1 Introduction

Breton is a Celtic language of the Brythonic branch largely spoken in Brittany in the north-west of France. Historically it was spoken in the north-west corner part of Brittany, Breizh-Izel (Lower Brittany). In Breizh-Izel (Higher Brittany), traditionally Romance languages are spoken.

2 Components

The system is based on Apertium (http://www.apertium.org/), a free/open-source rule-based machine translation platform.

2.1 Morphological analysis

The Breton morphological analyser returns, for every Breton word, the possible lexical forms (analyses) of the word. For the sentence ‘Gallouit a mon oder anv daore.’ (‘I can do that [thing].’)

Figure 5: Output of morphological analysis with a finite-state transducer

The analyser has a coverage of between 87–90% over two free corpora of Breton (Wikipedia and Breizh). The output in Figure 5 is generated by two rules. The first takes a sequence of an infinitive verb, followed by a verbal particle and a form of the auxiliary ‘to do’ and outputs the infinitive verb conjugated according to the auxiliary. The second rule (see Figure 9) takes a sequence of determiner followed by a noun and a demonstrative adverb and outputs a demonstrative determiner followed by the noun.

2.2 Part-of-speech tagging

The part-of-speech tagging module for the system is based on two technologies, the first is Constraint Grammar, which uses linguistic-rules to disambiguate morphologically ambiguous words based on sentence context. The second is a bigram HMM part-of-speech tagger.

2.2.1 Constraint grammar

The Breton constraint grammar has been written manually and contains 206 rules for disambiguating Breton sentences.

Figure 6: Morphological disambiguation with constraint grammar

An extract from the constraint grammar rule set:

# ex. “Dav e vije”
HUNCHTJ34: Part ID (58 VerbiFin);
# ex. “Ober a ra gwestell”
HUNCHTJ35: Part ID (58 VerbiFin);
...

2.2.2 HMM-based tagger

The HMM-based tagger was trained on a database dump of the Breton Wikipedia (http://br.wikipedia.org) and chooses a single analysis where the constraint grammar does not perform a complete disambiguation.

2.3 Bilingual dictionary

The bilingual dictionary, or transfer lexicon contains mappings between lemmas, parts-of-speech and other tags. For example, to indicate to the transfer stage that gender and number need to be inserted, or to indicate a change in a feature.

Figure 7: Extract from the bilingual dictionary

2.4 Transfer rules

The structural transfer process is split into three parts. Rules are written in XML (see example in Figure 9).

2.4.1 Chunker

Local transfer operations and chunking are performed by the first stage. The output in Figure 8 is generated by two rules. The first takes a sequence of an infinitive verb, followed by a verbal particle and a form of the auxiliary ‘to do’ and outputs the infinitive verb conjugated according to the auxiliary. The second rule (see Figure 9) takes a sequence of determiner followed by a noun and a demonstrative adverb and outputs a demonstrative determiner followed by the noun.

Figure 8: Output from the the first transfer stage

2.5 Morphological generator

The morphological generator for French takes a sequence of lexical forms and generates the appropriate surface-forms.

Figure 9: Rule to translate a demonstrative noun phrase

3 Evaluation

The evaluation used Word error rate (WER) and position-independent word error rate (PER). A corpus of 398 sentences (5,804 words) was extracted from the Breton-Bre archives. Sentences were extracted fitting the following conditions. No unknown words and between 5–30 words long.

<table>
<thead>
<tr>
<th>Word error rate and position-independent word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word error rate and position-independent word error rate</td>
</tr>
</tbody>
</table>

Table 2: Word error rate and position-independent word error rate

The big difference in the scores for WER and PER is because only local reordering is performed, constituent reordering is reserved for simple phrases.

4 Future work

The performance of the system could be improved by:

- Improving coverage
- Better source language disambiguation
- Lexical selection
- Deeper transfer

Acknowledgements

Funded by: Grup Transdicsne, Ofis ar Brezhoneg, and Prompt Language Engineering. Many thanks to the Ofis ar Brezhoneg, in particular its director Fulup Jakez for his work on the system.