A Method using Language Grid and Concept Base for Japanese-English Cross-language Information Retrieval

Pham Huy Anh[1]and Yukawa Takashi[2]

[1] Department of Information Science and Technology, Nagaoka University of Technology, Nagaoka-shi, 940-2188 Japan

[2] Department of Information Science and Technology, Nagaoka University of Technology, Nagaoka-shi, 940-2188 Japan

Abstract

This paper describes query translation using language resources and a concept base method for Cross-language Information Retrieval (CLIR). In the proposed method, queries are translated by multiple machine translation systems on the Language Grid. The queries are then expanded by using a bilingual dictionary to translate compound words or word phrases. In addition, documents related to the translated query are retrieved with a TF-IDF term weighting model. The top 100 retrieved documents are re-ranked by a specificity-considered concept base with the noun phrases and compound words extracted from the query. The re-ranked results are combined with the results retrieved by the probabilistic model. For evaluation of the proposed method, we use the average precision of the non-interpolated recall and precision to compare our method with the NTCIR1 participation systems. The proposed method achieved the highest precision.

Keywords: *Cross-language Information Retrieval, CLIR, Language resources, Concept base, Language Grid.*

1. Introduction

The number of electronic documents on the Internet has rapidly increased. As a result, documents containing the kinds of information required by a user are not limited to those written in the user's native language. Therefore, research on Cross-language Information Retrieval (CLIR), which uses a query in one language to retrieve documents in another language, is especially of interest. In the NTCIR (NII-NACSIS Test Collection for IR Systems) workshop, [1] CLIR is one task that is being specifically investigated by various organizations. To retrieve information in other languages, a query is translated by the machine translation system. Therefore, not only a retrieval model but also language resources and language processing functions are important factors of the CLIR system. Information retrieval models include the probabilistic model proposed by Robertson and Sparck Jones [2] and the vector space model proposed by Salton and McGill [3], both of which are used in CLIR.

Language resources include a dictionary, thesaurus, and bilingual corpus. Language processing functions include, for example, morphological analysis and machine translations.

To increase the performance of CLIR, a highly accurate query translation system is constructed by using existing multiple translation systems and improving the retrieval model. Limitations in the dictionary vocabulary and the presence of word ambiguity pose problems for query translation. Although numerous research studies have used the sentence translation system and the bilingual dictionary for query translation, the problems of such query translation remain unsolved.

We propose query translation using multiple machine translation systems on the Language Grid. Two machine translation systems are utilized in this method for translating a query, and a bilingual dictionary is used for translating the compound words and noun phrases of a query. To overcome the problem of mistranslated words appearing in the query, a filtering method using the concept base is proposed. In particular, to delete a mistranslated word, the similarity between the back translation of the word and the word in the source query is calculated. A word having a low similarity is considered to be the mistranslated word. In this paper, we describe "Using Language Grid for CLIR" method, "Deleting mistranslated queries using improved concept base" method and "Re-ranking retrieve results using concept base" method in detail, discuss its implementation, and present its experimental evaluation.

2. Related work

Cross-language information retrieval is divided into the query translation part and the information retrieval part. In the query translation part, many research studies have focused on a method using a sentence translation system and a bilingual dictionary [4] [5].

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

Atsushi Fujii and Tetsuya Ishikawa proposed a method integrating query and document translation using Machine Translation (MT) [6]. Aitao Chen et al. proposed a method combining multiple sources for short query translation in Chinese-English using two transfer dictionaries [7]. Wang et al. proposed a method of dictionary expansion using Wikipedia [8]. However, all of these methods had limitations. If only the bilingual dictionary was used for query translation. As a result, the ambiguity problem remained unsolved. In addition, if a query was expanded after translation, a mistranslated word could also be expanded. In addition, MT translates a query based on the context of the query. In MT, an input sentence is translated into an output sentence. If MT mistranslates the query, the mistranslated word appears in the query. The query then produces an inaccurate retrieval result.

