Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages

Sumithra Velupillai
DSV/KTH-Stockholm University
SE-164 40 Kista
Sweden
sumithra@dsv.su.se

Hercules Dalianis [1, 2]
[1)] DSV/KTH-Stockholm University
SE-164 40 Kista
Sweden
[2)] Euroling AB
Igeldammsgatan 22c
112 49 Stockholm, Sweden
hercules@dsv.su.se

Abstract

Hallå Norden is a web site with information regarding mobility between the Nordic countries in five different languages; Swedish, Danish, Norwegian, Icelandic and Finnish. We wanted to create a Nordic cross-language dictionary for the use in a cross-language search engine for Hallå Norden. The entire set of texts on the web site was treated as one multilingual parallel corpus. From this we extracted parallel corpora for each language pair. The corpora were very sparse, containing on average less than 80 000 words per language pair. We have used the Uplug word alignment system (Tiedemann 2003a), for the creation of the dictionaries. The results gave on average 213 new dictionary words (frequency > 3) per language pair. The average error rate was 16 percent. Different combinations with Finnish had a higher error rate, 33 percent, whereas the error rate for the remaining language pairs only yielded on average 9 percent errors. The high error rate for Finnish is possibly due to the fact that the Finnish language belongs to a different language family. Although the corpora were very sparse the word alignment results for the combinations of Swedish, Danish, Norwegian and Icelandic were surprisingly good compared to other experiments with larger corpora.

1 Introduction

Hallå Norden (Hello Scandinavia) is a web site with information regarding mobility between the Nordic countries and is maintained by the Nordic Council. Mobility information concerns issues such as how employment services, social services, educational systems etc. work in the different countries. The web site has information in five different languages; Swedish, Danish, Norwegian, Icelandic and Finnish. In this paper Nordic languages are defined as Swedish, Danish, Norwegian, Icelandic and Finnish. Scandinavian languages are defined as the Nordic languages excluding Finnish.

The texts on the web site were almost parallel and there were also ten minimal dictionaries with on average 165 words available for the different languages. The dictionaries consisted of domain-specific words regarding mobility information in the Nordic countries. The Nordic Council wanted to extend the dictionaries so they would cover a larger part of the specific vocabulary, in order to help the people in the Nordic countries to find and learn the concepts in their neighboring countries.

The entire set of texts on the web site was treated as one multilingual parallel corpus. From this we extracted parallel corpora for each language pair. We discovered, as expected, that the corpora were very sparse, containing on average less than 80 000 words per language pair. We needed to construct 10 different dictionaries and therefore we processed 10 pairs of parallel text sets. We have used the Uplug word alignment system (Tiedemann 2003a), for the creation of the dictionaries. The system and motivation for the choice of system is further discussed in Section 2.1.

We also discovered that the texts were not completely parallel. Therefore, we made a small experiment on attempting to enhance the results by deleting texts that were not parallel. Multilingual parallel corpora covering all Nordic languages are very rare. Although the corpora created in this work are domain-specific, they are an important contribution for further research on Nordic multilingual issues. Moreover, many large governmental, industrial or similar web sites that contain information in several languages may profit from compiling multilingual dictionaries automatically in order to enhance their search engines and search results.

In this project, our two main goals were to compile parallel corpora covering the Nordic languages, and to evaluate the results of automatically creating dictionaries using an existing tool with basic settings, in order to find out where more work would need to be done and where performance is actually acceptable. We have limited the work by only testing one system (Uplug) with basic settings. Our experiments and results are described in further detail in the following sections. Conclusions and future work are discussed in the final section.

2 Related Work

Word alignment systems have been used in previous research projects for automatically creating dictionaries. In Charitakis (2007) Uplug was used for aligning words in a Greek-English parallel corpus. The corpus was relatively sparse, containing around 200 000 words for each language, downloaded from two different bilingual web sites. A sample of 498 word pairs from Uplug were evaluated by expert evaluators and the result was 51 percent correctly translated words (frequency > 3). When studying high frequent word pairs (>11), there were 67 percent correctly translated words. In Megyesi & Dahlqvist (2007) an experiment is described where they had 150 000 words in Swedish and 126 000 words in Turkish that gave 69 percent correct translations (Uplug being one of the main tools used). In this work the need for parallel corpora in different language combinations is also discussed.

The ITools' suite for word alignment that was used in Nyström et al (2006) on a medical parallel corpus, containing 174 000 Swedish words and 153 000 English words, created 31 000 word pairs with 76 percent precision and 77 percent recall. In this work the word alignment was produced interactively.

