Shallow-transfer rule-based machine translation for Swedish to Danish

Francis M. Tyers
Dept. Lleng. i Sist. Informàtics,
Universitat d'Alacant,
Alacant. E-03070
ftyers@dlsi.ua.es

Jacob Nordfalk
Center for Videreuddannelse
Ingeniørhøjskolen i København,
DK-2750 Ballerup, Denmark
jano@ihk.dk

Abstract

This article describes the development of a shallow-transfer machine translation system from Swedish to Danish in the Apertium platform. It gives details of the resources used, the methods for constructing the system and an evaluation of the translation quality. The quality is found to be comparable with that of current commercial systems, despite the particularly low coverage of the lexicons.

1 Introduction

Both Swedish and Danish languages were standardised in the 12th to 15th centuries out of the Old Norse which was spoken across Scandinavia. Swedish was standardised on the speech of the zone around Stockholm, whereas Danish was standardised on the speech of Copenhagen and surrounding areas. The languages are largely mutually intelligible (Haugen, 1990).

Given this, a machine translation system between the two languages should largely focus on *dissemination*, that is the production of text to be post-editted and published, rather than the production of text for *assimilation*, or understanding.

This paper is laid out as follows, first a brief review is given of the design of a shallow-transfer rule-based machine-translation system in the Apertium platform. We then present a section describing how the data for this system was created. Following this, an evaluation is given of the quality of the output of the system and its suitability for the post-edition task, and finally we finish with a discussion and prospectives for future work.

2 Design

The Apertium platform follows a transfer-based machine translation model. A source language text is first morphologically analysed using finite-state transducers. It is then disambiguated for part of speech by a bigram HMM part-of-speech tagger.[1] Subsequently, lexical transfer is performed by the same module that performs structural transfer. Syntactic transfer consists of matching fixed-length patterns of lexical units[2] and performing operations such as insertions, removals and substitutions, along with concordancing. Finally, generation is performed by the same module that performs analysis. Figure 1 shows the main modules of a given system built upon the platform. A more complete description of the platform may be found in Armentano-Oller et al. (2006).

Two models of structural transfer are supported by the platform: a single-stage transfer, where only one set of transfer rules is used, and a three-stage transfer where transfer rules are also used to group words into *chunks*, on which later operations can be performed. The Swedish–Danish pair uses the original, single stage transfer as a result of the closeness of the languages.

[1] The part-of-speech tagger outputs a single disambiguated word, along with both part of speech and any extended morphological information.

[2] A lexical unit is a lemma and its part of speech and morphological information.

J.A. Pérez-Ortiz, F. Sánchez-Martínez, F.M. Tyers (eds.)
Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 27–33
Alacant, Spain, November 2009

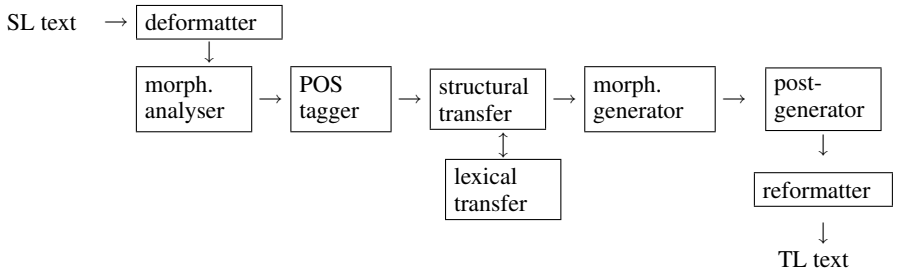

Figure 1: The eight modules of the shallow-transfer machine translation system

3 Development

3.1 Resources

We were able to make use of a number of freely available sources of information for constructing the system. Both Swedish and Danish have free, high coverage spell-checkers available in the *aspell*[3] project. These were used to provide lists of valid word forms for input into the *Extract* (Forsberg et al., 2006) program, which attempts to generate matches between lemmas and inflectional paradigms based on full-form lists and extraction paradigms.

The Swedish, Danish and English Wiktionaries have a fair amount of information regarding both inflectional forms and translations between the two languages.

3.2 Analysis and generation

For analysis and generation of Swedish and Danish, two morphological dictionaries are used, one for each language. Each dictionary was built in a slightly different way due to the differing amounts of information available. In both cases however, the closed categories were described manually.

For Swedish, the Swedish Wiktionary[4] has inflectional tables for many of the words. In order to make use of these, all of the pages in a particular category (for example Nouns) were downloaded in HTML form. The inflectional information was extracted using a variety of scripts and then all the possible paradigms were generated using the `speling-tools`.[5] That is, for each word, one paradigm was created. These were then merged using the same tools such that for each paradigm, any duplicates were removed.

