LLM-based Models for Transforming of Diachronic Bulgarian Spelling to Contemporary Bulgarian

Kiril Simov, Nikolay Paev, Petya Osenova

Artificial Intelligence and Language Technology Dept. , IICT-BAS

kivs@bultreebank.org, nikolay.paev@iict.bas.bg, petya@bultreebank.org

1. Bulgarian Spelling Systems (1878-1945)

From 1870 to 1945 eight spelling forms were proposed and applied either officially or non-officially:

- the Marin Drinov's spelling
- the spelling project by the philological committee from 1893
- the spelling project from 1895
- the Ivan Vazov's spelling system from 1899
- the Drinov-Ivanchevski spelling system from 1899
- the Omarchevski spelling system (1921 - 1923)
- the spelling project of the Historic-philological branch of the Bulgarian Academy of Sciences (1899)
- the Tsankov's spelling system (1923)

Excerpt from a newspaper article published in 1878:

На телеграммата отъ 10 Юлия Главнокомандующийтъ на войскитѣ
На телеграмата от∅ 10 юли∅ главнокомандващият∅ на войските
позволи изнасяньето на хранитѣ отъ България .
позволи изнасян∅ето на храните от∅ България .
'In a telegram from 10 July, the Commander-in-chief
gave permission to export the food from Bulgaria.'

2. Parallel Corpora for Fine-tuning

- The source data for the task are the collection of periodicals and books provided to us by the National Library "Ivan Vazov" (NLIV) in Plovdiv.
- The data were created in two ways: (1) through an alignment between books with diachronic spelling, and the digital form of the same book published several decades later according to the contemporary spelling; and (2) through a lexicon-based "translation" from the old to the new spelling. The lexicon (**OldNewLex**) was created on the basis of the Bulgarian inflectional lexicon. Each lexical entry in it contains the paradigms of nearly 80 000 lemmas.
- The first dataset - **DS-1** (about 200 000 running words) consists of novels and short stories published in the period of 1910-1944 for which we found also electronic versions published after 1945.
- The second dataset - **DS-2**, consists of newspaper articles (5 000 tokens).
- We consider *the task of mapping between the old and the new spelling norms* as a *spell checking task*.
- **DS-2** was used only for testing.

3. Pre-training Language Models

The mapping is a sequence to sequence task between texts with slight differences in the spelling. Classic subword token based language models are blind to characters.

Hypothesis 1:: Models that tokenize on character level are more suitable to the task.

Hypothesis 2: Encoder-Decoder models will perform better.

We pre-train character-based T5 and Llama models on 4B Bulgarian words as well as classic subword token based T5 and Llama models.

Tested models:

- Character level T5 - 353M
- Subword level T5 - 403M
- Character level Llama - 302M
- Subword level Llama - 371M
- Subword level EuroLLM - 1.7B

4. Experiments and Results

We fine-tuned the models on the dataset **DS-1** for 10 epochs with a peak learning rate of 1e-04. As a baseline we used **OldNewLex** substitutions.

The same models were tested on **DS-2** to assess whether they retain their abilities across a set with different spelling. We do not include a lexicon substitution baseline, because the dataset **DS-2** was constructed with the same lexicon, which makes the results artificially high.

Model type	Architecture	Tokenization level	Whole-Word BLEU
Lexicon substitution (baseline)			87.69%
Llama	Decoder	Character	93.01%
Llama	Decoder	Subword	89.07%
T5	Encoder-Decoder	Character	**95.32%**
T5	Encoder-Decoder	Subword	92.01%
EuroLLM 1.7B	Decoder	Subword	94.21%

Table 1: BLEU Scores of the tested fine-tuned models on the test split of **DS-1**

Model type	Architecture	Tokenization level	Whole-Word BLEU
Llama	Decoder	Character	72.17%
Llama	Decoder	Subword	72.76%
T5	Encoder-Decoder	Character	**87.02%**
T5	Encoder-Decoder	Subword	80.02%
EuroLLM 1.7B	Decoder	Subword	78.76%

Table 2: BLEU Scores of the tested fine-tuned models on **DS-2**

5. Post-editing of the Result

- If the goal is to produce 100 % correct result we need to manually read the whole result.
- In this experiment, we assume that when the two best models *Character-based T5* and *EuroLLM 1.7B* make the same prediction, it is the correct one.
- In order to find the problematic spans in the result we aligned three texts: the *input text* and the *results from the two models* - selecting the correct token (**DS-3**)
- The input is necessary when the models do not process some input tokens - they are inserted in the alignment. The results from the alignments are sequences of triples $\langle tok_{T5}, tok_{EuroLLM}, tok_{DS\text{-}3}\rangle$. We assume that a triple represents a correct transfer if tok_{T5} or $tok_{EuroLLL}$ is the same as tok_{DS3}:
  - The groups of inserted tokens are 147 = 962 tokens.
- In the rest of tokens (5 734) there are three cases:
  - $tok_{T5} = tok_{DS3}$ in 4 736 tokens — 82.59% correct correspondences
  - $tok_{EuroLLM} = tok_{DS3}$ in 4 470 tokes — 77.95% correct correspondences
  - $tok_{T5} = tok_{DS3}$ *or* $tok_{EuroLLM} = tok_{DS3}$ in 4 866 — 84.86% correct correspondences
- In this way we reduce the numbers of editings in order to check the correctness of the text.

6. Conclusions and Future work

We present a number of experiments for transforming the combined diachronic Bulgarian spelling system to contemporary Bulgarian. Directions for future work are as follows:

- Extending the datasets further not only in quantity but also in text variety
- Performing similar experiments for improving the OCR quality
- Testing the task with bigger language models
- Providing a deep enough error analysis

Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG