Indonesian-Japanese CLIR Using Only Limited Resource

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

Abstract

Our research aim here is to build a CLIR system that works for a language pair with poor resources where the source language (e.g. Indonesian) has limited language resources. Our Indonesian-Japanese CLIR system employs the existing Japanese IR system, and we focus our research on the Indonesian-Japanese query translation. There are two problems in our limited resource query translation: the OOV problem and the translation ambiguity. The OOV problem is handled using target language's resources (English-Japanese dictionary and Japanese proper name dictionary). The translation ambiguity is handled using a Japanese monolingual corpus in our translation filtering. We select the final translation set using the mutual information score and the TF×IDF score. The result on NTCIR 3 (NII-NACSIS Test Collection for IR Systems) Web Retrieval Task shows that the translation method achieved a higher IR score than the transitive machine translation (using Kataku (Indonesian-English) and Babelfish/ Excite (English-Japanese) engine) result. The best result achieved about 49% of the monolingual retrieval.

1 Introductions

Due to the various languages used by different nations in the world, the CLIR has been an interesting research topic. For language pair with a rich language resource, the translation in the CLIR can be done with a bilingual dictionary - based direct translation, machine translation - or a parallel corpus - based translation. For a rare language pair, there is an attempt to use a pivot language (usually English), known as transitive translation, because there is no ample bilingual dictionary or machine translation system available. Some studies have been done in the field of transitive translation using bilingual dictionaries in the CLIR system such as [Ballesteros 2000; Gollins and Sanderson 2001]. Ballesteros [2000] translated Spanish queries into French with English as the interlingua. Ballesteros used Collins Spanish-English and English-French dictionaries. Gollins and Sanderson [2001] translated German queries into English using two pivot languages (Spanish and Dutch). Gollins used the Euro Wordnet as a data resource. To our knowledge, no CLIR is available with transitive translation for a source language with poor data resources such as Indonesian.

Translation using a bilingual dictionary usually provides many translation alternatives only a few of which are appropriate. A transitive translation gives more translation alternatives than a direct translation. In order to select the most appropriate translation, a monolingual corpus can be used to select the best translation. Ballesteros and Croft [1998] used an English corpus to select some English translation based on Spanish-English translation and analyzed the co-occurrence frequencies to disambiguate phrase translations. The occurrence score is called the *em* score. Each set is ranked by *em* score, and the highest ranking set is taken as the final translation. Gao et al. [2001] used a Chinese corpus to select the best English-Chinese translation set. It modified the EMMI weighting measure to calculate the term coherence score. Qu et al. [2002] selected the best Spanish-English and Chinese-English translation using an English corpus. The coherence score calculation was based on 1) web page count; 2) retrieval score; and 3) mutual information score. Mirna [2001] translated Indonesian into English and used an English monolingual corpus to select the best translation, employing a term similarity score based on the Dice similarity coefficient. Federico and Bertoldi [2002] combined the N-best translation based on an HMM model of a query translation pair and relevant document probability of the input word to rank Italian documents retrieved by English query. Kishida and Kando [2004], used all terms to retrieve a document in order to obtain the best term combination and chose the most frequent term in

Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?, pages 1–8,
Sydney, July 2006.

each term translation set that appears in the top ranked document.

In our poor resource language – Japanese CLIR where we select Indonesian as the source language with limited resource, we calculate the mutual information score for each Japanese translation combination, using a Japanese monolingual corpus. After that, we select one translation combination with the highest TF×IDF score obtained from the Japanese IR engine.

By our experiments on Indonesian-Japanese CLIR, we would like to show how easy it is to build a CLIR for a restricted language resource. By using only an Indonesian (as the source language) – English dictionary we are able to retrieve Japanese documents with 41% of the performance achieved by the monolingual Japanese IR system.

The rest of the paper is organized as follows: Section 2 presents an overview of an Indonesian query sentence; Section 3 discusses the method used for our Indonesian-Japanese CLIR; Section 4 describes the comparison methods, and Section 5 presents our experimental data and the results.

