The Impact of Part-of-Speech Filtering on Generation of a Swedish-Japanese Dictionary Using English as Pivot Language

Ingemar Hjälmstad
DSV, Stockholm University
Stockholm, Sweden
iingemar@gmail.com

Martin Hassel
DSV, Stockholm University
Stockholm, Sweden
xmartin@dsv.su.se

Maria Skeppstedt
DSV, Stockholm University
Stockholm, Sweden
mariask@dsv.su.se

Abstract

A common problem when combining two bilingual dictionaries to make a third, using one common language as a pivot language, is the emergence of false translations due to lexical ambiguity between words in the languages involved. This paper examines if the translation accuracy improves when using part-of-speech filtering of translation candidates. To examine this, two different Japanese-Swedish lexicons were created, one with part-of-speech filtering, one without. The results show 33 % less translation candidates and a higher quality lexicon when using part-of-speech filtering. It also resulted in a free lexicon of Swedish translations to 40 716 Japanese entries with a 90 % precision, and the conclusion that part-of-speech filtering is an easy way of improving the translation quality in this context.

1 Introduction

Bilingual dictionaries are specialized dictionaries used to translate words or phrases from one language to another. They aid us in understanding other languages, people and cultures – something that becomes more and more important in today's high paced Internet connected society.

The manual creation of bilingual dictionaries is very time consuming and requires hard work. Several automatic methods have been proposed to aid this work. The more common ones are statistical analysis of parallel corpora or by translating from one language to another through a common third language – the pivot language. These methods are, however, not perfect and this paper aims to investigate how to improve on these methods, in particular how part-of-speech filtering affects automatic generation of translation candidates when using a pivot language.

2 Background

An alternative to corpus based methods was proposed by Tanaka and Umemura (1994) where they showed that it is possible to automate the translation from source language to target language through a common intermediate language. Here a Japanese-English and an English-French dictionary were used to automatically generate translation candidates of Japanese-French translations. They discovered some false translations, however, and in order to filter out these *inverse consultation*—a method for assessing the quality of a translation candidate—was proposed.

One time inverse consultation is done by taking each translation candidate in the target language and translating it back to the intermediate language. These translations are then compared to the translations of the word from the source language into the intermediate language. The larger the number of common translations in the intermediate language, the better candidate. In this way it can be measured how close the meaning of the original word is to the meaning of the translation candidate. In two times inverse consultation, also proposed by Tanaka and Umemura (1994), the method is taken one step further and the translations in the intermediate language are translated back to the source language and compared to the source word.

Shirai and Yamamoto (2001) proposed a variant of Tanaka and Umemura's method where they used one time inverse consultation to create a Korean to Japanese dictionary. Here, the degree of similarity for the translation candidates was calculated using the Dice coefficient for the two sets of words in the intermediate language. One of the sets consisted of the translations of the source word from Korean to English and the other set consisted of translations back from the Japanese candidate to English.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 98–105

Bond and Ogura (2008) combined several of the above methods when creating a Japanese-Malay dictionary. They also used matching of part-of-speech in the generation of translation candidates. Pairs were only accepted if they had the same part-of-speech which, according to Bond and Ogura, gave a marked reduction of false translations. It lowered the number of translation candidates with 15 %, out of which the majority of the candidates were wrong.

For Swedish, similar work on automated generation of bilingual dictionaries have been made by Sjöbergh (2005). This approach differed from earlier work by using a measure similar to inverse document frequency, and by allowing a source language word to be translated by a combination of two target language words.

Sjöbergh finally suggested an improvement in the method by examining the part-of-speech on the suggested translation candidate, to primarily distinguish between nouns and verbs. This was something which, according to Sjöbergh, gave rise to a number of erroneous suggestions for translations. Khanaraksombat and Sjöbergh (2007) used the same method as Sjöbergh (2005) with only a few small changes, including part-of-speech matching. However, many of the words had no part-of-speech marked, and they do not report on how the part-of-speech matching affected the results.

3 Problem

In aforementioned works the most common type of problems in automatic generation of bilingual dictionary with a pivot language are due to lexical ambiguity. Examples of different forms of lexical ambiguity are internal and external homographs and ambiguity in part-of-speech, referred to as polysemy.

