FACULTY OF MECHATRONICS,
INFORMATICS AND INTERDISCIPLINARY
STUDIES TUL

SKFM 2023
STUDENT CONFERENCE FM

ONLINE PUNCTUATION RESTORATION USING ELECTRA MODEL FOR STREAMING ASR SYSTEMS

Martin Poláček[1] (martin.polacek@tul.cz), Petr Červa[1], Jindřich Žďánský[1], Lenka Weingartová[2]
[1]Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
[2]NEWTON Technologies, Na Pankraci 1683/127, 140 00 Praha 4, Czech Republic

Summary
This contribution introduces a lightweight online approach to Automatic Punctuation Restoration (APR), designed for real-time speech transcription systems such as live captioning for TV or radio broadcasts. The approach uses textual input without prosodic features and employs a fine-tuned ELECTRA-Small model with a two-layer classification head. This allows for the restoration of question marks, commas, and periods with minimal inference time and a latency of just three words.

Proposed approach

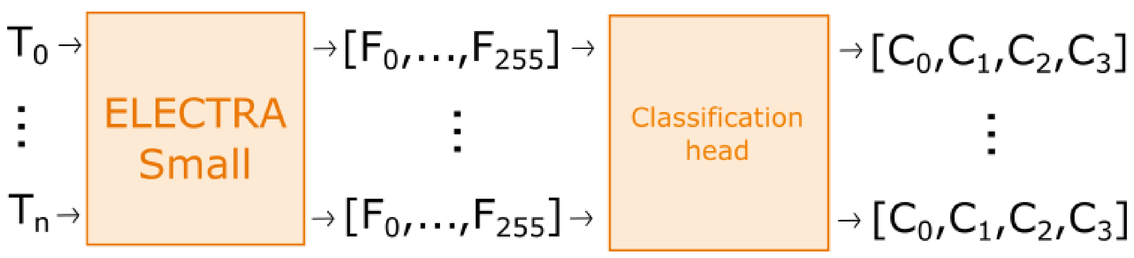

Fig. 1. Proposed APR module

Text preprocessing:
- SentencePiece tokenizer with 30522 tokens.

Pre-trained model:
- ELECTRA-Small architecture with embedding vector of size 256.

Classification head for APR:
- two feed-forward layers (512 and 4 neurons) with SELU;
- fine-tuned with a small learning rate (LR) for the pretrained model and higher LR for classification head;
- produces probability for question mark, comma, period and none.

Training & Development Data
The train dataset includes:
- 23 GB of Czech texts (i.e., 5 billion tokens);
  - newspaper articles, manually corrected ASR transcripts of Czech TV/R broadcasts, diploma theses and legal texts.
- Distribution of punctuation marks among tokens:
  - dots (4.5%), commas (4.5%) and question marks (0.2%).

The development dataset consists of manually corrected TV/R transcripts containing 259K tokens.

Evaluation Metrics
- precision (P), recall (R), F1-score (F1);

Evaluation was also performed in the one class scenario:
- all three punctuation marks were merged into one class;
- substitutions of individual punctuation marks were ignored.

Impact of Architecture Type
Different transformer architectures compared:
- BERT (pre-trained), ELECTRA-Small and ELECTRA-Base (trained from scratch) and GPT-3 based service in an edit mode.

ELECTRA-Small achieves the best results with the lowest inference time.

architecture	F1 [%]	F1 (one class) [%]	inf. time [ms]
BERT-Base	75.2	88.9	61
ELECTRA-Small	**76.0**	**90.7**	**11**
ELECTRA-Base	75.4	90.7	60
GPT-3 (in edit mode)	65.1	73.7	-

Table 1. Comparison of performance of various architectures in the offline block-processing mode

Block-processing vs streaming mode
Block-processing mode:
- punctuation is restored for the entire block of text.

Streaming mode:
- each forward pass determines the punctuation for only one word;
- it is not possible to use an input block of constant size;
- the left context is limited by the number of already recognized words (maximum 100 words);
- the right (future) context should be as short as possible.

Context of three words yields F1 values just by 1% smaller than those achieved in the block-processing mode.

max. left cont.	max. right cont.	P [%]	R [%]	F1 [%]
100	1	73.3	67.4	70.2
100	2	75.0	72.5	73.7
100	3	75.3	74.2	74.7
100	4	75.2	75.1	75.1
100	5	75.5	75.5	75.5
100	10	**76.0**	76.0	**76.0**
100	100	75.6	73.7	74.6
block-processing with no overlay		75.3	**76.8**	**76.0**

Table 2. Results [%] of the proposed APR module in the streaming mode

Results for streamed ASR transcripts
The test data represents a real output from E2E ASR system.

Comparison to RNN-based real-time APR module for Czech:
- It utilizes LSTM units, word embeddings, prosodic features and information about silence extracted from speech signal.

The proposed APR module:
- achieves comparable or better results without using prosodic features;
- just in spontaneous speech, dots and commas are more often confused.

Speech	Proposed APR F1 [%]	Proposed APR F1 (one class) [%]	RNN APR F1 [%]	RNN APR F1 (one class) [%]
Scripted	**71.2**	**84.3**	62.1	73.4
Spontaneous	69.4	**89.0**	**71.6**	73.3

Table 3. Comparison of the proposed APR module in online mode to the RNN APR module

Conclusions
The proposed lightweight APR module for Czech:
- uses the ELECTRA-Small transformer;
- operates online with a latency of just three-words;
- consumes pure text on its input;
- almost matches block regime accuracy;
- struggles with low-frequency question marks;
- has very low computation demands;
  - forward pass takes 11 ms on Intel i7-9700K processor.

Acknowledgements
- The research leading to these results has received funding from the Norway Grants and the Technology Agency of the Czech Republic within the KAPPA Program (project No. TO01000027).
- This work was supported by the Student Grant Competition (SGS) project of the Technical University of Liberec in 2023.