Automatic accent restoration with transformer model based neural machine translation for Hungarian

László János Laki[ab], **Zijian Győző Yang**[abc]

[a]MTA-PPKE Hungarian Language Technology Research Group

[b]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
{laki.laszlo,yang.zijian.gyozo}@itk.ppke.hu

[c]Eszterházy Károly University
yang.zijian.gyozo@uni-eszterhazy.hu

Abstract

In the last few years the text writing on mobile devices suddenly increased. People often type messages without accent, therefore more and more corpus are generated online that contains texts without accents, which made text mining task so hard. An accent restore application could be able to clean and prepare the corpus for research or help people to correct their texts.

In our study, we created an accent restore method based on the state-of-the-art neural machine translation techniques (transformer model and SPM). Our method can restore accent with about 99,8% relative accuracy, which means, that our system made wrong decision only in 31 cases out of more than 18,000 tokens. We compared our system with the previous best systems, where we could reach statistically significant quality gain, so our system become the state-of-the-art solution for accent restoration task.

Furthermore our system has tried out of different domain corpora and we also did experiments in other languages, where we could measure similar performance. We made a demo to represent our application.

Keywords: accent restoration, neural machine translation, NMT, transformer model