English-Japanese Cross-lingual Query Expansion Using Random Indexing of Aligned Bilingual Text Data

Magnus Sahlgren Preben Hansen Jussi Karlgren
Swedish Institute of Computer Science, SICS,
Box 1263, SE-164 29 Kista, Sweden
{mange, preben, jussi}@sics.se

Abstract

Vector space models can be used for extracting semantically similar words from the co-occurrence statistics of words in large text data. In this paper, we report on our NTCIR 2002 experiments using the Random Indexing vector space method for extracting an English-Japanese cross-lingual thesaurus from aligned English-Japanese bilingual data. The cross-lingual thesaurus has been used for automatic cross-lingual query expansion in the NTCIR patent retrieval task.

Keywords: *Vector-space model, bilingual thesauri, query expansion.*

1 Introduction

Retrieving relevant documents across languages presents an interesting problem for Information Retrieval (IR) systems. Not only does a cross-lingual IR system face the traditional retrieval problem of how to find documents that are relevant to a user's query, but there is also the additional translation problem of how to translate between the query and the documents. There are several possible ways for a cross-lingual IR system to deal with these problems [7]. One way is to use an existing dictionary to translate the words of the query, one by one, into the language of the document collection, and to use this translated query as input to a traditional retrieval system. Another methodology is to translate both query and documents into some form of common internal representation, or to use a machine translation system (if there exists one for the particular languages in question) to translate between query and documents.

We handle the cross-lingual retrieval task, and the related translation problem, by using a purely statistical method for automatic query translation and expansion. In short, the idea is to use aligned bilingual training data to automatically construct a bilingual thesaurus in which, to any given word, a number r (in these experiments, we used $r = 5$) of semantically related words in the other language is given. We then use this bilingual thesaurus to translate and expand every word in the original queries with the r words listed in the thesaurus. We thus perform automatic query translation and expansion, without the need for any previous knowledge of the langauges in question.

Note that this methodology also handles the *vocabulary* (or *synonymy*) problem in information retrieval, which consists in the fact that people might choose different words to express the same information. For example, one person might use the word "boat" to refer to water crafts, while another person might use the word "ship". If the IR system does not attempt to handle this vocabulary problem, it runs the risk of missing relevant documents. We handle the problem explicitly by semantically based query expansion.[1]

Note also that this methodology handles what could be called *the translation problem*, which consists in the fact that there need not exist (in fact, there rarely do) a one-to-one translation of words across languages. That is, a given word in a given language, for example, English word "bank", might have several possible translations in another language, and if the retrieval system only chooses one of the possible translations, it runs the risk of missing relevant documents (or, in the worst case, of retrieving totally irrelevant documents). Our attempt to handle this problem is, as above, to expand each word in the original query by its 5 nearest neighbors in the other language.

In what follows, we report on our NTCIR 2002 experiments in the English-to-Japanese patent retrieval task.

2 The Vector Space Methodology

We use a vector space model to automatically construct the bilingual thesaurus. The vector space model do this by observing word co-occurrence statistics in aligned bilingual corpora, and it assumes that two

[1] If "boat" is in the query, the system will (hopefully) expand the word with "ship," "vessel," "craft," "water" and so on.

words that occur with similar distribution in the training data (i.e. that occur in similar documents) are semantically similar. This means that if one English word occurs with the same (or similar) distribution (i.e. if it occurs in the same (or similar) documents) as one Japanese word, the model will rank these two words as semantically similar – in fact, the model will assume that they are translations of each other.

Traditionally, vector space models represent the text data as an $n \times m$ co-occurrence matrix [1], [10], where each row n represents a unique word and each column m represents a document (or a word). The cells of the matrix are the (normalized) frequency counts of a particular word in a particular document. The rationale for this form of representation is that the rows of the co-occurrence matrix can be interpreted as *context vectors* for the words in the vocabulary, making it straight-forward to express the distributional similarity of words in terms of vector similarity.

