ACCURATE QUERY TRANSLATION FOR JAPANESE-ENGLISH CROSS-LANGUAGE INFORMATION RETRIEVAL

Vitaly Klyuev[1] and Yannis Haralambous[2]
[1]*Software Engineering Lab, University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580, Japan*
[2]*Institut Télécom – Télécom Bretagne, Dép. Informatique, UMR CNRS 3192 Lab-STICC, Technopôle Brest Iroise, CS 83818, 29238 Brest Cedex 3, France*

Keywords: EWC, CLIR, Automatic Query Translation, Search.

Abstract: In this paper, a novel approach to translate queries from Japanese into English for the CLIR task is discussed. To get all possible English senses for every Japanese term, the online dictionary SPACEALC is utilized. The EWC semantic relatedness measure is used to select the most related meanings for the results of translation. This measure combines the Wikipedia-based Explicit Semantic Analysis measure, the WordNet path measure and the mixed collocation index. The preliminary tests of the proposed technique are done utilizing the NTCIR data collection. The performance of retrieval is compared with the variant of retrieval using queries generated by Google Translate.

1 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 by Egozi et al. (2011) and ESA combined with WordNet and collocations (EWC) proposed by Haralambous and Klyuev (2011) allow us to look at the problem from a different angle. In this paper, we propose a novel approach to translate queries for a Japanese-English CLIR task. We assume that terms in any query should be semantically related to each other. This is a basis assumption of our approach. After segmenting the queries, we obtain full sets of translations for each Japanese term. The final step is to select the most semantically related alternatives applying EWC.

The rest of the paper is organized as follows. Section 2 presents a review of the approaches to the CLIR task. Section 3 describes the key ideas of EWC. Section 4 introduces the details of the proposed technique to translate queries from Japanese into English. Section 5 gives details of experiments conducted. Section 6 discusses the results of the tests. Section 7 presents the concluding remarks.

2 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.

Book by Nie (2011) and by Voorness and Hartman (2005) 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.

Klyuev V. and Haralambous Y..
ACCURATE QUERY TRANSLATION FOR JAPANESE-ENGLISH CROSS-LANGUAGE INFORMATION RETRIEVAL.
DOI: 10.5220/0003905902140219
In *Proceedings of the 2nd International Conference on Pervasive Embedded Computing and Communication Systems* (PECCS-2012), pages 214-219
ISBN: 978-989-8565-00-6

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. Pinto et al. (2009) 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 by Xiaoning et al. (2008). They applied it to the Chinese-English task.

Wikipedia is gaining popularity as a source of knowledge to get translation of terms. Nguyen et al (2008) 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. Sorg and Cimiano (2008) 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.

3 EWC MEASURURE DESCRIPTION

Haralambous and Klyuev (2011) introduce a new measure of words relatedness. It combines the ESA measure μESA, the ontological WordNet path measure μWNP, and the collocation index C_ξ. This measure is called EWC (ESA plus WordNet, plus collocations) and is defined as follows:

$$\mu_{EWC}(w_1,w_2)= \mu_{ESA}(w_1,w_2)\cdot(1+\lambda_\sigma(\mu_{WNP}(w_1,w_2)))\cdot(1+\lambda_{\sigma'}(C_\xi(w_1,w_2)))$$

where λ_σ weights the WordNet path measure (WNP) with respect to ESA, and $\lambda_{\sigma'}$ weights the mixed collocation index C_ξ with respect to ESA. This index is defined as follows:

$$C_\xi=2\cdot f(w_1w_2) / (f(w_1)+f(w_2))+\xi\cdot 2\cdot f(w_2w_1) / (f(w_1)+f(w_2))$$

where $f(w_1,w_2)$, $f(w_2,w_1)$ are the frequencies of the collocations of w_1w_2 and w_2w_1 in the corpus, and $f(w_i)$ is the frequency of word w_i. The values for constants λ_σ, $\lambda_{\sigma'}$, and ξ were set to 5.16, 48.7, and 0.55, respectively.

