

What Types of Translations Hide in Wikipedia?

Jonas Sjöbergh¹, Olof Sjöbergh², and Kenji Araki¹

¹ Graduate School of Information Science and Technology
Hokkaido University

{js, araki}@media.eng.hokudai.ac.jp

² KTH CSC

olofsj@kth.se

Abstract. We extend an automatically generated bilingual Japanese-Swedish dictionary with new translations, automatically discovered from the multi-lingual online encyclopedia Wikipedia. Over 50,000 translations, most of which are not present in the original dictionary, are generated, with very high translation quality. We analyze what types of translations can be generated by this simple method. The majority of the words are proper nouns, and other types of (usually) uninteresting translations are also generated. Not counting the less interesting words, about 15,000 new translations are still found. Checking against logs of search queries from the old dictionary shows that the new translations would significantly reduce the number of searches with no matching translation.

1 Introduction

In the increasingly interconnected world, communication with people from different places all over the world is becoming more and more important. Since there are many languages being used in different parts of the world, tools that aid in communication between speakers of different languages, or in reading material produced in a language not one's own etc. can be very helpful. This is of course not a new problem, people have been dealing with the fact that it is often useful to communicate with people using other languages for a very long time, though the need is more prevalent in the modern globalized world. Many different tools have been developed, such as machine translation systems to automatically translate from one language to another, or computer programs that aid professional translators in their work.

One basic tool that has been used a lot is the bilingual dictionary. In its most basic form, a list of words in one language and their corresponding translations in another language. Such dictionaries can be useful in many situations, for instance as a reference for when not finding the appropriate word in conversation, as a support when reading in a foreign language, as one tool among many in a computer system for translation etc.

There are many ways to create bilingual lexicons. Traditionally they have been created by hand. A linguist assembles a set of words and their translations. This usually gives very high quality dictionaries, with high translation quality

and containing the most important or useful words. It is however also very time consuming, and thus expensive.

Since bilingual dictionaries are so useful and manual creation is expensive, different ways to create them by automatic means have been devised. While automatic methods normally have drawbacks such as including erroneous translations or missing many important words, they are still useful because of their huge time saving potential. Automatic methods can also be used to generate a first rough draft of a dictionary, that is later cleaned up and extended by manual work.

One useful resource for creating bilingual dictionaries is large bilingual corpora, i.e. collections of text with the same text written in two languages. Depending on how detailed the information in the bilingual corpus is different methods to extract bilingual dictionaries can be used. The granularity could be which document corresponds to which document in the other language, or which sentence or even word. For good results, the bilingual corpus needs to be fairly large. Most large available parallel corpora are government texts, which gives a somewhat slanted dictionary. With a lack of appropriate parallel corpora, methods have also been developed that use “comparable corpora”, i.e. texts that cover similar topics in different languages (such as international newspaper articles from roughly the same time in different countries), or even monolingual corpora for each language. An example where methods to create bilingual dictionaries from parallel corpora is compared to similar methods using only monolingual or comparable corpora can be found in [1].

Bilingual dictionaries have also been created by using other bilingual dictionaries. Given for instance a Swedish-English dictionary and a Japanese-English dictionary, a Japanese-Swedish dictionary can be generated by using the English translations as a pivot. Words with similar English translations have a high probability of having the same meaning. Ambiguous English words naturally cause problems, as does the fact that different English words can be used to describe essentially the same thing. Using English as a pivot language is common, since there are large dictionaries available between English and many other languages. If dictionaries are available with other possible pivot languages too, this can be used to achieve a higher quality result. The same ambiguity occurring in all pivot languages is unlikely, if the languages are not too closely related. For languages where large semantic lexicons, i.e. where the different meanings of the words and their relations to the meanings of other words are available, similar approaches can be used. For instance by mapping meanings from one such resource to meanings in another, or mapping translation candidates to semantic meaning representations to find words that mean the same in two different languages.

