Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignment

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

Abstract

Bilingual news article alignment methods based on multi-lingual information retrieval have been shown to be successful for the automatic production of so-called noisy-parallel corpora. In this paper we compare the use of machine translation (MT) to the commonly used dictionary term lookup (DTL) method for Reuter news article alignment in English and Japanese. The results show the trade-off between improved lexical disambiguation provided by machine translation and extended synonym choice provided by dictionary term lookup and indicate that MT is superior to DTL only at medium and low recall levels. At high recall levels DTL has superior precision.

1 Introduction

In this paper we compare the effectiveness of full machine translation (MT) and simple dictionary term lookup (DTL) for the task of English-Japanese news article alignment using the vector space model from multi-lingual information retrieval. Matching texts depends essentially on lexical coincidence between the English text and the Japanese translation, and we see that the two methods show the trade-off between reduced transfer ambiguity in MT and increased synonymy in DTL.

Corpus-based approaches to natural language processing are now well established for tasks such as vocabulary and phrase acquisition, word sense disambiguation and pattern learning. The continued practical application of corpus-based methods is critically dependent on the availability of corpus resources.

In machine translation we are concerned with the provision of bilingual knowledge and we have found that the types of language domains which users are interested in such as news, current affairs and technology, are poorly represented in today's publically available corpora. Our main area of interest is English-Japanese translation, but there are few clean parallel corpora available in large quantities. As a result we have looked at ways of automatically acquiring large amounts of parallel text for vocabulary acquisition.

The World Wide Web and other Internet resources provide a potentially valuable source of parallel texts. Newswire companies for example publish news articles in various languages and various domains every day. We can expect a coincidence of content in these collections of text, but the degree of parallelism is likely to be less than is the case for texts such as the United Nations and parliamentary proceedings. Nevertheless, we can expect a coincidence of vocabulary, in the case of names of people and places, organisations and events. This time-sensitive bilingual vocabulary is valuable for machine translation and makes a significant difference to user satisfaction by improving the comprehensibility of the output.

Our goal is to automatically produce a parallel corpus of aligned articles from collections of English and Japanese news texts for bilingual vocabulary acquisition. The first stage in this process is to align the news texts. Previously (Collier et al., 1998) adapted multi-lingual (also called "translingual" or "cross-language") information retrieval (MLIR) for this purpose and showed the practicality of the method. In this paper we extend their investigation by comparing the performance of machine translation and conventional dictionary term translation for this task.

2 MLIR Methods

There has recently been much interest in the MLIR task (Carbonell et al., 1997)(Dumais et al., 1996)(Hull and Grefenstette, 1996). MLIR differs from traditional information retrieval in several respects which we will discuss below. The most obvious is that we must introduce a translation stage in between matching the query and the texts in the document collection.

Query translation, which is currently considered to be preferable to document collection translation, introduces several new factors to the IR task:

- *Term transfer mistakes* - analysis is far from perfect in today's MT systems and we must con-

sider how to compensate for incorrect translations.

- *Unresolved lexical ambiguity* - occurs when analysis cannot decide between alternative meanings of words in the target language.
- *Synonym selection* - when we use an MT system to translate a query, generation will usually result in a single lexical choice, even though alternative synonyms exist. For matching texts, the MT system may not have chosen the same synonym in the translated query as the author of the matching document.
- *Vocabulary limitations* - are an inevitable factor when using bilingual dictionaries.

Most of the previous work in MLIR has used simple dictionary term translation within the vector space model (Salton, 1989). This avoids synonymy selection constraints imposed by sentence generation in machine translation systems, but fails to resolve lexical transfer ambiguity. Since all possible translations are generated, the correctly matching term is assumed to be contained in the list and term transfer mistakes are not an explicit factor.

Two important issues need to be considered in dictionary term based MLIR. The first, raised by Hull *et al* (Hull and Grefenstette, 1996), is that generating multiple translations breaks the term independence assumption of the vector space model. A second issue, identified by (Davis, 1996), is whether vector matching methods can succeed given that they essentially exploit linear (term-for-term) relations in the query and target document. This becomes important for languages such as English and Japanese where high-level transfer is necessary.