We have chosen a somewhat different methodology to construct the vector space. The technique, which we call *Random Indexing* [3], [4], uses *distributed* representations to accumulate the co-occurrence matrix. The methodology is as follows:

- First, we assign a unique high-dimensional sparse random *index vector* to each document in the text data.
- Then, every time a word occurs in the text data, we add the document's index vector to the row for the word.

Words are thus represented in the co-occurrence matrix by high-dimensional context vectors that contain traces of every document the word has occurred in. These context vectors can now be used to calculate similarity between words using some vector similarity measure, such as the cosine of the angles between the context vectors.

Using distributed representations makes the Random Indexing methodology more efficient and scalable than vector space methods that use local representations, such as Latent Semantic Analysis (LSA [1]) or Hyperspace Analogue to Language (HAL [6]). Also, the Random Indexing mehtod is very robust towards noisy data, and it is "brain-like" in the sense that the distributed representations have a certain cognitive veracity.

In the NTCIR patent retrieval task, we have used Random Indexing to construct a vector space from aligned English-Japanese bilingual corpora. We have then used the vector space to extract the nearest neighbors to a given target word. In effect, what we have produced is an automatically generated bilingual thesaurus.

3 Morphological Analysis of the Training Data

We used the Japanese patent abstracts with English translations provided by NTCIR for years 1995, 1996, and 1997 to train the system.

Morphological analysis of the English abstracts was done using the Functional Dependency Grammar (FDG) parser from Connexor.[2] We used the lemma forms of the English words and discarded words from a stoplist created by collection frequency analysis, retaining most nouns, adjectives, and verbs but discarding most other words. The queries were processed in the same way.

Morphological analysis of the Japanese abstracts was done using the freely available ChaSen analysis system.[3] ChaSen is highly configurable, but time constraints precluded us from delving into the niceties of the system and we ran the system using standard settings to lemmatize words and assign them part-of-speech tags. We then picked out the lemma forms of content words from the output based on their part of speech.[4]

4 Constructing the Bilingual Thesaurus

To construct the bilingual thesaurus using Random Indexing, we assigned a 1,000-dimensional sparse random index vector to each patent abstract in the training data. This means that a patent abstract has the same index vector in both the Japanese and English versions. The 1,000-dimensional random index vectors consisted of 12 randomly distributed -1s and $+1$s (6 of each), with the rest of the elements in the vector set to zero. These parameters are a rough approximation of what have been shown empirically in other experiments to be optimal for other, related tasks [4], [8]. Generally, the Random Indexing technique will perform better the higher the dimensionality of the vectors, but there is a trade-off between performance and efficiency. In the present experiment, we were forced to limit the dimensionality of the vectors to 1,000 dimensions because of computational reasons. This is a serious limitation of the technique, since 1,000 dimensions is probably near the absolute minimum for the technique to produce any kind of intelligible results.

As previously described, the 1,000-dimensional random index vectors were then used to accumulate a

[2] http://www.connexor.com

[3] http://chasen.aist-nara.ac.jp

[4] We noticed that with the settings we used ChaSen analyzed hyphens ("–") as a "symbol" even when inside technical terms written in katakana. We reassembled the technical terms by assuming that any pair of nouns surrounding a hyphen was a potential longer noun, and added it to the document. This addition process was done recursively. The sequence $N_1 - N_2 - N_3$ thus resulted in the nouns N_1, N_2, N_3, $N_1 - N_2$, $N_2 - N_3$, $N_1 - N_2 - N_3$ being added.

co-occurrence matrix by adding a patent's index vector to the row for a given word every time the word occured in that patent. We excluded words with a frequency of less than 5 occurrences, since low frequency words give unreliable statistical estimates. This produced 1,000-dimensional context vectors for the words in the training data that ocurred more than 4 times. We then used these context vectors to extract translations to any given word by computing the cosine of the angle between the context vector for the target word and the context vectors for the other words in the opposite language. The 5 words whose vectors were most similar (i.e. that had the highest correlation) to the context vector of the target word were chosen as translations and entered into the bilingual thesaurus. The cosine of the angles between two vectors is given by:

$$d_{cos}(x,y) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

5 Query Construction and Expansion

Generally, when performing search tasks (such as TREC and similar experiments), automatic procedures are used to select keywords/terms representing the information need. However, in order to get a realistic set of English keywords/terms for the search run, we adopted a more real-life approach [2]. The model is based on the representation of an information need as realistically as possible and is based on human assessors with medium and high level of expertise. In the mandatory run we used <article> and <supplement>, in the optional run we used <description> and finally, in the last run we used the <title>.

