Developing prototypes for machine translation between two Sámi languages

Francis Tyers
ftyers@prompsit.com
Universitat d’Alacant, Spain

Linda Wiezchek
linda.wiezchek@uit.no
Tromsø University, Norway

Trond Trosterud
trond.trosterud@uit.no

Rule-based machine translation

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

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)

\[\text{Anne råhkä biepmu. (N. Sámi)}\]
\[\text{Anne biebmov dähkä. (L. Sámi)}\, \text{‘Anne makes food.’}\]

The SOV rule captures the pattern (subject, verb, object) and outputs them in the order subject–object–verb by reordering the chunks.

Statistical machine translation

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

\[\text{Vx:0} \leftrightarrow \text{Vow} \text{ Cns:} + 1 \text{ (...) X5:} \text{; where Vx in (e o a);}\]

In the diphthong simplification rule, X5 marks that the second vowel (e 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.

Results

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)

The rich morphology and especially the paucity of parallel corpora for Sámi make SMT less suited for languages like the Sámi languages.

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 RBMT model.