A shared task on languages with sparse resources is described in Martin et al (2005). The language pairs processed were English-Inuktitut, Romanian-English and English-Hindi, where the English-Inuktitut parallel corpus contained around 4 million words for English and 2 millions words for Inuktitut. English-Hindi had less words, 60 000 words and 70 000 words respectively. The languages with the largest corpora obtained best word alignment results, for English-Inuktitut over 90 percent precision and recall and for English-Hindi 77 percent precision and 68 percent recall. One conclusion from the shared task was that it is worth using additional resources for languages with very sparse corpora improving results with up to 20 percent but not for the languages with more abundant corpora such as for instance English-Inuktitut.

2.1 Word Alignment: Uplug

We have chosen to use the Uplug word alignment system since it is a non-commercial system which does not need a pre-trained model and is easy to use. It is also updated continuously and incorporates other alignment models, such as GIZA++ (Och & Ney 2003). We did not want to evaluate the performance of different systems in the work presented here, but rather evaluate the performance of only one system applied on different language combinations and on sparse corpora. Evaluating the performance of different systems is an important and interesting research problem, but is left for future work. An evaluation of two word alignment systems Plug (Uplug) and Arcade is described in Ahrenberg et al (2000).

The Uplug system implements a word alignment process that combines different statistical measures for finding word alignment candidates and is fully automatic. It is also possible to combine statistical measures with linguistic information, such as part-of-speech tags. In the preprocessing steps the corpora are converted to an xml-format and they are also sentence aligned.

We have chosen to use basic settings for all corpora in the different language pairs, in order to evaluate the effect of this. The default word alignment settings in Uplug works in the following way:

- create basic clues (Dice and LCSR)
- run GIZA++ with standard settings (trained on plain text)

Language pair	No. texts	No. words	Word distribution, first language in language pair, %
sw-da	191	83871	49.2
sw-no	133	62554	49.7
sw-fi	196	73933	57.6
sw-ice	187	82711	48.5
da-no	156	68777	50.2
da-fi	239	84194	58.4
da-ice	232	97411	49.5
no-fi	156	58901	58.2
no-ice	145	64931	49.6
Average	182	75254	52.3

Table 1: General corpora information, initial corpora

- learn clues from GIZA's Viterbi alignments
- "radical stemming" (take only the 3 initial characters of each token) and run GIZA++ again
- align words with existing clues
- learn clues from previous alignment
- align words again with all existing clues[1]

This approach is called the *clue alignment* approach and is described further in Tiedemann (2003b). In the work presented here, we have not included any linguistic information, as we wanted to evaluate the performance of applying the system on sparse, raw, unprocessed corpora for different (Nordic) language pairs, using default settings.

3 Experiments and Results

For the project presented in this paper we wanted to see if it was possible to create domain-specific dictionaries on even smaller corpora. (compared to the ones described in Section 2) for all the Nordic language pairs. We did not have the possibility to evaluate the results for Icelandic-Finnish, since we did not find any evaluator having knowledge in both Icelandic and Finnish. Therefore we present the results for the remaining nine language pairs. In total we had four evaluators for the other language combinations. Each evaluator evaluated those language pairs she or he had fluent or near-fluent knowledge in. The domain was very restricted containing only words about mobility between the Nordic countries.

The Scandinavian languages are closely related. Swedish, Danish, and Norwegian are comprehensible for Scandinavians. A typical Swede will for instance understand written and to a certain degree spoken Danish, but is not able to speak Danish. Typical Swedes will, for instance, have a passive understanding of Danish (and vice versa for the other languages). Finnish on the other hand belongs to the Finno-Ugric group of the Uralic languages, while the Scandinavian languages are North-Germanic Indo-European languages. We wanted to investigate if, and how, these differences affect the word alignment results. We also wanted to experiment with different frequency thresholds, in order to see if this would influence the results.

The first step was to extract the web pages from the web site and obtain the web pages in plain text format. We obtained help for that work from Euroling AB,[2] our contractor.

In Table 1 we show general information about the corpora. We see that the distribution of words is even for the Scandinavian languages, but not for the combinations with Finnish. It is interesting to observe that Finnish has fewer word tokens than the Scandinavian languages.

All Nordic languages, both Scandinavian and Finnish, have very productive word compounding. In Finnish word length is longer, on average,

[1] Steps taken from the Quickstart guidelines for the Uplug system, which can be downloaded here: http://uplug.sourceforge.net/

[2] See: http://www.euroling.se/

and the number of words per clause lower, on average, due to its extensive morphology.