Part-of-speech matching has previously been suggested as a possible improvement of the method. Zhang et al. (2007), Khanaraksombat and Sjöbergh (2007) as well as Bond and Ogura (2008) have used part-of-speech matching with positive results. However, it has not been investigated fully with Swedish as one of the included languages, probably due to the absence of a dictionary in which all entries have been marked with the part-of-speech. Since the data sources used here have all entries marked with the part-of-speech, matching will be performed on all the suggestions of the translation candidates in this work.

4 Method

This survey has been carried out with Japanese as source language, English as intermediate (pivot) language and Swedish as target language. The Japanese WordNet (Isahara et al., 2008) and the People's English-Swedish dictionary (Kann and Hollman, 2011) are used for Japanese-English and English-Swedish translation, respectively. These were selected since they are the largest available dictionaries for the languages involved that also have all entries marked with part-of-speech, and that are available in digital format for free download and use. Since the lexicons use different notations for part-of-speech, e.g. the Japanese WordNet uses the abbreviation "nn" for nouns while the People's English-Swedish dictionary uses "n" for nouns, a mapping was done to a common part-of-speech notation for easier comparison in later stages.

4.1 Translation Candidate Generation

The method for generating translation candidates from Japanese to Swedish is based on the method by Tanaka and Umemura (1994) with a pivot language. Two sets of translation candidates were generated, one with part-of-speech matching and the other without.

Meta code to generate Japanese-Swedish translation candidates:

1. For each Japanese word in the Japanese WordNet, look up its English translations.
2. For each English translation, look up its Swedish translations in the People's lexicon.
3. For each Swedish translation, if it exists:
   (a) Perform part-of-speech filtering, that is compare part-of-speech in the Japanese and Swedish dictionaries and save as filtered translation candidate only if both words have the same part-of-speech.
   (b) Do not perform part-of-speech filtering, save as unfiltered translation candidate and continue with next.

This method differs from previous work in the same area by using part-of-speech filtering on all translation candidates, thus thoroughly examining the impact of the part-of-speech filtering step.

4.2 Translation Candidate Scoring

The method of scoring the automatically generated translation candidates is based on the method by Tanaka and Umemura (1994): one time inverse consultation. This method has been used successfully in a number of other works (Shirai and Yamamoto, 2001; Zhang et al., 2007; Bond and Ogura, 2008; Sjöbergh, 2005). One time inverse consultation requires an additional data source: a bilingual dictionary from the target language back to the intermediate language. For this the People's Swedish-English dictionary with 22 014 Swedish dictionary words has been used.

One time inverse consultation is carried out according to the following steps.

1. For each translation candidate, translate the word in the target language back to the intermediate language.
2. Translate the word in the source language into the intermediate language.
3. Count how many common translations there are in the intermediate language.

The more matches, the better translation candidate. To calculate the score for the proposal the formula from Shirai and Yamamoto (2001) is used. Points p for a translation candidate w are then, here, calculated using the following generalized formula, where s denotes the source language (in this case Japanese), t denotes the target language (here Swedish) and i denotes the intermediate language (here English).

$$\mathrm{p(w)} = 2 * \frac{\text{Common translations}}{\text{Translations}_{s \to i} + \text{Translations}_{t \to i}}$$

The resulting *p(w)* shows on a scale from 0 to 1 how good the proposed translation is, with 1 as the highest score.

Additional calculations of how good the translation candidate is can be done. Other methods are Sjöbergh's (2005) inverse document frequency, and Varga and Yokoyama's (2007) lexicological checks in WordNet or Bond and Ogura's (2008) matching through a second intermediate language. None of these methods have been used since they were not considered essential for the study of how part-of-speech filtering affects the outcome. One scoring is enough to show any differences in the quality of the resulting lexicon. Shirai and Yamamoto's method is also well proven (Zhang et al., 2007; Bond and Ogura, 2008).

4.3 Data

The People's English-Swedish dictionary (version 1.1) currently contains 46 762 English entries, which are carefully grouped by part-of-speech. The People's English-Swedish dictionary is freely available for use and download in XML format under the Creative Commons Attribution-ShareAlike 2.5 Generic license.[1] The Swedish-English part of the People's dictionary (version 2009-07-08) contains 22 014 Swedish entries. It has a similar breakdown of the part-of-speech groups as the English-Swedish part of the lexicon. It is, however, not yet available for free download.

The Japanese WordNet (version 0.92) is a semantic lexicon of the Japanese language. It is produced by the National Institute of Information and Communications Technology (NICT) in Japan and contains 87 133 unique Japanese entries with translations into English. All entries are marked by one of the following parts-of-speech: noun, verb, adjective or adverb. Bond and Ogura (2008) claim that they have reached a WordNet with reasonable coverage of most common Japanese words. They finish, however, with a caveat that 5 % of the lexicon's entries may contain errors, but that this is something they intend to correct manually while working with future versions. The Japanese WordNet is freely available for use, reproduction and distribution, and is available for download as an SQLite database.[2]

5 Evaluation

Precision and recall are the most common measures in evaluating the quality of automatically generated glossaries and variations of these have been used by Hara et al. (2008), Varga and Yokoyama (2007), Bond and Ogura (2008), Sjöbergh (2005) and Khanaraksombat and Sjöbergh (2007).

Precision is a measure used to evaluate systems for information retrieval and is defined as the proportion of retrieved relevant answers. The relevant answers in this case are all the suggestions for translations that are correct. Thus, the accuracy is the percentage of correct translation candidates compared to all translation candidates.

What is most interesting for this work, however, is the difference in quality of dictionaries pro-

[1] http://folkets-lexikon.csc.kth.se/folkets/folkets.en.html

[2] http://nlpwww.nict.go.jp/wn-ja/

duced without part-of-speech filtering compared to the lexicon produced using part-of-speech filtering. The precision p of a word w is calculated by the following formula:

$$\mathrm{p(w)} = \frac{\text{Correct translation candidates}}{\text{All translation candidates}}$$

Accuracy is calculated by performing a sample survey. In addition, a stratified sample was made based on the translation candidate's score—calculated using Tanaka and Umemura's one time inverse consultation—to get an idea of whether the precision varies with the score of the translation candidate. This also shows what threshold could be appropriate to use when presenting results to users of the lexicon. The translation candidates were divided into 10 strata, from 0.0 to 1.0 points, where each stratum corresponds to 0.1 points. From each stratum a random sample of 100 words was then drawn by systematic sampling, that is every n suggested translation was chosen, where n is calculated by all units in the population divided by the size of the sample.

To determine whether a translation candidate is relevant, that is correctly translated, you can use native speakers of both source and target languages and have them manually correct the translations. This method has been successfully used by Khanaraksombat and Sjöbergh (2007). Since access to such persons was missing, the samples were instead evaluated by manually checking the English translations of the translation candidate and whether the English translation is consistent with its proposed Swedish translation. Manually performing an exhaustive survey of this kind is not reasonable, so a sample survey was carried out instead. This method has been used successfully by Sjöbergh (2005).

Recall is another commonly used measure in the evaluation of systems for information retrieval. When evaluating an automatically generated dictionary, it is not reasonable to check all translation candidates. Instead one can compare with a baseline set to a selection of entries from a printed manually constructed lexicon in which all words are assumed to be correctly translated. Recall is, then, calculated as follows:

$$\mathrm{r(w)} = \frac{\text{Correct part-of-speech filtered candidates}}{\text{Correct baseline translation candidates}}$$

For this work, however, it is more interesting to examine the coverage of the method relative to earlier methods. Therefore the correct translation candidates produced without part-of-speech filtering was set as the baseline. It was then examined whether the translations were among the proposals of translations produced with part-of-speech filtering. Thus a measure of the method's recall relative to the baseline method is calculated, which shows to what extent the method using part-of-speech filtering catches all the correct translations generated by the method not using part-of-speech filtering.

6 Results

Table 1 shows the number of translation candidates generated for various points in each range, with and without the use of part-of-speech filtering, as well as the difference in the number of translation candidates created using each method. By using part-of-speech filtering the total number of translation candidates created decreased by 578 387 words, or 33.04 %. The reduction of translation candidates varies depending on the score and range from 9.43 % to 34.73 %, with a tendency of more translation candidates with different parts-of-speech in target and source language in the lower scoring ranges. This is probably because most of the translation candidates with different parts-of-speech are wrong, partly filtered out by the one time inverse consultation.

A large part of the translation candidates have not received any score at all. This applies to 1 341 391 translation candidates generated without part-of-speech filtering and 875 527 for those with part-of-speech filtering. This is mainly because the Swedish entries were missing in the People's Swedish-English dictionary, which has relatively few Swedish dictionary entries, and no look-up means zero score. This has the effect that the generated Japanese-Swedish lexicon contains fewer good words and translations. This can hopefully be addressed if more extensive versions of the People's Swedish-English Dictionary are released, rendering more Swedish entries available.

6.1 Quality

The quality of the automatically generated dictionaries is measured by calculating the precision of the suggested translation candidates. Table 2 shows the estimated quality of the translation candidates generated without part-of-speech filtering, while Table 3 shows the estimated quality of the translation candidates generated with part-of-

Score (p)	Without pos-filtering	With pos-filtering	Difference	Diff (%)
0	1 341 391	875 527	465 864	34.73
$0.0 < p \leq 0.1$	69 015	47 646	21 369	30.96
$0.1 < p \leq 0.2$	150 142	105 400	44 742	29.80
$0.2 < p \leq 0.3$	79 937	58 671	21 266	26.60
$0.3 < p \leq 0.4$	57 479	43 397	14 082	24.50
$0.4 < p \leq 0.5$	25 917	20 096	5 821	22.46
$0.5 < p \leq 0.6$	3 989	3 462	527	13.21
$0.6 < p \leq 0.7$	14 914	11 654	3 260	21.86
$0.7 < p \leq 0.8$	1 665	1 508	157	9.43
$0.8 < p \leq 0.9$	176	154	22	12.50
$0.9 < p \leq 1.0$	6 078	4 801	1 277	21.01
Total	1 750 703	1 172 316	578 387	33.04

Table 1: Score from one time inverse consultation, divided in intervals, for translation candidates generated without and with part-of-speech filtering.

Score (p)	Quantity	Precision	$\neq$
$0.9 < p \leq 1.0$	6 078	0.73	19
$0.8 < p \leq 0.9$	176	0.87	10
$0.7 < p \leq 0.8$	1 665	0.90	7
$0.6 < p \leq 0.7$	14 914	0.72	18
$0.5 < p \leq 0.6$	3 989	0.82	15
$0.4 < p \leq 0.5$	25 917	0.71	17
$0.3 < p \leq 0.4$	57 479	0.54	28
$0.2 < p \leq 0.3$	79 937	0.66	23
$0.1 < p \leq 0.2$	150 142	0.46	32
$0.0 < p \leq 0.1$	69 015		
0	1 341 391		

Table 2: Precision for translation candidates generated without part-of-speech filtering. Each sample is 100 words, $\neq$ represents the number of translation candidates where both source and target language have different part-of-speech.

Score (p)	Quantity	Precision
$0.9 < p \leq 1.0$	4 801	0.93
$0.8 < p \leq 0.9$	154	0.94
$0.7 < p \leq 0.8$	1 508	0.94
$0.6 < p \leq 0.7$	11 654	0.90
$0.5 < p \leq 0.6$	3 462	0.92
$0.4 < p \leq 0.5$	20 096	0.92
$0.3 < p \leq 0.4$	43 397	0.91
$0.2 < p \leq 0.3$	58 671	0.92
$0.1 < p \leq 0.2$	105 400	0.70
$0.0 < p \leq 0.1$	47 646	
0	875 527	

Table 3: Precision for translation candidates generated with part-of-speech filtering. Each sample is 100 words.

speech filtering.

Most interesting for this paper is the difference in quality between the two generated lexicons. Figure 1 illustrates the difference in the quality of translation candidates (y-axis) generated without part-of-speech filtering compared to translation candidates generated with part-of-speech filtering. Quality is the precision, that is the number of correctly translated translation candidates compared to all translation candidates, divided into strata (x-axis) based on the translation candidate's score to illustrate how the precision varies with the score.

Figure 1 shows a higher precision for the translation candidates generated with part-of-speech filtering than without for all tested strata. Each examined stratum has a sample size of 100 words. We also see a positive correlation (r = 0.77 without part-of-speech filtering, r = 0.62 with) that high values on the translation candidate's score correspond to high values for precision and that a low score equals low precision. All examined translation candidates where source and target languages are of different part-of-speech have been found incorrect.

