1 Introduction

2 Resources

As for many less-resource language pairs, there is very little aligned bilingual text.

- Parallel corpora for Breton-French anything are far and few between.
- Small parallel corpus from the Olis at Brezhonk
  - Breton and French.
  - Short segments from a translation memory
  - Large in the domains of tourism, localisation and education
- 31,000 lines
- Approximately 285,000 Breton words and 282,073 French words

3 Two machine translation systems

A rule-based MT system for Breton-French is currently being developed inside the Apertium project (http://www.apertium.org/). The current status of the Breton-French system is as follows: the system has a morphological analyser for Breton with approximately 11,000 lemmata (approx. 85% coverage), a bilingual dictionary with 10,797 part-of-speech tagged correspondences, and a very small number of transfer rules (e.g. for concordance and re-ordering within noun phrases, verb conjugation and preposition insertion). In the morphological analyser, there are two kinds of paradigms (parap), the first for specifying the initial consonant mutations, the second for listing all the morphological forms of a given word and their analyses. For verbs, one combination of lemma and paradigm can generate between 37 surface forms (for unmutating initial consonants) and 193 (for mutating initial consonants). Examples for the verbs labour ‘to work’ and kaenarn ‘to confirm’, including inflectional paradigm labourograf ‘to work’ and mutation paradigm (initial-ia) can be found below:

\[
\begin{align*}
\text{name} & \quad \text{parap} & \quad \text{parap} & \quad \text{name} \\
\text{labour} & \quad \text{labourograf} & \quad \text{labourograf} & \quad \text{labour} \\
\text{kaenarn} & \quad \text{kaenarn} & \quad \text{kaenarn} & \quad \text{kaenarn} \\
\end{align*}
\]

Figure 5 Example of morphological analyser entries for two verbs

Figure 6 shows two bilingual lexicon entries for verbs. The bilingual lexicon specifies correspondences between lemmata and parts of speech.

\[
\langle \text{part} \rangle \langle \text{parap} \rangle \langle \text{parap} \rangle \langle \text{name} \rangle \\
\text{mignon} & \quad \text{mignon} & \quad \text{mignon} & \quad \text{mignon} \\
\text{amis} & \quad \text{amis} & \quad \text{amis} & \quad \text{amis} \\
\end{align*}
\]

Figure 6 Bilingual dictionary entries

4 Extending the parallel corpus

To try and alleviate the problem of low coverage of the training data, the resources available in the nascent rule-based system were used. Two approaches were taken.

- Simulating the bilingual transfer lexicon from the system to the end of the training data. This consisted of 10,797 lemmata.
- Automatically generating appropriate mappings between all of the surface forms of the given lemmata in the dictionaries of this system. This consisted of 116,514 mappings of inflected Breton forms to inflected French forms.

In order to generate the surface-form mappings, an expansion of all possible surface forms was taken, along with analyses in the Breton morphological analyser. This was then passed through the rest of the Apertium pipeline in order to produce all of the translations of surface forms in French. This provides a bilingual inflected wordlist (see figure 7). Entries for verbs are generated, where appropriate (e.g. finite verb tenses) with the corresponding subject pronoun in French, and Breton tenses which are not found in French are converted into French tenses (e.g. past habitual is converted to imperfect)

\[
\begin{align*}
mignon,\text{amis} & \quad \text{mignon,amis} \\
mignon,\text{amis} & \quad \text{mignon,amis} \\
\text{dus,retarde} & \quad \text{dus,retarde} \\
\text{dus,retarde} & \quad \text{dus,retarde} \\
\end{align*}
\]

5 Evaluation and error analysis

As expected, the number of unknown words decreases when the bilingual lexicon is added to the training data, and even more so when the fully expanded bilingual lexicon is added. The rise in BLEU is probably also due to side effects such as a better word alignment and a better French context available to the language model scoring.

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
<th>BLEU</th>
<th>Phrases</th>
<th>Unknown words</th>
</tr>
</thead>
<tbody>
<tr>
<td>system 1</td>
<td>word “natural”</td>
<td>0.10</td>
<td>294</td>
<td>913</td>
</tr>
<tr>
<td>system 2</td>
<td>baseline SMT</td>
<td>0.29</td>
<td>800k</td>
<td>623</td>
</tr>
<tr>
<td>system 3</td>
<td>uninflected dictionary</td>
<td>0.30</td>
<td>807k</td>
<td>562</td>
</tr>
<tr>
<td>system 4</td>
<td>inflected dictionary</td>
<td>0.36</td>
<td>843k</td>
<td>531</td>
</tr>
</tbody>
</table>

Example 1

- mignon, amis
  - output de cleure par syllable
  - hyphenation tool

Example 2

- Vous devez vous connecter a toutes les sources de donnees.
  - Devez connexion de données : les daziurhi.
  - Vous devez se connecter tous les daziurhi de données.

6 Future work

Improving this work, it might also be possible to try to learn probabilities for the rule-based created phrase pairs as in Koehn and Knight (2000).

Another option would be to try and create “expanded” phrases based on chunks extracted from a bilingual corpus. For example if you have want nomen t’m, “sur le toit de la maison” (on the roof of the house), it would be fairly straightforward to generate all possible morphological combinations, viz. sur le toit de la maison, war nommen on t’em, “sur les toits des maisons”, war tenn on t’em, “sur le toit des maisons” respectively.

It is also worth noting that at present the Breton–French lexicon in Apertium has only one (generally the most frequent) translation per word. It would be feasible to generate more than one entry per word, and then score these on language models.

7 Conclusion

The method presented here is knowledge-light, requiring only a morphological analyser, bilingual dictionary and some very basic transfer rules (for verb conjugation) and could be applied to other under-resourced language pairs to improve the coverage of a statistical system where little parallel data is available.

Acknowledgements

Many thanks to the Olis at Brezhonk, in particular its director Filip Jakub.

Thanks also to everyone at Prompsit Language Engineering, Mikel L. Forcada, and two contributors who do not wish to be named.


Figure 7 ‘Expanded’ dictionary entries