The selection of the English keywords/terms was done in 2 steps as follows:

- In the first step, three sets of queries were prepared: one mandatory and 2 optional. An assessor was selected with medium knowledge in the patent domain. For the mandatory run, the assessor was assigned to read through the <article> and <supplement> of the 30 search topics in english and mark out the keywords/terms relevant to that search topic. As a result of this process, a set of query terms was identified. The same procedure was done for the first optional run based on the <description> part of the search topic, and finally, for the second optional run based on the <title> field.

- The next step of query construction involved an assessor with high level of patent domain knowledge and experience. The assessor then judged the query terms selected from the 1st round based on the different information given to the different runs (e.g. mandatory, optional). A set of 10 search topics was randomly selected for inspection/assessment. This second inspection resulted in small adjustments to the terms selected for each search topic. Finally, it must be stressed that the assessor was not a subject expert in any of the 30 search topics.

All in all, three sets of the English queries were prepared, all of them manually: the first conforming to the requirements of the mandatory run, the second set was based on the <description> field, and finally a third set with title words only. The second and third sets were used for the two submitted optional runs.

To construct our Japanese queries, we simply replaced each word in the English queries with the 5 Japanese expansion terms that were entered into the bilingual thesaurus. However, some of the words in the English queries did not have any Japanese expansion terms. The reason for this is that these words were excluded by the frequency threshold, and so were not included in the thesaurus. In one case (query number 21 in the optional run consisting of words from the title only), this meant that the query was left blank, since the only English query word ("tablecloth") occurred only once in the training data, and so was excluded by the frequency threshold. The expanded Japanese queries were finally re-edited into InQuery query syntax (we used the synonym operator for the expanded terms).

6 Retrieval and Results

The retrieval itself was done using an InQuery system set up at SICS. The Japanese abstracts for years 1998 and 1999 were indexed after morphological processing in six separate partitions to speed up the indexing process. Retrieval was then performed in parallel on the six separate indexes, with results merged by retrieval score. The results are shown in Table 1.

Table 1. Results in English-Japanese patent retrieval

Run name	Avg. precision	R-precision
Mandatory run	0.0001	0.0008
Optional run (description)	0.0014	0.0054
Optional run (title)	0.0020	0.0054

7 Failure Analysis

As can be seen in Table 1, the retrieval results can fairly be characterized as disastrously low. For most queries in our submitted runs our results are well under

the median. In fact, most queries did not return one single relevant document. There are many potential reasons for this:

- The ChaSen system performed some unwarranted segmentation of some of the technical terms. This was due to the previously mentioned problem that our set up of ChaSen treated some of the katakana characters (e.g. long vowels in English loan words) as hyphens, thus segmenting the words at the position of the character in question. This obviously reduced the quality of the Japanese data, introducing noise and removing valuable search terms.
- The sheer amount of data available in the patent retrieval task proved to be computationally problematic in various processing stages. For example, we experienced difficulties in running the Connexor FDG parser on the provided English data on our system. Furthermore, the alignment process was problematic due to the large quantity of data that had to be processed.
- In the application of the Random Indexing method, we were forced to limit the dimensionality of the vectors to 1,000 dimensions due to computational reasons. This proved to be a serious limitation of the methodology, since we know from previous experiments that the dimensionality of the vectors must be sufficiently high in order to produce viable results. Of course, insufficient dimensionality does not alone explain the calamitous results, but it certainly is a contributing factor.
- Our hasty indexing procedure in six parallel InQuery instances with post-retrieval merging introduced a large number of potential error sources that may have contributed to the disappointing end results. For instance, we noted that some of the queries did not retrieve the required 1000 documents. To solve this problem, we padded the result lists with random material to conform to experiment requirements. This obviously introduced noise into the results.
- Most importantly, our lack of Japanese precluded us from checking results during the various processing stages. Even simple errors during the preprocessing of both training and retrieval data may have passed unnoticed — only due to a chance re-check did we notice that a bug in the morphological analysis script first mis-aligned document id with document text. That bug we caught, but several may have remained. Also, since we could not inspect the queries or the data for sanity, we do not even know whether the encoding of the Japanese characters made any sense.
- There might also have been linguistic factors contributing to the disastrous results. Japanese is of course inherently very different from European languages, with a different structure and a different notion of context. Our inability to read Japanese precluded us from modifying the distributional methodology to conform to the contextual distinctiveness of the Japanese language. Furthermore, the fact that Japanese uses three different writing systems (katakana, hiragana and kanji) is highly problematic for our methodology, since it might be the case that some word occurs in more than one writing system in the data, and the queries only feature words in one of the possible writing systems, or that some word occurs in one writing system in the queries and another in the data.

