden, W.: Text Alignment in the Real World: Improving Alignment of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons, *Proc. Seventh European Conference of the Association of Computational Linguistics* (*ACL*), Dublin (1995).

12) Fung, P. and McKeown, K.: Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping, *AMTA-94*, Columbia, MD (1994).

13) Fung, P.: A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora, *Proc. Annual Conference of the Association for Computational Linguistics* (*ACL-95*) (1995).

14) Fung, P. and McKeown, K.: A Technical Word and Term Translation Aid using Noisy Parallel Corpora across Language Groups, *Machine Translation – Special Issue on New Tools for Human Translators*, pp.53–87 (1996).

15) Fung, P.: Finding Terminology Translations from Non-Parallel Corpora, *Proc. 5th Workshop on Very Large Corpora* (1997).

16) Gale, W., Church, K. and Yarowsky, D.: A Method for Disambiguating Word Senses in a Large Corpus, *Comput. Hum.*, Vol.26, pp.415–439 (1993).

17) Hull, D. and Grefenstette, G.: A Detailed Analysis of English Stemming Algorithms, Rank Xerox Technical Report MLTT-023, 6 chemin de Maupertuis, 38240 Meylan, France (1995).

18) Hull, D. and Grefenstette, G.: Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, pp.49–57 (1996).

19) Kaji, H., Kida, Y. and Morimoto, Y.: Learning Translation Templates from Bilingual Text, *COLING-92*, Nantes, pp.672–678 (1992).

20) Kaji, H. and Aizono, T.: Extracting Word Correspondances from Bilingual Corpora Based on Word Co-occurrence Information, *Proc. 16th International Conference on Computational Linguistics* (*COLING-96*), Copenhagen, pp.23–28 (1996).

21) Kitamura, M. and Matsumoto, Y.: Automatic Extraction of Word Sequence Correspondences in Parallel Corpora, *Fourth Workshop on Very Large Corpora*, University of Copenhagen, pp.79–87 (1996).

22) Kupiec, J.: An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, *31st Annual Meeting of the Association for Computational Linguistics*, Ohio, pp.17–22 (1993).

23) Melamed, I.: Automatic Detection of Omissions in Translations, *16th International Conference on Computational Linguistics*, Copenhagen (1996).

24) Porter, M.: An Algorithm for Suffix Stripping, *Program*, Vol.14, No.3, pp.130–137 (1980).

25) Salton, G.: Automatic Processing of Foreign Language Documents, *J. Am. Soc. Inf. Sci.*, Vol.21, pp.187–194 (1970).

26) Salton, G.: Experiments in Multi-lingual Information Retrieval, Technical Report 72-154, Cornell University, Ithaca, New York (1972).

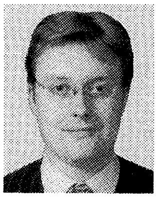
27) Salton, G.: *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA (1989).

28) Sheridan, P. and Ballerini, J.: Experiments in Multilingual Information Retrieval Using the SPIDER System, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, pp.58–65 (1996).

29) Singhal, A., Buckley, C. and Mitra, M.: Pivoted Document Length Normalization, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Infor-*

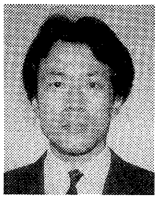
mation Retrieval, Zurich, pp.21–29 (1996).

(Received October 29, 1997)
(Accepted October 2, 1998)



Nigel Collier received his B.Sc. in computer science from Leeds University in the UK in 1992, an M.Sc. in machine translation and a Ph.D. in language engineering from UMIST in the UK in 1994 and 1996 respectively. From 1990 to 1991 he was a visitor at the University of Tokyo. From 1996 to 1998 he was a Toshiba Fellow at Toshiba's Research and Development Center in Kawasaki and is currently a research associate at Tokyo University investigating information extraction from genome domain texts. His current research interests are in machine translation, cross-language information retrieval, and knowledge acquisition. He is a member of the ACM and SIGART as well as the British Computer Society.

Hideki Hirakawa received the B.E. and M.E. degrees in electrical engineering from Kyoto University, Kyoto, in 1978 and 1980. He joined Toshiba Corp. in 1980, and is currently leading research and development of human interface and natural language processing at the Toshiba Research and Development Center as a senior manager. From 1982 to 1985, he was a researcher at the Institute for New Generation Computer Technology (ICOT). From 1994 to 1995, he was a visiting researcher at the Media Laboratory in M.I.T., USA. His current research interests are in natural language processing and human interface. He is a member of the IPSJ, the Japanese Society of Artificial Intelligence, the Association of Natural Language Processing and the Association for Computational Linguistics.



Akira Kumano received the B.E. degree in computer science from Tokyo Institute of Technology, Tokyo, in 1982. He joined Toshiba Corp. in 1982, and is currently researching natural language processing at the Toshiba Research and Development Center as a research scientist. From 1986 to 1989, he was a member of the research staff at the Japan Electronic Dictionary Research Institute, Ltd. (EDR). His current research interests are in machine translation and knowledge acquisition. He is a member of the IPSJ, the Japanese Society of Artificial Intelligence and the Association of Natural Language Processing.