Creating a free digital Japanese-Swedish lexicon

Jonas Sjöbergh
KTH Nada
SE-100 44 Stockholm, Sweden
jsh@nada.kth.se

Abstract

This paper describes the creation of a Japanese-Swedish dictionary. The well known technique of using English as an intermediate language was used. A few modifications of this method are presented, such as weighting words with an idf-like measure and allowing one Japanese word to be described by several Swedish words when there is no directly corresponding translation available. For about 60,000 entries good translations are found. The lexicon has been made freely available.

1 Introduction

There are many papers describing different methods to build a bilingual dictionary automatically. While automatic methods often have drawbacks, such as including noise in the form of erroneous translations of some words, they are still popular because the alternative, manually constructing a dictionary, is very time consuming. Automatic methods are often used to generate a first noisy dictionary which can then be cleaned up and extended by manual work.

Many methods to generate a bilingual dictionary from parallel corpora have been presented. Other approaches use existing bilingual dictionaries from the source and target language to some common intermediate language. Usually, English is used as the "interlingua", since bilingual dictionaries exist between many languages and English. This is the approach we used, since there is no available parallel corpus of Japanese and Swedish, but there are large available Japanese-English and Swedish-English dictionaries available (but no available Japanese-Swedish dictionary).

Some problems surface when using two dictionaries to build a new one. Many English words are ambiguous, which can lead to erroneous translations in the new dictionary. A similar problem is English translations with a wider meaning than the original word. Paraphrasing is another problem. The same meaning is often described in very different ways by different lexicographers, so even though two translations are both in English it can be hard to automatically match them. Is a small difference in translation indicative of a difference in nuance or is it just different lexicographers describing the same thing? This can lead to many "missing" translations in the new dictionary. Another problem with the same effect is that many words in the source language do not have directly corresponding words in the target language. The same meaning would instead be described using several words.

Previous work on automatic bilingual lexicon creation similar to our own include the work by Tanaka and Umemura (1994), Shirai et al. (2001) and Shirai and Yamamoto (2001). The problem of ambiguity can be mitigated by using several intermediate languages as done by Paik et al. (2001) and Bond et al. (2001), who also use part-of-speech and semantic categories for this. Hopefully the different intermediate languages will not be ambiguous in the same way. Paik et al. (2004) describe the impact of using bilingual dictionaries in different directions.

In short, our method works by matching all English descriptions of Japanese words to all English descriptions of Swedish words. Matches are basically word overlap, and the best matches are selected

as translation candidates. The main focus was on creating a dictionary with very large coverage, possibly at the cost of including less than ideal translations. Two new ideas helped in this regard: weighting words with a measure similar to idf (Inverse Document Frequency), useful when ranking several poor translation candidates; and allowing one source language word to be translated by a combination of two target language words, which gives many new translations.

We have made the 16,000 Japanese words with the most reliable translation candidates freely available on the Internet, at `http://www.japanska.se/`, where it is possible to search the dictionary and also correct the remaining errors or add new words. We plan to extend this with other parts of the resulting dictionary in the future.

2 Creating the dictionary

We used the Japanese-English dictionary EDICT (Breen, 1995), which is freely available for personal use. It contains about 110,000 Japanese index terms. The Swedish-English dictionary used contains about 160,000 Swedish index terms.

From the English descriptions we removed a few stop words, such as "the" or "an", and all words with only one letter. All characters that were neither letters, numbers or the characters ' or - were removed.

All remaining words had a weight calculated. This was basically the inverse document frequency used for instance in information retrieval, and will thus be called idf here.

$$idf(w) = \log(\frac{|S| + |J|}{S_w + J_w}) \qquad (1)$$

where w is the word we calculate the weight for, $|S|$ is the total number of dictionary entries in the Swedish-English dictionary, $|J|$ the same for Japanese, S_w is the number of descriptions in the Swedish-English dictionary this word occurs in and J_w similarly for Japanese.

Then all English descriptions in the Japanese-English dictionary were matched to all descriptions in the Swedish-English dictionary. Matches were scored by word overlap, weighted by the idf of the words. A word was only counted once, even if it occurred many times in the same description. So as not to give longer descriptions an unfair advantage the score was normalized by the lengths of the descriptions.

$$score = \frac{2\sum_{w \in S \cap J} idf(w)}{\sum_{w \in S} idf(w) + \sum_{w \in J} idf(w)} \qquad (2)$$

where J is the text in the Japanese-English lexicon and S is the text in the Swedish-English lexicon that we are trying to match it to.

