Design and Construction of a machine-tractable Japanese-Malay Lexicon

Francis Bond,* Takefumi Yamazaki,‡ Ruhaida Binti Sulong‡ and Kentaro Ogura*
* **NTT Communication Science Laboratories** <{bond, ogura}@cslab.kecl.ntt.co.jp>
‡ **NTT MSC** <{yamazaki,ruhaida}@nttmsc.com.my>

Abstract

We present a method for combining two bilingual lexicons to make a third, using one language as a pivot. In this case we combine a Japanese-English lexicon with a Malay-English lexicon, to produce a Japanese-Malay lexicon suitable for use in a machine translation system. Our method differs from previous methods in its use of semantic classes to rank translation equivalents: word pairs with compatible semantic classes are preferred to those with dissimilar classes. We have made a prototype lexicon of over 75,000 pairs, and are now in the process of removing inappropriate entries by hand.

1 Introduction

We present a method for combining two bilingual lexicons to make a third, using one language as a pivot. The aim of our research is to create a lexicon to be used in the machine translation system **ALT-J/M**: the Automatic Language Translator — Japanese-to-Malay (Ogura *et al.* 1999). We take the Japanese-to-English lexicon used in the machine translation system **ALT-J/E** (Ikehara *et al.* 1991) and cross it with a Malay-English lexicon to produce a Japanese-Malay lexicon.

The **ALT** systems are semantic transfer systems, and rely on having nouns marked with appropriate semantic classes (from our 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, バス *basu* "bus" ⟨`vehicle`⟩ vs バス *basu* "bath" ⟨`furniture`⟩. 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) are harder to distinguish using semantic classes. 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.

For example *dove* and *pigeon* are both potential translations of Japanese 鳩 *hato*, with the same basic meaning.[1] In the absence of other information, **ALT-J/E** will always translate 鳩 *hato* as *pigeon* because it is the preferred translation.

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, and indeed often impossible: different languages make different distinctions. Because of this, when building our lexicon, it is essential to distinguish homonyms correctly, and our method aims to do this.

1.1 Previous work

Tanaka and Umemura (1994) and 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 (ECs) 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". This can be extended further, by looking up all the Japanese translations of all the English translations of a given French word and seeing how many times the Japanese word appears; this is "two times inverse consultation".

Matching published Japanese-English and English-French dictionaries against each other, Tanaka *et al.* (1998) were able to find translation equivalents not found in equivalent Japanese-French

[1] *Pigeon*: "wild and domesticated birds having a heavy body and short legs" ; *dove* "any of numerous small pigeons" (WordNet 1997).

dictionaries. 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.

Lafourcade (1997) also used English as an intermediate language, in his case to construct a multi-lingual French-English/Malay/Thai on-line lexicon, aimed at human users.[2] Malay and Thai entries were added to an existing French-English dictionary by linking entries in Malay-English and Thai-English dictionaries. There was no automatic filtering of the crossed results; instead emphasis was placed on producing a productive environment for human dictionary editors. In particular, human editors were found to prefer word-processor based environments to data-base interfaces.

One shared characteristic of our work with the earlier work 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.

2 Crossing the lexicons

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

2.1 The Japanese-English lexicon: Goi-Taikei

For the Japanese-English lexicon, we are using the lexicons developed for the machine translation system **ALT-J/E** (Ikehara *et al.* 1991), a subset of which has been published as Goi-Taikei (**GT**) — Japanese lexicon (Ikehara *et al.* 1997).

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.

2.1.1 Ontology

GT's ontology classifies concepts to use in expressing relationships between words. The meanings of common nouns are given in terms of a semantic hierarchy of 2,710 nodes. Each node represents a semantic class. Edges in the hierarchy represent **is-a** or **has-a** relationships, so that the child of a semantic class related by an **is-a** relation is subsumed by it. For example, `nation` **is-a** `organization`.

2.1.2 Semantic Transfer Dictionary

The semantic transfer dictionary includes roughly 380,000 Japanese-English word-pairs.

Each record specifies an index form (Japanese), translation (English), preference ranking, English syntactic information and a set of semantic classes. Optionally there may be more detailed selectional restrictions, domain and genre information and so on.

In the noun dictionary, there are 63,926 Japanese index words. 90% have only one translation, 8.5% have two, 2% have three. The maximum number of translations is 12, the average is 1.12, for a total of 71,818 Japanese-English pairs. There is a tendency for many Japanese words to be translated into the same English translation, there are only 49,205 different English entries (many of them are multi-word expressions).

