LINGUA CUSTODIA

Lingua Custodia @ Covid-19 MLIA

Round1

Raheel Qader

14 January 2021

About Lingua Custodia

- Based in paris + an office in Luxemburg
- We are specialized in financial machine translation
- Supported languages: most western european languages + Chinese and Japanese
- R&D team: 3 Researchers in machine learning and NLP

LINGUA CUSTODIA

Our Participation

- **We participate in the translation task**
  - **Languages in Round 1: English => French, Spanish, German, Italian, Sweedish, Greek**
  - **Constrained Task**

LINGUA CUSTODIA

Our Participation

- **We participate in the translation task**
  - **Languages in Round 1: English => French, Spanish, German, Italian, Sweedish, Greek**
  - **Constrained Task**
- **Goal:**
  - **Use a multilingual network to translate to all the languages**
  - **No additional data, or tools**
  - **Test an in-house seq2seq toolkit**

LINGUA CUSTODIA

Machine translation models

Preprocessing

- **Remove very long sentences**
- **Keep top dev set sentences**
- **Apply SentencePiece for subword segmentation**
- **Split numbers character-by-character.**
- **Shared vocabulary: 50K for single and 70K for multilingual models.**

Machine translation models

Model architectures

- **We use the Seq2SeqPy toolkit: A lightweight and customizable toolkit for neural sequence-to-sequence modeling.** https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/seq2seqpytorch
- **Single-language models**
  - **One model per language direction**
- **Multilingual models:**
  - **a single model can translate between several language directions**
  - **add a token in the beginning of the source sequence (e.g., 2fr, 2de, 2es)**

LINGUA CUSTODIA

Machine translation models

You are here
Additional information:

What should I do?

limiting contacts between people

Vous êtes ici
Informations complémentaires:

¿Qué puedo hacer?

Einschränkung der Kontakte zwischen den Menschen

LINGUA CUSTODIA

Machine translation models

2fr You are here	Vous êtes ici
2fr Additional information:	Informations complémentaires:
2es What should I do?	¿Qué puedo hacer?
2de limiting contacts between people	Einschränkung der Kontakte zwischen den Menschen

LINGUA CUSTODIA

Experiments - round 1

Hyper-parameters

- Standard transformer architecture 6 encoder and 6 decoder layers.
- Embedding size: 512, FFN size: 2048 with 8 attention heads.
- Source and target embeddings are tied with the vocabulary projection layer.
- Adam optimizer, learning rate of 0.0002, a warmup step of 5000.
- Label smoothing of 0.1
- Beam size: 5
- Averaged 4 best checkpoints
- 5 RTX 2080 Ti gpus

Experiments - round 1

Results

	En-De		En-Fr		En-Es		En-It		En-Sv	
	BLEU	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**
Single model	26.7	0.556	48.9	0.703	-	-	-	-	-	-
Multilingual	29.5	0.584	49.0	0.705	47.6	0.698	28.4	0.572	30.4	0.589

LINGUA CUSTODIA

Experiments - round 1

Results

	En-De		En-Fr		En-Es		En-It		En-Sv	
	BLEU	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**	**BLEU**	**chrF**
Single model	26.7	0.556	48.9	0.703	-	-	-	-	-	-
Multilingual	29.5	0.584	49.0	0.705	47.6	0.698	28.4	0.572	30.4	0.589

- **Top3 in 3 language pairs**

LINGUA CUSTODIA

Future rounds

Transfer learning

- Train on multilingual data and finetune on single language direction
- Use pre-trained models like BERT.

Improve preprocessing

- Clean the data
- Try BPE-dropout.
- Use masking for urls, dates, etc.

France +33 1 80 82 59 70
Luxembourg +352 2 786 76 11

contact@linguacustodia.com

www.linguacustodia.finance

Lingua Custodia