A Bootstrapping Method for Extracting Bilingual Text Pairs

Hiroshi Masuichi [†] **Raymond Flournoy** [‡] **Stefan Kaufmann** [‡] **Stanley Peters** [‡]

[†] Fuji Xerox Co., Ltd.
Corporate Research Center
430 Sakai, Nakai-machi, Ashigarakami-gun,
Kanagawa 259-0157, Japan

[‡] Center for the Study of Language and Information
Stanford University
210 Panama Street, Stanford,
CA 94305-4115, U.S.A.

{masuichi, flournoy, kaufmann, peters}@csli.stanford.edu

Abstract

This paper proposes a method for extracting bilingual text pairs from a comparable corpus. The basic idea of the method is to apply bootstrapping to an existing corpus-based cross-language information retrieval (CLIR) approach. We conducted preliminary tests with English and Japanese bilingual corpora. The bootstrapping method led to much better results for the task of extracting translation pairs compared with a corpus-based CLIR method without bootstrapping, and the extracted translation pairs could be useful training data for improving results of the corpus-based CLIR method.

1 Introduction

A parallel corpus is an important resource for corpus-based approaches to CLIR. These approaches use parallel corpora as statistical training data and then retrieve documents written in a language different from that of the query. One disadvantage of these approaches is lack of resources. Parallel corpora are not always readily available and those that are available tend to be relatively small or to cover only a small number of subjects.

A bilingual comparable corpus is a set of texts in two different languages from the same domain or on the same topic. Unlike a parallel corpus it is composed independently in the respective language text sets. It can be more readily obtained from the Internet or CD-ROM resources than parallel corpora. Zanettin (1998) introduced several available bilingual comparable corpora such as news paper articles selected by dates and subject codes, medical articles from journals and textbooks, and articles for tourists from brochures and guides. Zanettin (1994) also reported that it is highly likely that much relevant information can be found across languages in a topic-related bilingual comparable corpus. In this paper, we propose a method for extracting bilingual text pairs which share the same information from a bilingual comparable corpus, and show the possibility that the resulting bilingual text pairs can be useful for corpus-based CLIR approaches when we use them as training data instead of a parallel corpus.

Sheridan (1998) also proposed an approach to building multilingual test collection from comparable corpora consisting of news articles. The idea is to reduce the work of manual relevance judgements by restricting news articles to be examined to a couple of days. Disadvantages to this approach are that it relies on time-sensitive texts, texts obtained by this approach are constrained to referencing specific events, and nontrivial work by humans is still necessary. On the other hand, our goal is to extract bilingual text pairs automatically from any kind of bilingual comparable corpora.

This paper is organized as follows: Section 2 introduces the basic idea for extracting relevant text pairs from a bilingual comparable corpus. Our method is based on a corpus-based CLIR method, so we overview previous corpus-based CLIR approaches in Section 3. Section 4 describes an experimental procedure, the results it produced, and an analysis of the results. The conclusion is given in Section 5.

2 The Basic Idea

As we will describe in Section 3, several CLIR approaches that rely on parallel corpora have been proposed and lead to successful retrieval results. In those approaches, a parallel corpus used as training data should be large enough to obtain good retrieval results. Although we use a CLIR method which relies on a parallel corpus, we begin with a very small parallel corpus. We retrieve bilingual text pairs from a bilingual comparable corpus using the small parallel corpus as training data. Then we concatenate the text pairs to the initial small parallel corpus and grow the parallel corpus by iterating the retrieval and concatenation processes (Figure 1).

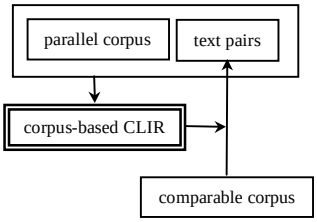

Figure 1: The bootstrapping method

This kind of bootstrapping method has a problem, however: It is highly sensitive to the accuracy of the text pairs obtained in the early stages of the iterations. In order to solve this problem, we concatenate only a small number of the most "reliable" text pairs to the initial parallel corpus in the early stages, then gradually increase the number of the text pairs which are concatenated to the initial parallel corpus. We will describe the details of the method in Section 4.