2 Indonesian Query Sentence

Indonesian is the official language in Indonesia. The language is understood by people in Indonesia, Malaysia, and Brunei. The Indonesian language family is Malayo-Polynesian (Austronesian), which extends across the islands of Southeast Asia and the Pacific [Wikipedia]. Indonesian is not related to either English or Japanese.

Unlike other languages used in Indonesia such as Javanese, Sundanese and Balinese that use their own scripts, Indonesian uses the familiar Roman script. It uses only 26 letters as in the English alphabet. A transliteration module is not needed to translate an Indonesian sentence.

Indonesian language does not have declensions or conjugations. The basic sentence order is Subject-Verb-Object. Verbs are not inflected for person or number. There are no tenses. Tense is denoted by the time adverb or some other tense indicators. The time adverb can be placed at the front or end of the sentence.

A rather complex characteristic of the Indonesian language is that it is an agglutinave language. Words in Indonesian, usually verbs, can be attached by many prefixes or suffixes. Affixes used in the Indonesian language include [Kosasih 2003] me(n)-, ber-, di-, ter-, pe(n)-, per-, se-, ke-, -el-, -em-, -er-, -kan, -i, -nya, -an, me(n)-kan, di-kan, memper-i, diper-i, ke-an, pe(n)-an, per-an, ber-an, ber-kan, se-nya. Words with different affixes might have uniform or different translation. Examples of different word translation are “membaca” and “pembaca”, which are translated into “read” and “reader”, respectively. Examples of same word translation are the words “baca” and “bacakan”, which are both translated into “read” in English. Other examples are the words “membaca” and “dibaca”, which are translated into “read” and “being read”, respectively. By using a stop word elimination, the translation result of “membaca” and “dibaca” will give the same English translation, “read”.

An Indonesian dictionary usually contains words with affixes (that have different translations) and base words. For example, “se-nya” affix declares a “most possible” pattern, such as “sebanyak-banyaknya” (as much as possible), “sesedikit-sedikitnya” (less possible), “sehitam-sehitamnya” (as black as possible). This affix can be attached to many adjectives with the same meaning pattern. Therefore, words with “se-nya” affix are usually not included in an Indonesian dictionary.

Query 1 Saya ingin mengetahui siapa yang telah menjadi peraih *Academy Awards* beberapa generasi secara berturut-turut (I want to know who have been the recipients of successive generations of *Academy Awards*)
Query 2 Temukan buku-buku yang mengulas tentang novel yang ditulis oleh *Miyabe Miyuki* (Find book reviews of novels written by *Miyabe Miyuki*)

Figure 1. Indonesian Query Examples

Indonesian sentences usually consist of native (Indonesian) words and borrowed words. The two query examples in Figure 1 contain borrowed words. The first query contains “Academy Awards”, which is borrowed from the English language. The second query contains “Miyabe Miyuki”, which is transliterated from Japanese. To obtain a good translation, the query translation in our system must be able to translate those words, the Indonesian (native) words and the borrowed words. Problems that occur in a query translation here include OOV words and translation ambiguity.

3 Indonesian - Japanese Query Translation System

Indonesian-Japanese query translation is a component of the Indonesian-Japanese CLIR. The query translation system aims to translate an

Indonesian query sentence(s) into a Japanese keyword list. The Japanese keyword list is then executed in the Japanese IR system to retrieve the relevant document. The schema of the Indonesian-Japanese query translation system can be seen in Figure 2.

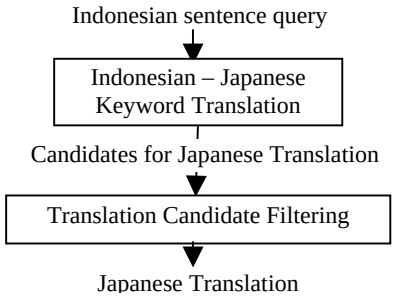

Figure 2. Indonesian-Japanese Query Translation Schema

The query translation system consists of 2 subsystems: the keyword translation and translation candidate filtering. The keyword translation system seeks to obtain Japanese translation candidates for an Indonesian query sentence. The translation candidate filtering aims to select the most appropriate translation among all Japanese translation alternatives. The filtering result is used as the input for the Japanese IR system. The keyword translation and translation filtering process is described in the next section.