We propose query translation using multiple machine translation systems and a bilingual dictionary on the Language Grid. In addition, to delete the mistranslated word in the query, the concept base is used. All language resources used in the method can be utilized in the Internet for CLIR system.

3. Background

To perfect the performance of CLIR, the accuracy of the query translation and the information retrieval must be improved.

The Language Grid is a new multilingual infrastructure on the Internet available for intercultural collaboration. This system can be used by freely combining the language resource and the language processing function in the Internet. By combining the multiple language resources on the Language Grid, it is possible to produce a translation result having high accuracy.

In the proposed method, the multiple machine translation systems and a bilingual dictionary open to the public in the Language Grid are combined, and highly accurate cross-language information retrieval is achieved. However, a mistranslated word in a query increases by using multiple language resources. Accordingly, the filtering method using a concept base is proposed for deleting the mistranslated word.

3.1 Language Grid

To satisfy the needs of users, the Language Grid allows users to easily develop new language services by combining existing ones.

The development of Semantic Web technologies enables the collaboration needed among language resources and language processing functions. The language resources include bilingual dictionaries, thesauruses and corpora, and the language processing functions include machine translation, morphological analysis and paraphrases. The Simple Object Access Protocol (SOAP) has been used for accessing the language resources of the Language Grid. SOAP specifies the exchange of structured information in the implementation of web services in computer networks. Web Services Description Language (WSDL) is a specification that describes networked XML-based services. WSDL provides a simple way for service providers to describe the basic format of requests to their systems regardless of the underlying SOAP protocol.

The Language Grid service layer includes the peer-to-peer (P2P) grid infrastructure, the language resource, the language service, and the intercultural collaboration tools. In our research, the retrieval service for CLIR is constructed in the layer of the P2P grid infrastructure and the language resource. Although multiple languages can be translated by the Language Grid, our data set is Japanese and English, so only a Japanese-English resource is needed.

3.2 Concept base

To delete the mistranslated word in a query, the concept base is used. In addition, concept base re-ranking is conducted with a noun phrase and a compound word extracted from the query.

The concept base was proposed by Schüetze and Pederson as a method of automatically constructing a thesaurus with the corpus and using a higher dimension vector space to express the relations between words appearing in a document [11]. Currently, the commonly utilized composition of the concept base is a word × word matrix.

First, in the traditional construction of the concept base, N words occurring with high frequency in the document for retrieval are selected to create a neighborhood co-occurrence matrix of a word with another word in the neighborhood (Figure1). Wij is the co-occurrence frequency between word i and word j. Before constructing the neighborhood co-occurrence matrix, it is necessary to perform a morphological analysis and remove the stopwords as a preprocessing step. Stopwords are words such as particles or auxiliary verbs that does not have an important role in the documents.

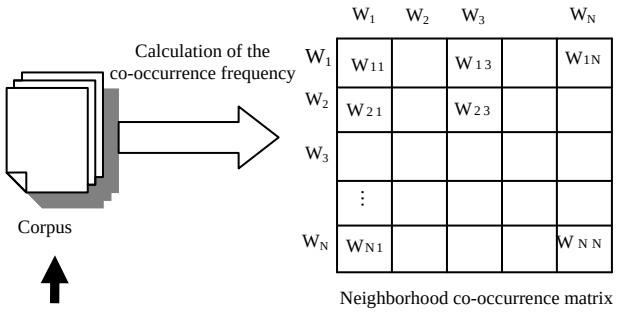

Fig. 1 Making the neighborhood co-occurrence matrix.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

The created neighborhood co-occurrence matrix can be considered as a word vector in which the number of words corresponds to the number of dimensions. However, there is the problem that the number of dimensions increases as the scale of the corpus grows because dimension depends on the number of words. Moreover, because each axis is a word, it is not easy to think of the axes as being mutually orthogonal. Therefore, to create the neighborhood co-occurrence matrix, singular value decomposition (SVD) is implemented. Under SVD, the neighborhood co-occurrence matrix is divided into three matrices: the transposed orthogonal matrix, the diagonal matrix, and the row orthogonal matrix. The row of 100~200 dimensions is extracted from the obtained row orthogonal matrix. The extracted matrix is the concept base (Figure 2).

