Comparing Writing Systems with Multilingual Grapheme-to-Phoneme and Phoneme-to-Grapheme Conversion

Julen Etxaniz
University of the Basque Country
jetxaniz007@ikasle.ehu.eus

Abstract

1 Introduction

The orthographic depth of an alphabetic orthography indicates the degree to which a written language deviates from simple one-to-one grapheme (letter) to phoneme (sound) correspondence.

Shallow (transparent) orthographies, also called phonemic orthographies, have a one-to-one relationship between its graphemes and phonemes, and the writing and reading of letters is very consistent. Spanish and Basque are examples of transparent orthographies.

In contrast, in deep (opaque) orthographies, the relationship is less direct, and the writer or reader must learn the arbitrarily irregular words. English and French are considered to have opaque orthographies.

Sequence-to-sequence models are very good at picking up on consistent patterns. Therefore, a seq2seq system should learn the regular writing systems much better than the irregular ones. That means that the performance of seq2seq models could be used as a measure of regularity of a writing system. The grapheme-to-phoneme (G2P) direction would suggest the difficulty of reading a language. Similarly, in the phoneme-to-grapheme (P2G) direction we would be looking at the inconsistencies of writing a language.

We use the transformer proposed in Wu et al. (2021) to compare the orthographies of 10 languages: Armenian (Eastern), Bulgarian, Dutch, French, Georgian, Serbo-Croatian (Latin), Hungarian, Japanese (Hiragana), Korean and Vietnamese (Hanoi). We use same data of the medium-resource languages from the SIGMORPHON 2021 G2P task (Ashby et al., 2021). We train a different model for each task and language.

The goal is not to build a perfect G2P and P2G system. Instead the goal is to build a translator which can indicate a degree of phonemic transparency and thus make it possible to rank orthographies. We do not have to translate a sequence of words into another sequence of words. Our model only requires translating a single word. Therefore, We can consider G2P and P2G as two translation tasks and apply the transformer at character level.

2 Related works

Sequence-to-sequence (seq2seq) models have been proven to be very successful on language translation tasks. Recently, attention (Bahdanau et al., 2016), transformers (Vaswani et al., 2017) and generative pre-training (GPT) (Radford and Narasimhan, 2018) have improved the performance of seq2seq models. Applying the transformer to character-level transduction tasks such as G2P has been proved to be a good option (Wu et al., 2021). It was one of the baselines in the SIGMORPHON 2020 G2P task (Gorman et al., 2020).

Many studies have discussed the transparency of orthographies based on the ease of reading and writing when learning a new language (Borleffs et al., 2017). However, there is not much work in NLP about measuring the level of transparency of an orthography. In an old work, three data-based algorithms were tested on three orthographies: Dutch, English and French (van den Bosch et al., 1994).

In a recent work, a minimalist GPT implementation has been used to compare 17 orthographies (Marjou, 2021) on the same G2P and P2G tasks. A unique multi-orthography model was trained to learn the writing and reading tasks on all languages at the same time. In other words, a single dataset containing samples of all studied orthographies was used.

On the one hand, the results show that Chinese, French, English and Russian are the most opaque regarding writing. English and Dutch are the most opaque regarding reading.

On the other hand, they indicate that Esperanto, Arabic, Finnish, Korean, Serbo-Croatian and Turkish are very shallow both to read and to write. Italian, Breton, German, Portuguese and Spanish are shallow to read and to write.

3 Material and methods

All the code and data is available on GitHub[1].

3.1 Data

3.2 Preprocessing

3.3 Training

The training process was monitored using TensorBoard[2].

3.4 Evaluation

The metric used to rank systems is word error rate (WER), the percentage of words for which the predicted sequence does not match the gold sequence. The average WER was also calculated for each task. This is the same metric that was used in SIGMORPHON 2021 Task 1. The evaluation was performed using the scripts provided in that task.

3.5 Baseline

The baseline for the SIGMORPHON 2021 G2P task is a neural transducer system using an imitation learning paradigm (Makarov and Clematide, 2018). Alignments are computed using ten iterations of expectation maximization, and the imitation learning policy is trained for up to sixty epochs (with a patience of twelve) using the Adadelta optimizer. A beam of size of four is used for prediction. Final predictions are produced by a majority-vote ten-component ensemble.

4 Results

4.1 Comparison with baseline

The G2P results are compared with the baseline results on the dev and test sets in Table 1 and Figure 1. The results are worse in all languages but the ranking between languages is maintained. As we said before, our aim is not to get perfect results, so our model is good enough for our aim.

[1] https://github.com/juletx/writing-systems

[2] https://tensorboard.dev/experiment/mcrHkdndR0ySobcxV1UxVA

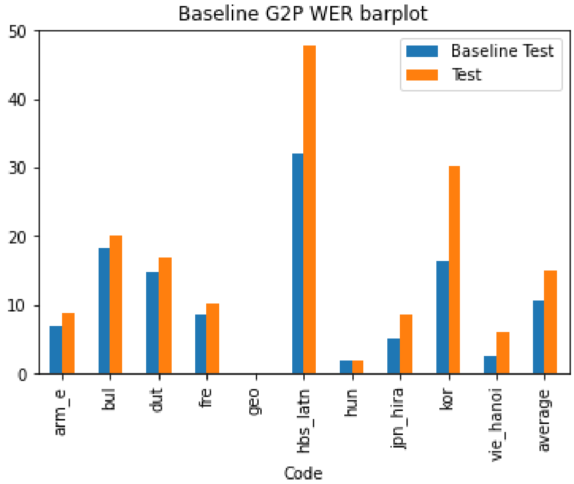

Figure 1: Comparison of our G2P results with the baseline

4.2 Comparison between languages

References

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Elisabeth Borleffs, Ben AM Maassen, Heikki Lyytinen, and Frans Zwarts. 2017. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and writing*, 30(8):1617–1638.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2018. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882, Brussels, Belgium. Association for Computational Linguistics.

Language	Code	Graphemes	Phonemes	Mean Gra	Mean Pho
Armenian (Eastern)	arm_e	38	41	7.09	7
Bulgarian	bul	29	45	8.44	8.64
Dutch	dut	30	45	7.71	6.99
French	fre	37	37	7.52	5.75
Georgian	geo	33	33	7.74	7.74
Serbo-Croatian (Latin)	hbs_latn	27	61	7.47	7.36
Hungarian	hun	34	61	7.65	7.18
Japanese (Hiragana)	jpn_hira	76	64	4.21	6.56
Korean	kor	559	60	2.58	6.54
Vietnamese (Hanoi)	vie_hanoi	89	49	5.81	7.51
Average	average	95.2	49.6	6.62	7.13

Code	Base Dev	Dev	Base Test	Test
arm_e	4.50	6.5	7.00	8.90
bul	8.30	14.9	18.30	20.10
dut	10.80	12.0	14.70	16.90
fre	7.40	9.5	8.50	10.20
geo	0.00	0.3	0.00	0.10
hbs_latn	34.70	49.2	32.10	47.70
hun	1.50	2.8	1.80	1.90
jpn_hira	6.20	8.4	5.20	8.50
kor	18.40	33.4	16.30	30.20
vie_hanoi	1.30	4.0	2.50	6.00
average	9.31	14.1	10.64	15.05

Table 1: Comparison of our G2P results with the baseline

Xavier Marjou. 2021. Oteann: Estimating the transparency of orthographies with an artificial neural network. *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

A.P.J. van den Bosch, A. Content, W.M.P. Daelemans, and B.L.M.F. de Gelder. 1994. Analysing orthographic depth of different languages using data-oriented algorithms. In *Proceedings of the Second International Conference on Quantitative Linguistics, QUALICO-94, Moscow, September 20-24, 1994*, pages 26–31. [s.n.]. Pagination: 6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction.

Code	G2P Train	G2P Dev	G2P Test	P2G Train	P2G Dev	P2G Test
arm_e	1.16	6.5	8.90	0.57	4.40	5.70
bul	0.34	14.9	20.10	16.25	19.70	27.30
dut	2.59	12.0	16.90	2.67	18.00	23.00
fre	2.04	9.5	10.20	32.96	53.50	54.00
geo	0.06	0.3	0.10	0.12	0.30	0.50
hbs_latn	22.56	49.2	47.70	0.09	0.20	0.60
hun	0.64	2.8	1.90	0.84	6.20	4.80
jpn_hira	2.74	8.4	8.50	0.59	2.40	3.90
kor	19.27	33.4	30.20	11.84	22.80	22.10
vie_hanoi	0.44	4.0	6.00	5.35	15.00	16.30
average	5.18	14.1	15.05	7.13	14.25	15.82