3.1 Indonesian – Japanese Key Word Translation Process

The keyword translation system is a process used to translate Indonesian keywords into Japanese keywords. In this research, we do transitive translation using bilingual dictionaries as the proposed method. Other approaches such as direct translation or machine translation are employed for the comparison method. The schema of our keyword transitive translation using bilingual dictionaries is shown in Figure 3.

The keyword translation process consists of native (Indonesian) word translation and borrowed word translation. The native words are translated using Indonesian-English and English-Japanese dictionaries. Because the Indonesian tag parser is not available, we do the translation on a single word and consecutive pair of words that exist as a single term in the Indonesian-English dictionary. As mentioned in the previous section dealing with affix combination in Indonesian language, not all words with the affix combination are recorded in an Indonesian dictionary. Therefore, if a search does not reveal the exact word, it will search for other words that are the basic term of the query word or have the same basic term. For example, the Indonesian word, “munculnya” (come out), has a basic term “muncul” with the postfix “-nya”. Here, the term “munculnya” is not available in the dictionary. Therefore, the searching will take “muncul” as the matching word with “munculnya” and give the English translation for “muncul” such as “come out” as its translation result.

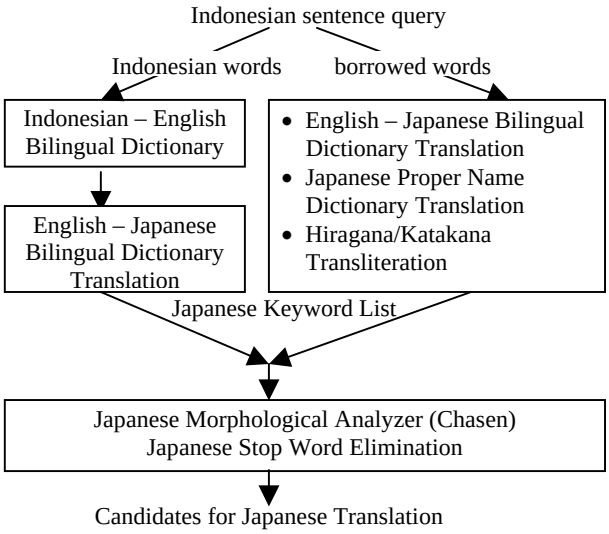

Figure 3. Indonesian-Japanese Keyword Translation Schema

In Indonesian, a noun phrase has the reverse word position of that in English. For example, “ozone hole” is translated as “lubang ozon” (ozone=ozon, hole=lubang) in Indonesian. Therefore, in English translation, besides word-by-word translation, we also search for the reversed English word pair as a single term in an English-Japanese dictionary. This strategy reduces the number of translation alternatives.

The borrowed words are translated using an English-Japanese dictionary. The English-Japanese dictionary is used because most of the borrowed words in our query translation system come from English. Examples of borrowed words in our query are “Academy Awards”, “Aurora”, “Tang”, “baseball”, “Plum”, “taping”, and “Kubrick”.

Even though using an English-Japanese dictionary may help with accurate translation of words, but there are some proper names which can not be translated by this dictionary, such as “Miyabe Miyuki”, “Miyazaki Hayao”, “Honjo Manami”, etc. These proper names come from Japanese words which are romanized. In the Japanese language, these proper names might be written in one of the following scripts: kanji (Chinese character), hiragana, katakana and romaji (roman alphabet). One alphabet word can

be transliterated into more than one Japanese words. For example, “Miyabe” can be transliterated into 宮部, 宮辺, みやべ or ミヤベ. 宮部 and 宮辺 are written in kanji, みやべ is written in hiragana, and ミヤベ is written in katakana. For hiragana and katakana script, the borrowed word is translated by using a pair list between hiragana or katakana and its roman alphabet. These systems have a one-to-one correspondence for pronunciation (syllables or phonemes), something that can not be done for kanji. Therefore, to find the Japanese word in kanji corresponding to borrowed words, we use a Japanese proper name dictionary. Each term in the original proper name dictionary usually consists of two words, the first and last names. For a wider selection of translation candidates, we separate each term with two words into two terms. Even though the input word can not be found in the original proper name dictionary (family name and first name), a match may still be possible with the new proper name dictionary.

