Developing prototypes for machine translation between two Sámi languages

The languages

![Map of Sámi showing the language areas](image)

North (N.) Sámi and Lule (L.) Sámi,
- belong to the Finno-Ugric language family
- are spoken in the north of Norway and Sweden, North Sámi also in Finland
- have 15,000 - 25,000 (N.) and < 2,000 (L.) Sámi speakers
- are heavily inflective and agglutinative
- are largely mutually intelligible

A machine translation system for this pair must be of a quality such that post-editing the output is faster than translating from scratch, since the users will prefer the original to a bad translation.

Existing resources

There are resources for both morphological analysis and disambiguation of N. and L. Sámi
- morpho(phono)logical transducers (lexc and twocl)
- a North Sámi Constraint Grammar parser

guolli AIGI "fish N"; has the accusative plural form guolli. Both consonant gradation (-lī vs. -lē) and diphthong simplification (oː > u) take place. The two-level compiler handles both.

Vx:0 ↔ Vow + Cns:+1 (...) X5: ;
where Vx in (o a) ;

In the diphthong simplification rule, X5 marks that the second vowel (o a) in a diphthong has to be simplified if the suffix contains an i.

Morphological and syntactic disambiguation is handled by the North Sámi Constraint Grammar parser. By means of context rules both morphological and syntactic analyses are removed except for the last reading.

Rule-based machine translation

<table>
<thead>
<tr>
<th>SL text</th>
<th>de-formatter</th>
<th>morph. analyser</th>
<th>POS tagger</th>
<th>structure trans</th>
<th>morph. gen.</th>
<th>post-generator</th>
<th>re-formatter</th>
<th>TL text</th>
</tr>
</thead>
</table>

Aside from the main modules in the Apertium system (morphological analysis, HMM part-of-speech tagger and structural/lexical transfer) the N-L. Sámi system also took advantage of:
- The N. Sámi constraint grammar – which made the HMM tagger largely redundant and including syntactic and semantic analysis for use in the structural transfer
- An bilingual transfer lexicon

Construction of the transfer lexicon

A transducer applied regular changes to turn N. Sámi lemmata into L. Sámi candidates (e.g. 8 vs. 9 in beavrus (‘sum’) -> beavrus). Words recognised by the L. Sámi morphological analyser with the same POS as the input word were accepted, words not recognised were revised.

Structural differences between North and Lule Sámi

- **Case differences**: N. Sámi locative -> L. Sámi inessive or elative
- **Negation**: the N. Sámi negation verb can inflect for tense, L. Sámi expresses tense by means of the main verb
- **Word order**: L. Sámi allows for a number of SOV (subject object verb) constructions whereas N. Sámi prefers SVO (1)

Example (L. Sámi)

dál smidá (both ‘now’) (3rd person plural)

Statistical machine translation

We used the
- Moses decoder,
- the word aligner GIZA++
- the srlm language model

Language models:
- For Lule Sámi we made both an unfactored and a factored (wrt POS) trigram language model on our Lule Sámi corpus, 278,000 words (120,000 words New Testament; 106,000 fact; 39,000 fiction).
- For N. Sámi we made both a factored and an unfactored model (NT corpus) and an unfactored model (the whole corpus)

Results

**RBMT:**

The translation of 16 Wikipedia test sentences were compared to a manual reference translation. Structural transfer is unproblematic with a few exceptions, the choice of lexical tags and the lexical choices are the bigger challenge.

<table>
<thead>
<tr>
<th>Type of deviation</th>
<th>Example (L. Sámi)</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-to-many relations</td>
<td>dália vs. dál (both ‘now’)</td>
<td>two different forms of one word</td>
</tr>
<tr>
<td>Tag inconsistencies</td>
<td>teşråddjáddje (‘self-governed’)</td>
<td>is analysed both as a deverbal form and a lexicalised adjective</td>
</tr>
<tr>
<td>POS asymmetries</td>
<td>gallådíddje (‘belonging’)</td>
<td>is analysed as a derived verb form</td>
</tr>
<tr>
<td>CT disambiguation error</td>
<td>lekkert (‘intuitive’ ‘to be’)</td>
<td>should be li (3rd person plural)</td>
</tr>
<tr>
<td>Lexical matters</td>
<td>tjåhpje vs. smidé (both ‘clever’)</td>
<td>can be used synonymously in certain contexts</td>
</tr>
<tr>
<td>Word order</td>
<td>määm i aymmnl vs. mantulmmn la (‘which is the purpose’)</td>
<td>SVO vs. SOV</td>
</tr>
</tbody>
</table>

**SMT:**

The same 16 test sentences were translated by
- a factored model (curriculum corpus)
- an unfactored model (curriculum corpus)
- an unfactored model (NT corpus)
- an unfactored model (the whole corpus)

Figures 1 and 2 show the BLEU score for each model and the word order.

Conclusions

The rich morphology and especially the paucity of parallel corpora for Sámi make SMT less suited for MT between North and Lule Sámi, despite the close relationship between the two languages. Therefore RBMT is the best approach for this language pair.

Apertium copes well with the structural transfer from North to Lule Sámi. Improving the lexicon and the coverage of the structural transfer rules will be the next steps forward for our RMBT model.