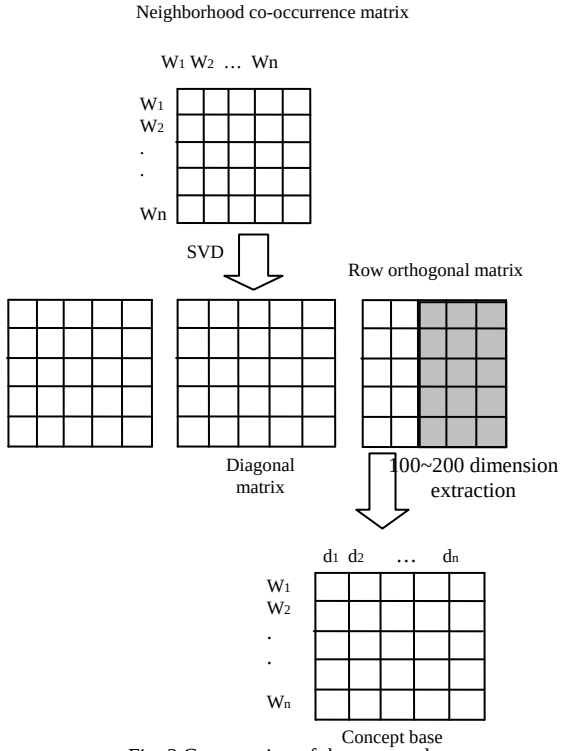

Fig. 2 Construction of the concept base.

4. CLIR System using the Language Grid and the Concept Base

4.1 System overview

Figure 3 depicts the overall design of our CLIR system. First, "Tokenizer E" processes the English language documents ("Doc in E") and "Tokenizer J" processes the Japanese language documents ("Doc in J"). If the documents are in English, the TreeTagger morphological analyzer and discard stopwords are utilized. In contrast, Japanese documents are segmented into lexical units using the ChaSen morphological analyzer and discard stopwords. Thereafter, the concept base is made from outputs of the tokenizer.

In the "Translator," a source language query ("Query in E") outputs the translation ("Query in J") by using the language resources of the Language Grid. We used two "Machine translations" to translate a query sentence and a "Dictionary" to translate the compound words and noun phrases of the query. Then, in the "Mistranslated word deletor," the "Concept base" is utilized for deleting the mistranslated words.

Finally, the "IR engine" outputs the top 1000 documents in descending order according to the similarity between the translated query and each document. The "Result combiner" combines the output of the two different IR engines. The first engine ("TF-IDF model") is a naive implementation of the vector space model with Term Frequency-Inverse Document Frequency (TF-IDF) term weighting, from which "Result re-rank" uses the concept base including the noun phrases for re-ranking the outputs of the top 100 documents. The second engine ("Probabilistic model") is a probabilistic model of the Lemur retrieval system.

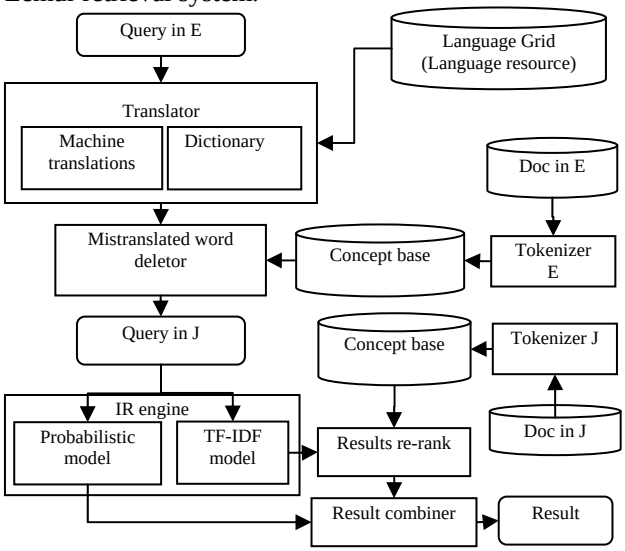

Fig. 3 The overall design of our CLIR system.

