Query Translation from Indonesian to Japanese Using English as Pivot Language

Ayu Purwarianti, Tsuchiya Masatoshi, Seiichi Nakagawa
Department of Information and Computer Science, Toyohashi University of Technology,
ayu@slp.ics.tut.ac.jp, tsuchiya@cc.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp

Abstract

In this paper, we propose a query translation method for Cross Lingual Information Retrieval (CLIR) system which works for Japanese (target language) documents with Indonesian (source language) queries. Because Indonesian-Japanese is an unfamiliar language pair, it is difficult to translate Indonesian queries to Japanese directly. Therefore, we use English as a pivot language for transitive translation using both Indonesian-English dictionary and English-Japanese dictionary. The experiment against NTCIR-3 Web Retrieval Task data shows that our proposed method is comparable to Indonesian-Japanese dictionary which contains 5000 most frequent words.

1. Introduction

With the fact that there are many languages available in the WWW, CLIR is effective to gain information widely. The most practical approach in building CLIR is to augment the Information Retrieval (IR) system with a query translation system. There are many cases where the language's pair in the query translation has a limited language's pair resources. For this need, a pivot language is commonly used in the translation (transitive translation) [2] [9] [14]. Transitive translation can be applied into many kinds of translation approach such as machine translation, bilingual dictionary translation, parallel corpus based translation. Because the availability of bilingual dictionary is better than other sources such as machine translation and parallel corpus, this technique is the most possible to be used for uncommon language pair.

This bilingual dictionary based transitive translation technique has some weakness, especially that there are many irrelevant translation candidates yielded. To overcome it, a translation filtering technique can be used. In this paper, we used term's co-occurrence score taken from monolingual corpus to select n-best translation pairs [2][8][15]. Other critical problem in query translation is OOV words most of which are proper noun. This problem is more important in languages that use Chinese characters, such as Japanese or Chinese. In this research, to get translation candidates of the proper noun, we employed a proper noun dictionary and transliteration.

The pivot language is usually used with two translations step, from source query to pivot language query, and then from pivot language query to target query. In this paper, we also tried to use pivot language in another way. We built a direct source-target bilingual dictionary from source-pivot and pivot-target bilingual dictionary with "one time inverse consultation technique" [16] to calculate word pair's score and WordNet to enhance intermediate translations.

We used Indonesian as query (source) language to retrieved Japanese (target) documents with English as pivot language. Our system has two components, the first is query translation system to prepare many Japanese candidates, and the second is filtering system to select appropriate words from candidates. Rest of the paper is organized as technique description, experiment result, and conclusion.

2. Query Translation System

2.1 Indonesian Query Characteristics

Words of Indonesian queries are classified into:

1. Indonesian words
2. Borrowed words
3. Japanese proper nouns (written in alphabet)

In order to solve this language combination, we did some strategies:

1. For Indonesian words, we did Indonesian-Japanese translation, by:
   - Using Indonesian-Japanese dictionary built from Indonesian-English and English-Japanese dictionaries
   - Do Indonesian-English-Japanese transitive translation
2. For borrowed words, we used English-Japanese dictionary
3. For Japanese proper nouns, we used Japanese proper noun dictionary which headwords are transliterated into alphabet.

Each strategy above is explained on the following sections.

2.2 Translation of Indonesian Words

2.2.1 Indonesian Japanese Direct Translation

Direct translation means directly translate source query into target query. In order to do direct translation, the source-target bilingual dictionary has to be built first. Here we tried to build source-target (Indonesian-Japanese) bilingual dictionary from source-pivot (Indonesian-English) and pivot-target (English-Japanese) bilingual dictionary.

When building Indonesian-Japanese dictionary from Indonesian-English dictionary and English-Japanese dictionary, explosion of possible translation pairs arises. To select the correct pair among them, we used "one time inverse consultation" [16]. We also used WordNet to get more English translation. The complete procedure is as follows:

1) Do word matching between English translation (from Indonesian-English dictionary) and English word in Japanese-English dictionary.
2) If the English term is a phrase and the matching word couldn't be found, then the English terms will be normalized by eliminating certain words ("to", "a", "an", "the", "to be", "kind of")
3) For every Japanese translation, "one time inverse consultation" score [16] is calculated. English translation for every Japanese candidate is matched with the English translation for Indonesian words. If the match word is more than one, then it is accepted as Indonesian-Japanese pair. But if it is not, then the English translation will be added by its synonym taken from WordNet. The "one time inverse consultation" score is recalculated by using the new English translation set. Figure 1 shows the advantage of using WordNet.

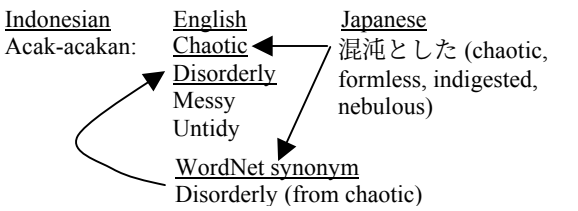

Figure 1. Using WordNet in Dictionary Construction

In Figure 1, if WordNet is not used, then "混沌とした" will not be selected as the translation of "acak-acakan". But if the

WordNet is used, then English translation set will be augmented by "disorderly" and "acak-acakan - 混沌とした" is selected.

2.2.2 Indonesian-English-Japanese Transitive Translation
Such as in [2], the query transitive translation system consists of 2 translation steps, from source query to pivot query and from pivot query to target query. In this research, each language processing step (either for Indonesian query, English query or Japanese query) includes morphological analyzer and stop word removal.

2.3 Translation of Borrowed Words
Because almost of borrowed words are tractable to English, we only concern borrowed English words in this paper. For example is "Academy Awards". It makes that for English sentence such as: "I want to know who have been the recipients of successive generation of *Academy Awards*", the Indonesian translation is "Saya ingin mengetahui siapa yang telah menjadi peraih *Academy Awards* beberapa generasi secara berturut-turut". This kind of term isn't available in Indonesian-Japanese or Indonesian-English dictionary, it is available in English-Japanese dictionary.

2.4 Translation of Japanese Proper Noun
To get the Japanese proper noun translation candidates, we used proper noun dictionary and transliteration. Before doing the proper noun matching, we processed the original proper noun dictionary. Its headwords are transliterated into alphabet and each first– last name pair is separated into two names.

The proper noun translation candidates are then grouped with other translation candidates to be selected by the term selection discussed in section 3.

3. Filtering Technique
In this research, a statistical filtering technique is used. It selected n-best term translation by computing coherence or similarity score for all translation sequences. We calculated terms similarity score with 2 equations: 1) mutual information between terms, such as in [15]; and 2) probability of term sequence. Mutual information between 2 terms is calculated by:

$$MI(t_1,t_2) = \log \frac{p(t_1,t_2)}{p(t_1)\,p(t_2)} \qquad (1)$$

Mutual information for all terms is calculated by:

$$MI(t_1..t_n) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} MI(t_i,t_j) \qquad (2)$$

In term sequence probability, sequence is assumed as a bag of words, where word order is not considered.

$$P(S) = P(w_1..w_n) = \prod_{i=1}^{n} P(w_i|C_i) \qquad (3)$$

Where C_i is other words in sentence except w_i.
Because frequency of many words (C_i) which occur together in a sentence mostly will be zero frequency, below is the definition of probability of a word related with other words in the sequence based on Bayesian law:

$$P(w_i|C_i) = \frac{P(C_i|w_i)}{P(w_i)} \approx P(C_i|w_i)\,P(w_i) = \prod_{j=1, j\neq i}^{n} P(w_j|w_i)\,P(w_i) \qquad (4)$$

$$P(w_i|C_i) \approx \left(\sum_{j=1, j\neq i}^{n} \log P(w_j|w_i)\right) + \log P(w_i) \qquad (5)$$

For all terms in sequence:

$$P(w_1 \ldots w_n) \approx \sum_{i=1}^{n}\left(\sum_{j=1, j\neq i}^{n} \log P(w_j|w_i) + \log P(w_i)\right) \qquad (6)$$

Problem arises when there are too many translation candidates. It will raise computational resource. Therefore, we use an iterative procedure in order to select some best translation. For example, if the query consists of 4 words, then at first we will select n-best translation sequences to get the translation of the first two words. The translation selected will be combined with the next translation set. From this combination set, some best translation sequence will be selected. It is repeated until all translation sets are processed.

4. Experiments
4.1 Experimental System
We conducted some experiments regarding to our proposed method:
1. transitive bilingual dictionary translation (Indonesian-English-Japanese)
2. direct bilingual dictionary translation (Indonesian-Japanese), in two schema:
   1) only using Indonesian-Japanese dictionary;
   2) using 2 dictionaries: Indonesian-Japanese and English-Japanese dictionary.

We also did some other experiments as comparison result:
- transitive machine translation (Indonesian-English-Japanese), taken from:
  - Kataku[10] (Indonesian-English translation)
  - Babelfish[1] and Excite[6] for English-Japanese translation
- using original Indonesian-Japanese dictionary (in direct translation), downloaded from [11]
- manually prepared Japanese keywords

In selecting translation candidates, we used some schemas:
- select all translation candidates (baseline).
- select 5 best term sequences, with 2 frequency matrixes (per word window frequency and per sentence frequency) and 2 term similarity score equations (mutual information and term sequence probability).

4.2 Experimental Data
Our CLIR experiments are done on NTCIR-3 Web Retrieval Task data. The Indonesian queries (47 queries) are manually translated from English queries. The IR system[7] is borrowed from Atsushi Fujii (Tsukuba University). A query example and its manually filtered keywords (underlined words) are shown in Figure 2.

Japanese:	サルサを踊れるようになる方法が知りたい
English:	I want to find out about methods for learning how to dance the salsa
Indonesian:	Saya ingin mengetahui metode untuk belajar bagaimana menari salsa

Figure 2. Query Example and Its Manually Filtered Keywords

The query translation system used resources below:
- Indonesian-English dictionary (KEBI[13], 29,054 Indonesian words)

- English-Japanese dictionary (Eijirou[4], 556,237 English words)
- English stop word list, combined from [6] and [18]
- English morphology rule, implement WordNet[17] description
- Indonesian morphology rule, restricted only for word repetition, posfix –nya and –i
- Chasen[3], Japanese morphological analyzer
- Japanese proper name dictionary (Jinmei Jisho[12], 61,629 Japanese words)
- Generic newspaper corpus

4.3 Experimental Result

Results of all methods that have been described above can be seen in Figures 3 and 4. Figure 3 describes experimental results of Indonesian-Japanese CLIR in Mean Average Precision (MAP) scores. In Figure 3, each query type has 4 average precision score: RL (highly relevant document as correct answer with hyperlink information used), RC (highly relevant document as correct answer), PL (partially relevant document as correct answer with hyperlink information used), and PC (partially relevant document as correct answer). Definition of query group labels on X-axis in Figure 3 can be seen in Table 1.

The highest CLIR score is achieved by transitive machine translation using excite machine (iej-mx). The highest bilingual dictionary based CLIR score is achieved by direct translation using original dictionary (ijo-e), which is comparable with the transitive machine translation using Babelfish machine translation (iej-mb). The main proposed method (iej-*, bilingual dictionary based transitive translation with 5-best term sequences selection method) achieved lower CLIR result than all comparison methods. It gets higher CLIR result than direct translation using original Indonesian-Japanese dictionary after the original dictionary being reduced into 5000 most frequent words.

We also compared bilingual dictionary based translation result with manually filtered keyword list (211 keywords). The comparison is shown in Figure 4. Using 5 best term sequences selection increases precision value for all type of queries, but decreases recall value. The highest recall score for bilingual dictionary based translation is gained by direct translation using original Indonesian-Japanese dictionary (ijo-e). The transitive translation (iej) achieved the second highest recall score. The direct translation with built in dictionary (ijn-e) achieved lower recall score than the transitive translation but it can be comparable with 8000 most frequent words original Indonesian-Japanese dictionary based translation (ijo8-e).

5. Conclusion

There are some points concluded from experiments:

1) Using a statistically corpus based filtering technique gives significant CLIR score improvement especially in transitive translation where there are many translation candidates;

2) Result of transitive translation with a statistical filtering technique can be comparable with 5000 words original Indonesian-Japanese dictionary based translation (direct translation with 2 dictionaries);

3) Using proper noun processing and English-Japanese dictionary significantly increases CLIR score.

Acknowledgement

We would like to give our appreciation to Mr. Atsushi Fujii (Tsukuba University) to allow us to use IR Engine in our research.