Along with this, we also used *Den stora svenska ordlistan* (DSSO),[6] a free full-form lexicon of Swedish, and entries extracted from a the `aspell` spelling dictionary of Swedish using the *extract* tool.

3.3 Disambiguation

For disambiguation we first chose to train a basic unsupervised bigram part-of-speech tagger using the `apertium-tagger`.[7] Although both tagged corpora and constraint grammars (see Karlsson et al. (1995)) exist for both Swedish and Danish, neither the constraint grammars nor the corpora are free. The training corpora used were the Danish[8] and Swedish[9] Wikipedias respectively.

3.4 Lexical transfer

Despite the closeness of the languages, one of the most labour intensive parts of the work on this pair was the creation of the bilingual dictionary (transfer lexicon). Swedish and Danish are largely mutually intelligible, so there is not much demand for general purposes bilingual dictionaries between the two.

In order to create a dictionary we used several methods. The closed categories, for ex-

[3] `http://aspell.net/`

[4] `http://sv.wiktionary.org`

[5] `http://wiki.apertium.org/wiki/Speling_tools`

[6] `http://dsso.se`

[7] A bigram tagger was chosen as during development there was no support for trigram tagging in the `apertium-tagger`.

[8] `http://da.wikipedia.org`; Access date: 17th September 2009; Filename: `dawiki-20090917-pages-articles.xml.bz2`

[9] `http://sv.wikipedia.org`; Access date: 8th February 2009, filename: `svwiki-20090208-pages-articles.xml.bz2`.

ample pronouns, determiners, prepositions were added by hand, along with some of the open categories. Then, the following semi-automatic methods were used:

- Cognates – The most obvious method for creating bilingual dictionary entries was to look at words which were the same in the two languages, or the same with different orthography. Frequent changes from Swedish to Danish include *ö* → *ø* and *ä* → *æ*. But also, non-orthographic changes such as verb endings in *-a* in Swedish changing to *-e* in Danish.

- Wordlists – We came across a number of free untagged Swedish–Danish wordlists. In order to reuse this information, we first tagged both sides, and created new bilingual dictionary entries where both the part-of-speech and (in the case of nouns) the gender matched up.

- Wiktionary – Along with the previously mentioned inflection tables, the Swedish and English Wiktionaries both have translations from Swedish to Danish. These were mined and treated in a similar way to the wordlists above.

- Wikipedia – For proper names, toponyms etc., we used the method described in Tyers and Pienaar (2008) to extract translations from Wikipedia.

- Probabilistic dictionary – Finally, we trained a statistical machine translation system using Moses (Koehn et al., 2007) on the Europarl (Koehn, 2005) parallel corpus. From this we took the probabilistic lexicon, and performed the same operation as with the wordlists above. In doing this we simply took the most probable translation that was in both the Swedish and Danish monolingual dictionaries.

All bilingual dictionary entries were manually checked and bad entries altered or discarded. It is worth noting that although many more entries were generated than eventually were included in the bilingual dictionary. This was motivated by the low number of entries that we were able to include in the morphological dictionary of Swedish. Bilingual entries which did not have corresponding entries in the Swedish and Danish dictionaries were not included.

3.5 Syntactic transfer

As Swedish and Danish are closely-related languages, there are few translation problems on the syntactic level. We created 17 transfer rules to deal with a number of divergences between the two languages. These were principally motivated by:

- Double definiteness – In most definite NPs in Swedish, both the determiner *den* and the definite form of the noun are used. In Danish when the determiner *den* is present, the definite form of the noun cannot be used. Compare in Swedish *Den stora utmaningen* 'The big challenge' with Danish *Den store udfordring* 'The big challege'.

- Swedish supine verb form – Swedish has a verb form called the supine which can be used with or without an auxiliary and functions somewhat like a past participle. Danish does not have this verbal form, and in its place, often just uses a past participle, for example in Swedish *Han hade blivit trott* 'He had been believed' translated to Danish *Han var blevet troet* 'He was being believed'.

- Changes in auxiliary verbs – There are some verbs in Swedish which do not take the same auxiliary verb in forming periphrastic verb forms as in Danish, for example in Swedish *Två personer har börjat* 'Two people has begun' translated to Danish *To personer er begyndt* 'Two people has begun' (literally, 'Two people are begun').

- Changes in passive formation – In Swedish, certain verbs in the passive (*slå* 'hit', *ligga* 'lie', *anta* 'suppose', ...) must be translated in Danish using an inflected form of the verb *blive* 'become' in the active voice and the past participle.