Each of the above translation processes also involves the stop word elimination process, which aims to delete stop words or words that do not have significant meaning in the documents retrieved. The stop word elimination is done at every language step. First, Indonesian stop word elimination is applied to a Indonesian query sentence to obtain Indonesian keywords. Second, English stop word elimination is applied before English keywords are translated into Japanese keywords. Finally, Japanese stop word elimination is done after the Japanese keywords are morphologically analyzed by Chasen (http://chasen.naist.jp/hiki/ChaSen).

The keyword transitive translation is used in 2 systems: 1) transitive translation to translate all words in the query, and 2) transitive translation to translate OOV (Indonesian) words from direct translation using an Indonesian-Japanese dictionary. We label the first method as the transitive translation using bilingual dictionary and the second method as the combined translation (direct-transitive).

3.2 Candidate Filtering Process

The keyword transitive translation results in many more translation candidates than the direct translation result. The candidates have a translation ambiguity problem which will be handled by our Japanese translation candidate filtering process, which seeks to select the most appropriate translation among the Japanese translation candidates. In order to select the best Japanese translation, rather than choosing only the highest TF×IDF score or only the highest mutual information score among all sets, we combine both scores. The procedure is as follows:

1. Calculate the mutual information score for all term sets. To avoid calculation of all term sets, we calculate the mutual information score iteratively. First we calculate it for 2 translation candidate sets. Then we select 100 sets with the highest mutual information score. These sets are joined with the 3rd translation candidate sets and the mutual information score is recalculated. This step is repeated until all translation candidate sets are covered.
   For a word set, the mutual information score is shown in Equation 1.

$$I(t_1 \ldots t_n) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(t_i;t_j) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\log P(t_i,t_j)}{\log P(t_i).\log P(t_j)} \quad (1)$$

   $I(t_1 \ldots t_n)$ means the mutual information for a set of words t_1, t_2,…t_n. $I(t_i,t_j)$ means the mutual information between two words (t_i,t_j). Here, for a zero frequency word, it will have no impact on the mutual information score of a word set.
2. Select 5 sets with highest mutual information score and execute them into the IR engine in order to obtain the TF×IDF scores. The TF×IDF score used here is the relevance score between the document and the query (Equation (2) from Fujii and Ishikawa [2003]).

$$\sum_{t} \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} . log \frac{N}{DF_t} \right) \quad (2)$$

   $TF_{t,i}$ denotes the frequency of term t appearing in document i. DF_t denotes the number of documents containing term t. N indicates the total number of documents in the collection. DL_i denotes the length of document i (i.e., the number of characters contained in i), and avglen the average length of documents in the collection.
3. Select the term set with the highest mutual information score among 3 top TF×IDF scores

Figure 4 shows an example of the keyword selection process after completion of the

keyword translation process. The translation combination and set rankings are for all words (4 translation sets) in the query. Actually, the translation combinations and sets for the query example are also ranked for 2 and 3 translation sets. All resulting sets (ranked by its mutual information score) are executed in the IR system in order to obtain the TF×IDF score. The final query chosen is the one with the highest TF× IDF score.

Query:
Saya ingin mengetahui metode untuk belajar bagaimana menari salsa (= I wanted to know the method of studying how to dance the salsa)

Keyword Selection:
Metode (method), belajar (to learn, to study, to take up), menari (dance), salsa

Japanese Keyword:
Metode: 規則正し,筋道,秩序,方法
Belajar: **調べる,勉強,研究,学ぶ,調査,検討,書斎,知る,わかる,暗記,覚える,確認,習う,突きとめる**
Menari: 舞踊,ダンス,パーティー,バレエ,舞う,踊る,踊ら
Salsa: サルサ

Translation Combination:
(規則正し,調べる,舞踊,サルサ)
(筋道,調べる,舞踊,サルサ)
(秩序,調べる,舞踊,サルサ), etc

Rank sets based on Mutual Information Score:
1. (**秩序，知る，　踊る，　サルサ**)
2. (**秩序，研究，　踊る，　サルサ**)
3. (**方法，わかる，ダンス，サルサ**)
4. (**方法，覚える，ダンス，サルサ**)
5. (**秩序，わかる，踊る，　サルサ**)

