Lang Resources & Evaluation (2008) 42:127–136
DOI 10.1007/s10579-007-9038-4

Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary

Francis Bond · Kentaro Ogura

Published online: 12 October 2007

Abstract We present a method for combining two bilingual dictionaries to make a third, using one language as a pivot. In this case we combine a Japanese-English dictionary with a Malay-English dictionary, to produce a Japanese-Malay dictionary. Our method differs from previous methods in its improved matching through normalization of the pivot language. We have made a prototype dictionary of around 76,000 Japanese-Malay pairs for 50,000 Japanese head words.

Keywords Bilingual lexicon · Lexicon construction · Japanese · Malay

1 Introduction

We present a method for combining two bilingual dictionaries to make a third, using one language as a pivot. It is an extension of Bond et al. (2001), with some improvements in the method for matching the pivot language. The original aim of our research was to create a dictionary to be used in the machine translation system **ALT-J/M**: the Automatic Language Translator—Japanese-to-Malay (Ogura et al. 1999). However, the resulting dictionary is potentially useful for human users, and has better cover than any currently published dictionary.

F. Bond
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kyoto, Japan

K. Ogura
NTT Software Corporation, Yokohama, Japan
e-mail: ogura-k@po.ntt.s.co.jp

Present Address:
F. Bond (✉)
Computational Linguistics Group, NICT, Kyoto, Japan
e-mail: bond@ieee.org

The reasons we wish to do this are 2-fold. First, there are no large-scale Japanese-Malay dictionaries available, either for human or machine use. The largest lexicons we could find had between 6,000 (Kasim and Jambi 1999) and 7,000 head words (Nagata 1994), and fewer than 15,000 translation pairs. This is too few for a large-scale machine translation system. There are also no significant aligned Japanese-Malay corpora, so we cannot induce a dictionary from aligned text.

Second, we need to build a dictionary that has not only Japanese words and their Malay equivalents, but also semantic and syntactic information. By using our existing Japanese-English dictionary, we can exploit the semantic information it contains, transferring as much as possible to the new dictionary. This rich dictionary can be used for a variety of tasks, in this paper we principally consider machine translation from Japanese to Malay.

The **ALT** systems are semantic transfer systems, and rely on having nouns marked with appropriate semantic classes (from an ontology of roughly 3,000 classes). These semantic classes are then used to describe the selectional restrictions of predicate-frames.

Clearly different senses of the same noun can be differentiated because they will appear in different semantic classes, for example, *seal* ⇔ あざらし *azarashi* ⟨animal⟩ vs *seal* ⇔ 印 *in* ⟨tool⟩. We will refer to such clearly distinct senses as *homonyms*. In a machine translation system, homonyms can be translated correctly if they have the correct semantic classes marked.

Finer grained *variations*, such as the difference between *doves* and *pigeons* (both 鳩 *hato* in Japanese, with the same basic meaning and the same semantic class bird) are harder to distinguish. Instead, collocation and usage information is necessary. Various methods exist to distinguish between such variants in machine translation, including the use of domain information, noun-modifier collocation, n-grams and other statistical information. The fall-back method for distinguishing between similar variants is frequency: which of a set of translation equivalents occurs most often. In our system, this is implemented as a preference value: if the semantic classes are the same, in the absence of other restrictions, choose the translation candidate with the highest preference.

When translating, it is essential to distinguish between homonyms, in order to faithfully convey the sense of a text. It is less important to distinguish between variations. Because of this, when building our dictionary, it is essential to distinguish homonyms correctly, and our method aims to do this. In practice it may be impossible to reliably distinguish variations: because different languages make different distinctions the source text may have insufficient information to disambiguate all the nuances in the target language.

1.1 Related work

Tanaka et al. (1998) used English as an intermediate language to link Japanese and French. Their method relies on inverse consultation. To find suitable equivalents for a given Japanese word, they first look up its English translations, and then the French translations of these English translations, giving a set of French equivalence

candidates of the original Japanese. For each French word, they then look up all of its English translations, and see how many match the English translations of the original Japanese word. The more matches there are, the better the candidate is. They call this "one time inverse consultation".

An example of one time inverse consultation, between Japanese and Malay, is given in Fig. 1. There are three translations of the Japanese word 印 *in "seal"*, and four translations of its equivalence candidate *tera "seal"*. There are two shared translations (underlined in the figure). To normalize the score, it is multiplied by two (thus if all words match the score will be one). This gives a score of $0.57 = 2 \times \frac{2}{3+4}$.