8 Lessons Learned

Our results are admittedly discouraging. However, we learned a few valuable lessons during the course of the NTCIR campaign. Most importantly, our lesson is not to try processing an unknown language with unfamiliar tools. Linguistic competence is absolutely necessary in order to be able to sanity check the different processing stages. As previously mentioned, even very small bugs may affect the results if they go unnoticed. Collaboration with a Japanese counterpart is a prerequisite for us to be able to participate in future NTCIR campaigns.

Although our results leave much to wish for, we believe that the overall methodology is viable and worth investigating further. We have reached satisfactory results using Random Indexing for query expansion in other evaluation campaigns — CLEF 2002[5] [9] and CLEF 2001 [8] — and we intend to pursue our query expansion experimentation further. We believe that it is important to look closer on query construction based on real-life situations [5] and to investigate term clustering, human term weighting, and human assessment of term-term relationships within a query. Retrieving patent documents involves a complex process of understanding the problem and the text in a specific situation [2].

Acknowledgements The work reported here is partially funded by the European Commission under contracts IST-2000-29452 (DUMAS) and IST-2000-25310 (CLARITY) which is hereby gratefully acknowledged.

[5] http://clef.iei.pi.cnr.it/

References

[1] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[2] Hansen, P. & Kalervo, J. The Information Seeking and Retrieval process at the Swedish Patent- and Registration Office. Moving from Lab-based to real life work-task environment. *Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval*, pp.43–53, 2000.

[3] Kanerva, P., Kristofersson, J. & Holst, A. Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p.1036, 2000.

[4] Karlgren, J. & Sahlgren, M. From Words to Understanding. In Kanerva et al. (eds.) *Foundations of Real World Intelligence.* CSLI publications, Stanford 2001, pp.294–308.

[5] Karlgren, J. & Hansen, P. Cross-language Relevance Assessment and Task Context. In Peters et al. (eds.) *Results of the CLEF 2002 Cross-Language System Evaluation Campaign*, pp.255–260.

[6] Lund, K. & Burgess, C. Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments and Computers*, 28(2):203–208, 1996.

[7] Oard, D. (ed.) Cross-Language Text and Speech Retrieval. Papers from the 1997 AAAI Spring Symposium, AAAI Technical Report SS-97-05.

[8] Sahlgren, M. & Karlgren, J. Vector-Based Semantic Analysis Using Random Indexing for Cross-lingual Query Expansion. In Peters et al. (eds.) *Evaluation of Cross-Language Information Retrieval Systems*, Springer 2002, pp.169–176.

[9] Sahlgren, M., Karlgren, J., Cöster, R. & Järvinen, T. SICS at CLEF 2002: Automatic Query Expansion Using Random Indexing. In Peters et al. (eds.) *Results of the CLEF 2002 Cross-Language System Evaluation Campaign*, pp.217–224.

[10] Schütze, H. Dimensions of Meaning. *Proceedings of Supercomputing '92*, pp.787–796, 1992.