In this paper we extend a Japanese-Swedish dictionary created by using English as an interlingua and connecting two different bilingual dictionaries [2]. We extend it by using the online encyclopedia Wikipedia³. Wikipedia consists of articles written by volunteers, and many articles are available in many different

³ <http://www.wikipedia.org>

languages. While not a parallel corpus it can at least give you comparable corpora in many languages. In our case we only use a very small set of Wikipedia that is easy to process when searching for translations. When an article on the same topic is available in another language this is indicated by a link in the source article. We automatically extract the article titles of articles available both in Japanese and Swedish. The title words used in the two languages can be assumed to be a good translation pair.

While this is a simple and useful method of extending the dictionary, the method is fairly trivial and not a very interesting research contribution. Indeed, for instance in [3] a bilingual dictionary is created in the same way, though the focus of the paper is on how to extract whole sentences that are similar in content but written in different languages (Dutch and English). Recently this method of finding translations has also been used in extending bilingual lexicons for information retrieval in multiple languages [4–6]. It has also been used in analyzing what people are currently interested in or concerned about, across different languages [7].

The main contribution of this paper is the evaluation of the generated translations. We examine what words are available to translation in this way, to see if the results are useful in our intended application.

2 Finding New Translations in Wikipedia

When extending the previously generated Japanese-Swedish dictionary we use Wikipedia, a multi-lingual freely available online encyclopedia. Wikipedia is a community project, created by volunteers from all over the world writing articles, editing other peoples writings or correcting mistakes etc. At the time of writing, there were about 2 million articles available in English (the largest language), and for Japanese and Swedish used in our experiments there were 403,000 and 245,000 articles respectively.

The new translation pairs for the dictionary were generated by checking each Swedish article in Wikipedia. If there was a link indicating that the same article was also available in Japanese, both these articles were fetched. The title was then extracted and the title words or expressions were considered as a new translation pair. We also did the same procedure using the Japanese articles, checking for links to Swedish translations. This gave essentially the same translation pairs, barring a few spelling variations.

This method is very simple, and more sophisticated methods that could find more translation pairs can be devised. Previously mentioned methods using comparable corpora are for instance applicable using the articles that are found to correspond to one another in the two languages. The method is of course also possible to use for other language pairs than Japanese and Swedish.

3 Evaluation

Using the method in the previous section, a list of 53,503 new translation pairs was generated. We evaluated the translations by manually going through a randomly selected subset of 3,000 of the translation pairs.

The first thing to be examined was the translation quality. The words were checked to see if the Japanese and Swedish expressions actually had the same meaning. This was grouped into three categories, good translations, erroneous translations and difference in scope. The last category was used for words where either the Japanese or Swedish expression had a much wider meaning than the suggested translation, but the suggested translation was a correct translation in some instances. All evaluators checking the translations were native speakers of Swedish and fairly fluent in Japanese.

The selected translation pairs were presented one pair at a time to an evaluator. The Wikipedia pages for the two translation candidates were also made available to use to determine if words the evaluator did not already know were good translations of each other. The evaluator then noted down the translation quality of the suggestion and some other information described below. The results are shown in Table 1.

Table 1. Translation quality of a sample of the generated translation pairs.

| Good Scope Wrong | | |
|------------------|----|------|
| 99% | 1% | 0.2% |

As can be seen, the translation quality is very high. There were very few words that were not correctly translated. One example was the Swedish word “hage”, which means roughly an enclosed pasture. The suggested Japanese translation was “Meadow” (i.e. written in English). This was caused by a link from the Swedish Wikipedia page that led to a Japanese page describing a software tool called Meadow, though it also explained that the word can also mean meadow in the sense of a field for cattle etc.

There are quite a few words that have a translation which sometimes overlap but has a wider meaning. Almost all such words are abbreviations.

Next, we examined what types of words are available for translation in this way, using the same randomly selected 3,000 translation pairs. Since they are taken from the title words of encyclopedia entries, one can expect mainly nouns and proper nouns. We also separated out dates and numbers, since there were many translations of the kind “1912” (Swedish) – “1912年” (Japanese), both meaning the year 1912. Such translations are generally uninteresting to include in a dictionary. The results are shown in Table 2.

As expected, most words are nouns or proper nouns. While there are many useless translations in the form of dates, they do not make up a very large part

Table 2. Distribution (%) between word types.

| Noun | Verb | Proper Noun | Date/Number | Other |
|------|------|-------------|-------------|-------|
| 33 | 0.03 | 62 | 4 | 1 |

of the total set of words. Words that were neither nouns, dates, nor proper nouns were mostly short phrases or multi-word expressions. Examples include “Do not threaten to sue” (from a page of rules for behavior on Wikipedia) and “two of a kind”. Since the majority of the words were proper names, we also examined what types of proper names occurred, see Table 3.

Table 3. Distribution (%) among proper names.

| Person | Place | Group | Title | Product | Event | Other |
|--------|-------|-------|-------|---------|-------|-------|
| 41 | 22 | 13 | 11 | 6 | 5 | 3 |

While translations of names are generally not what dictionaries are used for, there are applications where they are useful. While names of places, international organizations and famous events are often similar in Japanese and Swedish, there are also many examples of such words where the words are wildly different in the two languages. One example from the generated translations is the “United Nations”, “Förenta nationerna” (Swedish, literally “united nations”) – “国際連合” (Japanese, literally “international union”). Translations of titles of books, movies, etc. are also often quite different, but perhaps less important.

Translations of personal names seem less useful, though it could be used to find the correct transliteration in the other language. This is not always trivial, since the writing systems are very different.

Names from the “other” category include such things as names of stars (astronomical) and names that can signify for instance both organizations and people.

We also checked if the words were predominantly made up of for instance specialized terminology from specific fields, slang, or other types of words not often encountered in “normal” texts. “Normal” was taken to mean words that also fairly frequently occur outside their specialized domain (if any). These categories were quite loosely specified, and are only meant to give an overview of the general trends. Proper names and dates were not included in this classification.

As can be seen in Table 4, there are very many technical terms present. That an online encyclopedia written by interested volunteers contains many articles on for instance computer or network technology related subjects is perhaps not surprising.

Table 4. Classification of translations (%) according to style.

| “Normal” | Technology | Jargon | Species | Sports | Foreign | Word |
|----------|------------|--------|---------|--------|---------|------|
| 41 | 15 | 15 | 9 | 1 | 19 | |

Besides the technical jargon, there are also many words that are quite specific to a certain field. For instance there is medical jargon, chemistry jargon, architectural jargon etc. While there are many sports related words in Wikipedia, most of these are names of athletes, teams, or sports arenas. Sports related words that are not names are less common.

A large part of the generated translations are words that are foreign loan words in Japanese, generally coming from English. For a native speaker of Swedish, most of whom are fluent enough in English to know these loan words, such translations are not that helpful. It is useful to know when one can expect Japanese speakers to understand the English word, though, and to indicate such instances when the English word is used in Japanese with a meaning different from the original English word. It is of course also useful for speakers of Japanese who want to know the Swedish translations of such words.

Another related point is that in the Japanese Wikipedia names of flowers and animals are generally written with phonetic script in the title field. This is the usual way to write when specifying a certain species or so in a technical sense, but in a dictionary it would be more useful to have the normal (i.e. using a different alphabet) writing way. For example when writing about cherry trees in a non-technical sense, “桜” would normally be used instead of “サクラ” found in the generated translations, though they are both two written variants of the same word.

To see if the generated translations consisted mostly of new words or if there was a large overlap with the previously generated dictionary, we automatically checked what percentage of translations were found in the old list too.

Table 5. Availability of the new translations in the previously generated dictionary.

| Availability | Words | Proportion |
|-------------------|--------|------------|
| New Translations | 50,863 | 95% |
| Already Available | 2,640 | 5% |

As can be seen in Table 5, the majority of the words are new translations. This is natural since most of the translations are proper names, which the original dictionary did not contain to any large degree. However, most of the other translations are also new translations, with less than one in four of the new useful translations being available in the old dictionary.

The previously generated Japanese-Swedish dictionary is searchable on the Web⁴, where it is also possible to add new translations or to correct errors. There is also an independently developed search field plug-in for some Web browsers that directly sends translation queries to that Web interface. In our final evaluation we checked the search logs for the Web interface. We examined whether the words searched for were present among the new translations, the old translations, or both. Each query is counted, so if the same word is searched for by several different users or by the same user many times, it will be counted many times. While the Web interface is not very heavily used, at least it gives an indication of what kinds of words are interesting to users of a Japanese-Swedish dictionary. About 380,000 queries were found in the logs. The results are shown in Table 6.