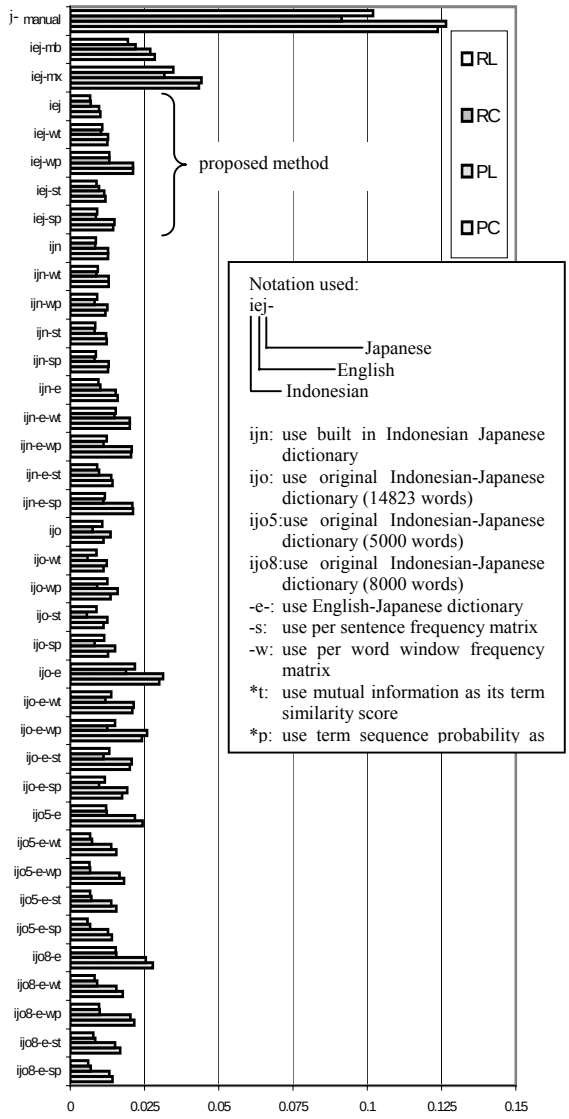

Figure 3. Indonesian-Japanese CLIR Result (MAP Score)

Table 1. Query Group Label Definition

Query	Translation System Description	RC Score
Manually Filtered Keyword		
j-manual	Monolingual Japanese query, filtered by human	0.0913
Machine Transitive Translation		
iej-mb	Using Babelfish for English-Japanese	0.0219
iej-mx	Using Excite for English-Japanese	0.0317
Dictionary Based Transitive Translation		
iej	All translation candidate	0.007
iej-wt	Word window frequency, mutual information score	0.0103
iej-wp	Word window frequency, sequence probability score	0.0131
iej-st	Sentence frequency, mutual information score	0.0096
iej-sp	Sentence frequency, sequence probability score	0.0087
Direct Translation		
1) Using built in Indonesian-Japanese Dictionary only		
ijn	All translation candidate-	0.0083
ijn-wt	Word window frequency, mutual information score	0.0089

Query	Translation System Description	RC Score
ijn-wp	Word window frequency, sequence probability score	0.0081
ijn-st	Sentence frequency, mutual information score	0.0082
ijn-sp	Sentence frequency, sequence probability score	0.0081
2) Using built in Indonesian-Japanese Dictionary and English-Japanese Dictionary		
ijn-e	All translation candidate-	0.0101
ijn-e-wt	Word window frequency, mutual information score	0.0148
ijn-e-wp	Word window frequency, sequence probability score	0.0113
ijn-e-st	Sentence frequency, mutual information score	0.0097
ijn-e-sp	Sentence frequency, sequence probability score	0.0131
3) Using Original Indonesian-Japanese Dictionary only		
ijo	All translation candidate-	0.0075
ijo-wt	Word window frequency, mutual information score	0.0058
ijo-wp	Word window frequency, sequence probability score	0.009
ijo-st	Sentence frequency, mutual information score	0.0057
ijo-sp	Sentence frequency, sequence probability score	0.0082
4) Using Original Indonesian-Japanese Dictionary and English-Japanese Dictionary		
ijo-e	All translation candidate	0.0187
ijo-e-wt	Word window frequency, mutual information score	0.0119
ijo-e-wp	Word window frequency, sequence probability score	0.0124
ijo-e-st	Sentence frequency, mutual information score	0.0112
ijo-e-sp	Sentence frequency, sequence probability score	0.0098
5) Using Original Indonesian-Japanese Dictionary (5000 Most Frequent Words) and English-Japanese Dictionary		
ijo5	All translation candidate	0.0123
ijo5-e-wt	Word window frequency, mutual information score	0.0073
ijo5-e-wp	Word window frequency, sequence probability score	0.0067
ijo5-e-st	Sentence frequency, mutual information score	0.0072
ijo5-e-sp	Sentence frequency, sequence probability score	0.0067
6) Using Original Indonesian-Japanese Dictionary (8000 Most Frequent Words) and English-Japanese Dictionary		
ijo8	All translation candidate	0.0156
ijo8-e-wt	Word window frequency, mutual information score	0.0091
ijo8-e-wp	Word window frequency, sequence probability score	0.0099
ijo8-e-st	Sentence frequency, mutual information score	0.0085
ijo8-e-sp	Sentence frequency, sequence probability score	0.0069