3 Corpus-based CLIR approaches

3.1 Previous Researches

As we mentioned in Section 2, we use a CLIR method which relies on a parallel corpus in our bootstrapping method. One approach to corpus-based CLIR is to use the Latent Semantic Indexing technique proposed by Furnas et al. (1988) on a parallel corpus to construct a language independent representation of queries and documents (Landauer and Littman, 1990). Another approach that relies on a parallel corpus has been suggested by Dunning and Davis (1993). Their method is based on the vector space model and involves the linear transformation of the representation of a query. A parallel corpus can also be used to enhance existing knowledge-based resources. The resources are used to translate the query and then classical IR matching techniques are applied to compute the similarity between the translated query and documents (Hull and Grefenstette, 1996).

3.2 Information Mapping for CLIR

For our bootstrapping method, we adopted a CLIR method which is based on the Information Mapping approach (Masuichi et al., 1999). Information Mapping is basically a variant of the vector space model, and is based on an approach first proposed by Schütze (1995). The approach is closely related to Latent Semantic Indexing, and the difference between these two is discussed in Schütze and Pedersen (1997). Note that our bootstrapping method does not depend on any particular properties of the Information Mapping approach, so it could employ other corpus-based CLIR methods such as Latent Semantic Indexing.

Information Mapping begins with a large word-by-word matrix. A list of *n* content-bearing words and *m* vocabulary words correspond to the columns and the rows of the matrix. The most frequently appearing *n* words in a training corpus are selected as content-bearing words and the most frequently appearing *m* words as vocabulary words. Each cell of the matrix holds the number of total cooccurrences between a content-bearing word and a vocabulary word in the training corpus. In this way, an *n*-dimensional vector which represents the word's distributional behavior is produced for each vocabulary word. Then the original *n*-dimensional vector space is converted into a condensed, lower-dimensional, real-valued matrix using Singular Value Decomposition (SVD) (Berry, 1992). The lower-dimensional vector space is called word space. A document vector and a query vector are calculated by summing the vectors corresponding to the vocabulary words in the document or the query, and the proximity between the two vectors is defined as the cosine of the angle between them.

To apply this method to CLIR, we regard each translation pair in a training parallel corpus of language L1 and L2 as a single compound document and create a word-by-word matrix and then a word space. The word space represents a language independent vector space for vocabulary words in both L1 and L2, and therefore query and document vectors in both L1 and L2 can be calculated and compared in the same word space.

4 Experimental tests and Results

4.1 Tests with complete-pair corpora

We used an English-Japanese bilingual patent text corpus for our experimental tests. For our first test, we prepared 1000 English-Japanese patent text pairs as a pseudo bilingual comparable corpus. For each Japanese patent text in the corpus, its English translation by humans exists[1], so this corpus could be regarded as an ideal bilingual comparable corpus. We also prepared 100 pairs as an initial parallel corpus (a training corpus) to create an initial word space. All the patents

[1] The quality of the translations varies greatly from word-for-word translations to short summaries.

in the two corpora were randomly selected from the Japanese patents issued in 1991, and the two corpora shared no patent. 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 corpora. 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 used all words which appeared in a training corpus as vocabulary words, and the most frequently appearing 3000 English words as content-bearing words and then reduced the dimension of the vectors from 3000 to 200 by SVD.

Hose for Transferring Fertilizer from Fertilizer Tanki of Mobile Farm Machine Abstract: PROBLEM TO BE SOLVED: To provide a mechanism to arrange a fertilizer transfer hose from a fertilizer tank without causing hindrance to the other mechanisms, etc. SOLUTION: A fertilizer transfer hose 38 to deliver a fertilizer from a fertilizer tank 31 placed at a side of a mobile machine body 1 to the downstream side of a fertilizing part 28 is laid along the outer circumference of a passage 23 placed along the back and a side of a driver's seat 8 and extending from the driver's seat 8 to a working machine 11.
移動農機における肥料タンクの送肥ホース【要約】【課題】肥料タンクからの送肥ホースを他の機構等の妨げにならないように配管する機構を提供する。【解決手段】走行機体1側に設けられた肥料タンク31から施肥部28下流側に肥料を繰り出す送肥ホース38を、運転席8の後方及び側方に設けた運転席8から作業機11側への通路23の外側周縁に沿わせて設けた移動農機における肥料タンクの送肥ホース。

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

We began with a word space created from the 100 English-Japanese translation pairs (the initial parallel corpus). Then using the word space, we calculated 1000 English patent vectors and 1000 Japanese patent vectors which correspond to the patent texts in the pseudo comparable corpus. Next we extracted English-Japanese patent pairs which satisfied the simple condition that the English patent vector in the pair has the highest proximity (the biggest cosine) with the Japanese patent vector in the pair among the 1000 Japanese patent vectors, and vice versa (hereafter we call these pairs mutual-proximity pairs). Note that mutual-proximity pairs are, of course, not always correct translation pairs. Then we selected the 10 most "reliable" mutual-proximity pairs, assuming that the higher the proximity between the two vectors of a mutual-proximity pair, the more reliable the mutual-proximity pair is. Finally we concatenated the 10 mutual-proximity pairs to the initial 100 translation pairs. This is the first stage of our bootstrapping method.

In the second stage, we created a new word space regarding the 110 English-Japanese pairs obtained in the first stage as a training corpus. Then we selected the 20 most reliable mutual-proximity pairs and concatenated them to the initial 100 patent translation pairs.

At the *N*th stage, we selected the *N**10 most reliable mutual-proximity pairs. If the number of the mutual-proximity pairs obtained in the stage is less than *N**10, all of the mutual-proximity pairs were concatenated to the initial 100 patent translation pairs.

We repeated this procedure up to the 100th stage. At the 100th stage, we obtained 727 mutual-proximity pairs and 721 pairs out of the 727 pairs were correct translation pairs. Therefore the recall of the obtained pairs was 72.1% (721/1000) and the precision was 99.2% (721/727) (see the column of Test1 and the row of the "bootstrapping method" of Table 2). On the other hand, we obtained 341 mutual-proximity pairs and 258 pairs out of the 341 pairs were correct translation pairs in the case of the normal Information Mapping method which corresponds to the first stage of our bootstrapping method. In this case, the recall was 25.8% and the precision was 75.7% (see the column of Test1 and the row of the "normal method" of Table 2).

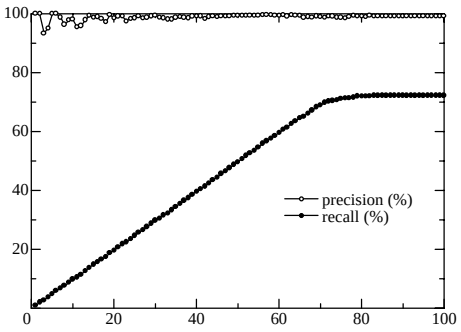

Figure 2: The change of precision and recall with complete-pair corpus

Figure 2 shows the change of the precision and the recall through the 100 stages. The precision was kept over 93.3% and the recall went up gradually. We could successfully grow the bilingual text pairs using bootstrapping.

		TEST1	TEST2	TEST3	TEST4	TEST5
normal method	Prec	75.7	75.6	76.6	78.2	72.8
	Rec	25.8	26.6	26.9	25.4	27.1
boot-strapping method	Prec	99.2	99.1	99.7	98.9	98.7
	Rec	72.1	74.0	73.0	71.0	70.6

Table 2: Results of extracting tests with complete-pair corpus

We prepared 4 more different sets of 1000 pairs for pseudo comparable corpora and different sets of 100 pairs for initial parallel corpora, and repeated the same test 4 more times. Table 2 shows results of the 5 tests of the bootstrapping method and the normal Information Mapping method. In each case the bootstrapping method could drastically improve both the precision and the recall.

We also conducted tests to see if the resulting text pairs obtained at the 100th stage in the previous tests are useful for the normal Information Mapping method. We prepared another 1000 English-Japanese patent translation pairs for each of the 5 previous tests as evaluation corpora. No same patents were shared between any two of all the corpora. We extracted mutual-proximity pairs from the new 1000 English-Japanese pair with the normal Information Mapping method, using (1) the initial parallel corpus in the previous test, (2) the initial parallel corpus + the mutual-proximity pairs obtained in the previous test, (3) the initial parallel corpus + the 1000 English-Japanese correct translation pairs in the pseudo comparable corpus of the previous test, as a training corpus respectively. For example, in Test 1, the number of pairs in the training corpus is 100 for (1), 827 with 6 error pairs for (2) and 1100 for (3).