In Dalianis et al (2007) lemmatizing the text set before the alignment process did not improve results. In the work presented here, we have also made some experiments on lemmatizing the corpora before the alignment process. We have used the CST lemmatizer[3] for the Scandinavian languages and Fintwol[4] for Finnish. Unfortunately, the results were not improved. The main reason for the decrease in performance is probably due to the loss of sentence formatting during the lemmatization process. The sentence alignment is a crucial preprocessing step for the word alignment process, and a lot of the sentence boundaries were lost in the lemmatization process. However, the resulting word lists from Uplug have been lemmatized using the same lemmatizers, in order to obtain normalized dictionaries.

The corpora were to some extent non-parallel containing some extra non-parallel paragraphs. We found that around five percent of the corpora were non-parallel. In order to detect non-parallel sections we have used a simpler algorithm than in for instance Munteanu & Marcu (2006). The total number of paragraphs and sentences in each parallel text pair were counted. If the total number for each language in some language pair differed more than 20 percent these files were deleted. The refined corpora have been re-aligned with Uplug and evaluated. In Table 2 we show the general information for the refined corpora.

3.1 Evaluation

Our initial plan was to use the manually constructed dictionaries from the web site as an evaluation resource, but the words in these dictionaries were rare in the corpus. Therefore we used human evaluators to evaluate the results from Uplug.

The results from the Uplug execution gave on average 213 new dictionary words (frequency > 3) per language, see Table 3. The average error rate[5] was 16 percent. We delimited the word amount by removing words shorter than six characters, and also multiword expressions[6] from the resulting word lists. The six character strategy is efficient for the Scandinavian languages as an alternative to stop word removal (Dalianis et al 2003) since the Scandinavian languages, as well

Language pair	No. parallel texts	Deleted files, %	No. words, parallel	Word distribution, first language in language pair, %
sw-da	179	6.3	78356	49.7
sw-no	128	3.8	59161	49.8
sw-fi	189	3.6	69525	58.1
sw-ice	175	5.9	76056	48.3
da-no	147	5.8	64946	50.2
da-fi	222	7.1	77849	58.6
da-ice	210	3.4	89093	49.0
no-fi	145	7.1	55409	58.3
no-ice	130	2.1	59622	49.0
Average	169	5.0	70002	52.3

Table 2: General corpora information, refined parallel corpora (non-parallel texts deleted)

[3] See: http://cst.dk/download/cstlemma/current/doc/
[4] See: http://www2.lingsoft.fi/cgi-bin/fintwol

[5] The error rate is in this paper defined as the percentage of wrongly generated entries compared to the total number of generated entries.
[6] A multiword expression is in this paper defined as words (sequences of characters, letters or digits) separated by a blank or a hyphen.

as Finnish, mostly produce compounds that are formed into one word (i.e. without blanks or hyphens). In Tiedemann (2008), a similar strategy of removing words with a word length shorter than five characters was carried out but in that case for English, Dutch and German.

Different combinations with Finnish had a higher error rate, 30 percent, whereas the error rate for the combinations of the Scandinavian languages only yielded on average 9 percent errors.

The high error rate for Finnish is possibly due to the fact that the Finnish language belongs to a different language family. We can see the same phenomena for Greek (Charitakis, 2007) and Turkish (Megyesi & Dahlqvist, 2007) combined with English and Swedish respectively, with 33 and 31 percent erroneously translated words.

However, one might expect even higher error rates due to the differences in the different language pairs (and the sparseness of the data). Finnish has free word order and is typologically very different from the Scandinavian languages, and the use of form words differs between the languages. On the other hand, both Finnish and the Scandinavian languages produce long, complex compounds somewhat similarly, and the word order in Finnish share many features with the word order in the Scandinavian languages. One important aspect is the cultural similarities that the languages share.

The main errors that were produced for the combinations of Finnish and the Scandinavian languages consisted of either errors with particles or compounds where the head word or attribute were missing in the Finnish alignment. For instance, the Swedish word *invånare* (inhabitant) was aligned with the Finnish word *asukasluku* (number of inhabitants). Another error which was produced for all combinations with Finnish was *lisätieto* (more information) which was aligned with *ytterligere* (additional, more) in Norwegian (and equivalent words in Swedish and Danish), an example of an error where the head word is missing. Many texts had sentences pointing to further information, which might explain this type of error.

