Translating the Swedish Wikipedia into Danish

Eckhard Bick
University of Southern Denmark
Rugbjergvej 98, DK 8260 Viby J
eckhard.bick@mail.dk

Abstract

Abstract. This paper presents a Swedish-Danish automatic translation system for Wikipedia articles (WikiTrans). Translated articles are indexed for both title and content, and integrated with original Danish articles where they exist. Changed or added articles in the Swedish Wikipedia are monitored and added on a daily basis. The translation approach uses a grammar-based machine translation system with a deep source-language structural analysis. Disambiguation and lexical transfer rules exploit Constraint Grammar tags and dependency links to access contextual information, such as syntactic argument function, semantic type and quantifiers. Out-of-vocabulary words are handled by derivational and compound analysis with a combined coverage of 99.3%, as well as systematic morpho-phonemic transliterations for the remaining cases. The system achieved BLEU scores of 0.65-0.8 depending on references and outperformed both STMT and RBMT competitors by a large margin.

1. Introduction

The amount of information available in Wikipedia differs greatly between languages, and many topics are badly covered in small languages, with short, missing or stub-style articles. This asymmetry can be found between Scandinavian languages, too. Thus, the Swedish Wikipedia has 6 times more text than its Danish equivalent. Robot-created articles have helped to increase the difference to 9:1 in terms of articles, but there are also 3.2 times more edits, indicating a substantial difference in human-authored material, too. In theory, Danes should read Swedish well enough to use both Wikipedias, but in practice this is problematic, especially for young people, and Danes buy Swedish books only in translation. Worse, Danes do not have any active command of Swedish, so they can't *search* in the Swedish Wikipedia. Translating not only the search term, but the whole Wikipedia, will allow in-text search hits, increase readability and permit true integration into the knowledge body of the Danish Wikipedia.

With human translators, even a one-time translation would cost billions of kronor, and it would be impossible to keep up with everyday edits and additions. Machine translation is therefore a sensible solution, if sufficient quality can be achieved. We believe that ordinary statistical machine translation is not the best solution for such a task, not just because of general quality concerns, but also because Wikipedia has a huge lexical spread and covers many subject domains making it difficult to acquire bilingual training data in sufficient quantities. In our own approach, GramTrans (Bick, 2007), we use a Constraint Grammar-based (CG) analysis and context-driven transfer rules. The underlying CG parser (SweGram, http://beta.visl.sdu.dk/visl/sv/parsing/automatic/) features a lexicon-based morphological analysis and ~8,500 tagging and disambiguation rules[1], providing syntactic function tags, dependency trees and a semantic classification of both nouns and named entities.

2. The Translation System (Swe2Dan)

In spite of the relatedness of Swedish and Danish, a one-on-one translation is possible in less than 50% of all tokens. Thus, though lexicon entries with transfer rules account for only 4% of the ca. 107,000 lexemes, but for 53% in *frequency* terms. Verbs stand for 40% of all contextual transfer rules. In the example below, 5 translations for the verb "fräsa" are distinguished by specifying daughter-dependents (D) or dependents of dependents (granddaughters, GD) as subjects (@SUBJ) or objects (@ACC) with certain semantic features, such as human <H>, vehicles <V> or <food>. For closed-class items such as prepositions or adverbs (here: "åt", "iväg", "förbi"), it often makes sense to refer directly to word forms. Negative conditions are marked with a '!'-sign, optional conditions with a '?'.

fräsa_V :hvæse *(to hiss like a cat);*
D1=("åt") GD1=(<H>) D2=(<H> @SUBJ) :vrisse *(to snap at sb)*;
D=(<[HV].*> @SUBJ) D=("(iväg|förbi)") D! =(@ACC) :rase *(tear/speed along)*;
D=(<food.*> @ACC) :stege, :brune, :brase, :lynstege *(to fry)*;
D=(@ACC) D=(<H> @SUBJ) :fræse *(to mill, to cut a material or tool);*

Where necessary, a separate translation CG can change or add tags on top of the SweGram parse. Examples are reflexivity, article insertion or the propagation of number, definiteness and the +human feature to

1 Both the parsing rules and the transfer rules represent a great deal of manual linguistic work, but - once written - are easy to maintain, amend or correct. Cross-language transfer of rules is possible between related languages at the parsing stage, but needs substantial tuning and lexicon support to achieve similar results.

under-specified heads or dependents, or from anaphoric referents to pronouns.

Finaly, compounding and affixation can help assign different translations depending on whether a lexical item is used as first, last (second) or middle part, if necessary together with further conditions:

lock_N (25) :lok, :hårlok *[curl]*; S=(<second>) :låg *[cover]*; S=(NEU) :låg; S=(< first>) :lokke *[luring]*

Due to the rich encyclopaedic lexicon of Wikipedia, out-of-vocabulary words are a particular issue. Because the underlying SweGram parser provides a productive compound analysis, the translator can perform part-for-part translations of coumpounds, using the above-mentioned rules for first and second parts. The second fallback is *transformation* rather than translation, exploiting the likelyhood of shared etymology. Thus, the definite plural noun ending '-orna' will be changed into Danish '-erne', the affixes '-ism' and '-skap' become '-isme' and '-skab', and Swedish 'ö'/'ä' will become Danish 'ø'/'æ'.