		TEST1	TEST2	TEST3	TEST4	TEST5
initial pairs	Prec	77.5	77.3	73.3	75.6	75.4
	Rec	23.8	29.3	26.1	25.1	25.8
initail + boot-strapping pairs	Prec	98.9	98.7	98.8	99.1	99.2
	Rec	74.5	75.0	75.1	75.0	73.5
initail + complete pairs	Prec	99.0	99.1	99.6	98.7	98.7
	Rec	77.5	79.3	79.0	77.4	78.6

Table 3: Results of evaluation tests for complete-pair corpus

Table 3 shows the results. The results of (3) can be considered as the ceilings of the precision and the recall, because we used all the correct translation pairs in the pseudo comparable corpus. In each case, both the precision and the recall of (2) is very close to the ceilings, so we think the bilingual text pairs obtained by our bootstrapping method is useful as a training corpus for the normal Information Mapping method.

4.2 Tests with incomplete-pair corpora

In the tests described above, we used the ideal pseudo comparable corpus. As described in the Introduction, it is highly likely that a real bilingual comparable corpus includes bilingual pairs which share the same information, but it also includes a lot of irrelevant texts. To simulate this, we replaced half of the Japanese patent texts in the pseudo comparable corpora of the previous tests with different Japanese patent texts which were randomly selected. Therefore the corpus included 500 English-Japanese translation pairs, and 500 English patents and 500 Japanese patents which were totally irrelevant to each other.

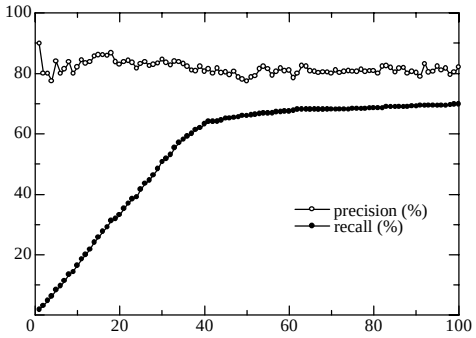

Figure 3: The change of precision and recall with 50%-error-pair corpus

		TEST1	TEST2	TEST3	TEST4	TEST5
normal method	Prec	55.3	50.0	52.8	46.6	53.5
	Rec	28.4	26.8	28.0	25.0	29.4
boot-strapping method	Prec	82.1	81.4	83.8	81.0	80.7
	Rec	69.8	70.8	67.4	67.4	69.2

Table 4: Results of extracting tests with 50%-error-pair corpus

		TEST1	TEST2	TEST3	TEST4	TEST5
initial pairs	Prec	77.5	77.3	73.3	75.6	75.4
	Rec	23.8	29.3	26.1	25.1	25.8
initail + boot-strapping pairs	Prec	96.2	95.7	93.4	93.3	95.9
	Rec	61.1	61.9	59.8	57.4	60.5
initail + complete pairs	Prec	98.4	98.7	97.9	98.0	98.9
	Rec	66.1	70.7	68.6	69.0	71.1

Table 5: Results of evaluation tests for 50%-error-pair corpus

Results are shown in Figure 3, Table 4 and Table5, which correspond to Figure 2, Table 2 and Table 3 respectively.

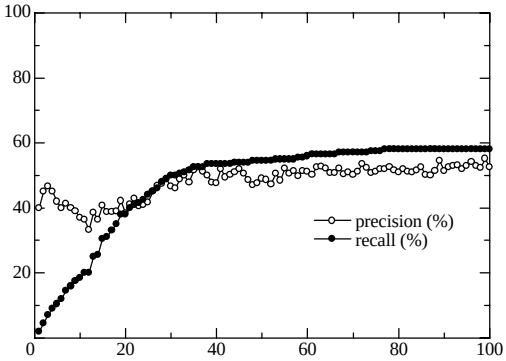

Figure 4: The change of precision and recall with 80%-error-pair corpus

		TEST1	TEST2	TEST3	TEST4	TEST5
normal method	Prec	23.4	17.8	20.7	21.1	27.0
	Rec	27.5	20.0	24.5	23.5	30.0
boot-strapping method	Prec	52.5	55.2	50.9	53.2	53.4
	Rec	58.0	53.0	55.0	53.5	50.2

Table 6: Results of extracting tests with 80%-error-pair corpus

		TEST1	TEST2	TEST3	TEST4	TEST5
initial pairs	Prec	77.5	77.3	73.3	75.6	75.4
	Rec	23.8	29.3	26.1	25.1	25.8
initail + boot-strapping pairs	Prec	82.9	85.5	81.1	83.5	85.4
	Rec	37.9	36.3	35.3	33.5	33.4
initail + complete pairs	Prec	96.1	96.4	96.0	94.0	95.7
	Rec	54.7	58.7	55.0	55.3	53.4

Table 7: Results of evaluation tests for 80%-error-pair corpus

Figure 4, Table 6 and Table 7 show results in the case that we replaced 80% of Japanese patent texts with irrelevant Japanese patent texts.

The results of these tests are not as good as the results of tests with the ideal pseudo comparable corpora. Figure 4 and 6 show, however, the bootstrapping method improved both the precision and recall of the extracted text pairs as compared to the normal method. Figure 5 and 7 also show that the bilingual text pairs obtained by the bootstrapping method are still useful as a training corpus for the normal method.

5 Conclusion

We proposed a method of extracting bilingual text pairs from a comparable corpus. The method is based on an existing corpus-based CLIR method and uses bootstrapping. Although our research is in the preliminary stage of development and tested with artificial corpora consisting of English and Japanese patent texts, the bootstrapping led to much better results for the task of extracting translation pairs than the results produced by a normal CLIR method, and the extracted translation pairs could be useful for improving the results of the normal CLIR when we used them as a training corpus.

References

Berry, M. W. (1992) *Large Scale Singular Value Computations.* International Journal of Supercomputer Applications, 6/1, pp. 13-49.

Dunning T. E. and Davis M. W. (1993) *Multi-lingual information retrieval.* Computational Memoranda in Cognitive and Computer Science MCCS-93-252, New Mexico State University, Computing Research Laboratory.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. and Lochbaum, K. E. (1988) *Information retrieval using a singular value decomposition model of latent semantic structure.* In proceedings of the 11th ACM International Conference on Research and Development in Information Retrieval, pp. 465-480.

Hull, D. and Grefenstette, G. (1996) *Querying across languages: A dictionary-based approach to multilingual information retrieval.* In Proceedings of SIGIR'96, pp. 49-57.

Landauer, T. K. and Littman, L. M. (1990) *Fully automatic cross-language document retrieval using latent semantic indexing.* In Proceedings of the 6th Conference of University of Wterloo Centre for the New Oxford English Dictionary and Text Research, pp. 31-38.

Masuichi, H., Flournoy, R., Kaufmann, S. and Peters, S. (1999) *Query Translation Method for Cross Language Information Retrieval.* In Proceedings of the Workshop on Machine Translation for Cross Language Information Retrieval, MT Summit VII, pp. 30-34.

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

Schütze, H. and Pedersen, J. (1997) *A coocur-rence-based thesaurus and two applications to information retrieval.* Information Processing & management, 33/3, pp. 307-318.

Sheridan, P., Ballerini, J. P. and Schäuble, P. (1998) *Building a large multilingual test collection from comparable news documents.* In "Cross-Language Information Retrieval", Kluwer Academic Publishers, pp. 137-150.

Zanettin, F. (1994) Parallel Words: Designing a Bilingual Database for Translation Activities. In "Corpora in Language Education and Research: a Selection of Papers from TALC 94", Lancaster University, UK, pp. 99-111.

Zanettin, F. (1998) *Bilingual comparable corpora and the training of translators.* In "META, XLIII, 4, Special Issue. The corpus-based approach: a new paradigm in translation studies", pp. 616-630.