Lingua Custodia @ Covid-19 MLIA

Round2

Raheel Qader

17 February 2022

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 2: English => French, Spanish, German, Italian, Swedish, Greek, Arabic
  - Constrained Task

LINGUA CUSTODIA

Our Participation

- **We participate in the translation task**
  - Languages in Round 2: English => French, Spanish, German, Italian, Swedish, Greek, Arabic
  - Constrained Task
- **Goal:**
  - Use a multilingual model to translate to all the languages
  - Oversample low and mid-resourced languages
  - Cluster languages based on different criteria
  - Find best fitting parameters for multilingual setting

LINGUA CUSTODIA

Language Clusters

- **Based on number of training data**

French	Italian	Arabic
Spanish	Swedish	
German	Greek	

- **Based on writing script**

French	Italian	Arabic
Spanish	Swedish	Greek
German		

LINGUA CUSTODIA

Machine translation models

- **Preprocessing**
  - Tokenization with Moses
  - Remove samples where length difference between source and target is larger than threshold
  - Remove consecutive spaces
  - Inline-casing (The is an ApPle => this **<T>** is an apple **<M>**)
  - Apply SentencePiece for subword segmentation (split numbers)
  - Add language token prefixes
  - Vocab size: 16K, 30K, 40K, 50K.

LINGUA CUSTODIA

Machine translation models

- **Model architectures**
  - We use the Sockeye toolkit:
  - Bilingual models
    - One model per language direction
  - Multilingual models:
    - a single model can translate between several language directions
    - Oversample IT, EL, AR, SV
    - Separate languages with different writing scripts

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 projection layer.
  - Beam size: 5
  - Trained on 3 RTX 2080 Ti GPUs

Experiments - round 2

- **Vocab size (dev set)**

Vocab size	Bleu score
16K	36.7
30K	38.4
40K	39
50K	39

LINGUA CUSTODIA

Experiments - round 2

- **Pre-process or not (dev set)**

	Bleu score
w/o pre-processing	39
w/ pre-processing	43.2

LINGUA CUSTODIA

Experiments - round 2

- **Experiments (dev set)**

	En-De	En-Fr	En-Es	En-It	En-El	En-Sv	En-Ar
Bilingual	41.2	**58.9**	**57.0**	45.4	42.4	16.1	23.8
Multi-5lang	**40.0**	58.1	56.1	47.6	-	17.9	-
Multi-5lang-ov	38.3	57.0	55.3	**48.1**	-	**21.9**	-
Multi-7lang	38.9	57.3	55.5	47.4	44.5	17.7	25.5
Multi-7lang-ov	37.6	56.3	54.5	47.6	**45.7**	18.8	**28.9**

LINGUA CUSTODIA

Experiments - round 2

— **Submitted results (test set)**

	En-De	En-Fr	En-Es	En-It	En-El	En-Sv	En-Ar
Bilingual	39.7	**57.2**	**56.6**	45.3	41.2	16.7	19.1
Multi-5lang	**40.3**	-	56	-	-	17.7	-
Multi-5lang-ov	-	-	-	**48.9**	-	**22**	-
Multi-7lang	38.6	55.8	55.3	-	43.2	-	22
Multi-7lang-ov	-	-	-	-	**44.7**	18.3	**25.1**

LINGUA CUSTODIA

Findings

- **Clustering languages**
  - For some languages it's better to train everything together
  - While others doesn't benefit always from the very different languages
- **Oversampling**
  - It’s very important to oversample less resourced languages
- **Bilingual models**
  - Some langauges like French still work better when trained alone

LINGUA CUSTODIA

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

contact@linguacustodia.com

www.linguacustodia.finance

Lingua Custodia