Developing and Evaluating a Searchable Swedish-Thai Lexicon

Wanwisa Khanaraksombat and Jonas Sjöbergh
KTH CSC
{wanwisa, jsh}@kth.se

Abstract

We present an automatically created Swedish-Thai lexicon. The lexicon was created by matching the English translations in a Thai-English and a Swedish-English lexicon. The search interface to the lexicon includes several NLP tools to help the target group: second language learners of Swedish. These include automatic generation of inflectional forms of words, automatic spelling correction, lemmatization and compound analysis of queries. A user study was performed and showed that while erroneous translations sometimes fool the users, they still find the lexicon good enough to be useful. They also like the NLP tools, though some grammatical information is presented in a hard to understand way. The lexicon and the interface tools were built using commonly available NLP tools.

1 Introduction

Since the world is rapidly becoming more and more interconnected, reading, writing and speaking foreign languages play important roles and have received much attention. Language technology can be an important tool to aid people when learning a new language and can improve efficiency of human activities and communication. One often used tool for understanding foreign languages is a bilingual lexicon.

This paper presents an automatically created large Swedish-Thai lexicon, a search interface for the lexicon with language tools and a user study evaluating the lexicon and the user interface. Everything was created from readily available tools, to see what results can be achieved with currently available NLP using very little manual work. A total of a few days of work was spent on creating the lexicon and the interface.

There are many ways to create bilingual lexicons. Traditionally it has been done by hand, which is time consuming and thus expensive if the desired lexicon is large, but it generally yields very high quality lexicons. Since manual work is expensive, automatic methods for creating lexicons have been devised. While they have drawbacks, such as including noise in the form of erroneous translations, they are still popular because of the enormous time saving potential. Automatic methods can be used to generate a first noisy lexicon which is then cleaned up and extended by manual work.

2 Related Research

There are many methods for generating bilingual lexicons from a parallel corpus, i.e. a corpus where the same text is available in different languages. Koehn and Knight (2001) discuss different methods using bilingual corpora, monolingual corpora and lexicon resources to extract bilingual dictionaries.

Other approaches use existing bilingual lexicons from the source and target language to some common intermediate language. Usually, English is used as the “interlingua”, since there exist large bilingual lexicons between English and many other languages. This is the approach we used, since there were lexicons available but no parallel corpora.

The fact that many English words are ambiguous is a problem that can lead to erroneous translations in the new lexicon. 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 lexicon. Another problem with the same effect is

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 324–328

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.

Work on automatic bilingual lexicon creation using existing bilingual lexicons and an intermediate language has been done before (Tanaka and Umemura, 1994; Shirai et al., 2001; Shirai and Yamamoto, 2001). The problem of ambiguity can be mitigated by using several intermediate languages (Paik et al., 2001) and using part of speech and semantic categories (Bond et al., 2001). Hopefully different intermediate languages will not be ambiguous in the same way. The impact of using lexicons in different directions, i.e. a source language-English or an English-source language lexicon, has also been examined (Paik et al., 2004).

The two biggest problems in our lexicon is that common words are often not translated and that there are erroneous translations in the resulting lexicon. Common words are often ambiguous, and thus hard to automatically translate with our method. Hopefully, learners of a new language will learn the common words quickly through some other means, and thus not need them in the lexicon.

Filtering out erroneous translations can as mentioned above be done using several intermediate languages. In our case, we only found lexicons with translations to English. It is also possible to use information from parallel corpora, but as mentioned before, we had no such corpora.

There are other methods that can be used too, for instance the monolingual frequencies of different translation candidates, or the "burstiness" of the words in monolingual news data (Schafer and Yarowsky, 2002). Words common in one language are often common in other languages, and words with the same meaning are often used in describing the same events.

While we could likely have found enough resources to use such methods, we did not. One problem that occurs is that Thai is non-trivial to segment into words, and we had very little experience in processing Thai.

3 Creating the Lexicon

The method used to create the lexicon was a fairly standard method. We used a program that was previously used in generating a Japanese-Swedish lexicon (Sjöbergh, 2005) with only slight modifications. All English descriptions of Thai words were matched to all English descriptions of Swedish words. Matches are basically word overlap, and the best matches are selected as translation candidates.

Quality	Words
All OK	66
Most OK	4
Some OK	3
Similar	11
Wrong	16

Table 1: Translation quality of a random sample of 100 Swedish words and their translation.

The English words were weighted with a measure similar to idf (Inverse Document Frequency), which is useful when ranking several poor translation candidates. Words marked as being a certain word class were not allowed to match words in the other language marked with another word class. Many words in the lexicons used have no word class markings at all, though.

The created lexicon consists of over 20,000 words, which is the largest Swedish-Thai lexicon known to us. The next largest machine searchable lexicon known contains about 2,000 words, though in book form there are lexicons of about 7,000 words available. The main drawback of the automatically created lexicon is of course that it contains erroneous translations.

