eTranslation’s Submissions to the COVID19-MLIA Translation Task (R2)

Csaba Oravecz, Katina Bontcheva, David Kolovratník, Bogomil Kovachev, Vilmantas Liubinas, Christopher Scott, Francois Thunus, Andreas Eisele

DG Translation, European Commission

MT Task Virtual Meeting, Feb 17, 2022

Outline

1. Introduction
2. English → German
3. English→Swedish
4. English→Greek, English→Spanish, English→French
5. English→ Italian
6. Conclusions

European Commission

Outline

1. Introduction

2. English → German

3. English→Swedish

4. English→Greek, English→Spanish, English→French

5. English→ Italian

6. Conclusions

European Commission

Dry facts

Participation in 6 language pairs

- English → German
- English → Swedish
- English → Greek
- English → Spanish
- English → Italian
- English → French

Categories

- constrained
- unconstrained (selected external health related data sets)

Development

Models

- use standard best practices
- find optimal architecture and parameters for the language pair
- experiment with available powerful general models in the unconstrained category

Data

- focus on cleaning of provided data
- use all accessible health related data for unconstrained systems

Outline

1. Introduction

2. English → German

3. English→Swedish

4. English→Greek, English→Spanish, English→French

5. English→ Italian

6. Conclusions

Data

Constrained

- significant increase of provided parallel data compared to R1
- noise in the data set → filtering (3% reduction):
  - language identification with FastText
  - deletion of segments where source/target token ratio exceeds 1:3 (or 3:1),
  - deletion of segments longer than 110 tokens,
  - exclusion of segments where the ratio between the number of characters and the number of words was below 1.5 or above 40,
  - exclusion of segments without a minimum number (4) of alphabetic characters.
- basic heuristic: mismatch in the numeric tokens between source and target → remove segment (6% reduction)
- final training data set: 2.25M segments

Data

Unconstrained

- TAUS Corona Crisis Corpora (610k segments)
- OPUS EMEA Corpus (760k segments)
- health related subset of the Euramis data set (1.1M segments)

Training

Constrained

- 1st step: base Transformer, Marian, SentencePiece, default set of hyperparameters
- 2nd step: big Transformer (doubled filter size and number of heads)
- joint vocabulary: 12k, 32k
- last step: 5 epoch fine-tuning on R1 and R2 validation and R1 test sets
- best submission: 4 member ensemble
- 2–4 NVIDIA V100 16GB GPUs, ≈ 30 epochs

European Commission

Training

Unconstrained

1. single big transformer trained on the extended data set (3.89M segments)
2. based on WMT 2021 News Task submission model
   - 4 member big transformer ensemble
   - each of the 4 big transformer models is fine tuned on the filtered training data for 3 epochs and ensembled together with equal weights

Results

		Test sets		
System	Data	Euramis (2k)	R2-V (2k)	R2 off. (4k)
Round 1 (c)	926k	32.7	29.2	27.7
Round 2 raw (c)	2.46M	35.3	39.9	39.7
Round 2 filt. (c)	2.25M	35.9	40.3	39.7
R 2 filt. big Tr. (c)	2.25M	35.3	39.0	38.3
Big tr. ensemble* (c)	2.25M	37.3	41.7	40.9
Big tr. ens. fine t.* (c)	2.25M+6.5k	37.3	–	**41.1**
Single big Tr.* (u)	3.89M	–	41.3	41.2
WMT21 (u)	430M	38.7	45.9	46.2
WMT21 fine t.* (u)	430+2.25M	39.9	47.4	**47.1**

Table: Results for En→De models. R2-V is the development test set extracted from Round 2 validation data. (c): constrained, (u): unconstrained model. Submissions are marked with an asterisk.

Outline

1. Introduction
2. English → German
3. English→Swedish
4. English→Greek, English→Spanish, English→French
5. English→ Italian
6. Conclusions

European Commission

Development

Data

- R1 plus R2 segments
- filter heuristics based on numeric and location tokens (-25%)

Training

- ≈ En→De, 36k vocabulary

Unconstrained

- general OPUS and ParaCrawl data filtered on basic medical terms; OPUS EMEA Corpus; additional in house health data

Results

		Test sets	
System	Data	R1+R2 (9k)	R2 off. (4k)
Base Tr. (c)	900k	51.8	20.3
Big Tr. (c)	900k	54.3	20.0
Big Tr. Ensemble (c)	900k	56.3	**22.7**
Base Tr. Euramis (u)	1.75M	52.2	20.9
Base Tr. multi (u)	2.5M	53.9	22.0
Big Tr. multi ensemble (u)	2.5M	56.6	**23.3**

Table: Results for En→Sv models. *multi* is the multiple source data described above. R1+R2 is the development test set we created ourselves from data from Round 1 and Round 2. (c): constrained, (u): unconstrained model.

Outline

1. Introduction
2. English → German
3. English→Swedish
4. English→Greek, English→Spanish, English→French
5. English→ Italian
6. Conclusions

European Commission

Development

English→Greek, English→Spanish

- ≈ other LPs
- data: R1 + R2 (+Euramis – unconstrained)
- models: base and big Transformer, 36k vocabulary
- results: best unconstrained En→Es (about the same score as the best constrained system)

En→Fr

- constrained: R1+R2 data (2.9M), models as above, winning submission (better than the unconstrained model)
- unconstrained: stock eTranslation general engine (237M)
- normalization of punctuation in postprocessing → huge difference (drop) in BLEU!

European Commission

Outline

1. Introduction
2. English → German
3. English→Swedish
4. English→Greek, English→Spanish, English→French
5. English→ Italian
6. Conclusions

European Commission

Development

Data

- R1 plus R2 segments filtered as in En→De(1.6M)
- subset extracted using a few keyword patterns → used for fine-tuning (100k)
- unconstrained: EMEA, TAUS Corona Crisis Corpora, Euramis exractions based on keywords and metadata (3.7M)

Training

- base and big Transformer, 4 member ensemble

Results

		Test sets	
System	Data	R1+R2 dev (10k)	R2 off. (4k)
Base Tr. (c)	R1+R2	43.3	45.1
Big Tr. (c)	R1+R2	46.3	44.0
Big tr. ens.* (c)	R1+R2	47.7	47.0
Big tr. ens. FT* (c)	R1+R2	47.8	46.7
Big tr. (u)	R1+R2+mtdata1	46.3	46.6
Base tr. (u)	R1+R2+mtdata1+mtdata2+var_med	45.9	46.8
Big tr. (u)	R1+R2+mtdata1+mtdata2+var_med	48.5	48.3
Big tr. ens.* (u)	R1+R2+mtdata1+mtdata2+var_med	49.4	**50.1**

Outline

1. Introduction

2. English → German

3. English→Swedish

4. English→Greek, English→Spanish, English→French

5. English→ Italian

6. Conclusions

Conclusion

- focus on data selection and filtering
- competitive systems ending up in first place in several categories
- large, robust systems perform best → diverse and noisy data sets?

The End

European
Commission