Haralambous and Klyuev (2011) demonstrated the superiority of this measure over ESA on the WS-353 test set. Results of tests on query expansion discussed in study Klyuev and Haralambous (2011) showed superiority of EWC over ESA and DFR (divergence from randomness) term weighting model. This measure was applied to evaluate semantic similarity of candidates in the English language for inclusion in the target query.

4 GENERATING ENGLISH QUERIES

The main assumption of the proposed approach is to segment the original Japanese queries, then translate each detected term collecting all senses, and finally select the sense of a term that is most related to all terms of the query. The final step is disambiguation. An unsupervised Word Sense Disambiguation (WSD) system discussed by Patwardhan et al. (2007) 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 neighbouring Japanese terms. The shortest path on this graph gives us the results of query translation.

To implement the approach, we apply a) (Mecab, 2011) to segment Japanese queries; b) the quite efficient online Japanese-English dictionary (SPACEALC, 2011) to obtain all English variants of the translation of every Japanese term, and collocations; 3) an experimental online service to calculate the EWC similarity between translated English terms; 4) 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.

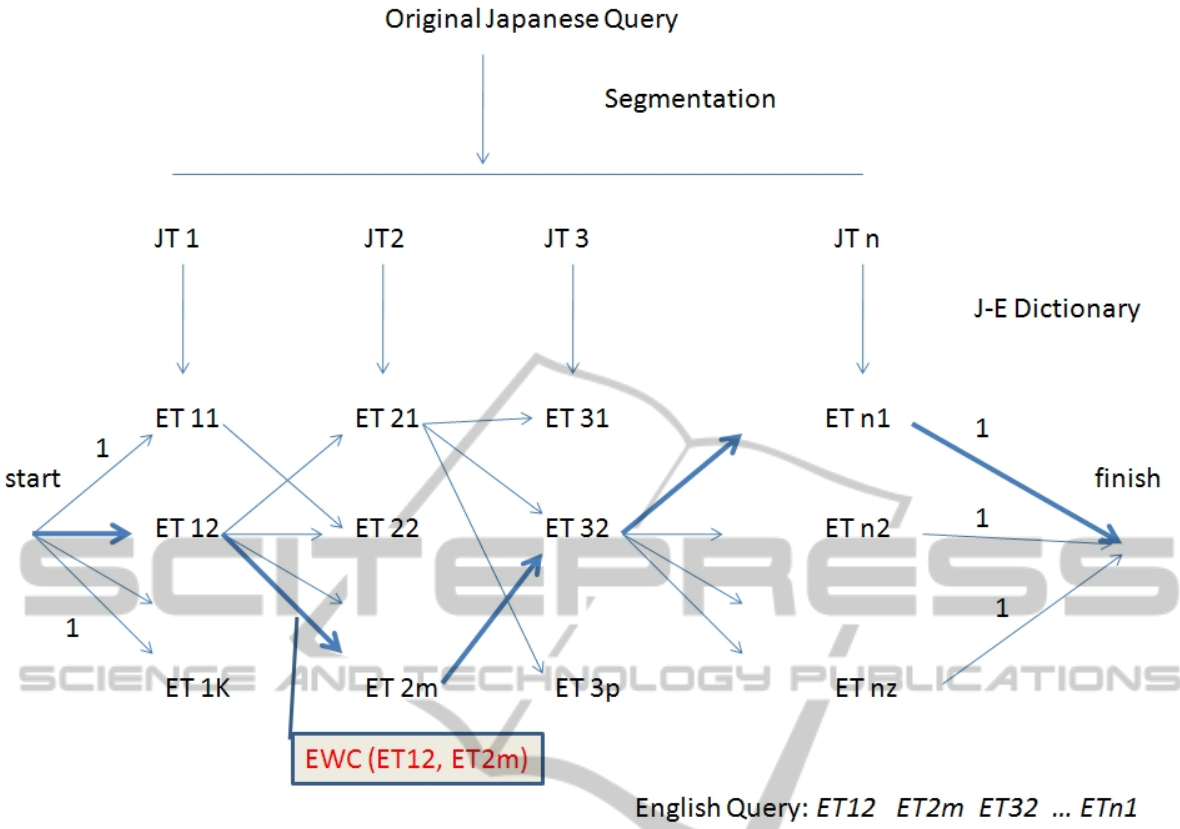

Figure 1: Generating an English query from the Japanese one.

5 PRELIMINARY EXPERIMENTS

We chose the open source search engine (Terrier, 2011) as a tool to index and retrieve data. It provides various retrieval approaches, among which TF-IDF and Okapi's B25 introduced by Robertson at al. (1995). The NTCIR CLIR data collection (NTCIR, 2011) 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 (TREC, 2011). 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. 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;

Table 1: 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 quality grip regulation
English query	information quality regulation
Google Translate	data quality control

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

Table 2: 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 collocations as values for terms
Number of queries	21	21	21	21
Retrieved	19087	18112	18234	19200
Relevant	1756	1756	1756	1756
Relevant retrieved	882	792	407	486
Average Precision	0.3017	0.2644	0.0997	0.1634
R Precision	0.3163	0.2658	0.0844	0.1760

Table 3: Original and translated queries.

Topic number	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; 3) queries generated according to the aforementioned technique applying the longest match strategy to segment Japanese texts, and 4) queries generated applying Mecab to segment Japanese texts and selecting collocations from the results of translation by SPACEALC.

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 by Chen at al. (1998). The longest mach 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.

6 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.

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, last column.

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 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: Average precision of each query.

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

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 (Word, 2011), 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 preliminary 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 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.

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

7 CONCLUSIONS

A novel approach to translate short queries from Japanese into English is introduced. It utilizes the EWC measure to calculate the similarity between translated terms and the shortest path idea to select terms from the list of candidates. It demonstrated the best performance when the longest match strategy is used to segment the original Japanese queries. 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.

To achieve the better accuracy in translation, the EWC implementation should be adjusted to take into account the Wikipedia articles with multiple word terms in the titles.

For short queries, word frequency information seems to be important. Additional experiments are needed to test this hypothesis and hypothesis about

the adjustment of EWC.

REFERENCES

Chen, A., Gey, F. C., Kishida, K., Jiang, H., and Liang, Q. (1999). 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*.

Egozi, O., Markovitch, S., and Gabrilovich, E. (2011). Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems, 29(2)*.

Haralambous, Y. and Klyuev, V. (2011). A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge. In *IJCNLP2011*, Thailand.

Klyuev, V., and Haralambous Y. (2011). Query Expansion: Term Selection using the EWC Semantic Relatedness Measure, In FedCSIS 2011, Poland.

MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Retrieved November 18, 2011, from http://mecab.sourceforge.net/

Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C., Song, R., Lin, Chuan, and Lee., C. (2010). 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.

Nie, J. (2011). *Cross-Language Information Retrieval*, Association for Computational Linguistics.

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

NTCIR-1 CLIR data collection. Retrieved November 18, 2011, from http://research.nii.ac.jp/ntcir/data/data-en.html.

Patwardhan, Banerjee, and Pedersen. (2007). UMND1: Unsupervised Word Sense Disambiguation Using Contextual Semantic Relatedness, In: Proc. *SemEval-2007: 4th International Workshop on Semantic Evaluations*, 390-393, Prague, Czech Republic.

Pinto1, D., Civera, J., Juan, A., Rosso, R., and Barron-Cedeno, A. (2009). A statistical approach to cross lingual natural language tasks. *Journal of Algorithms* Volume 64 Issue 1, 51 – 60.

Robertson, S., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. (1995). Okapi at TREC-4, in Proc. *TREC 4*.

Sorg., P., Cimiano, P. (2008). Cross-lingual Information Retrieval with Explicit Semantic Analysis. In CLEF 2008.

Terrier. Retrieved November 18, 2011, from http://terrier.net

TREC. Retrieved November 18, 2011, from http://trec.nist.gov/

SPACEALC. Retrieved November 18, 2011, from http://www.alc.co.jp/

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

Word frequency lists and dictionary. Retrieved November 18, 2011, from http://www.wordfrequency.info/

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