TANDO: A Corpus for Document-level Machine Translation
LREC 2022
Harritxu Gete\textsuperscript{1}, Thierry Etchegoyhen\textsuperscript{1}, David Ponce\textsuperscript{1}, Gorka Labaka\textsuperscript{2}, Nora Aranberri\textsuperscript{2}, Ander Corral\textsuperscript{2}, Xiaiber Saralegi\textsuperscript{3}, Igor Ellakuria Santos\textsuperscript{4} and Maite Martin\textsuperscript{5}
\textsuperscript{1}Vicomtech Foundation, \textsuperscript{2}IXA taldea, \textsuperscript{3}Elhuyar, \textsuperscript{4}ISEA, \textsuperscript{5}Ametzagaiña

Motivation
- Parallel corpora that include contextual information are scarce or non-existent for some languages (e.g., Basque-Spanish).
- Contrastive datasets are needed to overcome the lack of sensitivity of standard MT metrics to improvements at the extra-sentential level.
- TANDO: a multi-domain corpus for Basque-Spanish document-level NMT that includes parallel and contrastive datasets.

Base Corpora
- Parallel dataset with context-level information.
- Domains:
  - EHURAC: a collection of classic books aligned at sentence level.
  - EITB: a corpus of news produced by the Basque public broadcaster EITB.
  - OPENSUBS: a parallel corpus of subtitles, covering a large number of genres.
- Corpus preparation process:
  - Document-level alignment.
  - For each aligned document pair: sentence splitting, alignment and filtering.
  - Context block: a sequence of \( n \) contiguous aligned sentence pairs (\( n > 1 \)).

Contrastive Corpora
- Discursive phenomena in Basque to Spanish translation:
  - Pronouns and some adjectives and nouns are marked for gender in Spanish, but not in Basque.
  - In Basque, the use of formal expressions is widespread for both formal and informal registers.
- Domains:
  - LITERATURE: books collected from Gutenberg and Elejandria repositories.
  - TED TALKS: Basque-Spanish 2020 v1 dataset.
- Contrastive block:
  - Sentence in Basque containing an ambiguous word in terms of gender or register.
  - Context of up to 5 preceding sentences with relevant information for translation.
  - Reference translation in Spanish.
  - Contrastive translation, created by switching the gender/register of the target word.
- Two sets of 300 blocks:
  - GDR-SRC+TGT: gender-sensitive phenomena with disambiguating information in both source and target data.
  - COH-TGT: gender and register discursive coherence on the target side.
- Balanced domain, gender and register alternations.

Models
- Baseline Transformer-base models.
- Context-aware models:
  - Input extension: token-separated sentence concatenation.
  - Parameters are initialized with the trained baseline models.
- Variants:
  - Using source or target context.
  - Using different context sizes (1 vs. 5 sentences).

Table 1: Corpora statistics ([\#sentence pairs])

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>TRAIN (MIN/\MAX/AVG)</th>
<th>DEV (MIN/\MAX/AVG)</th>
<th>TEST (MIN/\MAX/AVG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHURAC</td>
<td>513.64 (1/190/90)</td>
<td>1009 (10/190/20)</td>
<td>2024 (10/190/20)</td>
</tr>
<tr>
<td>EITB</td>
<td>472.96 (2/198/3)</td>
<td>1027 (5/14/6)</td>
<td>2017 (5/14/6)</td>
</tr>
<tr>
<td>OPENSUBS</td>
<td>785.478 (10/50/42)</td>
<td>1057 (25/50/46)</td>
<td>2055 (10/50/42)</td>
</tr>
<tr>
<td>MERGED</td>
<td>1,535,126 (2/198/5)</td>
<td>3953 (5/49/12)</td>
<td>4663 (5/49/13)</td>
</tr>
</tbody>
</table>

Metrics Results

<table>
<thead>
<tr>
<th>MODEL</th>
<th>MERGED</th>
<th>LITERATURE</th>
<th>PARLIAMENT</th>
<th>TED</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:SRC</td>
<td>23.3%</td>
<td>54.8%</td>
<td>28.2%</td>
<td>60.3%</td>
</tr>
<tr>
<td>1:TGT</td>
<td>23.1%</td>
<td>54.8%</td>
<td>27.9%</td>
<td>60.1%</td>
</tr>
<tr>
<td>5:SRC</td>
<td>21.6%</td>
<td>52.9%</td>
<td>25.9%</td>
<td>58.1%</td>
</tr>
<tr>
<td>5:TGT</td>
<td>21.6%</td>
<td>53.5%</td>
<td>26.3%</td>
<td>58.4%</td>
</tr>
</tbody>
</table>

Table 2: Metrics results for Spanish to Basque translation

Contrastive Results

<table>
<thead>
<tr>
<th>MODEL</th>
<th>MERGED</th>
<th>LITERATURE</th>
<th>PARLIAMENT</th>
<th>TED</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:SRC</td>
<td>31.7%</td>
<td>50.0%</td>
<td>39.2%</td>
<td>62.2%</td>
</tr>
<tr>
<td>1:TGT</td>
<td>31.9%</td>
<td>52.3%</td>
<td>39.2%</td>
<td>62.2%</td>
</tr>
<tr>
<td>5:SRC</td>
<td>29.9%</td>
<td>51.4%</td>
<td>38.9%</td>
<td>61.8%</td>
</tr>
<tr>
<td>5:TGT</td>
<td>29.9%</td>
<td>51.4%</td>
<td>38.9%</td>
<td>61.8%</td>
</tr>
</tbody>
</table>

Table 3: Percentage of correct gender answers on GDR-SRC+TGT

Conclusions
- TANDO: a suitable corpus to train and evaluate document-level MT models.
- Simple context-aware variants outperform sentence-level baselines.
- Corpus shared under Creative Commons CC-BY-NC-SA 4.0 license: https://github.com/Vicomtech/tando