Apertium-IceNLP: A rule-based Icelandic to English machine translation system

Martha Dís Brandt, Hrafn Loftsson, Hlynur Sigurþórsson[1]
Francis M. Tyers[2]

[1]School of Computer Science
Reykjavik University

[2]Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant

EAMT, May 31st, 2011

Outline

Outline

Rule-based MT

Apertium-IceNLP: Why?

- No freely available large parallel corpus exists for Icelandic $\Rightarrow$ SMT not an option.
- Apertium: an open-source rule-based shallow-transfer MT platform
- NLP tools available for Icelandic, e.g. IceNLP (open-source)
- $\Rightarrow$ Apertium-IceNLP
- **Research question**: Can we obtain better translation quality by such an hybrid system, as opposed on relying solely on Apertium modules?

Outline

Apertium

License

The whole platform, both program and data, are licensed under GPL (`http://www.apertium.org`).

The Apertium pipeline (main modules)

- A morphological analyser (performs tokenisation and lemmatisation as well)
- A bigram/trigram Hidden Markov Model (HMM) PoS tagger
- A lexical selection module (based on Constraint Grammar)
- A transfer module: lexical and structural transfer
- A morphological generator

Apertium

Language pair specifics

- A monolingual source language (SL) dictionary
- A monolingual target language (TL) dictionary
- A bilingual SL-TL dictionary
- Transfer rules

Outline

IceNLP

License

The whole platform, both program and data, are licensed under LGPL (`http://icenlp.sourceforge.net`)

Main modules

- A tokeniser/sentence segmentiser
- IceMorphy: A morphological analyser/unknown word guesser
- IceTagger: A linguistic rule-based PoS tagger
- TriTagger: A bigram/trigram HMM PoS tagger
- Lemmald: A mixed-method lemmatiser
- IceParser: A shallow (finite-state) parser

PoS-tagging Icelandic

Tagset

A large detailed tagset consisting of 700 tags from the Icelandic Frequency Dictionary (IFD) corpus

Tagging accuracy

- Data-driven taggers, including state-of-the-art HMM taggers, have not achieved high accuracy
- IceTagger has achieved higher accuracy than the data-driven taggers
- IceTagger+TriTagger (HMM): The state-of-the-art PoS tagger for Icelandic (92.5%)

Outline

Apertium-IceNLP

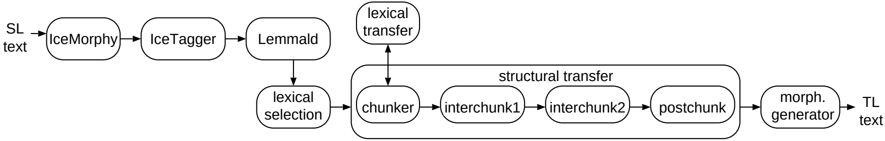
The idea

- To integrate existing tools from IceNLP into the Apertium pipeline in order to develop a prototype of an Icelandic-English MT system.
- (The "pure" Apertium version was developed in parallel)

Benefits

- Use state-of-the-art PoS tagger, IceTagger, in the pipe-line
- Information about syntactic functions can be of help in the translation process. IceParser can potentially be used.

Apertium-IceNLP



IceNLP enhancements

Which ones?

- Mappings from the output of IceTagger to the output generated by the Apertium tagger
- Daemonising IceNLP: Client-Server functionality was added to IceNLP.

IceNLP enhancements

Mappings – Examples

```
[TAGMAPPING]
nkfng <n><m><pl><nom><def>
[LEMMA]
vera <vblex><actv> <vbser>
[MWE]
að_einhverju_leyti <adv>
```

IceNLP enhancements

Client-Server

- **IceNLPServer**: A server which contains an instance of IceNLP in memory. A daemon which runs in the background. Can serve requests to the modules in IceNLP without any loading delay.
- **IceNLPClient**: A client for communicating with the server. Reads from standard input and writes to standard output (as the Apertium modules).
  - We replaced the Apertium tokeniser/morphological analyser/lemmatiser and PoS tagger with IceNLPClient.

IceNLP enhancements

Client-Server

```
echo "Hún er góð" | RunClient.sh

^Hún/hún<prn><p3><f><sg><nom>$
^er/vera<vbser><pri><p3><sg>$
^góð/góður<adj><pst><f><sg><nom><sta>$
```

The bilingual dictionary

is-en bilingual dictionary

No is-en dictionary publicly available when this project was initiated.

Built in three stages

- Populated with entries spidered from the Internet (Wikipedia, Wiktionary, Freelang, etc.). Around 5,000 entries.
- A wordlist acquired from an individual, Anton Karl Ingason. Around 6,000 entries.
- A wordlist obtained from the dictionary publishing company Snara. Around 18,000 entries.
- In total, the bilingual dictionary currently contains about 22,000 entries.

Transfer rules

Split into four stages

1. Local reordering and chunking (78 rules).
2. Produces chunks of chunks, e.g. chunking relative clauses into noun phrases (3 rules).
3. Longer distance reorderings, e.g. constituent reordering (26 rules).
4. Cleanup operations and insertion of the indefinite article (5 rules).

Outline

Evaluation

Competing systems

- **Apertium-IceNLP** (`http://nlp.cs.ru.is/is-en.htm`)
- **Apertium** (`http://www.apertium.org/`)
- **Tungutorg** – a proprietary rule-based system (`http://www.tungutorg.is`)
- **Google Translate** (`http://translate.google.com`)

Test data

- About 400 sentences from the Icelandic Wikipedia
- Run through the four systems, post-edited by a bilingual person to obtain reference sentences

Evaluation

Evaluation measures

- **Word error rate** (WER): Percentage of the TL words that require correction (substitutions, deletions and insertions)
- **Position-independent word error rate** (PER): Similar to WER, except that PER does not penalise correct words in incorrect positions
- Computed using the `apertium-eval-translator` tool

Evaluation

Results: Icelandic to English

MT system	WER	PER
Apertium-IceNLP	50.6%	40.8%
Apertium	45.9%	38.2%
Tungutorg	44.4%	33.7%
Google Translate	36.5%	28.7%

Table: Word error rate (WER) and position-independent error rate (PER) over the test sentences

Outline

Error analysis

Development data created

- From the largest online Icelandic newspaper *mbl.is*
- Translated about 1700 files with Apertium-IceNLP and randomly selected 50 for analysis
- 6 error catagories created

Error analysis

Error category	**Freq.**	**%**
Missing from the bidix	912	60.7%
Need further analysis	414	27.5%
Multiword expressions	90	6.0%
Abbreviations and initials	31	2.1%
More sophisticated patterns	31	2.1%
Other	24	1.6%
Total	**1502**	**100%**

Table: Error categories and corresponding frequencies.

Future work

Only use IceTagger from IceNLP?

- MWEs and lemmata are an integral part of the Apertium morphological analyser
- It may produce better results to only use IceTagger for PoS tagging but rely on Apertium for tokenisation, morphological analysis and lemmatisation.
- IceTagger then needs to be changed to receive the necessary information from the morphological component (lt-proc) of Apertium.

Outline

Conclusion

- *Apertium-IceNLP* is the first system in which the whole morphological and tagging component of Apertium is replaced by modules from an external system.
- *Apertium-IceNLP* does neither perform as well as *Google Translate* nor as *TunguTorg*.
- *Apertium-IceNLP* also performs worse than the is-en system based solely on Apertium modules.
  - May be due to the fact that MWEs and lemmata are an integrated part of the Apertium morphological analyser whereas IceNLP does not handle them adequatly.
  - This will be verified in future work.