There are quite a few words in the Japanese-English dictionary with no direct correspondence in the Swedish-English dictionary, or sometimes even in the Swedish language. These can often be described using two Swedish words though.

One example is "perpetual motion". There is no Swedish word with this meaning listed in the dictionary (though there is a similar word in Swedish). There are however words for "motion" and "perpetual" in the dictionary. Combining these two Swedish words gives a very good description of the meaning of the Japanese word.

To find this type of description all pairs of words were also treated as one word, with the translation being the concatenation of the respective descriptions. To favor a directly corresponding Swedish word, if there was one, over a combined description all such pairs had their matching score lowered by 5%.

When all matching descriptions had been found the translation candidates were ranked according to the score of their descriptions. The highest scoring Swedish word is hopefully the best translation. Of course this was not always the case, sometimes the best translation was not ranked as number one, and sometimes there was no correct translation available in the Swedish-English dictionary but other words partly match and were suggested instead, but in general the ranking worked well.

Our focus was on creating a dictionary with very large coverage. Preferably with high translation quality, but if the choice was between a poor but at least somewhat helpful translation and no translation we would prefer the poor translation. The two new contributions in our method both help in this regard. First, weighting by idf tends to give the best suggestion of several poor suggestions when no good suggestions are available. Second, allowing one word

to be matched by a combination of two words drastically increases the number of useful translations.

3 Evaluating the dictionary

The resulting dictionary was evaluated by randomly drawing words and classifying them into five categories, depending on the translation quality. This is a quite common way to evaluate automatically created bilingual dictionaries, though the classification is often quite coarse, for instance "good translation", "acceptable translation", or "bad translation".

Our first evaluation category is the best and most common case; that all top scoring suggested Swedish translations for the Japanese word are correct.

It is common to find many translation suggestions with the same score. If not all are correct but more are correct than incorrect a Swedish reader will still be able to understand what the word means. This is the second category.

The third category, that only a minority of the suggestions are correct, is still useful. A Swedish reader will (probably) understand the correct meaning in context, since it is (hopefully) the most likely of the suggested meanings in the text the reader is reading. It is also useful when manually improving the dictionary; since the correct translation is available the lexicographer only has to remove the bad translations.

Something that is quite common is suggestions that are not correct, but very similar to the correct translation, such as "broadcasting (usually radio or TV)" as the suggestion for "webcast / Internet broadcast" or "blue" as the suggestion for "light blue". While these translations are not correct they are helpful enough that the general meaning of a text is usually clear even with these erroneous suggestions, so they have their own category.

Finally, the last category is for when the suggestions are just plain wrong.

The evaluation was mainly performed by the author, a native speaker of Swedish, with some knowledge of Japanese and good knowledge of English. When evaluating translations the Japanese word, the candidate Swedish translations and the original English translation of the Japanese word was presented.

Another native speaker of Swedish, also with some knowledge of Japanese and good knowledge of English, also independently classified a subset of the evaluated words (300 words). This was done to see if there was large agreement in classification or bias from the author in the evaluations. Both evaluators agreed on almost all words, though in the few cases that differed it was usually the author that was more forgiving of the translations.

Type	Words	%
All correct	353	50
Majority correct	78	11
Some correct	107	15
Similar	116	17
Wrong	46	7

Table 1: Translation quality of 700 randomly selected words with $score \geq 0.2$. There are 104,439 words in this category.

Another way of evaluating bilingual dictionaries that has been used by others is to select translation pairs from some other dictionary and see how many of these are correctly matched in the new dictionary. Since the largest Japanese-Swedish dictionary we had available was smaller than our randomly selected sets of words we did not use this evaluation method.

4 Results

Of the 110,000 Japanese index terms in EDICT, 104,000 had a matching description from the Swedish-English dictionary with a score of at least 20%. Of these, about 75% had at least one correct translation among the top ranked suggestions, see Table 1. If we use a threshold on the overlap score the quality of the translations of the remaining words is high, but of course many correct translations are also removed. With a threshold of 90% overlap well over 90% of the 28,000 remaining words have a correct translation among the top ranked suggestions, see Table 2.

The scoring is generally quite good. When there is a correct translation available in the Swedish-English dictionary it is usually the suggestion with the highest score. When there is no correct translation available, available words similar in meaning, such as hyponyms, will normally have higher score

Type	Words	%
All correct	522	75
Majority correct	83	12
Some correct	59	8
Similar	24	3
Wrong	12	2

Table 2: Translation quality of 700 randomly selected words with $score \geq 0.9$ and at most 10 suggestions with top score. There are 28,178 words in this category.