4.2 Query translation and expansion using language resources

In query translation, three unresolved matters remain: mistranslated words, limitations in the dictionary vocabulary and the ambiguity problem. This paper proposes query expansion using multiple machine translation systems on the Language Grid with a bilingual dictionary to compensate for the limited vocabulary and ambiguity problems. Specifically, for translation of a given query in the source language, we use the two

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

machine translation systems in the Language Grid, one of which reduces the ambiguity problem. Moreover, the query is expanded to compensate for the vocabulary limitations of the dictionary. In a query, compound words and noun phrases play important roles in deciding the retrieval result. Therefore, compound words and noun phrases are used for query expansion, and they are translated by the bilingual dictionary. We extract compound words and noun phrases from the query in the following example:

"マルチキャスト通信における関連する複数データの品質制御手法について論じたものはないか。"

For this query, we extract only "マルチキャスト通信", "複数データ", and "品質制御手法".

To overcome the problem of mistranslated words appearing in the query, a filtering method using the concept base is proposed.

4.3 Deletion of mistranslated query words

By using the language resources for query expansion, mistranslated words can appear within the query. Therefore, after completing the back translation, mistranslated words are found through the similarity of words. It is thought that if the mistranslated words are deleted, a better query can be obtained. Accordingly, a filtering method using a concept base is proposed.

In particular, to delete a mistranslated word, the similarity in the back translation word and the word in the source query is calculated (Figure 4). The source query vector Wi and the back translation word vector Wj are obtained from the concept base. The scalar product in the two word vectors is a degree of similarity of the words.

A word having a low degree (0.8) of similarity is considered to be the mistranslated word, which is then deleted. To delete the mistranslated word, the concept base is improved, as follows.

The concept base is made by using a corpus. The corpus is segmented by the morphological analyzer and stopwords, nouns, noun phrases, and adjectives are extracted.

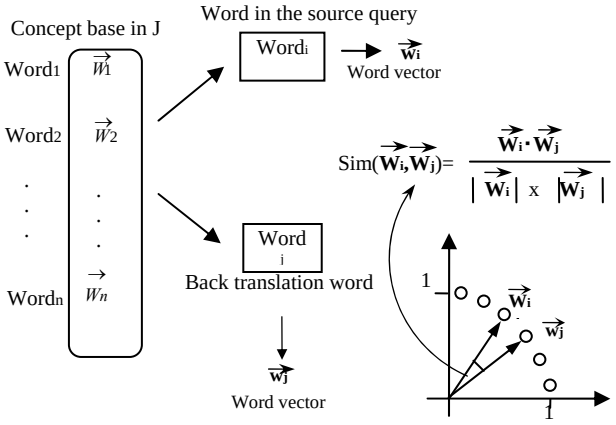

Fig. 4 Calculating method of similarity between a query and a back translated query

Since the concept base is composed of the co-occurrence frequencies between words, it does not consider the specificity element of a document containing a word. Consequently, the retrieval performance is decreased. We propose to construct a concept base in which this specificity is considered. The IDF is an index showing the specificity of a word. The concept base considers the frequency of a pair of words related to the co-occurrence as a single element. The IDF of a word pair is evaluated and used as a weighting term. For example, since the co-occurrence frequency of (computer, network) is greater than one, the IDF of the word pair is calculated as follows (eq. (1)):

$$idf\{pair(t_1,t_2)\} = \log \frac{N}{df\{pair(t_1,t_2)\}} \quad (1)$$

Here, ***pair*** (t_1,t_2) is a pair of the words t_1 and t_2 that exist in the co-occurrence relation, **N** is the total number of documents, and $df\{pair(t_1,t_2)\}$ represents the number of documents in which the word pair appears.