2.1.3 Semantic Structure Dictionary

The basic structure of a clause comes from the relationship between the main verb and nouns. **GT**'s structure transfer dictionary, designed for machine translation applications, provides this basic clause structure. **GT** has over 15,000 patterns.

2.2 The Malay-English lexicon

The Malay-English lexicon we used is one that we compiled ourselves, based on a dictionary produced by a translation company (Quah *et al.* 2001). The lexicon currently has 67,658 Malay words with English translations. 79% have only one translation, 14% have two, 4.1% have three; the average number of translations is 1.35, giving 91,426 Malay-English pairs.

Each entry in the lexicon consists of the following fields: (1) Malay index word; (2) Malay root word; (3) Malay part of speech (POS); (4) detailed syntactic features; (5) semantic classes; (6) English translation; (7) English gloss and comments; (8) Chinese translation. Detailed on-line documentation (in HTML) has been prepared for all fields. All entries have values for fields 1,2 and 3; most have syntactic features. Only 30.4% have semantic classes using the **GT** ontology, 25% have Chinese translations. We also use a variety of meta-codes, to show other relevant information such as honorific use, origin, and register. English and Chinese translations and comments are provided for use in a machine translation system, as well as an aid for non-Malay speakers.

When entering the index words, the wide variety of spelling variation in Malay was particularly problematic. To deal with this we have allowed a single record to have multiple index forms, with the preferred form, as judged by native speakers of Riau (standard) Malay, listed first. There are currently 1,039 such entries in our dictionary, for example *hasab;hisab* "calculation".

Semantic classes were entered in four ways: (1) the original dictionary we purchased had some syntactic-semantic codes. These were mapped to the **GT** semantic classes. (2) The CICC Indonesian dictionary has semantic classifications (CICC 1994a). 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 there. These were then mapped to the **GT** semantic classes. (3) For nouns, we listed the numeral classifier (or classifiers) most commonly used to count it, following the CICC Malay dictionary (CICC 1994b). At present these are marked for 18,303 nouns. Because some classifiers select for the meanings of their targets (Bond and Paik 2000), we could use the classifiers to

[2] An on-line version of the FEM dictionary can be found at `http://www-clips.imag.fr/geta/services/fem/`.

predict the semantic class of their targets. For example, anything counted by *orang* is `human`, anything counted by *ekor* is animal, anything counted by *pokok* is a plant and so on. Shape classifiers (such as *batang* "long thing") could not be used for this, as they select for physical shape, not semantic class. Finally, (4) we added semantic classes by hand.

2.3 Crossing

The crossing of the lexicons proceeds as follows:

- For each pair in the Japanese-English lexicon
  - Look up the Malay equivalent of the English if an entry with the same POS exists
    - create a Japanese-Malay pair
    - store the intermediate English
    - Calculate a one time inverse consultation score
    - Calculate a semantic matching score
  - else mark the Japanese-English pair
- For each Japanese index in the Japanese-English lexicon
  - Output any Japanese-Malay pairs ranked by total score
  - Output marked Japanese-English pairs ranked by preference

The result is a dictionary with as many Japanese-Malay entries as possible, followed by English entries as a last resort. We deliberately kept the English entries, both as a guide to the lexicographers to identify possibly missing senses; and as default translations: most Malaysians speak more English than Japanese, so it is better to translate to English than to leave unknown words in Japanese.

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, auxiliary, preposition`). This cut down greatly on the number of false matches.

The scores were calculated as follows: The one time inverse consultation score for Japanese word J and Malay word M is given in Equation (1), where $E(W)$ is the set of English translations of W:

$$\text{inverse consultation score} = \frac{2 \times (|E(J) \cap E(M)|)}{|E(J)| + |E(M)|} \quad (1)$$

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 one semantic class subsumes the other, or visa versa. For example, `animal` is compatible with `living-thing`.

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 semantic matches come first, followed by high ranked pairs; within the same ranking, pairs are ordered by one time inverse consultation score. There is no mechanism in our algorithm for deleting candidates, that is left to the lexicographers.

Consider the following simplified example.