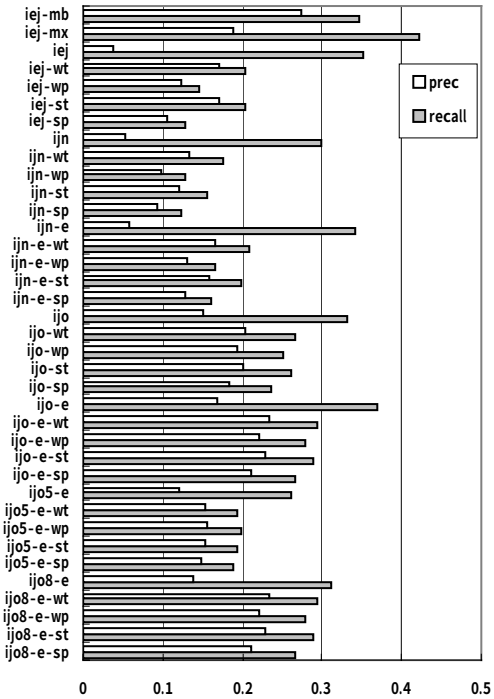

Figure 4. Keyword Comparison of Query Translation Result

References

[1] Babelfish English-Japanese Online Machine Translation, http://www.altavista.com/babelfish/, April 2004.

[2] Ballesteros, Lisa A., "Cross-Language Retrieval via Transitive Translation", *Advances in Information Retrieval*, Kluwer Academic Publisher, 2000, pp. 203-230.

[3] Chasen, http://chasen.naist.jp/hiki/ChaSen/, February 2004.

[4] Eijirou, Alc Co., http://www.alc.co.jp/, 2002

[5] Excite English-Japanese Online Machine Translation, http://www.excite.co.jp/world/, April 2004.

[6] Fox, Christopher, "A Stop List for General Text", *ACM Sigir*, Vol 24, Issue 2 Fall 89/Winter 90, pp. 19-21

[7] Fujii, Atsushi and Katunobu Itou, "Evaluating Speech-driven Web Retrieval in the Third NTCIR Workshop", in *Proc. AAAI Spring Symposium: Intelligent Multimedia Knowledge Management*, 2003.

[8] Gao, Jianfeng, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang, "Improving Query Translation for Cross-Language Information Retrieval using Statistical Model", in *Proc. Sigir*, 2001.

[9] Gollins, Tim and Mark Sanderson, "Improving Cross Language Information Retrieval with Triangulated Translation", in *Proc. Sigir*, 2001.

[10] Indonesian-English Online Machine Translation, http://www.toggletext.com/kataku_trial.php, May 2004.

[11] Indonesian-Japanese Online Dictionary, http://ml.ryu.titech.ac.jp/~indonesia/tokodai/dokumen/kamusjpina.pdf, May 2004

[12] Jinmei Jisho, Nichigai Associates Co., 2003

[13] KEBI, Kamus Elektronik Bahasa Indonesia, http://nlp.aia.bppt.go.id/kebi/, February 2004.

[14] Kishida, Kazuaki and Noriko Kando, "Two-Stage Refinement of Query Translation in a Pivot Language Approach to Cross-Lingual Information Retrieval: An Experiment at CLEF 2003", CLEF 2003, LNCS 3237, pp. 253-262, 2004.

[15] Qu, Yan, G. Grefenstette, D. A. Evans, "Resolving Translation Ambiguity using Monolingual Corpora", *Advanced in Cross-Language Information Retrieval*, vol. 2785 of LNCS, pages 223-241. Springer Verlag, 2002.

[16] Tanaka, Kumiko, and Kyoji Umemura, "Construction of a Bilingual Dictionary Intermediated by a Third Language", 15th *International Conference on Computational Linguistic: COLING-94*, pg. 297-303, Kyoto, 1994.

[17] WordNet, http://wordnet.princeton.edu/, February 2004.

[18] Zu, Guowei, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura, "Automatic Text Classification Techniques", *IEEJ Trans EIS*, Vol. 124, No. 3, 2004.