Select query with highest TF×IDF score
方法．わかる．ダンス．サルサ

Figure 4. Illustration of Translation Filtering Method

4 Compared Methods

In the experiment, we compare our proposed method with other translation methods. Methods for comparing Indonesian-Japanese query translation include transitive translation using MT (machine translation), direct translation using existing Indonesian-Japanese dictionary, direct translation using a built-in Indonesian-Japanese dictionary, transitive translation with English keyword selection based on mutual information taken from English corpus, and transitive translation with Japanese keyword selection based on mutual information only.

4.1 Transitive Translation using Machine Translation

The first method compared is a transitive translation using MT (machine translation). The Indonesian- Japanese transitive translation using MT has a schema similar to Indonesian-Japanese transitive translation using a bilingual dictionary. However, machine transitive translation does not use Indonesian-English and English-Japanese dictionaries. Indonesian queries are translated into English queries using an online Indonesian-English MT (Kataku engine, http://www.toggletext.com). The English translation results are then translated into Japanese using 2 online MTs (Babelfish engine, http://www.altavista.com/babelfish and Excite engine, http://www.excite.co.jp/world).

4.2 Direct Translation using Existing Indonesian-Japanese Bilingual Dictionary

The second method compared is a direct translation using an Indonesian-Japanese dictionary. This direct translation also has a schema similar to the transitive translation using bilingual dictionary (Figure 2). The difference is that in translation of an Indonesian keyword, only 1 dictionary is used, rather than using 2 dictionaries; in this case, an Indonesian-Japanese bilingual dictionary with a fewer words than the Indonesian-English and English-Japanese dictionaries.

4.3 Direct Translation using Built-in Indonesian-Japanese Dictionary

We also compare the transitive translation results with the direct translation using a built-in Indonesian-Japanese dictionary. The Indonesian-Japanese dictionary is built from Indonesian-English, English-Japanese and Japanese-English dictionaries using “one-time inverse consultation” such as in Tanaka and Umemura [1998]. The matching process is similar with that in query translation. A Japanese translation is searched for an English translation (from every Indonesian term in Indonesian-English dictionary) as a term in the Japanese-English dictionary. If no match can be found, the English terms will be normalized by eliminating certain stop words (“to”, “a”, “an”, “the”, “to be”, “kind of”). These normalized English terms will be checked again in the Japanese-English dictionary. For every Japanese translation, a “one-time inverse consultation” is calculated. If the score is

more than one (for more than one English term), then it is accepted as an Indonesian-Japanese pair. If not, the WordNet is used to find its synonym and recalculate the "one-time inverse consultation" score so as to compensate for the poor quality of Indonesian-English dictionary (29054 words).

5 Experiments

5.1 Experimental Data

We measure our query translation performance by the IR score achieved by a CLIR system because CLIR is a real application and includes the performance of key word expansion. For this, we do not use word translation accuracy, as for the CLIR, since a one-to-one translation rate is not suitable, given there are so many semantically equivalent words.

Our CLIR experiments are conducted on NTCIR-3 Web Retrieval Task data (100 Gb Japanese documents), in which the Japanese queries and translated English queries were prepared. The Indonesian queries (47 queries) are manually translated from English queries. The 47 queries contain 528 Indonesian words (225 are not stop words), 35 English borrowed words, and 16 transliterated Japanese words (proper nouns). The IR system (Fujii and Ishikawa [2003]) is borrowed from Atsushi Fujii (Tsukuba University). External resources used in the query translation are listed in Table 1.

Table 1. External Resource List

Resource	Description
KEBI	Indonesian-English dictionary, 29,054 words
Eijirou	English-Japanese dictionary, 556,237 words
Kmsmini2000	Indonesian-Japanese dictionary, 14,823 words
ToggleText Kataku	Indonesian-English machine translation
Excite	English-Japanese machine translation
Babelfish	English-Japanese machine translation
[Fox, 1989] and [Zu et al., 2004]	English stop words (are also translated into Indonesian stop words)
Chasen	Japanese morphological analyzer
Jinmei Jisho	Japanese proper name dictionary, 61,629 words
Mainichi Shinbun & Online Yomiuri Shinbun	Japanese newspaper corpus

5.2 Experimental Result

In the experiments, we compare the IR score of each translation method. The IR scores shown in this section are in Mean Average Precision (MAP) scores. The evaluation metrics is referred to [Fujii and Ishikawa 2003b]. Each query group has 4 MAP scores: 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). The documents hyperlinked from retrieved documents are used for relevance assessment.