The lemmatizers produced some erroneous word forms. In Dalianis & Jongejan (2006) the CST lemmatizer was evaluated and reported an average error rate of nine percent. Moreover, since the lemmatization process is performed on the resulting word lists, and not within the original context in which the words occur, the automatic lemmatization is more difficult for the two lemmatizers used in this project. These errors have not been included in our evaluation since they are not produced by the Uplug alignment procedure.

We can also see in Table 3 that deleting non-parallel texts using our simple algorithm did not improve the overall results significantly. Perhaps our simple algorithm was too coarse for these corpora. The texts were in general very short and simple frequency information on paragraph and sentence amounts might not have captured non-parallel fragments on such texts.

	Initial		Deleting non-parallel	
Language pair	No. dictionary words	Erroneous translations, %	No. dictionary words	Erroneous translations, %
sw-da	322	7.1	305	7.2
sw-no	269	6.3	235	9.4
sw-fi	138	29.0	133	34.6
sw-ice	151	18.5	173	16.2
da-no	322	3.7	304	4.3
da-fi	169	34.3	244	33.2
da-ice	206	6.8	226	10.2
no-fi	185	27.6	174	30.0
no-ice	159	14.5	181	14.4
Average	213	16.4	219	16.1

Table 3: Produced dictionary words and error rate

The produced dictionary words were of high domain-specific quality. The majority of the correct and erroneous word pairs were covered by both the initial and the refined corpus. Deleting non-parallel texts produced some new, valuable words that were not included in the initial results. However, since these dictionaries were generally smaller, this did not improve the overall results, and the error rate was somewhat higher for most language pairs. Improved dictionary in this work means as many word pairs as possible with domain-specific significance.

Since the texts were about different country-specific issues they could contain sections in another language (names of ministries, offices etc). This produced some errors in the alignment results. These errors might have been avoided by applying a language checker while processing the texts.

The errors for the Scandinavian languages were also mainly of the same type, and mostly due to the fact that the texts were not completely parallel, or due to form words or compounds. For instance, the Swedish word *exempelvis* (for example) was aligned with the Norwegian word *eksempel* (example), which was counted as an error, but which, in its context, is not completely erroneous.

Even at a relatively low frequency threshold the results were very good for the Scandinavian languages. We tried to increase the frequency threshold in order to see if this would improve the results for Finnish, which it unfortunately did not. However, as stated above, the errors were mainly of the same type, and probably constant over different frequencies. We also see that for Icelandic, unlike the other languages, deleting non-parallel fragments yielded larger dictionaries. Uplug produced more multiword units for the initial corpora containing Icelandic, single word pairs were more frequent in the refined corpus. However, the overall results were not improved.

4 Conclusions and Future Work

Although the corpora were very sparse the word alignment results for Swedish-Danish, Swedish-Norwegian and Danish-Norwegian were surprisingly good with on average 93.1 percent correct results. The results for Finnish were worse with on average only 67.4 percent correct results.

However, as discussed above, the main errors were of the same type. Creating dictionaries for non-related languages might need more elaborate alignment approaches. In the special case of Finnish combined with one (or several) of the Scandinavian languages, simple preprocessing steps might improve the results. For instance, removing stop words before running the corpora through a word alignment system might handle the errors where particles and form words are included. Also, tagging the corpora with part-of-speech tags and lemmatizing as a preprocessing step might improve results.

An important aspect of automatically creating multilingual dictionaries is the need for preprocessing tools covering all languages. This is often difficult to obtain, and different tools use different formatting and tagging schemes. Moreover, they might differ in robustness, which also affects the end results. In this project, we encountered such problems during the lemmatization process for instance, but we did not have the opportunity to explore and evaluate alternative tools. In the future, evaluating the performance of the preprocessing steps might be desirable.

Evaluating translated words is not easy. Many words may be related without being direct translations. Manual evaluation has the advantage of taking such issues into account, but this also means that the results might differ depending on the evaluator. Furthermore, evaluating translations without contextual information is problematic. Also, the criteria for judging a translation as correct or not depend on the goal for the use of the word lists. For instance, the errors for the combinations with Finnish might not be problematic in a real-world search engine setting, depending on which demands there are on the search results. The errors produced in the work presented here would probably yield acceptable search results. Such user and search engine result aspects have not been evaluated here, but are interesting research questions for future work.