- Japanese-English pair (Input)

Japanese	あざらし *azarashi*
English	*seal*
Sem Classes	⟨`animal`⟩

- Malay-English pairs (Input)

Malay	*anjing laut*
English	*seal*
Classifier	*ekor*
Sem Classes	⟨`animal`⟩

Malay	*tera*
English	*seal*
Sem Classes	⟨`stationary`⟩

Malay	*mohor*
English	*seal*
Classifier	*buah*
Sem Classes	⟨`tool`⟩

- Japanese-Malay pairs (Output)

Japanese	あざらし *azarashi*
Malay	*anjing laut*
Rank	1
English	*seal*
Sem Classes	⟨`animal`⟩

Japanese	あざらし *azarashi*
Malay	*tera*
Rank	2
English	*seal*
Sem Classes	⟨`---`⟩

Japanese	あざらし *azarashi*
Malay	*mohor*
Rank	3
English	*seal*
Sem Classes	⟨`---`⟩

In this small example, there are three potential translations for あざらし *azarashi* "seal". English *seal* is homonymous, the correct sense here is the "marine animal" sense, which corresponds to *anjing laut*. The translations *tera* and *mohor* are variations of the sense that means "a stamp used to authenticate documents".

The semantic class of *anjing laut* matches with あざらし *azarashi* "seal", so it is listed first: it is the only correct translation. The other two translations are listed according to their one time inverse consultation scores.

3 Results and Discussion

In this section we will report on crossing the Japanese-English common-noun dictionary with the Malay-English dictionary.

22,658 out of 63,926 Japanese words were linked to 16,974 Malay words. There were 32.7% with one translation, 19.5% with two and 11% with three. The average number of translations was 3.4 for a total of 75,872 pairs. Clearly, we have introduced many spurious translations: the average number of translations is almost triple that of the original dictionaries.

However, we do not consider this a serious problem for the following reasons. The main reason is that, most of the time, only the first translation is output by the machine translation system. 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. Our lexicographers prefer to be presented with a large list to be whittled down, rather than having to add translations from scratch.

A preliminary evaluation of 66 randomly selected Japanese index words with 222 translations gave the following result: 67% of translations were acceptable. Concentrating only on the highest ranked translation (the translation most likely to be used), 77% of the translations were acceptable.

We are now concentrating on improving the environment for our lexicographers. As the dictionary has grown to several tens of thousands of entries, many of them will be unfamiliar, even to an educated native speaker. It is thus useful to make it easy to look up monolingual dictionaries with as few key strokes as possible, allow browsing of the semantic classes, and present examples of words in context.

3.1 Further Work

Our Malay-English lexicon also has Chinese entries for 21,190 of its entries. We plan to use Chinese as a second pivot (Japanese-Chinese-Malay). We assume that anything that matches through two languages (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 would also like to extend the number of matches by adding a British/American spelling converter. Our Malay-English dictionary uses mainly British spelling, but our Japanese-English dictionary uses mainly American spelling, so currently words such as *armor/armour* don't match.

4 Conclusion

Our method differs from previous methods in its use of semantic classes to rank translation equivalents: word pairs with close semantic classes are preferred to those with dissimilar classes. We have made a prototype lexicon of over 75,000 words, and are now in the process of removing inappropriate entries by hand.[3]

References

Bond, Francis, and Kyonghee Paik. 2000. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, 90–96, Saarbrücken.

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

CICC. 1994b. Research on Malaysian dictionary. Technical Report 6—CICC—MT54, Center of the International Cooperation for Computerization, Tokyo.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Tokyo: Iwanami Shoten. 5 volumes/CDROM.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, 101–106, Washington DC. (http://xxx.lanl.gov/abs/cmp-lg/9510008).

Lafourcade, Mathieu. 1997. Multilingual dictionary construction and services. In *3rd Pacific Association for Computational Linguistics Conference: PACLING-97*, 173–181. Meisei University, Tokyo.

Ogura, Kentaro, Francis Bond, and Yoshifumi Ooyama. 1999. **ALT-J/M**: A prototype Japanese-to-Malay translation system. In *Machine Translation Summit VII*, 444–448, Singapore.

Quah, Chiew Kin, Francis Bond, and Takefumi Yamazaki. 2001. Design and construction of a machine-tractable Malay-English lexicon. In *Asialex-2001*, Seoul. (to appear).

Tanaka, Kumiko, and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *15th International Conference on Computational Linguistics: COLING-94*, 297–303, Kyoto. (http://xxx.lanl.gov/abs/cmp-lg/9410020).

——, ——, and Hideya Iwasaki. 1998. Construction of a bilingual dictionary intermediated by a third language. *Transactions of the Information Processing Society of Japan* 39.1915–1924. (in Japanese).

WordNet, 1997. *WordNet - a Lexical Database for English*. Cognitive Science Laboratory, Princeton University, 221 Nassau St., Princeton, NJ 08542. Version 1.6, http://www.cogsci.princeton.edu/~wn/.

[3] Slightly different versions of the dictionaries discussed in this paper are on-line at http://sangenjaya.arc.net.my/.