Query Translation Method for Cross Language Information Retrieval

Hiroshi Masuichi
Center for the Study of Language and Information
Stanford University

Raymond Flournoy
Department of Computer Science
Stanford University

Stefan Kaufmann
Department of Linguistics
Stanford University

Stanley Peters
Department of Linguistics
Stanford University

Abstract

This paper proposes a method of query translation for Cross Language Information Retrieval. The method uses a parallel bilingual corpus to produce word vectors and can readily be applied to monolingual vector-retrieval models. The Cross Language Information Retrieval system produced with the method showed 97.4% accuracy in our preliminary tests of finding the counterparts in a parallel corpus of English and Japanese documents.

1 Introduction

A cross language information retrieval (CLIR) system is a system for retrieving documents across language boundaries. A query written in one language should be translated into a representation for finding documents in another language. In this paper, we propose a method for the translation which uses a parallel corpus and can readily be combined with the **information mapping** approach[1][2].

This paper is organized as follows: In Section 2, we overview the information mapping approach. Section 3 introduces our basic idea for the query translation across languages. Section 4 describes an experimental procedure, the results it produced, and an analysis of the results. The conclusion is given in Section 5.

2 Overview of Information Mapping

Information mapping is basically a variant of the vector-retrieval method [3], and is based on an approach first proposed by Hinrich Schütze [2]. The approach is closely related to Latent Semantic Indexing [4], and the difference between these two is discussed in [5]. In the information mapping approach, we use a multidimensional vector space, called **word space** for defining the association of words. To represent the word space, we begin with a large word-by-word matrix. We prepare a list of *n* **content-bearing words** and *m* **vocabulary words**, and for each vocabulary word we scan all documents in the training corpus and count the total cooccurrences between the vocabulary word and each content-bearing word. It is typical that the most frequently appearing *n* words in the training corpus are selected as content-bearing words and the most frequently appearing *m* words as vocabulary words. In this way, we produce for each vocabulary word an *n*-dimensional vector which represents the word's distributional behavior. This simple representation has problems, however: high dimensionality and sparseness. In order to solve these problems, the original *n*-dimensional vector space is converted into a condensed, lower-dimensional, real-valued matrix using Singular Value Decomposition (SVD) [6][7].

We call the lower-dimensional vector space word space. Using this word space, we define the proximity of two word vectors as the cosine of the angle between them. If two words' vectors have high proximity, then the words tend to occur in similar contexts and, in our terms, are associative. By clustering word vectors on the basis of proximity, the word space can be used for word sense disambiguation and thesaurus construction [2][8][9].

For IR, a document vector and a query vector are calculated by summing the vectors corresponding to the words in the document or the query, and the proximity between the two vectors is defined in the same way as the word proximity (i.e. the cosine of the angle between their vectors) [1].

3 The Basic Idea for Cross Language Information Retrieval

One advantage of information mapping is that a query can lead to retrieval of a relevant document even if the two have no words in common. This feature is critical in CLIR because direct word matching is of little use to retrieve documents written in a different language than the query.

Our basic idea for applying the information mapping approach to CLIR is to produce a vector for a word in one language L with content-bearing words in another language L'. We use a bilingual parallel corpus to produce the word vectors. In the parallel corpus, each document written in language L has its translation written in L'. For CLIR, we consider this pair of corresponding documents as a single compound document, and produce a word space representation of the words in both L and L'. However, we limit the

content-bearing words to words in L' only; in other words, all words in L and L' alike are represented by words in L' . Now all word vectors, and therefor a query, in L can be located in the word space represented by L' ; thus all the word vectors, and therefore all document vectors, in L' are located in the same word space. We can retrieve documents written in L' from a query written in L using this word space. The word space based on the content-bearing words in L' is identical to the word space for monolingual information mapping on L' , so it is possible to add the CLIR mechanism for a query in L to an existing information mapping system on L' without modifying the system.

4 Experimental Tests and Analysis

4.1 Tests of Finding Counterparts in Parallel Corpus

We used an English-Japanese bilingual patent abstract corpus for our experimental tests. Each Japanese patent abstract in the corpus has an English translation by humans. We used only the title and abstract texts and removed all other information, such as author, patent ID and issue date. Table 1 shows an example of an English-Japanese pair in the corpus. We randomly chose 1000 English-Japanese pairs as a training corpus (called A-type corpus). Four distinct sets of 100 English-Japanese pairs were chosen at random as test corpora (B-type corpora). No patent appears in any two of these five corpora (i.e. one A-type corpus and four B-type corpora). We constructed a word space from the A-type corpus. Because all the fields of patents - ranging from Medicine and Agriculture to Computers - are randomly included in the corpora, we selected a larger corpus for training than for testing in order to obtain word vectors for a broad enough vocabulary.