Type	Words	%
All correct	622	89
Majority correct	38	5
Some correct	21	3
Similar	9	1
Wrong	10	1

Table 3: Translation quality of 700 randomly selected words with $score = 1$. There are 16,843 words in this category.

than unrelated words.

The idf helps in giving good ranking among suggestions, especially for words with longer descriptions in English. These have many translation candidates, since there are many words in their descriptions that can match the description of a Swedish word. The idf orders these matches so that suggestions matching the more important words are preferred over matches on for instance prepositions. The idf also allows the stop word list to be very short, since words which should be stop words but are not included in the list will tend to have a very low idf and thus not have a great impact on the matching.

Having a good ranking is very helpful when manually cleaning up the dictionary. This allows the inclusion of words with only very weak matches as suggestions for the lexicographer, which otherwise would perhaps be thought of as too noisy, but still includes many words with correct translations.

Allowing word pairs as translations increases the number of correct translations drastically. The EDICT includes many words which have no direct translation in Swedish, at least not one that is available in the other dictionary. The coverage thus would be very low using just a one to one matching of the index terms from the two dictionaries. Of course, there are also some words and expressions in EDICT that would require more than two of the available Swedish index terms.

Since it is generally better to have a match on one entry in the Swedish-English dictionary than on two, the ranking score of pairs was reduced by 5%. During the evaluation it was found that it might be better to reduce them even further, perhaps as much as 25%. Examples where this would be better include many colors, such as "light green" which is translated as "light" + "green", with perfect overlap from a pair of Swedish words. While this is quite good it is not as good as the Swedish word for light green, which is available. The reason this does not rank higher is that the Swedish word is translated as "light or pale green", thus only scoring 76% overlap. Then again, "heavyweight" also scores 76% as a translation for "light heavyweight" and would thus replace the current translation "light" + "heavyweight", but a better value than 5% could likely be found.

Finally, here is a simple example of the impact of the ranking methods: The word *"horoyoi"* is translated as "slightly drunk, tipsy" in the Japanese-English dictionary. Since no Swedish word has this exact translation, there are only partial matches. The top scoring matches are all Swedish words for drunk or tipsy, ranked as 52% overlap (matching "tipsy"). Next comes the Swedish word for "slightly", with 50% overlap. This is followed by more Swedish words matching "drunk". When allowing pairs of words to match one word, the top suggestions all consist of "slightly" and different words for "drunk", with an overlap of 76%.

One future possible improvement is checking the word class of suggestions, mostly disambiguating between the noun and verb sense of many English words. Many erroneous translations include the related verb form for a noun and vice versa. Another problem is that the Japanese-English dictionary uses American spelling (e.g. "honor") while the Swedish-English dictionary uses British spelling (e.g. "honour"). Harmonizing the spelling would give better translations, since currently some words that should match will not be considered equal.

5 Conclusions

Our method produces a large dictionary, with some noise but generally the quality is good. It is possible to retrieve from 100,000 words with quite a lot of noise to 16,000 words with very little noise, or something in between.

We would also have liked to use other intermediate languages to improve the quality of the translations, but there was no other language with sufficiently large dictionaries available to us. Mainly this bottleneck was on the Swedish side; for Japanese there are other quite large available dictionaries. Future plans include using the smaller available dictionaries for other languages to improve the quality for the covered vocabulary.

The highest quality translations have already been made available on the Internet, at `http://www.japanska.se`, and other parts of the results are available on request.

Acknowledgments

We thank Viggo Kann for contributing useful ideas and helpful suggestions.

This work has been funded by The Swedish Research Council (VR).

References

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proceedings of MT Summit VIII*, pages 53–58, Santiago de Compostela, Spain.

Jim Breen. 1995. Building an electronic Japanese-English dictionary. In *Japanese Studies Association of Australia Conference*, Brisbane, Australia.

Kyonghee Paik, Francis Bond, and Shirai Satoshi. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *Proceedings of NLPRS-2001*, pages 63–70, Tokyo, Japan.

Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. 2004. Automatic construction of a transfer dictionary considering directionality. In *Proceedings of MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources*, pages 31–38, Geneva, Switzerland.

Satoshi Shirai and Kazuhide Yamamoto. 2001. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of ICCPOL-2001*, pages 174–179, Seoul, Korea.

Satoshi Shirai, Kazuhide Yamamoto, and Kyonghee Paik. 2001. Overlapping constraints of two step selection to generate a transfer dictionary. In *Proceedings of ICSP-2001*, pages 731–736, Taejon, Korea.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of COLING-94*, pages 297–303, Kyoto, Japan.