The $idf\{pair(t_1,t_2)\}$ value becomes the origin of the concept base. The weight is calculated by multiplying this value by each element of the neighborhood co-occurrence matrix. Therefore, element ***W*** of the neighborhood co-occurrence matrix is determined as follows (eq. (2)).

$$W_{t_1t_2} = F_{t_1t_2} \times idf\{pair(t_1,t_2)\} \quad (2)$$

Here, F_{t1t2} is the co-occurrence frequency of words t_1 and t_2. The element of the neighborhood co-occurrence matrix in Fig. 1 replaces **Wi (i=1,…,n)** with the value of the above expression. The concept base, composed of the neighborhood co-occurrence matrix with the element of eq. (2), is the specific concept base.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

The evaluation of how many mistranslated words deleted by the filtering using the concept base was performed. The words matching with the requirement of the retrieval is subjectively evaluated by human. Human examines the word in the query one by one. The fact that whether the word is appropriate to information retrieval is judged. The evaluation results are shown in the Table 1. Before filtering, the average number of mistranslated word of all queries is 6. The average number of mistranslated word of all queries after filtering is 0.417. Numbers of mistranslated word were deleted. The mistranslated word rate deleted by the concept base is 93.05%.

Table 1: Mistranslated word rate is deleted by concept base

Average number of mistranslated word of all queries before filtering	6.0
Average number of mistranslated word of all queries after filtering	0.417
Mistranslated word rate is deleted by concept base	**93.05%**

4.4 Re-ranking using a concept base

By using multiple machine translation systems, a query can contain many words that are the keywords of relevant documents. However, multiple systems also increase the number of words that are keywords of non-relevant documents, and, consequently, the retrieval performance decreases. We carry out re-ranking using a concept base with the noun phrases and compound words extracted from the query. Generally, when information retrieval is performed by the concept base, calculation of the similarity of the query vector and the document vector is carried out. However, if the number of words in the document is much larger than the number of words in the query, the similarity influence and the retrieval performance decrease. In our method, the document vector ($\mathbf{D}$) is divided into sentence vectors ($\mathbf{S}i$). The number of words in the query vector ($\mathbf{Q}$) and the number of words in the sentence vectors are almost the same. The similarity of the query vector and the sentence vector is then calculated. The highest similarity is assumed to be the similarity of the query vector and the document vector. The formula is shown below (eq. (3), (4)).

$$\vec{\mathbf{D}} = \vec{\mathbf{S}}_1 + \vec{\mathbf{S}}_2 + \ldots + \vec{\mathbf{S}}_n \quad (3)$$

$$\mathbf{Sim}(\vec{\mathbf{Q}}, \vec{\mathbf{D}}) = \max_{i=[1:n]} (\mathbf{Q}, \vec{\mathbf{S}}i) \quad (4)$$

where n is the number of sentences in the document.

In addition, the re-rank system is evaluated in comparison with the system which is not re-ranked. Table 2 shows the evaluation results. The non-interpolated average precision values of averaged over 39 queries are compared with 50, 100, 200, 400, 600, 800, 1000 of retrieved documents. All the comparison results of the re-ranking system are exceeded. Re-ranking using a concept base method was effective.

Table 2: Comparison of the results of the re-rank system and the result combined with no re-rank (Non-interpolated average precision values, averaged over the 39 queries)

Method		Result combined with no re-rank	Re-rank
Number of retrieved documents(N)	50	0.1740	0.1954
	100	0.1829	0.2058
	200	0.1876	0.2106
	400	0.1896	0.2128
	600	0.1905	0.2137
	800	0.1907	0.2140
	1000	0.1909	0.2142

5. Performance Evaluation

5.1 Baseline system

The proposed system is evaluated in comparison with the baseline system, which corresponds to the single machine translation system shown in Figure 5.