Other changes made in the transfer rules include changing a passive followed by an infinitive

	Number entries
Monolingual dict. (sv)	5,230 lemmas
Bilingual dict.	6,854 lemmas
Monolingual dict. (da)	10,694 lemmas
Transfer rules (sv → da)	17 rules

Table 1: Status of pair as of version 0.5.0, 9th October 2009

Corpus	WER	PWER	Free rides
Wikipedia	30 %	28 %	38 %

Table 3: Evaluation results for the post-edition task. Free rides are those words which are identical in both the source and target language. Thus although they do not cause a degradation in translation quality, it is relevant to take them into account when evaluating the system.

in Swedish to passive followed by full infinitive in Danish, for example in Swedish *Tros ha dödat* 'Believed to have killed' is expressed in Danish as *Menes at have dræbt*.

3.6 Status

Table 1 gives details of the current status of the system in terms of the number of lemmas in each of the dictionaries and the number of transfer rules. The number of lemmas in the Danish dictionary is greater than the number of lemmas in the Swedish dictionary as a result of a more lax process taken to adding automatically generated entries. It can be expected that some will be erroneous.

4 Evaluation

The evaluation was split into four parts, the first is an evaluation of the coverage of the system with respect to a number of available corpora of Swedish. The second provides a quantitative evaluation using post-edition word error rate (WER) which gives an indication as to how much work a post-editor needs to do in order to achieve an adequate target language translation. The third is a qualitative evaluation which looks at some of the major deficiencies of the system with respect to disambiguation, and lexical and structural transfer. Finally we provide a short comparative evaluation of our system against two proprietary systems.

4.1 Coverage

The vocabulary coverage of the system is calculated over two available corpora. Here coverage is defined as *naïve coverage*, that is for any given surface form at least one analysis is returned. This may not be complete. The first corpus is a database dump of the Swedish Wikipedia,[10]. the second is the Swedish sentences from the EuroParl corpus Koehn (2005). The results are presented in table 2.

4.2 Quantitative

The quantitative evaluation involved the post-edition of 65 machine translated sentences (1,151 words) from the Swedish Wikipedia. The sentences were selected from an article on history, run through the Apertium machine translation system and then a human post-editor corrected the resulting Danish translation.

Both word error rate (WER) and position-independent error rate (PWER) were calculated by counting the number of insertions, substitutions and deletions between the post-editted text and the original translation. The tool used for calculating both WER and PWER was the freely available `apertium-eval-translator`.[11]

The results of this evaluation are shown in table 3 and indicate that the system is still not ready for being used in a post-edition environment.

4.3 Qualitative

Currently the auxiliary, required in Danish but sometimes omitted in Swedish supine verb form is not being added. Thus *Han blivit trott* 'He had been believed' is incorrectly being translated to *Han blevet troet* instead of the correct *Han er blevet troet*.

The pair currently uses `lttoolbox` (Ortiz-Rojas et al., 2005) for both morphological analysis and generation. The package does not currently support productive compounding and as both Swedish and Danish are compounding languages this causes problems for coverage – even if both constituent parts of the compound are in the dictionary they will not be analysed. For ex-

[10] `http://sv.wikipedia.org`; Access date: 8th February 2009; Filename: `svwiki-20090208-pages-articles.xml.bz2`

[11] The package can be downloaded from Apertium SVN, for details see `http://www.apertium.org/`.

Corpus	Running tokens	Known tokens	Coverage
Wikipedia	30,662,861	22,030,690	71.84%
EuroParl	15,531,107	12,499,971	80.48%

Table 2: Naïve coverage for two corpora

ample, the word *universitetsbibliotek* 'university library' is not found, but both *universitet* 'university' and *bibliotek* 'library' are in the dictionaries.

A large part of the errors in the Apertium output are due to the coverage of the dictionaries. This is either directly a result of a word not being translated, or indirectly as in the case of an unknown word causing a transfer rule not to apply, for example *det baltiska havet* 'the Baltic Sea+DEF' being translated as *det *baltiska havet* instead of *det baltiske hav*, with the double definiteness being removed.

4.4 Comparative

There are two existing proprietary machine translation systems online which translate between Swedish and Danish, Gramtrans[12] and Google Translate[13]. Gramtrans is a rule-based system built on top of constraint grammar, and Google Translate is a statistical machine translation system. To compare the results of the three systems, we used the selected 65 sentences and translated them with all of the systems, they were then post-editted and the WER and PWER calculated as in the quantitative evaluation.

Because of time constraints, and because both the Gramtrans and the Apertium translations were very similar, we considered the post-edited Apertium text to be valid also as post-edited Gramtrans text. This gives a small bias against Gramtrans which must be taken into account when making conclusions.