Tanaka et al. (1998) were able to find translation equivalents not found in equivalent Japanese-French dictionaries by matching published Japanese-English and English-French dictionaries against each other. Evaluating the results for one time inverse consultation gave recall of 44% and precision of 76% for nouns, down to 15% and 65% for adjectives.

Shirai and Yamamoto (2001) also use one time inverse consultation to create a Japanese-Korean Dictionary, using English as the pivot language. By limiting the types of matching allowed, they were able to increase precision to as high as 82.6%, but at the cost of greatly reducing the number of pairs found. Paik et al. (2001) extended this work by using Chinese characters (used in both Japanese and Korean) as a second pivot. Chinese characters were also used as a second pivot by Zhang et al. (2005) to create Japanese-Chinese lexicons.

One shared characteristic of these approaches is the use of English as the pivot language. This is because, in general, there are more bilingual resources available with English as one of the languages. None of the previous work uses semantic information or matches through two or more languages.

2 Creating a Japanese-Malay dictionary

In this section we first describe the Japanese-English and Malay-English dictionaries we use, and then how we combine them.

2.1 The Japanese-English dictionary: Goi-Taikei

For the Japanese-English dictionary, we use the dictionaries developed for the machine translation system **ALT-J/E** (Ikehara et al. 1991).

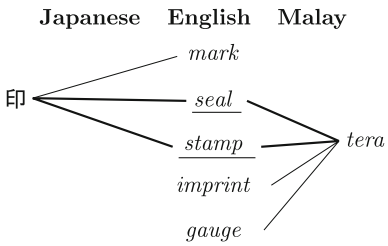

Fig. 1 One time inverse consultation score

GT consists of three main components: (i) an ontology, (ii) a semantic word dictionary, and (iii) a semantic clause structure dictionary which includes subcategorization frames for predicates.

Each record specifies an index form (Japanese), translation (English), preference ranking, English syntactic information and a set of semantic classes from a semantic hierarchy of 2,710 nodes. Optionally there may be more detailed selectional restrictions, domain and genre information and so on. English translations follow American spelling conventions.

There are 343,901 unique Japanese head word/part of speech (POS) entries, linked by 427,918 translations into 293,140 unique English head words. On average each Japanese word has 1.2 translations. There is a tendency for many Japanese words to be translated into the same English translation: there are fewer unique English entries than Japanese, and many of them are multi-word expressions.

2.2 The Malay-English dictionary: KAMI

We use the Malay-English Dictionary **KAMI**: KAmus Melayu-Inggeris. This dictionary was compiled by NTT-MSC, based on a dictionary produced originally by a translation company (Quah et al., (2001). The dictionary currently has 67,670 Malay words with English translations. 69% have only one translation, 19% have two, 7% have three; the average number of translations is 1.57, giving 106,558 Malay-English pairs. Each entry in the dictionary consists of the following fields: (1) Malay index word; (2) Malay root word; (3) Malay POS; (4) detailed syntactic features; (5) semantic classes; (6) English translation; (7) English comments; (8) Chinese translation. All entries have values for fields 1,2 and 3; most have syntactic features. Only 28% have semantic classes from the **GT** ontology, 22% have Chinese translations. English and Chinese translations and comments are provided for use in a machine translation system, as well as an aid for non-Malay speakers. English translations follow British spelling conventions. Semantic classes were entered in several ways: (1) The original dictionary we purchased had some syntactic-semantic codes. (2) The CICC Indonesian dictionary has semantic classifications (CICC 1994). As Malay and Indonesian share much of their vocabulary, we looked up Malay-English pairs in the CICC Indonesian-English dictionary, and took the semantic classes from the matching Indonesian pairs (14,784 entries). (3) Because some classifiers select for the meanings of their targets, we could use the classifiers to predict the semantic class of their targets (18,303 entries). For example, anything counted by *orang* is `human`, anything counted by *ekor* is `animal` and so on. (4) We added semantic classes by checking against known word lists such as the ISO 639 language names and the ISO 4217 currency names (a few hundred entries). Finally, (5) we added some semantic classes to some words by hand, although not in any systematic way. Because of the overlap between the five classes described above, we only have semantic classes for around 29,900 entries (28%).