The Nordic languages are highly inflectional. Combining compound splitting and lemmatizing before the alignment process might improve the results. Especially compound splitting could probably handle the errors produced for the combinations of Finnish with the Scandinavian languages. Cross-combining the different language pairs might enhance the results and create more specific and errorless dictionaries. Other word alignment systems should also be tested, in order to compare different approaches and their results. Perhaps results from different systems could also be combined, in order to produce more extensive dictionaries. Furthermore, other approaches to

detect non-parallel fragments should be investigated.

Finding the boundary for the minimum size of parallel corpora in order to obtain acceptable dictionaries is also an interesting research issue which should be explored.

Automatically creating multilingual dictionaries is not trivial. Many aspects need to be considered. Especially, the final use of the produced results influences both the preprocessing steps required and the evaluation of the results. Also, the languages in consideration affect the steps that need to be made. However, in this paper we have shown that using state-of-the-art tools on sparse, raw, unprocessed domain-specific corpora in both related and non-related languages yield acceptable and even commendable results. Depending on the purposes for the use of the dictionaries, simple adjustments would probably yield even better results.

In a real-world setting, parallel (or near-parallel) corpora covering several (small) languages are difficult to obtain and compile. Most resources are found on the Internet, and the quality of the corpora may vary depending on many aspects. Formatting, translations, text length and style may differ considerably depending on the type of texts. Freely available text sets for small languages are often sparse. Despite this, we have shown that it is possible to compile valuable resources from available data.

There are very few sources of dictionaries covering the Nordic language pairs. The created corpora will be made publicly available for further research and evaluation.

Acknowledgements

We would like to thank Pernilla Näsfors at Euroling AB and SiteSeeker for her work with producing the Hallå Norden corpora, Eija Jacklin at DSV, Stockholm University, and Jussi Karlgren, at SICS, for their help in evaluating Finnish in various combinations with Swedish, Norwegian and Danish and Óðinn Albertsson at Norræna félagið for his help in evaluating Icelandic in various combinations with Swedish, Norwegian and Danish.

References

Ahrenberg, L., M. Merkel, A. Sågvall Hein and J. Tiedemann 2000. Evaluation of word alignment systems. In the Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000), Athens, Greece, 31 May - 2 June, 2000, Volume III: 1255-1261.

Charitakis, K. 2007. Using parallel corpora to create a Greek-English dictionary with Uplug, in the Proceedings of the 16th Nordic Conference on Computational Linguistics - NODALIDA ‘07.

Dalianis, H. and B. Jongejan 2006. Hand-crafted versus Machine-learned Inflectional Rules: the Euroling-SiteSeeker Stemmer and CST's Lemmatiser, in Proceedings of the International Conference on Language Resources and Evaluation, LREC 2006.

Dalianis, H., M. Rimka and V. Kann 2007. Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian. Workshop: The Automatic Treatment of Multilinguality in Retrieval, Search and Lexicography, Copenhagen, April 2007.

Dalianis, H., M. Hassel, J. Wedekind, D. Haltrup, K. de Smedt and T.C. Lech. 2003. Automatic text summarization for the Scandinavian languages. In Holmboe, H. (ed.) Nordisk Sprogteknologi 2002: Årbog for Nordisk Språkteknologisk Forskningsprogram 2000-2004, pp. 153-163. Museum Tusculanums Forlag.

Martin, J and R. Mihalcea and T. Pedersen. 2005. Word Alignment for Languages with Scarce Resources. Proceedings of the ACL 2005 Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, MI, June 2005.

Megyesi, B. and B. Dahlqvist, 2007. The Swedish-Turkish Parallel Corpus and Tools for its Creation, in Proceedings of the 16th Nordic Conference on Computational Linguistics - NODALIDA ‘07.

Munteanu, D.S. and D. Marcu 2006. Extracting Parallel Sub-sentential Fragments from Non-parallel Corpora. ACL ’06: Proceedings of the 21st International Conference on Computational Linguistics, pp. 81-88, Sydney, Australia.

Nyström, M., M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson and H. Åhlfeldt. 2006. Creating a Medical English-Swedish Dictionary using Interactive Word Alignment, in BMC medical informatics and decision making, 6:35.

Och, F. J. and N. Hermann. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Tiedemann, J. 2003a. Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Acta Universitatis Upsaliensis: Studia linguistica upsaliensia, ISSN 1652-1366, ISBN 91-554-5815-7.

Tiedemann, J. 2003b. Combining clues for word alignment. In the Proceedings of the Tenth Conference on European Chapter of the Association For

Computational Linguistics - Volume 1, Budapest, Hungary. .

Tiedemann, J. 2008. Synchronizing Translated Movie Subtitles. In the Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco.