A North Sámi to South Sámi machine translation prototype

Lene Antonsen,
Giellatekno,
Tromsø
lene.antonsen@uit.no

Francis Tyers,
Giellatekno,
Tromsø
ftyers@dlsi.ua.es

Trond Trosterud,
Giellatekno,
Tromsø
trond.trosterud@uit.no

Abstract

This paper describes the development of a rule-based machine translation system from North Sámi to South Sámi, and its evaluation in a setting where North Sámi functions as pivot translation via manual translation from Norwegian North Sámi, and thereafter MT to South Sámi.

1 Introduction

The paper presents a machine translation system from North Saami to South Saami. The system is intended to work in a translation setting where North Saami acts as a pivot language, with manual translation from Norwegian into the largest of the Saami languages, and thereafter to offer machine translation service, with postediting, to the other, smaller, Saami languages. On the one hand, the Saami languages are closely related, and therefore lend themselves better to MT than a system translating from Norwegian directly, but on the other hand, the classical problems of translating via a pivot language apply in this case as well.

The paper is structured as follows. Afther looking at previous work and at the languages themselves, we give a presentation of the actual machine translation system. Thereafter comes an evaluation of the system. We presented translated text to translators, alongside with the Norwegian original, without revealing the fact that the texts were actually translated from North Saami.

finally comes a conclusion.

Tyers et al. (2009) Wiechetek et al. (2010) Trosterud and Unhammer (2012)

Babych et al. (2007)

2 Languages

As a Germanic language, Norwegian differs from the Saami languages in numerous respects. It has definiteness as a morphological category, no case, prepositions and verb / particle constructions, and a relatively strict word order, with V2 in main clauses. Norwegian and the Saami languages also show Sprachbund phenomena, though. Most notably, they have the same tense system. Whereas the Saami languages also possess a number of infinite verb construction, the Norwegian embedded clause pattern is in most cases a possible option for the Saami languages as well.

The Saami languages constitute the westernmost branch of the Uralic language family. They possess most of the classical Uralic characteristica: A rich verbal morphology (3 numbers and persons, a rich repertoire of infinite forms), and a medium-size case system with both grammatical and adverbial functions, and no gender distinction. They also show extensive contact with their Germanic neighbours. Many grammatical structures are head-initial constructions, as compared to the classical Uralic pattern, and the tense system is compatible with the North Germanic one.

North and South Saami are not mutually intelligible, due both to linguistic distance and to radically

different prinicples for their literary languages. An analogy would be the difference between English and Frisian.

Grammatical similarities between North and South Saami include the following: Both languages have 3 persons and numbers, they have a similar system for postpositions and verb derivation. Noun phrase syntax is similar, and the case systems are almost identical.

Taking a closer look, there are still differences. In both North and South Saami, negation is expressed with a negation verb which inflects for person and number, but in South Saami, this verb is inflected for Tense as well. South Saami has OV word order and lacks a copula in predicative constructions, whereas North Saami uses VO in simple sentences, and requires a copula. The possession construction is different from the Norwegian one, normally using copula rather than possessor subject + a verb *to have*. In North Saami, the construction is *locative possessor* + *copula* + *possessed in nominative*, whereas the possessor is in the genitive in South Saami.

Although the case systems are similar, North Saami has a locative case that covers the semantic field of both inessive and elative in South Saami, the locative must thus be split into two cases. There is also some differences in case usage, plural objects are accusative in North Saami, but nominative for indefinite and accusative for definite objects in South Saami.

For a machine translation system the orthographic differences imply that apart from person names and acronyms there will be no free rides during the conversion process. The core vocabulary is distinct, even recent loans from the same donor languages are different, and the vocabulary coverage for a working system must thus be very good.

An overview over the differences between the two languages can be found in Sammallahti (1998).

3 Implementation

The translator was implemented using the Apertium platform Forcada et al. (2011). Apertium provides a highly modular set of tools for building rule-based machine translation systems. Apertium language pairs are set up as Unix pipelines, where the typical pipeline consists of:

- deformatting (encapsulating formatting/markup from the engine),
- source-language (SL) morphological analysis with a finite-state transducer (FST),
- disambiguation using a Hidden Markov Model (HMM) and/or Constraint Grammar (CG),
- lexical transfer (word-translation on the disambiguated source),
- lexical selection (choosing the appropriate word out of a set of possible translations),
- one or more levels of finite-state based structural transfer (reordering, and changes to morphological features),
- target-language (TL) generation with an FST
- reformatting (unencapsulating format information)

See Figure 1 for an overview of the modules used in this particular language pair.

3.1 Analysis

Morphological analysis is done on the input using the Helsinki Finite-State Toolkit Lindén et al. (2011). For each surface form, a finite-state transducer returns a set of the possible analyses, where an analysis is a combination of lemma, and a sequence of tags which describe the morphological structure of the surface form.

