PROMT Submissions for Covid-19 MLIA Shared Translation Task Round 2

Alexander Molchanov
Alexander.Molchanov@promt.ru

DATA

Constrained

- all provided train data
- no filtering
- 8k SentencePiece model
- language tags on the English side
- **synthetic data**

Unconstrained

- **same as last year**
- OPUS, statmt.org, private data
- no fine-tuning (no Covid-19 MLIA Data used)
- case-insensitive BPE, 8 to 16k for different models

PROMT

DATA

Constrained synthetic data

- intermediate multilingual X->EN model for back-translations built on all data from Rounds 1 and 2
- monolingual data from the MSS Task (MEDISYS 2020)
- simple handcrafted rules for filtering
- score translations with Round1 model (except for AR)

PROMT

MODELS

Constrained

- baseline transformer, Marian
- a single multilingual model, all language pairs (EN to ES, DE, IT, EL, FR, SV, **AR**)
- shared vocabulary
- vocabulary filtering (removing tokens with frequency < 10)
- **fine-tuning for each language pair**

Unconstrained

- **same as last year**
- baseline transformers, Marian
- all language pairs, multilingual models for IT (EN to IT, PT) and SV (EN to SV, DA, NO)
- shared vocabularies except for EN to EL and AR

RESULTS and FUTURE WORK

Results

- Rank top in English-Greek

Future Work

- Domain selection for synthetic data
- Check fine-tuned models for possible overfitting
- Fine-tune general unconstrained models

PROMT

QUESTIONS?

Alexander Molchanov
PROMT LLC

Alexander.Molchanov@promt.ru

PROMT