Figure 1 also shows the threshold that is appropriate to use when presenting results to users. If you want a dictionary of good quality, you might choose precision 0.9 as threshold, which corresponds to a score of $>$0.7 with translation candidates generated without part-of-speech filtering, while you can go down to a score of $>$0.2

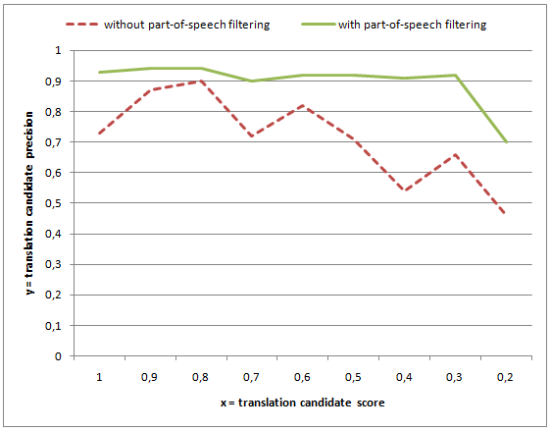

Figure 1: Score and precision for translation candidates generated without and with part-of-speech filtering.

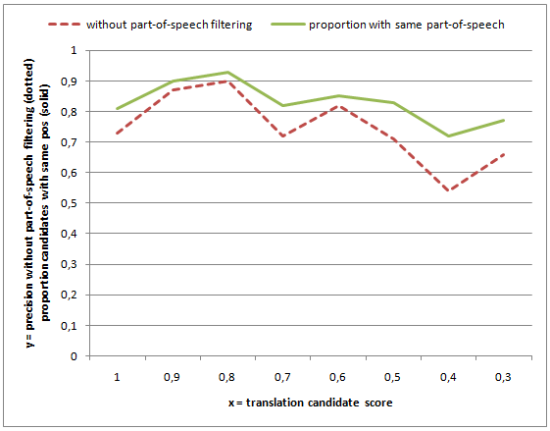

Figure 2: Precision of translation candidates generated without part-of-speech filtering and the proportion of these with same part of speech.

with translation candidates generated with part-of-speech filtering. This would provide a dictionary with a maximum of 10 % false translation candidates, and with 143 743 translation candidates of 40 716 unique Japanese entries with an average of 3.5 candidates per entry. Without part-of-speech filtering a dictionary of comparable quality would only contain 7 919 translation candidates of 5 244 unique Japanese entries with an average of 1.5 candidates per entry. Summarized there is a ratio of 20 between the two configurations.

Because of declining precision, no translation candidates with a precision below 0.1 have been checked. A too poor precision is not interesting, therefore the manual survey was terminated before all translation candidates were checked.

Review of data from Table 2 also showed a positive correlation between proposal precision and the percentage of translation candidates that had the same part-of-speech in both source and target language (r = 0.99). This applies for all investigated strata, where each sample stratum is 100 words.

To illustrate this more clearly, the two graphs are overlaid on each other in Figure 2. The dashed curve is from Figure 1 and shows the quality, i.e. precision, of translation candidates (y-axis) generated without part-of-speech filtering. The solid curve shows the percentage of the checked translation candidates that have the same part-of-speech in both source and target language. The proportion of such proposals (y-axis) are shown for each stratum (x-axis).

The positive correlation between the percentage of translation candidates with same part-of-speech and the candidate's accuracy may also explain the irregular jagged curve, since the random sample selected by systematic sampling appears to have an uneven distribution of proposals with different parts-of-speech. A better idea might be to ensure that each stratum has the same proportion of translation candidates with and without the same part-of-speech as is the case in the complete population. For example the score range $0.9 < p < 1.0$ had 21 % words with different parts-of-speech, which should then also be the case in the sample.

A review of the translation candidates that were filtered away, i.e. the candidates that have different parts-of-speech in the source and target languages, showed that they were all wrong. This also suggests that the more candidates that have different parts-of-speech in the source and target languages, the lower the precision of the generated dictionary.

Recall has been calculated for all investigated strata. The correct translation candidates generated without part-of-speech filtering have been set as the baseline, then it was checked whether they were among the translation candidates generated with part-of-speech filtering. All the evaluated candidates were included. This shows that the recall of the method is 100 % relative to the baseline method, that is, all correctly translated translation candidates produced without part-of-speech filtering were among the candidates generated with part-of-speech filtering.

7 Discussion

The most common cause of erroneous suggestions for translations were lexical ambiguity and specifically due to homographs. This problem has been effectively reduced by using part-of-speech filtering of the translation candidates. The second most common cause is ambiguity within the same part-of-speech. These may be filtered by inverse consultation, but not always, which then requires manual checking afterwards.