2.3 Crossing the dictionaries

Building the Japanese-Malay dictionary involves two steps: creating Japanese-Malay equivalence candidates, and then filtering and ranking the candidates. The overall flow is shown below:

- For each pair in the Japanese-English dictionary
  - Look up the Malay equivalent of the normalized English (normalize by case, US/GB spelling variant, number) if an entry with the same POS exists
    - Create a Japanese-Malay pair (with English link)
    - Calculate match scores
  - else mark the Japanese-English pair
- For each Japanese index word in the Japanese-English dictionary
  - Output any Japanese-Malay pairs ranked by total score
  - Output marked Japanese-English pairs ranked by preference

English entries are normalized, in particular articles (*a/an, the*) and infinitival *to* are stripped from the beginning of noun and verb entries respectively. If no match is found, the English is treated further, by normalizing case, then British/US spelling, then number. Case is normalized by downcasing the index word.

Spelling is normalized using the VarCon tables (Rev 2) of American, British, and Canadian spellings and vocabulary (Atkinson). These consists of triples such as "*labor, labour, labour*". Words are matched against British (column two) and converted to American (column one).

Number is normalized by attempting to convert the index entry first to singular (sg) using simple regular expressions (`s/ses $/s/`, `s/ies$/y/`, `s/s$//`), then if no match is found to plural (pl) using the `Lingua :: EN :: Inflect` perl module (Conway 2000). For case and spelling conversion, each word in an entry is checked, for number, only the final word. Some examples of normalization are given in Table 1. Case conversion is not done for proper nouns, as case is informative, and singular/plural conversion is only done for nouns, as other parts of speech do not inflect for number in English.

Our crossing process is opportunistic: taking immediate advantage of any circumstance of possible benefit. Ideally we will only apply it once, and then check all entries by hand. Because of this, we do minimal filtering, preferring instead to

Table 1 Matching through normalization

Japanese	English	English	Malay	Type
石器時代	Stone Age	Stone age	zaman batu	case
付近	Neighborhood	Neighbourhood	kejiranan	var
色温度	Color temperature	Colour temperature	suhu warna	var
おとし卵	Poached eggs	Poached egg	telur rebus carak	sg
石炭層	Coal seam	Coal seams	jaluran arang batu	pl
定期航空路	Air line	Air lines	penerbangan awam	pl

maximize the number of equivalence candidates. However, we wish to use the dictionary immediately, as thorough checking may take several person-years. Therefore, it is important to get as good a translation as possible in the top ranked position.

Pairs were only crossed if they had the same part of speech (using a small set of coarse categories: `common−noun`, `proper−noun`, `verb`, `adjective`, `adverb`, `pronoun`). We used the English part of speech in the J-E dictionary, and the Malay part of speech in the M-E dictionary. Ideally we would like to use English parts of speech for both lexicons, if available. Matching only compatible parts of speech cut down greatly on the number of false matches. Crossing to different parts of speech only increased the number of new Japanese matches by 2%, at the cost of increasing the number of equivalence candidates by 15%, most of which were spurious.

We combine three scores. The one time inverse consultation score is the same as Tanaka et al. (1998) (§ 1.1). The semantic matching score was the number of times a semantic class of *J* was compatible with a semantic class of *M*, where two classes are compatible if either semantic class subsumes the other. For example, `animal` is compatible with `living−thing`. Only nouns have semantic classes in our lexicons, so this score is only applicable to nouns.

The second-language matching score used Chinese as a second intermediate language. Our Malay-English dictionary also has Chinese entries for 21,190 of its entries (25%). If a matched Malay entry had a Chinese translation, then we checked to see whether the Japanese and Chinese pair could be found in a Japanese-Chinese dictionary of some 83,000 entries (Shogakukan and Peking Shomoin sho kan, 1987). We assume that anything that matches through two different languages (Japanese to Malay through English and Chinese) should be a good match. In particular, we expect different homonyms in different languages, so using two pivot languages should be effective in distinguishing between them. We give an example of a match through two languages in Fig. 2. Here *tera "seal"* matches through both English and Chinese, so is a good match. The entry *mohor "seal"* matches through two English words, so is a reasonable match, and *anjing laut "seal"* matches through only one word, so is a bad match.

