2012 IEEE International Conference on Information Science and Technology
Wuhan, Hubei, China; March 23-25, 2012

Query Translation for CLIR: EWC vs. Google Translate

V. Klyuev and Y. Haralambous

***Abstract*—A new approach to find accurate translation of search engine queries from Japanese into English for the CLIR task is proposed. The Mecab system and online dictionary SPACEALC are utilized to segment Japanese queries and to get all possible English senses for every term detected. To disambiguate terms, the idea of the shortest path on an oriented graph is applied. Nodes of this graph symbolize word senses and edges connect nodes representing neighboring Japanese terms. The EWC semantic relatedness measure is used to select the most related meanings for the translation results. This measure combines the Wikipedia-based Explicit Semantic Analysis measure, the WordNet path measure and the mixed collocation index. The proposed technique is tested on the NTCIR data collection. Queries generated by Google Translate were used to evaluate the quality of translation.**

I. INTRODUCTION

CROSS-LANGUAGE Information Retrieval (CLIR) can be used to retrieve documents in one language in response to a query given in another. The usual approach consists of two steps: 1) translation of the user query into the target language and then 2) retrieval of documents in this language by using a conventional mono-lingual information retrieval system. There is abundant literature on the CLIR task, using several approaches and different pairs of languages.

Recent findings in information retrieval such as explicit semantic analysis (ESA) introduced in study [2] and ESA combined with WordNet and collocations (EWC) proposed in article [3] allow us to look at the problem from a different angle.

In this paper, we continue to discuss an approach to translate queries for a Japanese-English CLIR task [20]. Our key assumption is as follows: Terms in any query should be semantically related to each other. After segmenting the queries applying different tools, we obtain full sets of translations for each Japanese term. The EWC measure helps us to select appropriate values for each term.

The rest of the paper is organized as follows. Section 2 reviews the approaches to the CLIR task. Section 3 describes shortly EWC. Section 4 discusses the proposed technique to translate queries from Japanese into English and to select word meanings. Section 5 describes experiments conducted. Section 6 discusses the results of the tests. Section 7 presents the concluding remarks.

Manuscript received December 3, 2011.

V. Klyuev is with the University of Aizu, Aizu-Wakamatsu Fukushima –ken 965-8580, Japan (phone: +81-242-37-2603; fax: +81-242 -37-2733; e-mail: vkluev@u-aizu.ac.jp).

Y. Haralambous is with Institut Télécom –Télécom Bretagne, Dép. Informatique, UMR CNRS 3192 Lab-STICC Technopôle Brest Iroise, CS 83818, 29238 Brest Cedex 3, France (e-mail: yannis. Haralambous @ teleco m -bretagne.eu).

II. RELATED WORK

Cross-language information retrieval has a long history. There is a lot of research done in this area. In this section, we review shortly the several studies to show the key current tendencies.

Books [7, 17] and study [6] introduce the key approaches and techniques used in CLIR. The most popular approaches include:

- Machine translation,
- Query translation using bilingual dictionaries,
- Transaction models derived from parallel texts, and
- Similarity thesaurus-based translation.

The main goal of these techniques is to convert the CLIR task into the Ad-hoc retrieval task and then use the methods for the monolingual task to do retrieval.

Bilingual dictionaries do not reflect the dynamic nature of the languages: They do not include the latest words, and phrases appeared in them. This is a main disadvantage of their usage. An approach to use parallel texts from huge volumes in order to obtain a statistical bilingual dictionary became popular. The authors of study [11] applied this approach to the Spanish-English cross-language information retrieval task and achieved improvements in the retrieval results.

Another way to get full translation of the queries is to use advanced translation systems. Google translation was proposed to use in the CLIR task [19]. The authors applied it to the Chinese-English task.