Another problem is the different categorizations of part-of-speech in the different dictionaries. An example found when searching the database for translation candidates with another part-of-speech than the original word is the following: The Japanese character for one (1) is categorized by the Japanese WordNet as a noun, while the People's English-Swedish dictionary categorized the Swedish translation as a cardinal numeral. This appears to be due to differences in the languages, where the Japanese language categorizes numerals as nouns. By part-of-speech filtering this correct translation candidate was errornously purged. However, no such entries were discovered in the systematic sample evaluation of the translation candidates, which implies that they are rather uncommon in this language combination.

One problem with using English as the intermediate language is the difference between British and American English. During the manual evaluation it was found that the Japanese WordNet had both British and American spelling of some English translations. One way to solve this, which has been tried by Bond and Ogura (2008), is to use some sort of British / American English dictionary for finding alternate spellings, if you can not find a direct translation.

8 Conclusion

The purpose of this study was to examine the impact of part-of-speech filtering on automatic generation of a bilingual dictionary by means of a pivot language. For this purpose two Japanese-Swedish lexicons were created, one without part-of-speech filtering and the other with part-of-speech filtering.

A comparison of these two dictionaries showed that the method with part-of-speech filtering gave 33 % fewer translation candidates. The manual evaluation of the quality of the candidates showed a higher precision of the candidates generated with part-of-speech filtering for all investigated strata. The results also showed a positive correlation (r = 0.99) between the percentage of translation candidates that have the same part-of-speech in both source and target language and proposal precision. Method recall is 100 %, according to the systematic manual evaluation, but later searches in the database uncovered that certain types of words still can be filtered out incorrectly, for example, numerals which appear to have different parts-of-speech in Japanese and Swedish.

From these results it is concluded that part-of-speech filtering is a useful method that reduces the number of erroneous suggestions for translations, at least for the current language of the trio (Japanese, English and Swedish). Part-of-speech filtering effectively eliminates problems stemming from external homographs in the intermediate language. Given the data, it is a simple step to add to the automatic generation of suggestions for translations, resulting in clear improvements.

As a result of the study 143 743 Japanese-Swedish translation candidates were created for 40 716 unique Japanese entries. Through using the precision curve in Figure 1, a precision of 0.9 was suggested as an appropriate threshold, which corresponds to >0.2 in score for translation candidates generated with part-of-speech filtering. These candidates have an estimated 10 % false translations, therefore it is important to conclude by pointing out that methods to automatically generate bilingual dictionaries are not perfect. They are great as preliminary and highly time-saving work, which should be followed by manual checks and cleaning of the resulting material. The result is also largely dependent on the source material used for one time inverse consultation to work properly.

The resulting Japanese-Swedish lexicon and the Java code used to generate it will be released under a Distributed Creative Commons Attribution-ShareAlike 2.5 Generic license[3] for free usage, sharing and remixing of the work.

References

Francis Bond and Kentaro Ogura. 2008. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2):127–136.

[3] http://creativecommons.org/licenses/by-sa/2.5/

Takahiro Hara, Maike Erdmann, and Shokiro Nishio. 2008. Extraction of bilingual terminology from a multilingual web-based encyclopedia. *Journal of Information Processing*, 16:68–79.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *LREC*. European Language Resources Association.

Viggo Kann and Joachim Hollman. 2011. People's English-Swedish dictionary, version 1.1. http://folkets-lexikon.csc.kth.se/folkets/om.en.html. [Online; accessed 21-January-2011].

Wanwisa Khanaraksombat and Jonas Sjöbergh. 2007. Developing and evaluating a searchable Swedish–Thai lexicon. In *Proceedings of Nodalida 2007*, pages 324–328, Tartu, Estonia.

Satoshi Shirai and Kazuhide Yamamoto. 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*, pages 174–179.

Jonas Sjöbergh. 2005. Creating a free digital Japanese-Swedish lexicon. In *Proceedings of PACLING 2005*, pages 296–300, Tokyo, Japan.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, COLING '94, pages 297–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Istvan Varga and Shoichi Yokoyama. 2007. Japanese-Hungarian dictionary generation using ontology resources. In *Machine Translation Summit XI*, pages 483–490.

Yujie Zhang, Qing Ma, and Hitoshi Isahara. 2007. Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In *Machine Translation Summit XI*, pages 551–557.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955