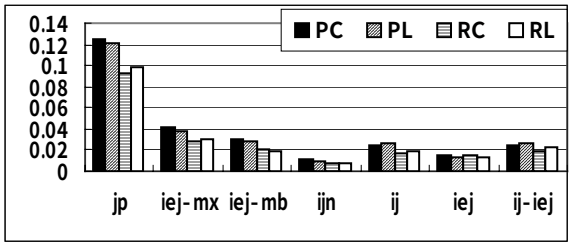

Figure 5. Baseline Indonesian-Japanese CLIR

Figure 5 shows the IR scores of queries translated using basic translation methods such as the bilingual dictionary or machine translation, without any enhanced process. The labels used in Figure 5 are:

- jp (monolingual translation), where "jp" denotes Japanese query
- iej (transitive translation using bilingual dictionary), where "i", "e", "j" denote Indonesian, English and Japanese, respectively,
- iej-mx (transitive machine translation using Kataku and Excite engines), where "m" denotes machine translation,
- iej-mb (transitive machine translation using Kataku and Babelfish engines),
- ijn (direct translation using the built in Indonesian-Japanese dictionary),
- ij (direct translation using Indonesian-Japanese dictionary),
- ij-iej (combination of direct (ij) and transitive (iej) translation using bilingual dictionary).

The highest CLIR score in the baseline translation (without the enhancement process) achieves 30% of the performance achieved by the monolingual IR (jp).

IR results in Figure 6 shows that OOV translation does improve the retrieval result. Here, our proposed methods (iej and ij-iej) achieve lower score than the comparison methods.