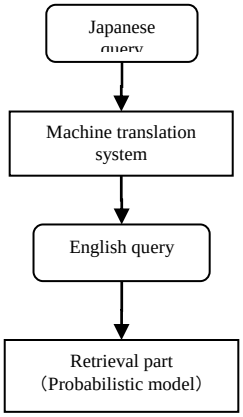

Fig. 5 The baseline system.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

5.2 Experimental data and environment

We used the test collection of Japanese-English science documents used in the NTCIR1 CLIR task (Japanese documents: 330,000, English documents: 190,000, Japanese queries: 39). To retrieve information, an English document set, from which only a verb and a noun were used for calculating the TF-IDF value, was used for the TreeTagger morphological analysis. The composition of the Japanese queries of NTCIR1 is as follows.

<TOPIC q=0038>
<TITLE>TCP/IP通信のスループット特性</TITLE>
<DESCRIPTION>ATM網を用いたTCP/IP通信のスループット特性について述べた論文はないか。
</DESCRIPTION>
<NARRATIVE>新しいネットワーク技術として登場したATM。これをバックボーンとして既存のTCP/IP通信を行なうことができる。ATM網を用いたTCP/IP通信のスループット特性についてのシミュレーションによる評価や解析を行なっている論文が欲しい。ATM網へのTCP/IP通信の接続だけではなく、そのスループット特性についての考察がなければ要求を満たさない。最新の研究動向を知りたい。
</NARRATIVE>
<CONCEPT>
<J.CONCEPT>a. ATMバックボーン, c. スループット, d. 品質評価, e. セル損失</J.CONCEPT>
<E.CONCEPT>a. ATM Backbone, c. Throughput, d. Quality Evaluation, e. Cell Loss</E.CONCEPT>
<A.CONCEPT>a. ATM, b. TCP/IP</A.CONCEPT>
</CONCEPT>
<FIELD>1. 電子・情報・制御</FIELD>
</TOPIC>

<TITLE>: Simple expression of the main concept of the retrieval request.
<DESCRIPTION>: Description of the retrieval request.
<NARRATIVE>: Explanation of the retrieval request, background, details, definition of terms, standard of correct answer judgment, and retrieval purposes.
<CONCEPT>: Presentation of concepts related to the retrieval request, synonym, hypernyms, and hyponyms.

For the experiment, only <DESCRIPTION> was used because it is the most standard retrieval request. As is already known, if the words in <CONCEPT> are added to the retrieval request, the precision increases. However, since the retrieval request most used in information retrieval systems is <DESCRIPTION> in NTCIR1, only <DESCRIPTION> was used in this experiment.
Table 3 shows experimental environment.

Table 3: Experimental environment

Computer	Pentium4 2.0GHz, Memory 2GB, HDD 40GB, Fedora Core 4.
Language resource	Cross Language WEB-Transer(Machine translation 1), Copyright Cross Language Inc.
	KODENSHA J-Server (Machine translation 2), Copyright Kodensha Co., Ltd. With provider Language Infrastructure Group, National Institute of Information and Communications Technology
	Online Dictionary of Academic Terms, Copyright National Center for Informatics, Aizawa Laboratory, Digital Content and Media Sciences Research Division, National Institute of Informatics
Concept base	Japanese concept base with 200000 words and 98 dimension
	English concept base with 100000 words and 151 dimension

5.3 Evaluation

The interpolated recall and precision, the average precision (non-interpolated) for all relevant documents and the precision for 50, 100, 200, 400, 600, 1000 documents were calculated using the TREC evaluation program, which uses the following formula to evaluate the average precision (eq. (5)):

$$\textit{Average Precision} = \frac{1}{D} \sum_{1 \le k \le N} r_k \; x \; \textit{Precision(k)} \quad (5)$$

$$r_k \begin{cases} 1 \text{ (Document of order k represents the correct answer)} \\ 0 \text{ (Document of order k represents incorrect answers)} \end{cases}$$

D: number of relevant documents in the retrieval result.
N: the relevant document appearing at the end order.
Precision (k): precision at the time of order *k*.

5.4 Evaluation results and comparison

For the proposed system, the average precision was evaluated and compared with that of the baseline system and the NTCIR1 participation system. Table 4 shows the evaluation results of the proposed system and the baseline system. The proposed system gives significantly higher precision than does the baseline system. Table 5 shows the evaluation results of the proposed system and the NTCIR1 participation system. The average precision of the proposed system is 0.2142. The best average precision of the NTCIR1 participation system is 0.2109. Again, the proposed system reaches the highest ranking.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

Table 6 shows the evaluation results of the proposed system, the query and document translation system of Atsushi Fujii and Tetsuya Ishikawa proposed [6]. They used the same test collection of Japanese-English. Under their system, query and document is translated by MT and human. The average precision values of the proposed system are better than those of the query and document translation system. As for the number of retrieved documents over 200, the average precision values of the query and document translation system that was ideally translated by human are higher than those of the proposed system. As for the number of retrieved documents from 50 to 100, however, the average precision values of the query and document translation system are lower than those of the proposed system. The number of retrieved documents from 50 to 100 is enough for users.

Table 4: Comparison of the results of the proposed system and the Baseline system

System ID	r-prec	Ave prec
Baseline system	0.1753	0.1654
Proposed system	**0.2396**	**0.2142**

Table 5: Comparison of the results of the proposed system and the NTCIR1 participation system (Non-interpolated average precision values, averaged over the 39 queries)

	System ID	r-prec	ave prec
Top 8 Run	BKEBDDS	0.2225	0.2109
	TSB4	0.2453	0.2090
	TSB3	0.2296	0.2084
	1KE3	0.2223	0.2062
	1KE	0.1950	0.1940
	1KE_ij	0.1976	0.1713
	TSB1	0.1816	0.1617
	TSB2	0.1872	0.1524
	Proposed system	**0.2396**	**0.2142**
Monolingual	BKEBDDS	0.2826	0.2618

Table 6: Comparison of the results of the proposed system and the query and document translation methods (Non-interpolated average precision values, averaged over the 39 queries)

Method		Query and Document translation methods (MT)	Query and Document translation methods (ideal translation by human)	**Proposed system**
Number of retrieved documents (N)	50	0.1690	0.1814	0.1954
	100	0.1766	0.1968	0.2058
	200	0.1901	0.2142	0.2106
	400	0.1946	0.2242	0.2128
	600	0.1958	0.2301	0.2137
	800	0.1967	0.2319	0.2140
	1000	0.1986	0.2356	0.2142

6. Discussion

In a query for information retrieval that includes many keywords of relevant documents, a good retrieval result is obtained. If multiple language resources are utilized, the number of keywords of relevant documents is increased in the query. In the proposed method, a query is translated by two machine translation systems. In the query, compound words and noun phrases play important roles in deciding the retrieval result. Therefore, compound words and noun phrases are used for query expansion and for translation by the bilingual dictionary. Other language resources of the two machine translation systems and the bilingual dictionary are not used because the utilization of these other resources does not help to increase the number of keywords of relevant documents, but instead can increase the number of mistranslated words. In this study, the concept base is used to delete mistranslated words from the query. Accordingly, high-precision retrieval results are obtained. That is, the query can be translated with high accuracy. The retrieval performance is improved by re-ranking with the concept base and combining the retrieval results. In this case, the utilization of language resources can enable a query to contain many words that are keywords of relevant documents, and thus, the retrieval results include various relevant documents. However, the language resources also increase the number of words that are keywords of non-relevant documents. Consequently, the rank of relevant documents is decreased and so is the retrieval performance. To increase the retrieval performance, we carry out re-ranking using a concept base with the noun phrases and compound words extracted from the query.

7. Conclusions

We set up and implemented a system for improving the performance of cross-language information retrieval by

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

combining the language resources of the Language Grid. By using multiple machine translation systems, the ambiguity problem was reduced. In addition, by using the bilingual dictionary for translating compound words and noun phrases, the word ambiguity problem was further reduced and the queries were expanded. By implementing a filtering method using the concept base, the mistranslated words in queries were reduced. For information retrieval we used the vector space model with TF-IDF term weighting. The outputs of the top 100 documents were re-ranked by a specifically considered concept base. For the second engine, we used the probabilistic model of the Lemur retrieval system. The final result came from the combined outputs of the two IR engines. Under the proposed method, the cross-language information retrieval system was implemented at a high rate of precision for a NTCIR1 dataset. In comparison with NTCIR1 participation systems, the proposed system attained the highest ranking.

