POST-EDITING AND RESCORING OF ASR RESULTS WITH OPENNMT-APE

BY MGR DOMINIKA WNUK & DR INŻ. KRZYSZTOF WOŁK

TASK 1:

- refine the results of automatic speech recognition (ASR)
- transcriptions of oral utterances
- better reflect the actually spoken phrases
- simple sequence of words, without capitalization, digits, punctuation, symbols nor abbreviations
- not a global solution

CORPORA

Clarin-PL studio corpus

Polish Parliament corpus

TRAINING FILES

reference

1-best

n-best

lattice output

OpenNMT-APE

- open source
- transfer learning (BERT pre-trained model)
- BERT as both encoder & decoder in seq2seq

TRAINING PARAMETERS

- Validation steps: 1000
- Checkpoint: 30
- Warmup steps: 5000
- Learning rate: 0.00005
- Average decay: 0.0001
- Source sequence length: 200
- Target sequence length: 200

Train steps and start decay steps adjusted per each experiment: 1 k, 10 k, 20 k, 40 k & 50 k

TRAINING PARAMETERS

- Self-attention shared between encoder and decoder
- Context attention & self-attention - the same weights
- Dropout rate: 0.1
- Label soothing: 0.1

PRE-PROCESSING

Dimensionality configured:

- 12 self-attention layers
- 12 attention heads
- RNN & word vector size of 768
- feed-forward inner layer of 3072

TRANSLATION

- Application of trained APE to test data.
- 2 additional stages: clean-up of missing lines and reprocessing of missed lines

TRAINING

5 experiments with Clarin-PL corpus & 2 experiments with joint Clarin-PL and parliament corpora

EVALUATION

- Word Error Rate
- NIST SCLITE (Score lite) package

BEST SCORE

Clarin-PL dataset - optimal result with 20 thousand iterations - 31.8% WER & 9.9% changes to the original

In case of Clarin_Sejm dataset, higher number of iterations improved the result.

EXPERIMENT	WER (%)	CHANGES (%)
PJA_CLARIN_1k	33.5	9.1
PJA_CLARIN_10k	32.0	9.6
PJA_CLARIN_20k	**31.8**	**9.9**
PJA_CLARIN_40k	31.8	10.3
PJA_CLARIN_50k	31.8	10.2
CLARIN_SEJM_40k	33.7	19.1
CLARIN_SEJM_50k	32.5	17.7

FINAL OBSERVATIONS

With the use of open source APE system, the optimal score of **31.8% WER** was achieved.
It falls behind the average word error rate of the ASR system (27.6% WER).

NOVELTY

Application of **encoder-decoder architecture** with **BERT** language model.

REFERENCES

Correia, Gonçalo & Martins, André. (2019). *A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning*. 3050-3056.10.18653/v1/P19-1292.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Klein, Guillaume & Kim, Yoon & Deng, Yuntian & Senellart, Jean & Rush, Alexander. (2017). *OpenNMT: Open-Source Toolkit for Neural Machine Translation*.

Levis, John & Suvorov, Ruslan. (2012). *Automatic Speech Recognition*. 10.1002/9781405198431.wbeal0066.

Lopes, António & Farajian, Amin & Correia, Gonçalo & Trenous, Jonay & Martins, André. (2019). *Unbabel's Submission to the WMT2019 APE Shared Task: BERT-based Encoder-Decoder for Automatic Post-Editing*.

Marasek, Krzysztof; Koržinek, Danijel; Brocki, Łukasz; et al., 2015, *Clarin-PL Studio Corpus (EMU)*, CLARIN-PL digital repository, http://hdl.handle.net/11321/236.

Negri, Matteo & Turchi, Marco & Bertoldi, Nicola & Federico, Marcello. (2018). *Online Neural Automatic Post-editing for Neural Machine Translation*. 10.4000/books.aaccademia.3534.

Ogrodniczuk, Maciej. *Polish Parliamentary Corpus*. In Darja Fišer, Maria Eskevich, and Franciska de Jong, editors, Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora, pp. 15–19, Paris, France, 2018. European Language Resources Association (ELRA).

Ogrodniczuk, Maciej. *The Polish Sejm Corpus*. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, pp. 2219–2223, Istanbul, Turkey, 2012. ELRA

Popovic, Maja & Ney, Hermann. (2007). *Word error rates*. 48-55. 10.3115/1626355.1626362.

Ziółko, M. & Gałka, Jakub & Ziółko, B. & Jadczyk, Tomasz & Skurzok, Dawid & Mąsior, Mariusz. (2011). *Automatic Speech Recognition System Dedicated for Polish*. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 3315-3316.