Comparative analysis of grapheme-to-phoneme models for the Russian-Chinese parallel corpus[1]

Alexandra Konovalova and Alena Tsvetkova
National Research University Higher School of Economics, Moscow

Introduction[2]

The Russian-Chinese parallel corpus of Ruscorpora (henceforth the Corpus) is an online corpus of texts provided with linguistic markup and meta-information. The Corpus contains 1070 text samples of literary works, news, and others. So far, it lacks a proper annotation of pinyin (Chinese transcription). Now the Grapheme-to-Phoneme (G2P) model applied to the Corpus is based on the dictionary CEDICT (Luo, Xu, Zhang, Ren and Sun, 2019). Thus, all the possible transcriptions are ascribed to each character without disambiguation. Therefore, the purpose of this study is to compare six G2P models to improve the quality of pinyin markup of the Corpus. Chinese G2P conversion seems to be challenging because of homophones and polyphones, which means that a character may have multiple pronunciations. Besides, the Chinese texts in the Corpus contain many phonetic borrowings from Russian which commonly are not included in dictionaries.

Related Work

Approaches to Chinese G2P conversion can be divided into rule-based (Wang, Chen and Yeung, 2004) and data-driven. Generally, rule-based methods suggest searching transcriptions of the words in a dictionary and mapping them to the text in accordance with the context. Although a set of rules is efficient for processing the majority of data, it faces problems with ambiguous characters. Data-driven approaches use statistical methods. Park and Lee (2020) proposed a developed dataset for polyphone disambiguation and trained a Bi-LSTM model on it. Chen, Zhao and Wang, (2015) argued that converting neural network language models into back-off n-gram language models helps to reduce computational cost. Following this approach, Cai, Yang, Zhang, Qin, and Li (2019) explored a bidirectional recurrent neural network. Recently, in Huang, Li, Zhang and Zhao (2018) and Zhang, Huang and Zhao (2019), an attention-based model which translates from pinyin sequence to Chinese character sequence was implemented.

Method

For this work, we explored the following models: A Context-aware Grapheme-to-Phoneme for Chinese (G2pC[3]) (2019), A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset (G2pM[4]) (2020), xpinyin[5] (2002) and Chinese Pinyin Conversion Tool for Python (pypinyin[6]). All the methods tested are data-driven: xpinyin is based on stochastic decision lists and pypinyin uses n-gram statistics. G2pM is entirely built on sequence translating. Among those, the only model to use grammar information is G2pC which supports POS-tagging. One of the main challenges for disambiguating polyphones is correct word segmentation and POS-tagging. Therefore, we fine-

[1] The project was supported by the Commission of the Support of Educational Initiatives of the Faculty of Humanities within the framework of the Competition of Project Groups for Students (the name of the project is «Linguistic Markup of Chinese Texts in the Russian-Chinese Parallel Corpus of Ruscorpora»)

[2] Code is available on https://github.com/vydra-v-getrax/pinyin_annotation

[3] https://github.com/Kyubyong/g2pC

[4] https://github.com/kakaobrain/g2pM

[5] https://github.com/lxneng/xpinyin

[6] https://github.com/mozillazg/python-pinyin

tuned G2pC model with different state-of-the-art tools for parsing Chinese texts: FastHan[7] (Geng, Yan, Qiu, and Huang, 2020), UDPipe[8] (Straka and Straková, 2017, August) and pkuseg[9] (Luo, Xu, Zhang, Ren, and Sun, 2019). For the test, we used a small dataset of 20 human-annotated sentences which were randomly selected from the Corpus and contained polyphones and proper names.

Results

Table 1 presents accuracy scores on the test dataset for each model.

Table 1. Metrics on evaluation dataset

Model	Accuracy
G2pC-pkuseg	0.903
G2pC-FastHan	0.899
G2pC-Udpipe	0.880
Xpinyin	0.861
Pypinyin	0.831
G2pM	0.824

(suffix *pkuseg/FastHan/UDPipe* refers to the tool for POS-tagging in G2pC annotator)

One of the typical mistakes is provoked by character 了 *le* or *liǎo*. For example, in the phrase 巴扎罗夫瞅了他一眼 'Bazarov looked at him' it is pronounced as *le*, but algorithms annotated it as *liǎo.* As for the Russian-Chinese corpus, specific mistakes in the output of G2P models were not found. We plan to test this hypothesis on a larger dataset. Names and loan words are transcribed correctly because phonetic borrowings from Russian are translated using a limited set of Chinese characters. Nevertheless, linguistic issues common for written Chinese G2P should be addressed.

Based on our analysis, we revealed that the best algorithm for our Corpus is G2pC annotator on texts preprocessed with pkuseg package as it seems that for interpreting Chinese characters correct word segmentation and POS-tagging are crucial. Unlike UDPipe and FastHan, pkuseg includes multiple domain-specific segmentation CRF models. Therefore, superiority of this model is mainly due to pre-training on a large-scale, multi-domain dataset. In the future, we plan to proceed with experiments, by applying other tools for word segmentation, fine-tuning the algorithms to our data, and providing an extended evaluation dataset.

References

Cai, Z., Yang, Y., Zhang, C., Qin, X., & Li, M. (2019). Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features.

Chen, S., Zhao, H., & Wang, R. (2015). Neural Network Language Model for Chinese Pinyin Input Method Engine. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 455–461. https://www.aclweb.org/anthology/Y15-1052.

Geng, Z., Yan, H., Qiu, X., & Huang, X. (2020). fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP. arXiv preprint arXiv:2009.08633.

Huang, Y., Li, Z., Zhang, Z., & Zhao, H. (2018). Moon IME: Neural-based Chinese Pinyin Aided Input Method with Customizable Association. Proceedings of ACL 2018, System Demonstrations, 140–145. https://doi.org/10.18653/v1/P18-4024.

Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. (2019). Pkuseg: A toolkit for multi-domain chinese word segmentation. arXiv preprint arXiv:1906.11455.

[7] https://github.com/fastnlp/fastHan

[8] http://ufal.mff.cuni.cz/udpipe

[9] https://github.com/lancopku/pkuseg-python

Park, K., & Lee, S. (2020). g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset. ArXiv:2004.03136 [Cs]. http://arxiv.org/abs/2004.03136.

Semenov, K., Durneva, S., & Kuznetsova, Y. (2020). The Russian-Chinese Parallel Corpus of Ruscorpora: Achievements and Challenges. DHN2020: Parallel Corpora as Digital Resources and Their Applications. Retrieved from https://parallelcorporadhn2020.github.io/talks/Durneva_Kuznetsova_Semenov.html.

Straka, M., & Straková, J. (2017, August). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (pp. 88-99).

Wang, X., Chen, Q., & Yeung, D. S. (2004). Mining Pinyin-to-Character Conversion Rules From Large-Scale Corpus: A Rough Set Approach. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), 34(2), 834–844. https://doi.org/10.1109/TSMCB.2003.817101.

Zhang, Z. R., Chu, M., & Chang, E. (2002). An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese. In International Symposium on Chinese Spoken Language Processing.

Zhang, Z., Huang, Y., & Zhao, H. (2019). Open Vocabulary Learning for Neural Chinese Pinyin IME. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1584–1594. https://doi.org/10.18653/v1/P19-1154.

Alexandra Konovalova: askonovalova@edu.hse.ru
Alena Tsvetkova: a.tsvetkova@hse.ru