For creating the Swedish-Thai lexicon, the Thai-English lexicon Lexitron (Palingoon et al., 2002) was used. It is a freely available dictionary from NECTEC (downloadable from http://www.nectec.or.th/) which includes not only translations but also word class information, example sentences and pronunciation. It contains over 40,000 words.

The Swedish-English lexicon used contains about 160,000 Swedish words or expressions with their corresponding translations in English. Many of these Swedish words are old words that are somewhat rare in modern Swedish though.

A Swedish-Thai lexicon was created by for each Swedish word in the Swedish-English lexicon taking the top scored suggestions in Thai. If the top suggestion for a word had a score lower than 0.75, the word was not translated. A Thai-Swedish lexicon was also created, by taking the top scoring Swedish suggestions for each Thai word in the same way. When searching the lexicon in Swedish, the Swedish-Thai lexicon can be used, and then when searching in Thai the Thai-Swedish lexicon. Of course, for very many words the closest matching Swedish word for a Thai word will have the same Thai word as its closest match.

A small evaluation of the translation quality was done by having a native speaker of Thai who is quite fluent in Swedish manually check translations. For 100 Swedish words, how well the translations matched the actual meaning was checked. The results are shown in Table 1.

Since there are often many translation suggestions for each word, the quality was classified into the following classes: *All OK*, meaning all suggested translations are correct translations of the Swedish word. *Most OK*, if there are more correct translations suggested than incorrect translations. *Some OK*, if there is at least one correct suggestion. If the word occurs in a context, the correct meaning would likely still be understood using these suggestions. *Similar* is the case that the translations at least give enough information so that the general idea will come across in most contexts. An example would be "vehicle" instead of "car", or "radio broadcast" instead of "TV broadcast". Finally, the class *Wrong* is used when the translation is simply wrong.

4 NLP Tools for Searching the Lexicon

A simple web interface for looking up words in the Thai-Swedish lexicon from the previous section was created. To help users of the lexicon interface, especially the main target group which was second language learners of Swedish, some readily available language technology tools were added. These include spelling checking (Domeij et al., 1994), lemmatization (Domeij et al., 2000), and compound analysis (Sjöbergh and Kann, 2006). These tools were only implemented for searches in Swedish. In the spirit of believing that the user is probably right if the system understands the query, these tools are only used if the search fails to return any matches.

The first tool is spelling correction. Experiences from other popular lexicon services on the Internet indicate that a substantial part of all queries are misspelled, even by native speakers. Thus, if a query returns no results, the word is put through a spelling checker. If there are suggested corrections from the spelling checker, all such suggestions are automatically used as search queries and the resulting translations are shown.

The second tool is a lemmatizer. When there are no results, the lemma form of the word is used instead, since some word classes have quite rich inflection in Swedish but only the lemma forms are listed in the lexicon. A possible improvement that was discovered during the evaluations is that verbs were generally not listed in their lemma form but in their present tense in the Swedish-English lexicon used, so for verbs the present tense form would likely be a better choice than the lemma.

Since Swedish has very productive compounding where the compound components are concatenated into one word, the third tool is compound splitting. Search queries that return no results can be automatically split into their compound components. The translations of each component are presented to give an indication of the meaning of the whole compound.

It is also possible to search the lexicon using words in Thai (or even English), though the language tools only work for Swedish. The interface also allows for choosing which of the lexicons, the Thai-Swedish or the Swedish-Thai, to search in.

Since there is a possibility of erroneous translations, mainly caused by ambiguous English words, it is also possible to view the original English translations, color coded to show which parts have a matching word in the corresponding translation.

The lexicon also presents other helpful information, such as example sentences, pronunciation information and inflectional forms for some words. Inflections are automatically generated by the same NLP tool that performed the lemmatization of queries (Domeij et al., 2000), example sentences and pronunciations are taken from another lexicon[1] when available there.

5 User Study

To evaluate if the resulting lexicon is useful for learners of Swedish a small user study was done. For a detailed account of this study written in Swedish, see (Khanaraksombat, 2007). Five native speakers of Thai with eight to twelve months of studies of Swedish were observed while doing two exercises using the lexicon and then interviewed. The first task was to create a story, in Swedish, from a series of eight pictures. The second task was to translate a short Swedish text into Thai. Thus, one task was intended to make the users search the lexicon using Thai words to find the wanted Swedish word and the other to make them look up Swedish words they did not know. These tasks were done in two groups, one with two persons and one with three.

5.1 Results from Observing the Users

During the exercises, there were quite a few examples of problems related to the somewhat poor quality of the translations in the lexicon:

Sometimes the word that was sought after was not found by the users despite it being present in

[1] Swedish-English Lexin, http://lexin.nada.kth.se/

the lexicon, because there were too many suggested translations of a word. This was mainly a problem of English ambiguous word giving many too many mappings between Thai and Swedish words.

Sometimes a word was mistranslated, confusing the users. This was usually caused by translations such as *"kort"* (Swedish word corresponding to the English word "card"), translated as "card, for example playing card, post card etc". This type of translations made the Swedish word for post card match the translation for the Thai word for playing card. There were also examples where the users understood that a translation was wrong, though this of course did not help them in understanding the correct translation.

Sometimes the sought after word was not available in the lexicon. Especially troublesome is that many common words are highly ambiguous, and thus very hard to translate automatically. Sometimes the word itself was not translated, but a short phrase containing the word was. This also confused the users, for instance when searching for the Swedish word *"finns"* (is/are/exists) which only matched a phrase meaning "there are some friends", which matched the translation of the Thai word for "friend". So if one does not look too carefully at the results, it appears that "exists" should be translated as "friend".

There were also problems with understanding the grammatical information displayed by the search interface. For instance for verbs, the uninflected form, the past tense and the present perfect is shown. This determines what inflectional pattern a verb belongs to, but the users generally wanted the present tense, which was not available, and were not familiar enough with inflectional rules to have much use of the presented information. The same was true for nouns, where the inflectional pattern was shown but the much sought after gender information was not explicitly stated, because it is implicitly available in the inflectional information and normally not presented to native speakers (the target group of the resource this information was taken from).

5.2 Results from the Interviews

On the whole, the users believed that lexicons are very important to them. They found the automatically created lexicon to be fairly useful, despite the erroneous translations. They did however think that there were too many results, fewer translations would be better as long as the important ones are present, i.e. less commonly used meanings of a word should preferably be removed from the lexicon. When it comes to the NLP tools the users were satisfied, and thought that the option to turn them off should be removed from the interface to make it cleaner. As explained in the previous section, the presented grammatical information was not what the users wanted. Similar but slightly different information would be much appreciated though. The users also found the layout of the interface to be a bit messy, making it hard to find the most important information.

6 Discussion

The problems related to the presentation of the grammatical information can be easily improved. The information the users wanted can easily be generated automatically from the resources available in the same way that the currently displayed information is.

Some of the other problems can be mitigated by using smaller Thai-English and Swedish-English lexicons, or by filtering out only important words from some list of what is considered important. Automatically determining which words are important is likely quite hard, though a rough estimate could be gained by concentrating on highly frequent words, for instance determined from some topically relevant corpus.

The translation quality was evidently good enough for the users to find the resource useful, but since there were quite a few problems for the users related to erroneous translations, it should preferable be improved. Having smaller lexicons to start with removes some erroneous matches, but probably not enough. Using more than one intermediate language is easy and generally also removes many faulty translations, but no other large lexicons are available for either Swedish or Thai, as far as we know. Monolingual corpora could be used to filter out translation suggestions were the monolingual frequencies are very different, and it would also likely be a good idea to remove longer phrases from the Swedish-English lexicon.

7 Conclusions

We automatically created a large Thai-Swedish lexicon and built a search interface for the lexicon including NLP tools. In a user study it was found that the users were generally satisfied with the NLP tools. Though some of the automatically generated grammatical information presented was not what the users wanted, the wanted information could easily be generated and the interface changed to display the more relevant information.

Regarding translations, the users ran into some problems caused by the translation quality of the lexicon being poor, though the they still found the lexicon useful. The users would prefer a lexicon containing only important words to a lexicon with a large coverage, so as not to be confused by rare translations.

In conclusion, it seems that it is possible to create a quite useful tool for second language learners with minimal amounts of manual work.

Acknowledgements

We would like to thank Viggo Kann, Ola Knutsson and our anonymous reviewers for valuable comments.

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.

Rickard Domeij, Joachim Hollman, and Viggo Kann. 1994. Detection of spelling errors in Swedish not using a word list en clair. *Journal of Quantitative Linguistics*, 1:195–201.

Richard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska – an efficient hybrid system for Swedish grammar checking. In *Proceedings of Nodalida '99*, pages 49–56, Trondheim, Norway.

Wanwisa Khanaraksombat. 2007. Utvärdering av ett svensk-thailändskt elektroniskt lexikon. Master's thesis, KTH CSC, Stockholm, Sweden.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of EMNLP 2001*, Pittsburgh, USA.

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.

Pornpimon Palingoon, Pornchan Chantanapraiwan, Supranee Theerawattanasuk, Thatsanee Charoenporn, and Virach Sornlertlamvanich. 2002. Qualitative and quantitative approaches in bilingual corpus-based dictionary. In *Proceedings of SNLP-O-COCOSDA 2002*, pages 152–158, Hua Hin, Thailand.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and a bridge lexicon. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 146–152, Taipei.

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.

Jonas Sjöbergh and Viggo Kann. 2006. Vad kan statistik avslöja om svenska sammansättningar? *Språk och Stil*, 16:199–214.

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 COLING-94*, pages 297–303, Kyoto, Japan.