The total score is a combination of the semantic matching score, the original preference of the Japanese-English pair, and the one time inverse consultation score, combined so that the Chinese matches come first, followed by the semantic matches, followed by high ranked pairs; within the same ranking, pairs are ordered by one time inverse consultation score. Candidates are never deleted, that is left to the lexicographers.

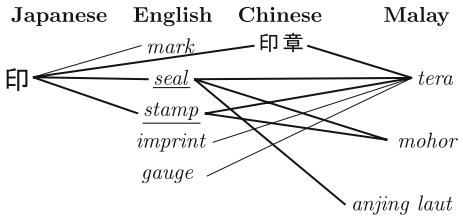

Fig. 2 Matching through two languages

3 Results and evaluation

3.1 Results

In this section we report on crossing the Japanese-English common-noun dictionary with the Malay-English dictionary. 50,034 out of 343,901 Japanese words were linked to 44,157 Malay words. Excluding proper nouns, for which there were only 770 matches, this is 49,283 out of 154,680 or 31%. There were 342,166 Japanese-Malay pairs, with an average ambiguity of 6.8. Clearly, we have introduced many spurious translations: the average number of translations is almost five times that of the original dictionaries.

We do not consider this a serious problem. In a machine translation system, most of the time, only the first translation is output. Therefore, as long as our ranking is correct, the spurious translations will be invisible to the user. Another important reason is that it is far quicker to delete a spurious entry than add a new one. Lexicographers prefer to be presented with a large list to be whittled down, rather than having to add translations from scratch.

In order to make the results more manageable, we flag the entries into three classes: *Accept* is words with a score above 1, that is they have a perfect inverse consultation score or matched through Chinese or semantics. *First* is the first ranked entry for those words with no acceptable translation: in that case we want to use it anyway. *Rest* are the remainder of the entries, we expect them to include many erroneous entries. However, they may also include good entries, so we flag them rather than deleting them. For machine translation with the uncorrected dictionaries, we would use a prototype lexicon made up of *accept* and *first* giving 75,932 pairs for 50,034 entries. The effects of the normalization are relatively small. There were around 1,200 new entries created by the normalization, roughly 0.5% of the total. Most were from number normalization (660), and equal numbers from US/GB spelling and case (270 each). However, the normalization itself is cheap, so it is worth doing. In particular, without checking for British/American spelling there would be a strange gap in the coverage.

3.2 Evaluation

We conducted two evaluations: a lexical sample of nouns and a comparison with existing lexicons. We also did a small check of those pairs which matched through both English and Chinese.

8,006 pairs matched using both English and Chinese as the intermediate language. We checked a sample of 100 pairs and found 84 good translations, 13 acceptable translations and three errors: 97% were good. This shows clearly that matching through two languages improves accuracy, as predicted. Thenumber of pairs is greatly reduced: only 8,006 out of 342,116. However, these still cover almost one in six of the 50,034 Japanese index words matched.

3.2.1 Evaluation by lexical sample

65 Japanese nouns were randomly selected for evaluation. They had 232 translations in all. 65% of translations were useful (good or acceptable). The results are summarized in Table 2. Concentrating only on the highest ranked translation (the translation most likely to be used), 80% of the translations were useful.

Ninety-three (40%) of the translations were judged to be good translations, usable in any context. 58 (25%) were judged to be usable in some contexts, and thus acceptable as dictionary entries, but not ideal as translation equivalents. 81 (35%) were judged to be inappropriate translations. Of these, just over a third (28) were due to errors in KAMI, the Malay-English dictionary. If the dictionary were perfect, the results would be around 77%.

The ranking successfully increased the percentage of good pairs to 46%, and acceptable pairs to 34%. This means that 80% of the translations provided by the machine translation system will be good, even with no manual revision.

Twenty-four of the entries had a single equivalence candidate (that is there was a single Japanese-English pair matching a single English-Malay pair with the same part of speech). In this case, 11 (46%) were good, 12 (50%) were acceptable, and only one was bad (due to an error in the ME lexicon). In applications which want to avoid any erroneous translations, one strategy would be to only take such single matches.