A Constraint Grammar-based disambiguator then selects the most appropriate analysis for each surface

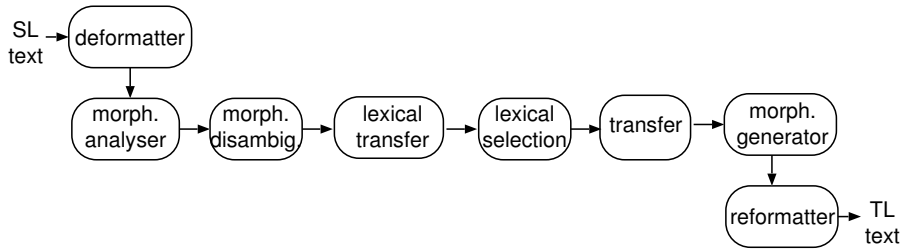

Figure 1: Overview of the translation pipeline.

form according to the context, and assigns to each analysis a syntactic tag denoting its syntactic function (subject, object, main verb, …).

3.2 Transfer

3.2.1 Lexical transfer

Some numbers – kladd

- 3568 general word pairs + 61 proper nouns
- added: 873 special domain word pairs
- added: 10 721 proper nouns

There is no North Saami-South Saami dictionary, so we made one by making word pairs by combining the Norwegian words in a North Saami-Norwegian dictionary and a South Saami-Norwegian dictionary. The word pairs were manually edited and the work revealed many incorrect pairs because of nob-words with more than one meaning. Example is the Norwegian verb *regne*, which means both 'to rain' and 'to calculate', which are two different verbs both in North Saami, *arvit* and *rehkenastit*, and in South Saami *abrodh* and *ryöknedh*. The result of this work was 3568 general word pairs and 61 proper nouns.

Large syntactic differences also have consequences for the bidix, because there are many examples of sme-verb has go be translated into object+verb og adverb+verb in sma. ex: nob 'å trekke lodd' sme *vuorbádit* sma *vuerpiem giesedh*, nob 'å presentere' sme *ovdanbuktit* sma *åvtese buektedh*. I mange tilfeller er grunnen til denne forskjellen at man på sme har fått utviklet flere termer som matcher de norske teremene, mens man på sma fremdeles må forklare innholdet ved hjelp av flere ord.

The bilingual dictionary
general dictionary
general words
special domain

Problems caused by this model: ex from two terms to one term to two terms

Example: nob-1,2 sme-x sma-1,2 example the verbs 'read, count, say', nob: *lese, telle, uttale*, sme: *lohkat* can be used about all three, sma: *lohkedh, ryöknehtidh, jiehtedh*

Little nob-sma lexical resources, and total lack of sme-sma, we have to find words by comparing translated texts (nob-sme with nob-sma)

3.2.2 Lexical selection

3.2.3 Structural transfer

The syntactic differences between North Saami and South Saami are greater than are usually dealt with between related language pairs in Apertium. In order to be able to transfer VO structures to OV more reliably, the transfer phase is split into five parts instead of the three more typically used in Apertium:

- chunker: Chunk input words into groups, e.g. noun groups, verb chains
- interchunk1: Merge chunks which have local coordination, e.g. the sequence [NP x] [CC and] [NP y] is merged into [NP x and y]
- interchunk2: Merges relative clauses and adpositional phrases.

Bilingual dictionary (sme→sma)	15,204
Transfer rules (sme→sma)	57

Table 1: Number of bilingual dictionary entries and transfer rules

Corpus	Tokens	Coverage (%)
Schoolbooks	468,697	92.4 ± 0.58
General		0.0 ± 0.0

Table 2: Vocabulary coverage of the system

- interchunk3: Reorder constituents, e.g. SVO → SOV.
- postchunk: Cleanup

3.2.4 Generation

Generation is done using a finite-state transducer, also compiled using HFST. Each lexical unit which is output by the postchunk module is looked up in the generation transducer and the surface form is output.

4 Evaluation

translation nob-sma which actually is nob-sme-sma (should somehow be in the title?) the evaluators evaluate nob-sma

4.1 Statistics

4.2 Quantitative

4.3 Qualitative

4.4 User feedback

5 Future work

6 Conclusions

reaction from sma linguist in the normgiving organ: she sees the possibility of using MT as a help for discussing terminology and getting the best ones into use

the language socity is used to the impact from the mojority languages, a new controversial(?) thought is the linguistic impact from another minority lanugage

Acknowledgements

Linda Wiechetek, Divvun

References

Babych, B., Hartley, T., and Sharoff, S. (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 29–35.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., and Pirinen, T. (2011). *HFST—Framework for Compiling and Applying Morphologies*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85.

Sammallahti, P. (1998). *The Saami Languages. An Introduction*. Davvi Girji.

Trosterud, T. and Unhammer, K. B. (2012). Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*, number 2013:03, pages 13–26.

Tyers, F., Wiechetek, L., and Trosterud, T. (2009). Developing prototypes for machine translation between two sámi languages. In *Proceedings of the*

13th Annual Conference of the European Association for Machine Translation, EAMT09, pages 120–128.

Wiechetek, L., Tyers, F., and Omma, T. (2010). Shooting at flies in the dark: Rule-based lexical selection for a minority language pair. *Lecture Notes in Artificial Intelligence*, 6233:418–429.