3. Evaluation

Lexical coverage was evalutated on a chunk of 144,456 non-punctuation tokens, where the parser classified 7,120 unknown non-name words as "good compounds" and 1,245 as outright heuristic analyses. Swe2dan came up with non-heuristic translations for 99.1% of the compounds and had ordinary lexicon entries for 62.1% of the heuristicals, leaving the rest to the transformation module. For ordinary, parser-sanctioned words[2] the translation lexicon had a coverage of 99.71%, missing out on only 368 words, half of which were left as-is and worked in Danish, too (typically foreign word). Including correct transformations and as-is translations, overall translation coverage was 99.62%.

We evaluated the system on 100 sentences (~1,500 words) from the Leipzig Wortschatz corpus (http://corpora.informatik.uni-leipzig.de/), comparing GramTrans three other systems, Google Translate, Bing Translator and Apertium, all of which maintain open-access user interfaces (accessed 15 March 2014). While Google Translate and Bing Translator rely on statistical machine translation (STMT), open-source Apertium (Tyers et al. 2010) uses rule-based machine translation (RBMT) like GramTrans itself. However, where Apertium uses corpus-trained HMM taggers, GramTrans is rule-based also in its analysis modules.

We measured all systems against both an independent manual translation and best-case edited system translations, using the BLEU (Papineni et al. 2002) and NIST metrics. The external systems were used through their online interfaces, and there was no resource sharing.

In this comparison, GramTrans clearly outperformed all other systems. Apertium performed slightly better than the statistical systems, when measured against one manual translation, but came out last when measured against "self-edit" or "all others".

	Manual reference (1)	edited system reference	multi-reference (all minus self)
GramTrans	**0.645** / 8.515	**0.838** / 9.817	**0.757** / 10.050
Google	0.387 / 6.300	0.645 / 8.361	0.539 / 8.150
Apertium	0.390 / 6.391	0.516 / 7.361	0.468 / 7.418
Bing	0.342 / 6.006	0.600 / 8.064	0.492 / 7.793

Table 1: BLEU/NIST scores

The statistical systems profited relatively more from the inclusion of self-edits, and in relative terms the difference between GramTrans and Google was bigger for BLEU than for NIST in all runs. Since NIST downplays the importance of short/common words and of few-letter differences, it is function words, definiteness and inflexion/agreement that will be affected - all of which are strong areas for rule-based systems.

4. WikiTrans

For the Wikipedia translation we use html-rendered articles, passing on formatting and other meta-information as tags through the translation pipe and reassembling text chunks into full Wikipedia layout at the other end. The translated articles retain a link to the original article version and its date, but look and read just like ordinary Wikipedia articles. Where Danish already has an article on the topic, this will be shown together with the translated Swedish article, with an integrated content index. Thus, users have easy access to subsections, pictures and graphics from both articles. The WikiTrans site (dan.wikitrans.net) uses a modified Lucene search index and presents weighted hits in both titles and text, with content snippets.

Special problems were caused by internal links and names. Thus, links may differ from their actual target titles in both contextual translation and inflexion, so internally the original Swedish link names are used. Names are problematic because it is impossible to achieve good lexicon coverage, and difficult to decide which names to translate. Using NER classification, we translated compound place names and certain types of institutions and organizations, but left person names untouched. For works of art etc., we experiment with a compromise, where the original is maintained, but with a translation in parenthesis.

WikiTrans continually monitors changes in the Swedish Wikipedia, adding new articles as they appear, and retranslating edited ones, at a speed of about 40.000 articles a day on a 40-core computer cluster shared with an English-esperanto sister WikiTrans.

2 I.e. words that the parser could find a lemma and inflection for without the use of heuristics

5. References

E. Bick. 2007. Dan2eng: Wide-Coverage Danish-English Machine Translation. In: B. Maegaard (ed.), *Proceedings of Machine Translation Summit XI, 10-14. Sept. 2007, Copenhagen, Denmark*. pp. 37-43

E. Bick. 2011. WikiTrans, The English Wikipedia in Esperanto. In: *Constraint Grammar Applications*, *Workshop at NODALIDA 2011*. NEALT Proceedings Series, Vol.14, pp.8-16. Tartu: Tartu University Library

F. Karlsson. 1990. Constraint Grammar as a Framework for Parsing Running Text. In: H. Karlgren (ed.), *Proceedings of COLING-90*, Vol. 3, pp.168-173

F. M. Tyers, F. Sánchez-Martínez, S. Ortiz-Rojas, M. L. Forcada. 2010. Free/open-source resources in the Apertium platform for machine translation research and development. In: *The Prague Bulletin of Mathematical Linguistics* No. 93, pp 67-76