Acknowledgments

We received permission to use the test collection of Japanese-English science documents from the NTCIR1 task of the National Institute of Informatics (NII). In addition, we used the language resources of the Language Grid Project during the research process. We would like to take this opportunity to express our sincere thanks to NII and the Language Grid Project.

References

[1] T. Ishida. “Language Grid: An Infrastructure for Intercultural Collaboration” IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), pp. 96-100, 2006.
[2] K. Sparck-Jones. “A Statistical interpretation of term specificity and its application in retrieval” Journal of Documentation. vol.28, no.1, pp. 11-21, 1972.
[3] G.Salton. “Introduction to Modern Information Retrieval” McGraw-Hill. 1983.
[4] C. Lin, W. Lin, G Bian, H. Chen “Description of the NTU Japanese-English Cross-Lingual Information Retrieval System for NTCIR Workshop”, NTCIR Workshop 1, 1999.
[5] DW. Oard, J. Wang. “NTCIR CLIR Experiments at the University of Maryland” NTCIR Workshop 1, 1999.
[6]Atsushi Fujii and Tetsuya Ishikawa. “Japanese-English Cross-Language Information Retrieval Integrating Query and Document Translation Methods” IEICE, J84-D-II(2), pp.362-369, 2001.
[7]Aitao Chen et al. “Combining multiple sources for short query translation in Chinese-English cross-language information retrieval.
Proceedings of the fifth international workshop on Information retrieval with Asian languages”, 2000.
[8] Y. Wang, C. Lee, R.Tsai, W.Hsu. “IASL System for NTCIR-6 Korean-Chinese Cross-language information retrieval” NTCIR Workshop 6, pp. 26-30, 2007.
[9] R. Huamg, L. Sun, J. Li, L. Pan, J. Zhang. “ISCAS in CLIR at NTCIR-6: Experiments with MT and PRF” NTCIR Workshop 6, pp. 26-30, 2007.
[10] Atsushi Fujii and Tetsuya Ishikawa. “Japanese/English Cross-language Information Retrieval: Exploration of Query Translation and Transliteration” Computers and the Humanities Vol.35, No.4, pp. 389–420, Nov 2001.
[11] H.Schüetze, J.Pederson. “Information retrieval Based on Word Senses, In Fourth Annual Symposiumon Document Analysis and Information Retrieval”, pp.161-175, 1994.
[12] A.Yasumune, H.Taiichi, T.Takenobu, T.Hozumi. “Research on cross- language information retrieval using vocabulary extension” Natural Language Processing 10, D3-1, 2004..
[13] T.Tokunaga. “Information retrieval and Language processing”, Junichi Tsujii [edit], Foundation the University of Tokyo publication association, Tokyo, 1999.
[14] P. Huy Anh, T. Yukawa. “Cross Language Information Retrieval Based on Concept Base and Language Grid”. ESAIR’10, October 30, 2010, Toronto, Ontario, Canada. ACM 978-1-4503-0372-9/10/2010.

Pham Huy Anh received a B.S. in Electrical Engineering from the National Defense Academy and an M.S. in Electrical Engineering from the Nagaoka University of Technology. He is currently a student of doctor course in the Department of Electrical Engineering at the Nagaoka University of Technology.

Takashi Yukawa Membership Number of IEEE : [870-2110]
Takashi Yukawa received a Master of Engineering degree from the Nagaoka University of Technology in 1987 and a Doctor of Informatics degree from Kyoto University in 2001. He has been involved in the research and development of a parallel computer for expert systems, a concept-sensitive information retrieval system and its application systems, knowledge management systems and an intelligent course management system for e-Learning. He is currently an associate professor in the Department of Electrical Engineering at the Nagaoka University of Technology.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.