Wikipedia is gaining popularity as a source of knowledge to get translation of terms. Authors of article [8] used it to translate the initial queries into the language of the documents. They tested this approach for queries in Dutch, French, and Spanish and an English data collection. The method proposed in [13] utilized the idea of mapping ESA vectors of the queries with respect to the Wikipedia query space into vectors with respect to the Wikipedia article space. They used the cross-language links of Wikipedia to map the ESA vectors between different languages. This technique was applied to the German – English and French – English language pairs. The authors reported that they did not gain the advantages in the performance compared to the other approaches utilized at CLIR 2008.

Approaches to translate queries do not preserve the semantics of the original queries. This results in the relatively low retrieval performance of the systems utilizing them.

In study [20], an idea of the shortest path on the oriented

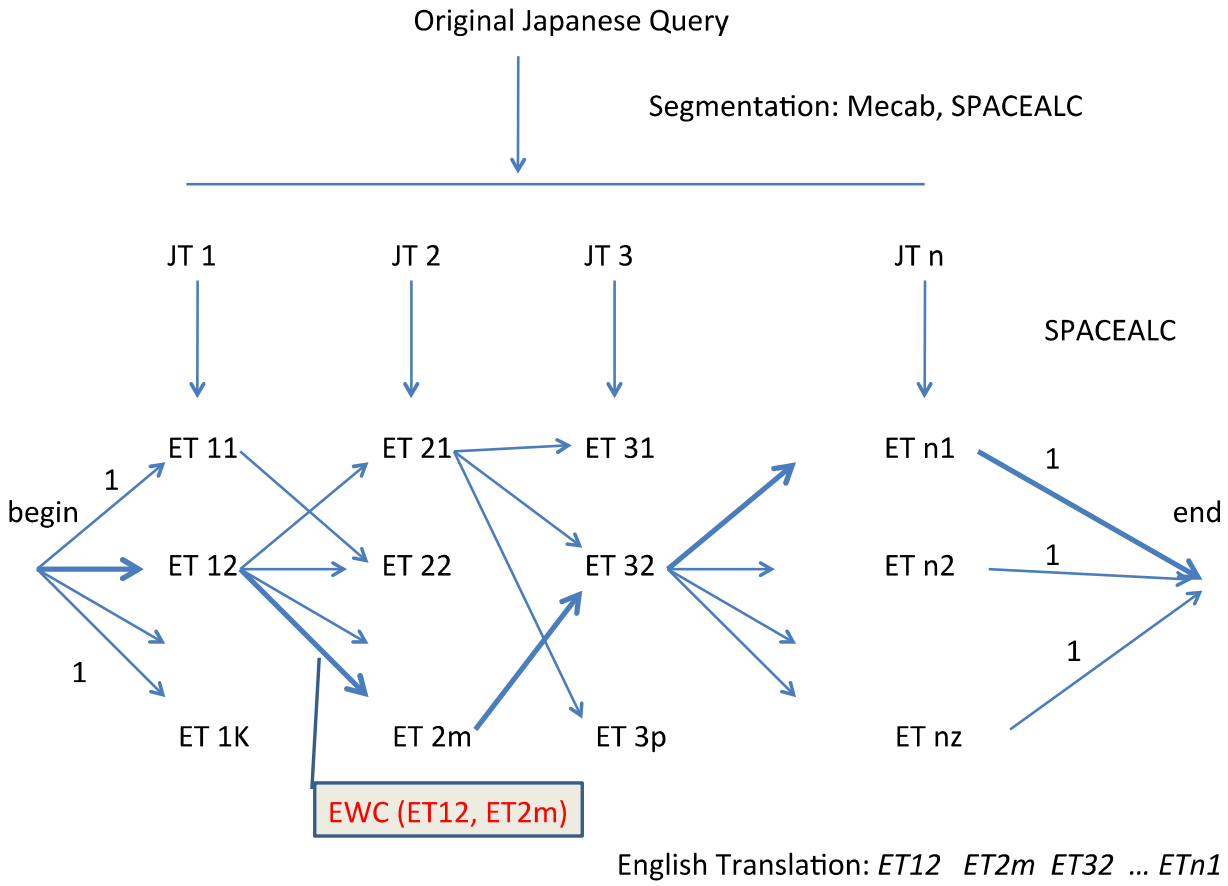

Fig. 1. Processing a Japanese query

graph to disambiguate terms was introduced. The nodes of the graph are different meanings of the terms. The authors reported pros and cons of this view. We found this view promising and analyse it in more details.