Japanese words were segmented using the Chasen morphological analyzer [10] and English words were also tokenized and stemmed [11][12]. All characters in the English texts are 1-byte characters and all characters, including alphabetical and numerical characters, in the Japanese texts are 2-byte, so there is no word which is shared by both English and Japanese texts. We selected the most frequently appearing 6000 English and 6000 Japanese words in the A-type corpus for vocabulary words and the most frequently appearing 1000 English words, a subset of the English vocabulary words, for content-bearing words. (Stop words were ignored.) Initially, a matrix of 12000 rows (for 6000 Japanese vocabulary words plus 6000 English vocabulary words) and 1000 columns (for content-bearing words) was produced. Each element of the matrix is the total number of cooccurrences between a vocabulary word and a content-bearing word in the scope of an English-Japanese pair. We treated each pair of patents as a bag of words; no other information such as word order was used. This 12000*1000 matrix was then reduced to 12000*200 matrix by means of SVD. Using this word space, 200 document vectors corresponding to 100 Japanese patents and 100 English patents were calculated for each B-type corpus by simply summing the vectors corresponding to the vocabulary words in each patent.

METHOD AND EQUIPMENT FOR PROTECTING CONCRETE PIER ETC. Abstract:PURPOSE: To improve the aseismicity of a pier etc. made of concrete. CONSTITUTION: The heating apparatus 5 of an induction heating coil and a cooling device 6 for quenching are disposed around a metallic material 2 such as an iron material covered on a pier 1etc. made of concrete. The metallic material 2 is inductively heated by the heating apparatus 5 of the induction heating coil, and quenched by the cooling device 6, and clearances 4 between concrete 3 and the metallic material 2 are removed, thus improving the aseismicity of the pier 1 etc. made of concrete.
コンクリート橋脚等の保護方法および装置【要約】【目的】コンクリート製の橋脚等の耐震性を高めることにある。【構成】コンクリート製の橋脚１等に被覆した鉄材等の金属材２まわりに、誘導加熱コイルの加熱装置５と急冷用の冷却装置６を配設し、金属材２を誘導加熱コイルの加熱装置５で誘導加熱し、冷却装置６で急冷して、コンクリート3と金属材２との間の間隙４をなくし、コンクリート製の橋脚１等の耐震性を高めるようにしている。

Table 1: An example of an English-Japanese patent pair

Our initial test task was to use a Japanese patent as a query and see the retrieved rank of its corresponding English patent. Ideally the translation of the query patent would be retrieved in the first rank. This procedure was repeated 4 times (TEST1-4), once for each B-type corpus. Table 2 shows the results: for each rank the number of Japanese patents which retrieved its corresponding English patent at that rank. Table 3 shows results for the opposite case, when query patents are English and retrieved patents are Japanese. In both tests, we used English words as content-bearing words.

rank	TEST1	TEST2	TEST3	TEST4
1st	98	100	96	100
2nd	1	-	3	-
3rd	1	-	1	-

Table 2: Test results of English content-bearing words and Japanese queries

rank	TEST1	TEST2	TEST3	TEST4
1st	98	98	96	99
2nd	2	2	1	1
3rd	-	-	3	-

Table 3: Test results of English content-bearing words and English queries

Tables 4 and 5 show the results when the most frequently appearing 1000 Japanese words are used as content-bearing words. The query patents are Japanese in Table 4 and English in Table 5. All four Tables show good results; a total of 1559 out of 1600 (97.4%) queries retrieved their counterpart patents at the first rank.

rank	TEST1	TEST2	TEST3	TEST4
1st	98	100	95	96
2nd	2	-	4	4
3rd	-	-	1	-

Table 4: Test results of Japanese content-bearing words and Japanese queries

rank	TEST1	TEST2	TEST3	TEST4
1st	99	94	95	97
2nd	1	4	5	3
3rd	-	2	-	-

Table 5: Test results of Japanese content-bearing words and English queries

We also conducted tests in which we used 2000 mixed-language, English and Japanese, words as content-bearing words. In these tests, the most frequently appearing 1000 English words, which were used in the tests of Tables 2 and 3, plus the most frequently appearing 1000 Japanese words, used in the tests of Tables 4 and 5, were used for content-bearing words. We produced 12000*2000 matrix then reduced it to 12000*200 matrix by means of SVD. Table 6 shows the results when the query patents are Japanese and Table 7 shows the results when English. A total of 735 out of 800 (91.9%) queries retrieved their counterpart patents at the first rank. In addition, tests in which we used 1000 mixed-language words (the most frequently appearing 500 English words plus the most frequently appearing 500 Japanese words) as content-bearing words led to worse results than the results of Tables 6 and 7.

The results of Tables 2-5 are better than the results of Tables 6 and 7. Therefore we believe these results show the word space based on single-language content-bearing words is a more direct and accurate representation across a language boundary than a word space based on mixed-language content-bearing words.

rank	TEST1	TEST2	TEST3	TEST4
1st	94	96	90	98
2nd	1	3	5	1
3rd	1	-	1	1
4-10th	4	1	3	-
11-20th	-	-	1	-

Table 6: Test results of English-Japanese content-bearing words and Japanese queries

rank	TEST1	TEST2	TEST3	TEST4
1st	87	92	90	88
2nd	7	7	6	8
3rd	2	1	2	1
4-10th	3	-	2	3
11-20th	1	-	-	-

Table 7: Test results of English-Japanese content-bearing words and English queries

4.2 Short Query Tests

The other additional tests we conducted were tests for short queries which are used in real world information retrieval systems. In the tests described above, the queries were long; a patent abstract was regarded as a query. For the tests with short queries, we randomly selected one sentence from each Japanese patent abstract in the B-type corpora and regarded the sentence as a query instead of the patent abstract. Note that the sentence was randomly selected, so it is not necessarily an appropriate query to represent the whole patent abstract from which the sentence was extracted.

Table 8 shows the results of short query tests in which we replaced all the Japanese patent abstract queries in the tests of Table 2 with the Japanese sentences extracted from the patent abstracts, while all the other settings such as content-bearing words and vocabulary words remained the same as in the tests of Table 2. A total of 334 out of 400 (83.5 %) counterpart English patent abstracts were correctly retrieved at the first rank and 361 out of 400 (90.3 %) were retrieved in the top 3 by the Japanese sentences.

rank	TEST1	TEST2	TEST3	TEST4
1st	83	85	81	85
2nd	3	5	3	6
3rd	1	3	4	2
4-10th	10	4	5	5
11-20th	2	-	4	-
21-80th	1	3	3	2

Table 8: Test results of English content-bearing words and Japanese sentence queries

rank	TEST1	TEST2	TEST3	TEST4
1st	74	75	79	82
2nd	14	9	7	8
3rd	4	2	4	2
4-10th	4	5	7	6
11-20th	3	7	2	-
21-80th	1	2	1	2

Table 9: Test results of English content-bearing words and English sentence queries

Table 9 shows the results of short query tests in which we replaced all the English patent abstract queries in the tests of Table 3 with the English sentences, while all the other settings remained the same as in the tests of Table 3. A total of 310 out of 400 (77.5 %) counterpart Japanese patent abstracts were correctly retrieved at the first rank and 360 out of 400 (90.0 %) were retrieved in the top 3 by the English sentences. The number of Japanese patents which were correctly retrieved at first rank by English sentences (in Table 9) is smaller than the number of English patents by Japanese sentences (in Table 8). The reason is that 5.99 sentences are included in an English patent abstract on average, as opposed to 3.98 sentences in a Japanese patent abstract, therefore a Japanese sentence has more information than an English sentence.

5 Conclusion

We proposed a method of query translation for CLIR. The method is based on single-language content-bearing words and can be easily applied to a monolingual information mapping system. The CLIR system produced with the method showed 97.4% accuracy in our preliminary tests of finding the counterparts in a parallel corpus of English and Japanese patents. The results were better than a method with mixed-language content-bearing words.

References

[1] Flournoy, R., Masuichi, H. and Peters, S. (1998). "Cross-Language Information Retrieval: Some Methods and Tools". In Proceedings of 14th Twente Workshop on Language Technology.

[2] Schütze, H. (1995). "Ambiguity Resolution in Language Learning: Computational and Cognitive Models". PhD thesis, Stanford University, Department of Linguistics.

[3] Salton, G. and MacGill, M. (1983). "Introduction to Modern Information Retrieval". McGrowHill.

[4] Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshamn, R. (1990). "Indexing by latent semantic analysis". Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391-407.

[5] Schütze, H. and Pedersen, J. (1997). "A coocurrence-based thesaurus and two applications to information retrieval". Information Processing & management, Vol. 33, No. 3, pp. 307-318.

[6] Berry, M. (1992). "Large Scale Singular Value Computations". International Journal of Supercomputer Applications, Vol. 6, No. 1, pp. 49.

[7] Berry, M., Do, T., O' Brien, G., Krishna, V. and Varadhan, S. (1993). "SVDPACKC USER' S GUIDE". Tech. Rep. CS-93-194. University of Tennessee, Knoxville, TN.

[8] Schütze, H. (1998). "Automatic Word Sense Discrimination". Computational Linguistics, Vol. 24, Issue 1, pp. 97-123.

[9] Kikui, G. (1998). "Term-list Translation using Mono-lingual Word Co-occurrence Vectors". Project Note, COLING-ACL '98.

[10] Matsumoto, Y., Kitauchi, K., Yamashita, T., Imaichi, S. and Imamura, T. (1997). "Japanese Morphological Analysis System: Chasen User' s Manual". NAIST Technical Report, NAIST-IS-TR97007.

[11] Frakes, W. and Baeza-Yates, R. (1992). "Information Retrieval, Data Structures and Algorithms". Englewood Cliffs, NJ, Printice Hall.

[12] Porter, M. (1980). "An algorithm for suffix stripping". Program, Vol. 14, pp. 130-137.