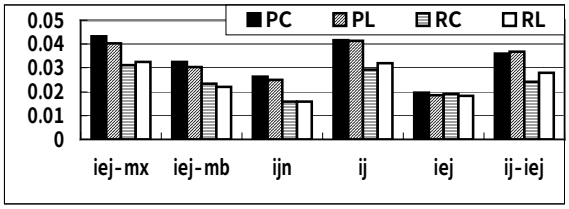

Figure 6. Indonesian-Japanese CLIR with OOV Translation

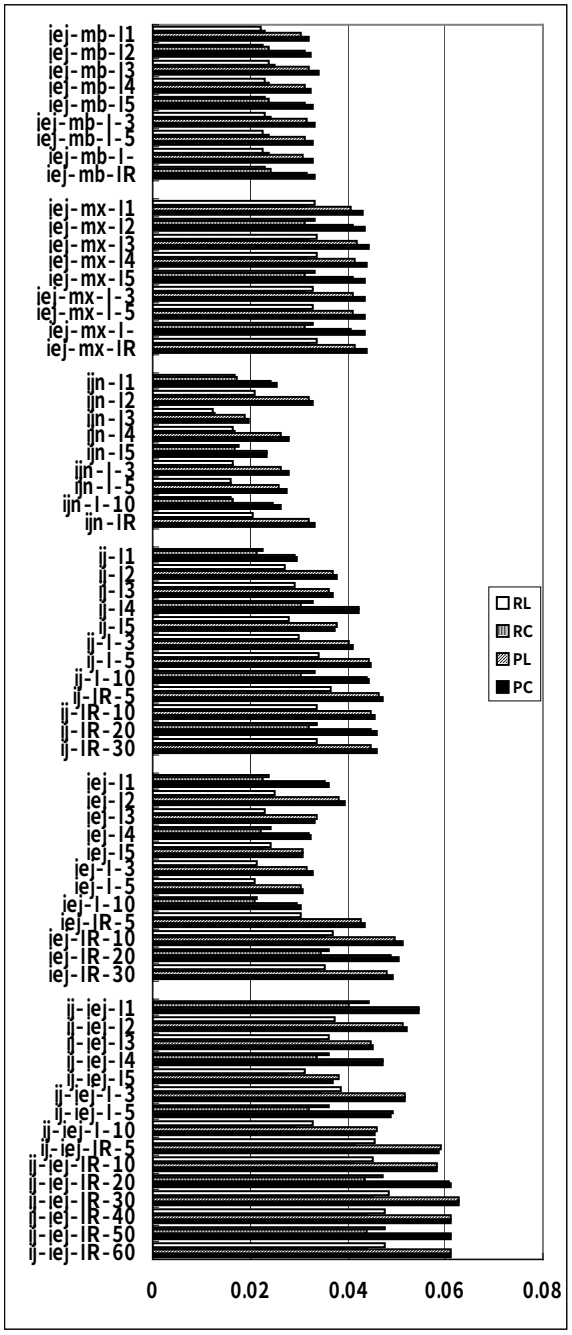

Figure 7. Indonesian-Japanese CLIR with OOV Translation and Keyword Filtering

Figure 7 shows the MAP score on the proposed Indonesian-Japanese CLIR. The keyword selection description of each query label follows:

- In (n = 1 .. 5): one query candidate based on mutual information score; example: I2 means the 2nd ranked query by its mutual information score.
- I-n (n = 3,5,10): combination of the n-best query candidates based on mutual information score; example: iej-3 (disjuncture of the 3-best mutual information score candidates).
- IR: the 1-best query candidate based on combination of mutual information score and TF × IDF engine score. X in IR-X shows number of combinations. For example, IR-5 means the highest TF × IDF score among 5 highest mutual information score sets.

Figure 7 shows that the proposed filtering method yields higher IR score on the transitive translation. We achieve 41% of the performance achieved by the monolingual IR. The proposed transitive translation (iej-IR-10) improves the IR score of the baseline method of transitive translation (iej) from 0.0156 to 0.0512. The *t*-test shows that iej-IR-10 significantly increases the baseline method (iej) with a 97% confidence level, $T(68) = 1.91$, $p<0.03$. *t*-test also shows that, compared to other baseline systems, the proposed transitive translation (iej-IR-10) can significantly increase the IR score at 85% ($T(84) = 1.04$, $p<0.15$), 69% ($T(86) = 0.49$, $p<0.31$), 91% ($T(83) = 1.35$, $p<0.09$), and 93% ($T(70) = 1.49$, $p<0.07$) confidence level for iej-mb, iej-mx, ij and ij-iej, respectively. Another proposed method, a combination of direct and transitive translation (ij-iej), achieved the best IR score among all the translation methods. The proposed combination translation method (ijiej-IR-30) improves the IR score of the baseline combination translation (ij-iej) from 0.025 to 0.0629. The *t*-test showed that the proposed combination translation improves IR score of the baseline ij-iej with a 98% confidence level, $T(69) = 2.09$, $p<0.02$. Compared to other baseline systems, *t*-test shows that the proposed combination translation method (ijiej-IR-30) improves the IR score at 95% ($T(83) = 1.66$, $p<0.05$), 86% ($T(85) = 1.087$, $p<0.14$), 97%, ($T(82) = 1.91$, $p<0.03$) and 99% ($T(67) = 2.38$, $p<0.005$) confidence level for iej-mb, iej-mx, ij and iej, respectively.

6 Conclusions

We present a translation method on CLIR that is suitable for language pair with poor resources, where the source language has a limited data resource. Compared to other translation methods

such as transitive translation using machine translation and direct translation using bilingual dictionary (the source-target dictionary is a poor bilingual dictionary), our transitive translation and the combined translation (direct translation and transitive translation) achieve higher IR scores. The transitive translation achieves a 41% performance of the monolingual IR and the combined translation achieves a 49% performance of the monolingual IR.

The two important methods in our transitive translation are the borrowed word translation and the keyword selection method. The borrowed word approach can reduce the number of OOV from 50 words to 5 words using a pivot-target (English-Japanese) bilingual dictionary and target (Japanese) proper name dictionary. The keyword selection using the combination of mutual information score and TF×IDF score has improved the baseline transitive translation. The other important method, the combination method between transitive and direct translation using bilingual dictionaries also improves the CLIR performance.

