TANDO: A Corpus for Document-level Machine Translation

LREC 2022

Harritxu Gete[1], Thierry Etchegoyhen[1], David Ponce[1], Gorka Labaka[2], Nora Aranberri[2], Ander Corral[3], Xabier Saralegi[3], Igor Ellakuria Santos[4] and Maite Martin[5]

[1]Vicomtech Foundation, [2]IXA taldea, [3]Elhuyar, [4]ISEA, [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:
  - EHUHAC: 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 ($n > 1$).

DOMAIN	TRAIN (MIN/MAX/AVG)	DEV (MIN/MAX/AVG)	TEST (MIN/MAX/AVG)
EHUHAC	513,613 (2/160/9)	1009 (10/49/20)	2024 (10/49/19)
EITB	472,963 (2/198/3)	1027 (5/14/6)	2017 (5/14/6)
OPENSUBS	785,478 (10/50/42)	1037 (25/50/46)	2085 (10/50/42)
MERGED	1,753,726 (2/198/8)	3051 (5/50/12)	6078 (5/50/13)

Table 1:Corpora statistics (#sentence pairs)

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.
  - PARLIAMENT: proceedings of the Basque Parliament.
  - 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 initialised with the trained baseline models.
  - Variants:
    - Using source or target context.
    - Using different context sizes (1 vs. 5 sentences).

Metrics Results

MODEL	MERGED		EITB		EHUHAC		OPENSUBS	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
BASELINE	22.7	54.4	27.5	59.8	15.6	48.0	21.8	50.5
1:SRC	**23.3**†	**54.8**†	**28.2**†	**60.3**†	15.9	**48.4**†	22.3	**51.0**†
1:TGT:RF	**23.1**†	**54.8**†	27.9	60.1	**16.1**†	**48.6**†	**22.6**†	50.9
1:TGT:MT	22.9	54.3	27.9	59.7	15.6	47.9	21.9	50.2
5:SRC	21.6†	52.9†	25.9†	58.1†	15.0†	46.8†	21.4	49.6†
5:TGT:RF	22.2†	53.5†	26.5†	58.7†	15.5	47.3†	22.5	50.5
5:TGT:MT	22.0†	53.5†	26.3†	58.7†	15.3	47.3†	21.7	50.1

Table 2:Metrics results for Spanish to Basque translation

MODEL	MERGED		EITB		EHUHAC		OPENSUBS	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
BASELINE	31.2	54.6	38.5	61.7	22.7	47.1	25.5	47.2
1:SRC	**31.7**†	**55.0**†	**39.2**†	**62.2**†	23.1	47.5†	25.4	47.4
1:TGT:RF	**31.9**†	**55.2**†	**39.2**†	**62.2**†	**23.4**†	**47.8**†	**26.1**†	**47.7**†
1:TGT:MT	31.5†	54.8	**38.9**†	61.8	22.9	47.4	25.1	47.0
5:SRC	29.9†	53.4†	36.8†	60.2†	21.9†	46.1†	24.3†	46.1†
5:TGT:RF	29.4†	52.9†	36.0†	59.6†	21.6†	45.7†	24.5†	46.0†
5:TGT:MT	29.1†	52.6†	35.8†	59.3†	21.4†	45.4†	23.7†	45.6†

Table 3:Metrics results for Basque to Spanish translation

Contrastive Results

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		MASC	FEM	MASC	FEM	MASC	FEM
BASELINE	53.67	68.00	32.00	84.00	36.00	76.00	26.00
1:SRC	71.00	80.00	**64.00**	**94.00**	**56.00**	88.00	44.00
1:TGT	**71.33**	**82.00**	**64.00**	92.00	54.00	82.00	**54.00**
5:SRC	69.67	78.00	60.00	90.00	50.00	**90.00**	50.00
5:TGT	66.00	76.00	52.00	90.00	50,00	80.00	48.00

Table 4:Percentage of correct *gender* answers on GDR-SRC+TGT

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		MASC	FEM	MASC	FEM	MASC	FEM
BASELINE	51.33	76.00	28.00	92.00	12.00	72.00	28.00
1:SRC	52.67	88.00	28.00	**100.00**	4.00	56.00	40.00
1:TGT	60.67	88.00	44.00	96.00	**28.00**	68.00	44.00
5:SRC	56.00	76.00	20.00	**100.00**	4.00	80.00	**56.00**
5:TGT	**64.67**	**100.00**	**48.00**	96.00	12.00	**84.00**	48.00

Table 5:Percentage of correct *gender* answers on COH-TGT

MODEL	MERGED	LITERATURE		PARLIAMENT		TED	
		FORM	INFORM	FORM	INFORM	FORM	INFORM
BASELINE	56.67	24.00	88.00	56.00	64.00	48.00	60.00
1:SRC	62.00	32.00	84.00	76.00	60.00	56.00	64.00
1:TGT	75.33	64.00	88.00	68.00	**84.00**	64.00	84.00
5:SRC	50.00	24.00	60.00	68.00	40.00	56.00	52.00
5:TGT	**84.00**	**80.00**	**92.00**	**88.00**	**84.00**	**68.00**	**92.00**

Table 6:Percentage of correct *register* answers on COH-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