The results of this evaluation are shown in table 5 and show that, in spite of the small bias against Gramtrans, it scores better than Apertium and Google Translate scores produces worse translations than Apertium in terms of WER, but offers a slight improvement in terms of PWER.

As can be seen from table 4, Gramtrans has a much higher-coverage lexicon than the Apertium system, as can be seen from the lower number of unknown words. This leads to fewer errors overall. That being said, the Gramtrans and the Apertium output is very similar, and seem to be following a similar *direct* translation strategy. The Google Translate output, however, is quite different from the other two, and contains artefacts from English in the translation, such as the Swedish word *överfart* (in Danish *overfart*) translated as the less precise English word 'passage'.

[12] `http://www.gramtrans.com`

[13] `http://translate.google.com`

System	Edit distance	WER	PWER
Apertium	350	30	28
Gramtrans	304	26	20
Google	415	35	22

Table 5: Comparative evaluation results for 65 sentences

5 Discussion

We have presented results from the first free-software translator of Swedish to Danish. This is also the first translator between two Germanic languages to be released as part of the Apertium platform. The performance of the system for the postedition task is somewhere in between two commercial competitors.

In terms of future work, we intend to reverse the direction to also translate from Danish to Swedish, to improve the vocabulary coverage, and to improve part-of-speech disambiguation. There is a free constraint grammar for Norwegian (Hagen et al., 2000) available, that could with some conversion work be altered to work as a constraint grammar for Danish (Norwegian Bokmål is even closer to Danish than Swedish is to Danish). Finally, the transfer rules could be expanded to deal with the cases where a supine is used without auxiliary, and a method of handling compound words could be implemented.

	Translation	Gloss
Original	*Det finns en kort överfart vid det baltiska havet vid Helsingborg, på vilket ställe Själland kan ses från Skåne, ett vanligt tillhåll för vikingar.*	There exists a short passage by the Baltic Sea by Helsingborg, on which place Sjælland can be seen from Skåne, a common hangout for Vikings.
Apertium	Det findes en kort *överfart* ved det *baltiska* havet ved Helsingborg, på hvilket *ställe Själland* kan ses fra Skåne, et *vanligt* tilhold før vikinger.	It exists a short *överfart* by the *baltiska* Sea by Helsinborg, on which *ställe Själland* can be seen from Skåne, a *vanligt* order before Vikings.
Gramtrans	Der findes en kort overfart ved det baltiske hav ved Helsingborg, på hvilket sted *Själland* kan ses fra Skåne, et sædvanligt tilhold for vikinger.	There exists a short passage by the Baltic Sea by Helsingborg, on which place *Själland* can be seen from Skåne, a common order for Vikings.
Google	Der er en kort *passage* i Østersøen i Helsingborg, i hvilken plads *Zealand* kan ses fra *Scania*, en regelmæssig tilholdssted for vikingerne.	There is a short *passage* in the Baltic Sea in Helsingborg, in which space/place/seat *Zealand* can be seen from *Scania*, a regular hangout for the Vikings.

Table 4: Comparison of the three systems for a single sentence. Unknown words are marked with *emphasis* and incorrect translations are underlined.

Acknowledgements

Development was funded as part of the Google Summer of Code[14] programme for Michael Kristensen. Thanks to Michael Kristensen for his big work. Many thanks to Thyge Larsen for his assistance with post-edition and evaluation, and to the anonymous reviewers for their useful comments.

References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese–Spanish machine translation. In *Computational Processing of the Portuguese Language, Proceedings of PROPOR 2006*, volume 3960 of *LNCS*, pages 50–59.

Forsberg, M., Hammarström, H., and Ranta, A. (2006). Morphological lexicon extraction from raw text data. *FinTAL 2006*, pages 488–499.

Hagen, K., Johannessen, J., and Nøklestad, A. (2000). A Constraint-Based Tagger for Norwegian. In Lindberg, C.-E. and Lund, S. N., editors, *17th Scandinavian Conference of Linguistics*, volume 19 of *Odense Working Papers in Language and Communication*, pages 31–48. Syddansk Universitet, Odense.

Haugen, E. (1990). *The World's Major Languages*, chapter Danish, Norwegian and Swedish, pages 157–179. OUP.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of MT Summit 2005*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. *ACL Demonstration Session*.

[14] `http://code.google.com/soc/`

Ortiz-Rojas, S., Forcada, M. L., and Ramírez-Sánchez, G. (2005). "Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas". *Procesamiento del Lenguaje Natural*, 35.

Tyers, F. M. and Pienaar, J. (2008). Extracting bilingual word pairs from Wikipedia. *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC08*, pages 19–22.