Next section reviews the technique to find English variants for Japanese queries.

III. EWC Measure Description

In study [3], the new measure of words relatedness is introduced. It combines the ESA measure μ_{ESA} [2], the ontological WordNet path measure μ_{WNP} [21], and the collocation index C_ξ. This measure is called EWC (ESA plus WordNet, plus collocations) and is defined as follows:

$$\mu_{EWC}(w_1, w_2) = \mu_{ESA}(w_1, w_2) * \alpha$$
$$\alpha = (1 + \lambda_\sigma(\mu_{WNP}(w_1, w_2))) * \gamma$$
$$\gamma = (1 + \lambda'_\sigma(C_\xi(w_1, w_2)))$$

where λ_σ weights the WordNet path measure (WNP) with respect to ESA, and λ'_σ weights the mixed collocation index with respect to ESA. This index is defined as follows:

$$C_\xi = \frac{2 * f(w_1, w_2)}{f(w_1) + f(w_2)} + \xi \frac{2 * f(w_2, w_1)}{f(w_1) + f(w_2)}$$

where $f(w_1, w_2)$, $f(w_2, w_1)$ are the frequency of the collocations of $w_1 w_2$ and $w_2 w_1$ in the corpus, and $f(w_i)$ is the frequency of word w_i. The values for constants λ_σ, λ'_σ, and ξ are set to 5.16, 48.7, and 0.55, respectively.

Study [3] demonstrated the superiority of this measure over ESA on the WS-353 test set. Results of tests on query expansion discussed in study [4] showed superiority of EWC over ESA and DFR (divergence from randomness) term weighting model.

The current implementation of EWC does not take into account Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). As a result, the proposed technique cannot distinct multiple term

items from collocations and give them the highest score.

IV. Query Translation

The first step of the technique to translate a Japanese query is segmentation. Individual Japanese terms are the outcome. Mecab [5] and SPACEALC [16] are the tools.

The next step is to translate each term and collect all meanings. SPACEALC is the main tool.

The step, following that, is disambiguation of the meanings of the terms. An unsupervised Word Sense Disambiguation (WSD) system discussed in study [10] is based on the hypothesis that the intended sense of an ambiguous word is related to the words in its context. We use the same hypothesis in the proposed approach. To achieve the goal, we create an oriented graph, which nodes are word senses. Edges connect nodes representing neighboring Japanese terms. The shortest path on this graph gives us the results of query translation. We take into account the all-possible permutations of the terms because the term order in different languages, expressing the same meaning, may be different.

The aforementioned tools Mecab and SPACEALC are key components of our prototype. The other components are an experimental online service to calculate the EWC similarity between translated English terms and software based on the Dijkstra algorithm to select the English variants of translations for each term of the original Japanese query.

Figure 1 illustrates our scheme for processing Japanese queries and obtaining English translations. This scheme is shown for one permutation.

V. Experiments

We chose the open source search engine Terrier [14] as a tool to index and retrieve data. It provides various retrieval approaches. TF-IDF and Okapi's B25 are among them [12]. The NTCIR CLIR data collection [9] consisting of 187,000 articles in English was used as a data set for experiments. These articles are summaries of papers presented at scientific conferences hosted by Japanese academic societies. The collection covers a variety of topics, such as chemistry, electrical engineering, computer science, linguistics, library science, and etc. The size of the collection is about 275.5 MB. 83 topics are in Japanese. We used topics 0001 to 0030. A structure of the dataset and topics is similar to that of TREC [15]. A Porter's stemmer was applied to documents and queries. A standard stop word list provided by Terrier was also utilized. We took into account only the title fields as a source of queries. They are relatively short: Each query consists of a few keywords.

Table 1 presents an original and segmented Japanese query (it consists of three terms; segmentation is done by Mecab), obtained variants of translation from SPACEALC, a generated English query and a query produced by Google Translate, and a query generated as the most appropriate permutation consisting of a combination of terms. The variant with the highest score is selected. A set of analyzed variants is as follows: データ　品質　制御; 品質 データ 制御; データ　制御　品質, etc. The first line in the SPACEALC row corresponds to the first Japanese term; the second line (the term *quality*) is translation of the second Japanese term; and the third one includes senses of the last Japanese term.

In the experiments, we submitted to Terrier 1) queries generated according to the aforementioned technique applying Mecab to segment Japanese texts; 2) queries obtained from Google Translate service; 3) queries generated according to the aforementioned technique applying the longest match strategy to segment Japanese texts, 4) queries generated applying Mecab to segment Japanese texts and selecting collocations from the results of translation by SPACEALC, and 5) queries with the highest semantic score.

TABLE I
Example of Translation

Procedure	Query
Original query	データ品質制御
Mecab	データ　品質　制御
SPACEALC	data figures information input
(one permutation)	quality
	grip regulation
English query	information quality regulation
Google Translate	data quality control

The longest match technique matches the initial string of characters against the dictionary entries and takes the initial string that matches the longest entry in the dictionary as a word. This technique was introduced in [1]. The longest match strategy was implemented utilizing SPACEALC: The original query was initially submitted to SPACEALC. If SPACEALC failed to translate, we cut the last character from the query and tried to translate it again. In the case of success, we retrieved the all senses for the detected term and repeated this process for the remaining part of the query.

VI. Discussions

Retrieval performance with queries generated utilizing Mecab was very low. See Table 2. The reason for this is as follows: The dictionary of Mecab does not include a big enough number of technical terms. As a result, there is no way to reconstruct the terms (to segment queries) correctly. The accurate segmentation gives the highest possible precision. The technique to detect a variant with the highest possible semantic score does not help much because of the same reason.

Our efforts to reconstruct technical terms and collocations from the information provided on the first page of SPACEALC did not help much: For value of a segmented Japanese term, the first collocation was selected, or the first meaning as a single term. See Table 2, column 6.

The longest match strategy gave the following results: SPACEALC translated successfully full queries without segmentation for topics: 4, 6, 12, 14, 15, 17, 19, 22, 24, 26, and 28. There were several variants of translations for topics 14 and 22. Results of translation applying Google Translate

TABLE II

RESULTS OF RETRIEVAL

	Queries generated by Google Translate	Queries generated utilizing the proposed technique and longest much segmentation	Queries generated utilizing the proposed technique and segmentation by Mecab	Queries generated utilizing segmentation by Mecab and selection of the bast permutation	Queries generated utilizing segmentation by Mecab and selection of collocations as values for terms
Number of queries	21	21	21	21	21
Retrieved	19087	18112	18234	18000	19200
Relevant	1756	1756	1756	1756	1756
Relevant retrieved	882	792	407	454	486
Average Precision	0.3017	0.2644	0.0997	0.1514	0.1634
R Precision	0.3163	0.2658	0.0844	0.1640	0.1760

TABLE III

ORIGINAL AND THE BEST VARIANTS OF TRANSLATED QUERIES

Topic num-ber	Japanese queries	Translation by Google Translate	Translation utilizing the proposed technique and the longest much segmentation
1	ロボット	Robot	automaton bot golem iron man robot
2	複合名詞の構造解析	Structural analysis of compound nouns	compound noun localization of structures
3	サンプル複雑性	Sample complexity	pattern complexity
5	特徴次元リダクション	Feature dimension reduction	point plane reduction
7	認知的側面	Cognitive aspects	cognitive interface side
9	インターネットトラヒック統計	Internet traffic statistics	web traffic side
10	キーワード自動抽出	Keyword extraction	keyword automatic extracting
11	連結全域グラフ	Connected spanning graph	combination entire area graphic
13	ループ領域解析	Analysis of the loop region	loop region mapping
16	最大共通部分グラフ	Maximal common subgraph	maximum common substructure graph
18	通信品質保証	Communication quality assurance	connection quality assurance
20	カタカナ外来語	Katakana foreign words	katakana loanword
21	機械翻訳の評価	Evaluation of machine translation	computer interpreter marks
27	シソーラスを用いた検索	Thesaurus search using	thesaurus search
29	位置計測	Position measurement	point measuring
30	データ駆動画像処理	Data-driven image processing	data driving image enhancement

service and SPACEALC were same for topics: 4, 6, 15, 17, 24, 26, and 28. These topics are omitted in Table 3 and Table 4. Table 3 presents the original Japanese queries and variants of translations by Google Translation service and by the proposed technique. Japanese queries initially were segmented utilizing the longest match technique and then they were passed through the procedure of translation. From Table 3, one can see that the proposed approach generates queries similar to queries produced by Google without segmentation done in advance.

Table 4 presents the average precision of each query. The right answers of retrieval are provided by organizers of the NTCIR Workshop only for 21 topics out of 30. N/A marks the topics without right answers. For the queries consisting of only one term, the all senses were selected: SPACEALC does not provide information about the frequency of terms, and terms are arranged in alphabetical order (See Table 3, topic 1). According to the word frequency list [18], the rank of word robot is equal to 4564. This is the most frequent term compared to “automaton”, “bot”, “golem”, and “iron man”. It seems that Google Translate service uses this information.

One can see from Table 4 that on queries generated according to the proposed technique, the system performed better only on topic 20.

Our experiments showed the superiority of the longest much technique applying SPACEALC over Mecab: Segmentation of Japanese texts is much more accurate. On the other hand, the current implementation of EWC does not take into account Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). As a result, the proposed technique cannot distinct multiple term items from collocations and give them the highest score. To illustrate this point, we consider the results of segmentation of topic 13 (See Table 3). Applying SPACEACL, we receive two terms with the following

possible values: 1) *loop region; looped domain*; 2) *analysis; deconvolution;* mapping; and *observational study*. There is the entry for the term of *observational study* in Wikipedia. If we take into account this knowledge, then the result query consisting of *loop region observational study* and submitted

TABLE IV

AVERAGE PRECISION OF EACH QUERY: BEST VARIANT OF TRANSLATION

Topic number	Precision for the queries generated by Google Translate	Precision for the queries generated utilizing the proposed technique
1	0.1076	0.0346
2	N/A	N/A
3	N/A	N/A
5	0.1997	0.0001
7	N/A	N/A
9	N/A	N/A
10	0.5187	0.3846
11	N/A	N/A
13	0.0526	0.0293
16	0.3327	0.1949
18	0.0123	0.0090
20	0.8762	1.0000
21	N/A	N/A
27	N/A	N/A
29	0.1042	0.0021
30	N/A	N/A

to the search system gives the precision of the retrieval as of 0.3515. This result is much more precise compared to the value of 0.0526 obtained for the query generated by Google Translate service.

As we mentioned earlier, articles of the collection are summaries of papers presented at scientific conferences hosted by Japanese academic societies. Their authors are non-native speakers. This issue may influence the results of retrieval: Topic 10, for example. Google translate generated the query *keyword extraction.* The proposed technique gave a more accurate variant: *keyword automatic extracting*. Retrieval results for this topic are better for Google: 0.5187 vs. 0.3846.

The EWC implementation should be enhanced in order to take the aforementioned issues into account. We strongly believe that this feature may improve the accuracy of translation significantly.

VII. CONCLUSION

The enhancement of the approach to translate short queries from Japanese into English is discussed. It utilizes the EWC measure to calculate the similarity between translated terms, the shortest path idea, and the idea of the best permutation to select terms from the list of candidates. It demonstrated the best performance when the longest match strategy is used. The NTCIR CLIR data collection was used to test the proposed approach. Results of preliminary experiments showed that queries generated are similar to queries obtained from Google Translate service. The performance of the retrieval system for queries generated by the proposed approach is slightly worse compared to the performance for the queries obtained from Google Translate service.

REFERENCES

[1] A. Chen, A., F.C. Gey, K. Kishida, H. Jiang, and Q. Liang, "Comparing Multiple Methods for Japanese and Japanese-English Text Retrieval", *In Proc. The First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition,* 1999.

[2] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-Based Information Retrieval using Explicit Semantic Analysis," *ACM Transactions on Information Systems,* vol. 29, no. 2, 2011.

[3] Y. Haralambous and V. Klyuev, "A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge," in *IJCNLP2011*, Thailand, 2011.

[4] V. Klyuev and Y. Haralambous, "Query Expansion: Term Selection using the EWC Semantic Relatedness Measure," in *FedCSIS 2011*, Poland.

[5] MeCab: Yet another Part-of-Speech and Morphological Analyzer. [Online]. Available: http://mecab.sourceforge.net/

[6] T. Mitamura, H. Shima, T. Sakai, N. Kando, T. Mori, K. Takeda, C. Lin, R. Song, C. Lin, and C. Lee, "Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access,". in *Proc. The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, Japan, 2010.

[7] J. Nie, *Cross-Language Information Retrieval*, Association for Computational Linguistics, 2011.

[8] D. Nguyen, A. Overwijk, C. Hauff, D. Trieschnigg, D. Hiemstra, and M.G. de Jong Franciska. "WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia", *CLEF 2008, LNCS 5706*, pp. 58–65, 2009.

[9] *NTCIR-1 CLIR data collection*. [Online]. Available: http://research.nii.ac.jp/ntcir/data/data-en.html.

[10] Patwardhan, Banerjee, and Pedersen, "UMND1: Unsupervised Word Sense Disambiguation Using Contextual Semantic Relatedness," in *Proc. SemEval-2007: 4th International Workshop on Semantic Evaluations,* Prague, Czech Republic, 2007, pp. 390-393.

[11] D. Pinto1, J. Civera, A. Juan, R. Rosso, and A. Barron-Cedeno, "A statistical approach to cross lingual natural language tasks". *Journal of Algorithms,* vol. 64 ,pp. 51 – 60, 2009.

[12] S. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne, "Okapi at TREC-4, " in *Proc. TREC 4,* 1995.

[13] P. Sorg and P. Cimiano, "Cross-lingual Information Retrieval with Explicit Semantic Analysis," in *CLEF* 2008.

[14] Terrier. [Online]. Available: http://terrier.net

[15] *TREC*. [Online]. Available: http://trec.nist.gov/

[16] *SPACEALC*. [Online]. Available: http://www.alc.co.jp/

[17] E. Voorness and D. Hartman, (eds.), *TREC: experiment and evaluation in information retrieval*. The MIT Press, 2005.

[18] Word frequency lists and dictionary. [Online]. Available: http://www.wordfrequency.info/

[19] H. Xiaoning, W. Peidong, Q. Haoliang, Y. Muyun, L. Guohua, and X. Yong, "Using Google Translation in Cross-Lingual Information Retrieval" in *Proc. NTCIR-7 Workshop Meeting*, Tokyo, Japan, 2008.

[20] V. Klyuev and Y. Haralambous, "Accurate Query Translation for Japanese-English Cross-Lingual Information Retrieval", in *Proc. of PECCS*, Italy, 2012.

[21] WORDNET. [Online]. Available: http://www.wordnet.org/