3.2.2 Evaluation by comparison to existing lexicons

We also compared our results to one of the existing Japanese-Malay Lexicons: the KAMUS Japun-Malaysia-Ingerris (henceforth JMI: Nagata 1994). We took the second word on each page, following cross references but ignoring numbers and phrases. This gave us a sample of 346 entries. We then looked up these entries in Kamus Makna: Jepun-Melayu (henceforth JM: Kasim and Jambi 1999), a similarly sized lexicon. There was surprisingly little overlap: only 138 entries were found in both, and of these only 74 had exactly the same translation, less than one in four. Around a quarter of the differences were due to variation in citation form. For example, as the translation of 閉める *shimeru "close"*, JMI had the root *tutup*

Table 2 Results (all pairs)

Evaluation	All pairs		Highest rank	
	Number	Percentage	Number	Percentage
Good translation	93	40.1	30	46.2
Acceptable translation	58	25.0	22	33.8
Bad (error in ME dic)	28	12.1	6	9.2
Bad (link mismatch)	53	22.8	7	10.8
Total	232	100.0	65	100.0

"close" where JM had *menutup "close"*, an inflected form. In comparison, our newly created lexicon found 80% of the entries, with 43% getting the same translation as in JMI. We analysed the remaining 37%, looking only at those ranked *accept* or *first* and found 69% of them were good translations (with 5% better than in JMI!), 18% were good translations for a specialized sense and only 13% were bad translations. Therefore, we can claim with confidence that our lexicon has better cover than existing published lexicons and a high precision. One of the arguments against transfer-based machine translation systems has been that it is hard to add new language pairs. However, as we show here, new pairs can be effectively bootstrapped from existing resources.

4 Conclusion

By using all the information we could, we have been able to automatically build a reasonably accurate large-scale dictionary Japanese-Malay dictionary, useful not only for humans, but with the information required by a semantic transfer-based machine translation system. This shows that information intended for one purpose (semantic classes in **ALT-J/E** and CICC, classifiers in KAMI) is often useful for other tasks (in this case linking lexicons). While creating rich lexical resources is expensive, they are useful in many different tasks.

Acknowledgments We thank Chooi-Ling Goh for help in the final evaluation.

References

Atkinson, K. Kevins word list page. http://www.wordlist.sourceforge.net/. Accessed 2 Jan 2003.

Bond, F., Sulong, R. B., Yamazaki, T., & Ogura, K. (2001). Design and construction of a machine-tractable Japanese-Malay dictionary. In *MT Summit VIII* (pp. 53–58). Santiago de Compostela, Spain.

CICC (1994). Research on Indonesian dictionary. Technical report 6—CICC—MT53, Center of the International Cooperation for Computerization, Tokyo.

Conway, D. (2000). Lingua-EN-Inflect. Perl Module (Vo1. 86). (cpan.org).

Ikehara, S., Shirai, S., Yokoo, A., & Nakaiwa, H. (1991). Toward an MT system without pre-editing – Effects of new methods in ALT-J/E —. In *Third machine translation summit: MT summit III* (pp.101–106). Washington DC.

Kasim, Z. A., & Jambi, J. (1999). *Kamus Makna: Jepun-Melayu*. Pernebit Universit Malaya.

Nagata, H. (Ed.) (1994). *Japanese-Malay-English dictionary*. Tokyo: TK Kenkyusha.

Ogura, K., Bond, F., & Ooyama, Y. (1999). ALT-J/M: A prototype Japanese-to-Malay translation system. In *Machine translation summit VII* (pp. 444–448). Singapore.

Paik, K., Bond, F., & Shirai, S. (2001). Using multiple pivots to align Korean and Japanese lexical resources. In *Workshop on Language Resources in Asia* (pp. 63–70). Tokyo.

Quah, C. K., Bond, F., & Yamazaki, T. (2001). Design and construction of a machine-tractable Malay-English Lexicon. In *Asialex 2001 Proceedings* (pp. 200–205). Seoul.

Shirai, S., & Yamamoto, K. (2001). Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001* (pp. 174–179). Seoul.

Shogakukan, & Shomoinshokan, P. (Eds.) (1987). *Ri-Zhong Cidian [Japanese-Chinese Dictionary]*. Shogakukan.

Tanaka, K., Umemura, K., & Iwasaki, H. (1998). Construction of a bilingual dictionary intermediated by a third language. *Transactions of the Information Processing Society of Japan, 39*(6), 1915–1924 (in Japanese).

Zhang, Y., Ma, Q., & Isahara, H. (2005). Automatic construction of a Japanese-Chinese translation dictionary using English as an intermediary. *Journal of Natural Language Processing, 12*(2), 63–85 (in Japanese).