Comparing Rule-Based and Data-Driven approaches to Spanish-to-Basque Machine Translation

Gorka Labaka1, Nicolas Stroppa1, Andy Way1, Kepa Sarasola1
1 - IXA NL group
University of the Basque Country
(jibiang, kepa.sarasola@ehu.es)
2 - National Centre for Language Technology
Dublin City University
{nstroppa, away}@computing.dcu.ie

**Basque**
- Basque is a morphologically-rich agglutinative language.
- Translating to Basque requires a huge generation of morphological information.
- Some Spanish words, like prepositions or articles, correspond to Basque suffixes. In case of ellipsis more than one suffix can be added to the same word.
- There are few corpora available and they are smaller than corpora available for other languages.

**MATREx SYSTEM**
- MATREx system is a Data-Driven system.
- MATREx has a language-independent design.
- SMT aligned phrases are enriched with linguistically motivated aligned chunks.
- Spanish is divided in chunks using Freeling shallow parser.
- Basque is divided in chunks using Eusung.

**MATRIN SYSTEM**
- MATRIN is a classical transfer system.
- Divided in three main steps: analysis, transfer and generation.
- It has been specifically designed for Spanish-Basque language pair.

**Training and development corpora**:
- Both are extracted from Consumer corpus.
- A collection of 1036 Spanish articles from the Consumer Eroski magazine and their translation into Basque.

**Automatic evaluation**
- For automatic evaluation we have calculated the BLEU and NIST metrics, on both in-domain (ConsumerTest) and out-domain (ElibTest) test corpora.
- Because of the specific nature of Basque, as pointed above, we perform two types of evaluation:
  - Word-based evaluation (WB)
  - Morpheme-based evaluation (MB)

**Human evaluation**
- For human evaluation we have calculated the edit-distance metric (Przybocki et al., 2006), also called HTER or Human-targeted Translation Error Rate (Snover et al., 2006), which is defined as the number of modifications a native Basque translator has to make so that the resulting edited translation is an easily understandable Basque sentence that contains the complete meaning of the source sentence.
- As in automatic evaluation we have calculated the metrics word based (WB) and morpheme based (MB).

**Some statistics from these corpora:**

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Sentences</th>
<th>Spanish words</th>
<th>Basque words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training corpus</td>
<td>51,940</td>
<td>976,720</td>
<td>786,700</td>
</tr>
<tr>
<td>Development</td>
<td>1,392</td>
<td>24,755</td>
<td>19,978</td>
</tr>
<tr>
<td>ConsumerTest</td>
<td>1,301</td>
<td>34,231</td>
<td>27,278</td>
</tr>
<tr>
<td>ElibTest</td>
<td>1,503</td>
<td>35,783</td>
<td>28,857</td>
</tr>
<tr>
<td>ConsumerTestHuman</td>
<td>50</td>
<td>746</td>
<td></td>
</tr>
<tr>
<td>ElibTestHuman</td>
<td>50</td>
<td>692</td>
<td></td>
</tr>
</tbody>
</table>

**MatrEx Hybrid System Architecture**
- Tuning MatrEx to Basque: due to Basque peculiarities slight changes are included.
  1) Basque text is segmented into morphemes, for example: etxeko
- In the one of the house/ /house-of=house/ /of=one-of household/ /the one of the household/ /towards the one of the house/.
  2) After translation a postprocessing stage is necessary. To carry out the generation we have reused Matrin’s lexical generation module.

**Corpora**
- Training and development corpora: Both are extracted from Consumer corpus.
- A collection of 1036 Spanish articles from the Consumer Eroski magazine and their translation into Basque.
- Contriadictory results for automatic and human evaluations, consistent with the findings of (Callison-Burch et al., 2006):
- The automatic metrics indicate that the data-driven system outperforms the rule-based system on the in-domain data.
- HTER “in-domain”: 14.3 points for the word-based evaluation and 10.5 points for the morpheme-based evaluation.
- HTER “out-domain”: 31.4 points for the word-based evaluation and 41.4 points for the morpheme-based evaluation.
- However, let us to stress that
- Matrin has been specifically developed and designed to translate from Spanish to Basque.
- MatrEx is generic and the cost of adapting it to Spanish-Basque translation is orders of magnitude lower.
- Next steps:
  - Building a hybrid system upon the respective strength of both approaches.
  - Investigating automatic evaluation metrics that would be more suited to the evaluation for morpheme-based translation.

**Conclusions**
- A new translation scheme based on morphemes instead of words.
- The automatic metrics indicate that the data-driven system outperforms the rule-based system on the in-domain data.
- BUE: difference of 1.68 points (27% relative increase) for word-based and 2.47 points (21% relative increase) for morpheme-based.
- On the contrary, the human evaluation indicates that rule-based system outperforms the data-driven approach for both corpora, irrespective of the corpus.
- HTER “in-domain”: 14.3 points for the word-based evaluation and 10.5 points for the morpheme-based evaluation.
- HTER “out-domain”: 31.4 points for the word-based evaluation and 41.4 points for the morpheme-based evaluation.
- However, let us to stress that
- Matrin has been specifically developed and designed to translate from Spanish to Basque.
- MatrEx is generic and the cost of adapting it to Spanish-Basque translation is orders of magnitude lower.
- Next steps:
  - Building a hybrid system upon the respective strength of both approaches.
  - Investigating automatic evaluation metrics that would be more suited to the evaluation for morpheme-based translation.