Machine translation of the query on the other hand, uses high level analysis and should be able to resolve much of the lexical transfer ambiguity supplied by the bilingual dictionary, leading to significant improvements in performance over DTL, e.g. see (Davis, 1996). We assume that the MT system will select only one synonym where a choice exists so term independence in the vector space model is not a problem. Term transfer mistakes clearly depend on the quality of analysis, but may become a significant factor when the query contains only a few terms and little surrounding context.

Surprisingly, to the best of our knowledge, no comparison has been attempted before between DTL and MT in MLIR. This may be due either to the unreliability of MT, or because queries in MLIR tend to be short phrases or single terms and MT is considered too challenging. In our application of article alignment, where the query contains sentences, it is both meaningful and important to compare the two methods.

3 News Article Alignment

The goal of news article alignment is the same as that in MLIR: we want to find relevant matching documents in the source language corpus collection for those queries in the target language corpus collection. The main characteristics which make news article alignment different to MLIR are:

- Number of query terms - the number of terms in a query is very large compared to the usual IR task;
- Small search space - we can reduce the search to those documents within a fixed range of the publication date;
- Free text retrieval - we cannot control the search vocabulary as is the case in some information retrieval systems;
- High precision - is required because the quality of the bilingual knowledge which we can acquire is directly related to the quality of article alignment.

We expect the end product of article alignment to be a noisy-parallel corpus.

In contrast to clean-parallel texts we are just beginning to explore noisy-parallel texts as a serious option for corpus-based NLP, e.g. (Fung and McKeown, 1996). Noisy-parallel texts are characterised by heavy reformatting at the translation stage, including large sections of untranslated text and textual reordering. Methods which seek to align single sentences are unlikely to succeed with noisy parallel texts and we seek to match whole documents rather than sentences before bilingual lexical knowledge acquisition. The search effort required to align individual documents is considerable and makes manual alignment both tedious and time consuming.

4 System Overview

In our collections of English and Japanese news articles we find that the Japanese texts are much shorter than the English texts, typically only two or three paragraphs, and so it was natural to translate from Japanese into English and to think of the Japanese texts as 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.

4.1 Dictionary term lookup method

DTL takes each term in the query and performs dictionary lookup to produce a list of possible translation terms in the document collection language. Duplicate terms were not removed from the translation list. In our simulations we used a 65,000 term

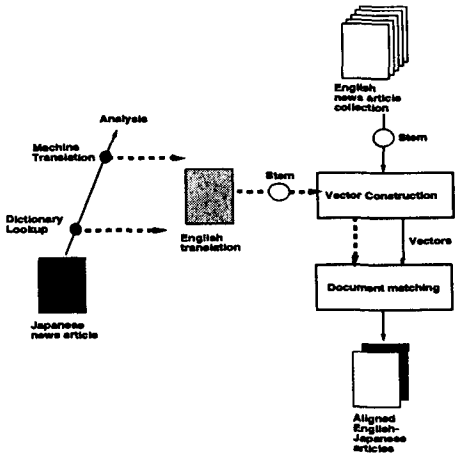

Figure 1: System Overview

common word bilingual dictionary and 14,000 terms from a proper noun bilingual dictionary which we consider to be relevant to international news events.

The disadvantage of term vector translation using DTL arises from the shallow level of analysis. This leads to the incorporation of a range of polysemes and homographs in the translated query which reduces the precision of document retrieval. In fact, the greater the depth of coverage in the bilingual lexicon, the greater this problem will become.

4.2 Machine translation method

Full machine translation (MT) is another option for the translation stage and it should allow us to reduce the transfer ambiguity inherent in the DTL model through linguistic analysis. The system we use is Toshiba Corporation's ASTRANSAC (Hirakawa et al., 1991) for Japanese to English translation.

The translation model in ASTRANSAC is the transfer method, following the standard process of morphological analysis, syntactic analysis, semantic analysis and selection of translation words. Analysis uses ATNs (Augmented Transition Networks) on a context free grammar. We modified the system so that it used the same dictionary resources as the DTL method described above.

4.3 Example query translation

Figure 2 shows an example sentence taken from a Japanese query together with its English translation produced by MT and DTL methods. We see that in both translations there is missing vocabulary (e.g. "スティーブ・フォーセット" is not translated); since the two methods both use the same dictionary resource this is a constant factor and we can ignore it for comparison purposes.

As expected we see that MT has correctly resolved some of the lexical ambiguities such as '世界 → world', whereas DTL has included the spurious homonym terms "earth, universe, world-wide, universal, international".

Original Japanese text:
気球による単独世界一周を目指す米国人、スティーブ・フォーセットさんは１９日、インド上空を飛行しているが、初の世界一周は達成できないもよう。

Translation using MT:
Although the American who aims at an independent world round by the balloon, and Mr. スティーブ フォーset are flying the India sky on 19th, it can seem to attain a simple world round.

Translation using DTL:
independent individual singlehanded single separate sole alone balloon round one round one revolution world earth universe world-wide international base found ground depend turn hang approach come draw drop cause due twist choose call according to based on owing to by by means of under due to through from accord owe round one round one revolution go travel drive sail walk run American スティーブ aim direct toward shoot for have direct India Republic of India Rep. of India フォー Mr. Miss Ms. Mis. Messrs. Mrs. Mmes. Ms. Mses. Esq. American sky skies upper air upper regions high up in the sky up in the air an altitude a height in the sky of over set arrangement arrange world earth universe world-wide universal international simple innocent naive unsophisticated inexperienced fly hop flight aviation round one round one revolution go travel drive sail walk run seem appear encaustic signs sign indications attain achieve accomplish realise fulfill achievement attainment

Figure 2: Cross method comparison of a sample sentence taken from a Japanese query with its translation in English

In the case of synonymy we notice that MT has decided on "independent" as the translation of "単独", DTL also includes the synonyms "individual, singlehanded, single, separate, sole,...",etc.. The author of the correctly matching English text actually chose the term 'singlehanded', so synonym expansion will provide us with a better match in this case. The choice of synonyms is quite dependent on author preference and style considerations which MT cannot be expected to second-guess.

The limitations of MT analysis give us some selection errors, for example we see that "インド上空を飛行している" is translated as "flying the India sky", whereas the natural translation would be 'flying over India", even though 'over' is registered as a possible translation of '上空' in the dictionary.

5 Corpus

The English document collection consisted of Reuter daily news articles taken from the internet for the December 1996 to the May 1997. In total we have 6782 English articles with an average of about 45 articles per day. After pre-processing 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 Reuter's, are very short. The Japanese is a translated summary of an English article, but considerable reformatting has taken place. In many cases the Japanese translation seems to draw on multiple sources including some which do not appear on the public newswire at all. The 1488 Japanese articles cover the same period as the English articles.

6 Implementation

The task of text alignment takes a list of texts $\{Q'_0, .., Q'_n\}$ in a target language and a list of texts $\{D_0, .., D_m\}$ in a source language and produces a list of aligned pairs. A pair $< Q'_x, D_y >$ is in the list if Q'_x is a partial or whole translation of D_y. In order to decide on whether the source and target language text should be in the list of aligned pairs we translate Q'_x into the source language to obtain Q_x using bilingual dictionary lookup. We then match texts from $\{Q_0, .., Q_n\}$ and $\{D_0, .., D_m\}$ using standard models from Information Retrieval. We now describe the basic model.

Terminology

An index of t terms is generated from the document collection (English corpus) and the query set (Japanese translated articles). Each document has a description vector $D = (w_{d1}, w_{d2}, .., w_{dt})$ 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 $Q = (w_{q1}, w_{q2}, .., w_{qt})$.

6.1 Model

We implemented the standard vector-space model with cosine normalisation, inverse document frequency *idf* and lexical stemming using the Porter algorithm (Porter, 1980) to remove suffix variations between surface words.

The cosine rule is used to compensate for variations in document length and the number of terms when matching a query Q from the Japanese text collection and a document D from the English text collection.

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

We combined term weights in the document and 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 (*tf+idf*) score:

$$w_{xk} = tf_{xk} \times log(|N|/n_k) \quad (2)$$

Since $log(|N|/n_k)$ favours rarer terms *idf* is known to improve precision.

7 Experiment

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. Although it is a significant effort to evaluate alignments by hand, this is possibly the only way to obtain an accurate assessment of the alignment performance. Once alignment has taken place we compared the threshold filtered set of English-Japanese aligned articles with the judgement set to obtain recall-precision statistics.

The judgement set consisted of 100 Japanese queries with 454 relevant English documents. Some 24 Japanese queries 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.

This set was then given to a bilingual checker who was asked to score each aligned article pair according to (1) the two articles are translations of each other, (2) the two articles are strongly contextually related, (3) no match. We removed type 3 correspondences so that the judgement set contained pairs of articles which at least shared the same context, i.e. referred to the same news event.

Following inspection of matching articles we used the heuristic that the search space for each Japanese query was one day either side of the day of publication. On average this was 135 articles. This is small by the standards of conventional IR tasks, but given the large number of distractor queries, the requirement for high precision and the need to translate queries, the task is challenging.

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

$$recall = \frac{\text{no. of relevant items retrieved}}{\text{no. of relevant items in collection}} \quad (3)$$

$$precision = \frac{\text{no. of relevant items retrieved}}{\text{no. of items retrieved}} \quad (4)$$

Results for the model with MT and DTL are shown in Figure 3. We see that in the basic *tf+idf* model, machine translation provides significantly better article matching performance for medium and low levels of recall. For high recall levels DTL is better. Lexical transfer disambiguation appears to be important for high precision, but synonym choices are crucial for good recall.

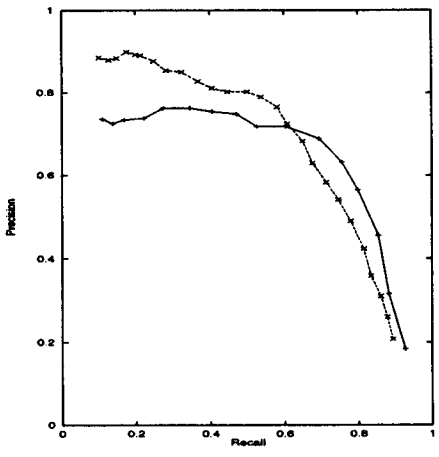

Figure 3: Model 1: Recall and precision for English-Japanese article alignment. +: DTL ×: MT.

Overall the MT method obtained an average precision of 0.72 in the 0.1 to 0.9 recall range and DTL has an average precision of 0.67. This 5 percent overall improvement can be partly attributed to the fact that the Japanese news articles provided sufficient surrounding context to enable word sense disambiguation to be effective. It may also show that synonym selection is not so detrimental where a large number of other terms exist in the query. However, given these advantages we still see that DTL performs almost as well as MT and better at higher recall levels. In order to maximise recall, synonym lists provided by DTL seem to be important. Moreover, on inspection of the results we found that for some weakly matching document-query pairs in the judgement set, a mistranslation of an important or rare term may significantly bias the matching score.

8 Conclusion

We have investigated the performance of MLIR with the DTL and MT models for news article alignment using English and Japanese texts. The results in this paper have shown surprisingly that MT does not have a clear advantage over the DTL model at all levels of recall. The trade-off between lexical transfer ambiguity and synonymy implies that we should seek a middle strategy: a sophisticated system would perhaps perform homonym disambiguation and then leave alternative synonyms in the translation query list. This should maximise both precision and recall and will be a target for our future work. Furthermore, we would like to extend our investigation to other MLIR test sets to see how MT performs against DTL when the number of terms in the query is smaller.

Acknowledgements

We gratefully acknowledge the kind permission of Reuters for the use of their newswire articles in our research. We especially thank Miwako Shimazu for evaluating the judgement set used in our simulations.

References

J. Carbonell, Y. Yang, R. Frederking, R. Brown, Y. Geng, and D. Lee. 1997. Translingual information retrieval: A comparative evaluation. In *Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan*, 23rd – 29th August.

N. Collier, A. Kumano, and H. Hirakawa. 1998. A study of lexical and discourse factors in bilingual text alignment using MLIR. *Trans. of Information Processing Society of Japan (to appear).*

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

S. Dumais, T. Landauer, and M. Littman. 1996. Automatic cross-language retrieval using latent semantic indexing. In G. Grefenstette, editor, *Working notes of the workshop on cross-linguistic information retrieval ACM SIGIR.*

P. Fung and K. McKeown. 1996. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation - Special Issue on New Tools for Human Translators*, pages 53–87.

H. Hirakawa, H. Nogami, and S. Amano. 1991. EJ/JE machine translation system ASTRANSAC - extensions towards personalization. In *Proceedings of the Machine Translation Summit III*, pages 73–80.

D. Hull and G. Grefenstette. 1996. 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.

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

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