Table 6. Distribution of search queries from the Web interface to the dictionary.

| Type of Query | Words Proportion | |
|--------------------------|------------------|-----|
| Available in Both | 96,558 | 26% |
| Only in Old Dictionary | 111,078 | 30% |
| Only in New Translations | 22,425 | 6% |
| Missing from Both | 142,566 | 38% |

Many words that users search for are not available in either set of translations. A very large part of these is made up from spelling mistakes and encoding problems, so they are not related to low coverage of the dictionary vocabulary. Searches using inflected forms are also common, and usually give no results. There are also (surprisingly) quite a lot of searches in English, Arabic and other languages, which naturally also fail to return results. It is clear that while the original dictionary covers more of the words users are interested in, the new translations also contribute with many sought after words that are missing.

4 Discussion

The previously generated dictionary that is used in the online search interface contains about 18,000 words. The newly generated translations numbered over 50,000, though many were not particularly useful. A rough estimate of how many new useful translations were generated can be calculated based on 33% of the new translations being interesting and about 5% of the translations already being in the old dictionary. Since the words that are in the original dictionary are almost exclusively interesting, this gives about 15,000 new interesting translations (28% of 53,000). Since the translation quality is very high and the method is very simple, this is a good method to almost double the size of the dictionary.

⁴ <http://www.japanska.se>

While many of the words searched for by users were already covered by the original dictionary, the new translations will significantly reduce the number of searched for but missing words. Reducing the failed searches from 44% to 38% is quite good, especially considering that many of the remaining failures are caused by spelling mistakes, inflected words, faulty character encodings etc. and are thus not related to lack of coverage of the dictionary.

For Japanese and Swedish, no really large electronic dictionaries are available as far as we know. There are dictionaries in printed form, with sizes ranging from 6,000 words to the order of 30,000 words. So a dictionary of the size generated can likely be quite useful, though naturally there are some important words missing, and some erroneous translations.

5 Conclusions

We generated over 50,000 new translation pairs between Japanese and Swedish from the online encyclopedia Wikipedia using a very simple method. Analyzing the generated translations we found that the translation quality is very high. Almost no actually erroneous translations were generated, though about 1% of the translations were not ideal.

There were many translations that seem fairly useless, such as translations of personal names. Even discarding all names, dates, and numbers, about 15,000 new translations were generated. Most of the generated translations were not available in the previously generated dictionary that we wanted to extend. Evaluating against the search logs from the interface to the old dictionary, a significant number of the searches that failed to return a match in the old dictionary would have resulted in a matching translation using the newly generated translations.

References

1. Koehn, P., Knight, K.: Knowledge sources for word-level translation models. In: Proceedings of EMNLP 2001, Pittsburgh, USA (2001)
2. Sjöbergh, J.: Creating a free digital Japanese-Swedish lexicon. In: Proceedings of PACLING 2005, Tokyo, Japan (2005) 296–300
3. Adafre, S.F., de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia. In: EACL 2006 Workshop on New Text – Wikis and Blogs and Other Dynamic Text Sources, Trento, Italy (2006)
4. Wang, Y.C., Lee, C.W., Tsai, R.T.H., Hsu, W.L.: IASL system for NTCIR-6 Korean-Chinese cross-language information retrieval. In: Proceedings of NTCIR-6 Workshop, Tokyo, Japan (2007)
5. Su, C.Y., Wu, S.H., Lin, T.C.: Using Wikipedia to translate OOV terms on MLIR. In: Proceedings of NTCIR-6 Workshop, Tokyo, Japan (2007)
6. Mori, T., Takahashi, K.: A method of cross-lingual question-answering based on machine translation and noun phrase translation using web documents. In: Proceedings of NTCIR-6 Workshop, Tokyo, Japan (2007)
7. Fukuhara, T., Murayama, T., Nishida, T.: Analyzing concerns of people from Weblog articles. *AI & Society* **In press** (2007)