Acknowledgements

We would like to give our appreciation to Dr. Atsushi Fujii (Tsukuba University) to allow us to use the IR Engine in our research. This work was partially supported by The 21st Century COE Program "Intelligent Human Sensing"

References

Adriani, Mirna. 2000. *Using statistical term similarity for sense disambiguation in cross language information retrieval*. Information Retrieval: 67-78.

Agency for The Assessment and Application of Technology: KEBI (Kamus Elektronik Bahasa Indonesia). http://nlp.aia.bppt.go.id/kebi/. Last access: February 2004.

Babelfish English-Japanese Online Machine Translation. http://www.altavista.com/babelfish/. Last access: April 2004.

Ballesteros, Lisa A. and W. Bruce Croft. 1998. *Resolving ambiguity for cross-language retrieval*. ACM Sigir.

Ballesteros, Lisa A. 2000. *Cross Language Retrieval via Transitive Translation*. Advances in Information Retrieval: 203-230. Kluwer Academic Publisher.

Chasen. http://chasen.naist.jp/hiki/ChaSen/. Last access: February 2004.

Chen, Kuang-hua, et,al. 2003. *Overview of CLIR Task at the Third NTCIR Workshop*. Proceedings of the Third NTCIR Workshop.

Excite English-Japanese Online Machine Translation. http://www.excite.co.jp/world/. Last access: April 2004.

Federico, M. and N. Bertoldi. 2002. *Statistical cross language information retrieval using n-best query translations*. Proc. Of 25th International ACM Sigir.

Fox, Christopher. 1989. *A stop list for general text*. ACM Sigir, Vol 24:19-21, Issue 2 Fall 89/Winter 90.

Fujii, Atsushi and Tetsuya Ishikawa. 2003. *NTCIR-3 cross-language IR experiments at ULIS*. Proc. Of the Third NTCIR Workshop.

Fujii, Atsushi and Katunobu Itou. 2003. *Building a test collection for speech driven web retrieval.* Proceedings of the 8th European Conference on Speech Communication and Technology.

Gao, Jianfeng, et, al. 2001. *Improving query translation for cross-language information retrieval using statistical model*. Proc. Sigir.

Gollins, Tim and Mark Sanderson. 2001. *Improving cross language information retrieval with triangulated translation*. Proc. Sigir.

ToggleText, Kataku Automatic Translation System. http://www.toggletext.com/kataku_trial.php. Last access: May 2004.

Information Retrieval Resources for Bahasia Indonesia. Informatics Institute, University of Amsterdam. http://ilps.science.uva.nl/Resources/. Last access: Jan 2005.

Kishida, Kazuaki and Noriko Kando. 2004. *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: 253-262.

Kosasih, E. 2003. *Kompetensi Ketatabahasaan dan Kesusastraan, Cermat Berbahasa Indonesia*. Yrama Widya.

Mainichi Shinbun CD-Rom data sets 1993-1995, Nichigai Associates Co., 1994-1996.

Michibata, H., ed.: Eijirou, Alc. Last access:2002.

Qu, Yan and G. Grefenstette, D. A. Evans. 2002. *Resolving translation ambiguity using monolingual corpora*. Advanced in Cross-Language Information Retrieval, vol. 2785 of LNCS: 223-241. Springer Verlag.

Sanggar Bahasa Indonesia Proyek: Kmsmini2000. http://ml.ryu.titech.ac.jp/~indonesia/tokodai/dokumen/ kamusjpina.pdf. Last access: May 2004.

Tanaka, Kumiko and Kyoji Umemura. *Construction of a bilingual dictionary intermediated by a third language*. COLING 1994, pages 297-303, Kyoto.

Wikipedia on Indonesian Language. http://en.wikipedia.org/wiki/_Indonesian_language. Last access: May 2005.

WordNet. http://wordnet.princeton.edu/. Last access: February 2004.

Zu, Guowei, et, al. 2004. *Automatic Text Classification Techniques*. IEEJ Trans EIS, Vol. 124, No. 3.