MG2P: An Empirical Study Of Multilingual Training for Manx G2P

Shubhanker Banerjee
ADAPT Centre
University of Galway
shubhanker.banerjee@adaptcentre.ie

Bharathi Raja Chakravarthi
ADAPT Centre University of Galway
bharathiraja.chakravarthi@adaptcentre.ie

John P. McCrae
ADAPT Centre University of Galway
john.mccrae@adaptcentre.ie

Abstract

Neural networks have achieved state of the art results on grapheme-to-phoneme (G2P) conversion. In this paper we focus on the development of a G2P system for Manx, an extremely low-resourced language of the Goidelic branch of the Celtic family of languages. We preprocess the data using two different data augmentation techniques which we call DA1 and DA2 and carry out experiments with various model architectures to answer the question *What is the optimal choice of data augmentation, training strategy and model architecture for building G2P systems in extremely low-resourced scenarios?* The results demonstrate that multilingual training of the Transformer with DA1 augmented Manx dataset along with data from orthographically similar English and Welsh improve upon the phoneme error rate of Phonetisaurus, LSTM and IBM model 2 by 10.25%, 14.42% and 24.05% respectively.

1 Introduction

Grapheme-to-phoneme (G2P) conversion is the task of generating a phoneme sequence representative of the pronunciation of a given input word. This conversion can be thought of as a sequence mapping task where graphemes in the input word are mapped to phonemes in the output sequence. In recent years, there has been tremendous increase in the efficiency and sophistication of computer aided tools. As a result these tools have increasingly been utilized in all spheres of life. Specifically, Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) tools have improved the accessibility of technology, more so for the disabled and the elderly.

G2P conversion is a critical component of TTS and ASR systems (Kim et al., 2002; Elias et al., 2021; Masumura et al., 2020). Pronunciation dictionaries can be used for building G2P systems, however such dictionaries have a limited coverage over the vast vocabulary of any language. This necessitates the development of G2P systems that can map written language to its phonemic transcription.

The problem statement defined in this paper is closely related to the work done by Jyothi and Hasegawa-Johnson (2017). They propose the use of recurrent neural networks (RNNs) for tackling G2P conversion in low-resourced scenarios and devise three different alignment strategies which are used to align the grapheme and phoneme sequences. These aligned sequences are then used to train a sequence-to-sequence model composed of RNNs (Rumelhart et al., 1985). The proposed model is evaluated on three low-resourced languages Pashto, Tagalog and Lithuanian. In order to understand the impact of size of the dataset on performance they carry out experiments with datasets of three different sizes: 250, 500 and 1000 samples and as expected they show that larger datasets improve the performance of the model. The main difference between the problem proposed in this paper and their problem statement is the size of the dataset; the size of our Manx dataset (refer to Section 4) is approximately 60% smaller than their smallest dataset (250 samples), thus making the development of a G2P system for Manx more difficult.

Zhao et al. (2022) propose a noise controlled G2P system wherein they inject noisy data during the training phase to develop models that less sensitive to orthographic noise in the data. They report significant significant improvements in the word error rate (WER) on dict-based sources.

Li et al. (2022) propose a zero-shot G2P model that uses data from related languages during training. The related languages are selected using a k-nearest neighbour approach on a phylogenetic tree of the language family.

G2P systems are usually language specific and are dependent on the orthographic properties of the language in consideration (Ager, 2008). There are challenges associated with the application of the rule-based or deep-learning-based G2P conversion methods for extremely low-resourced languages such as Manx. In these scenarios the linguistic expertise necessary to curate the grapheme-to-phoneme rules is often missing and this in turn makes the development of rule-based systems challenging. Furthermore, the development of deep learning based systems is dependent on annotated datasets which are also not available in extremely low-resourced scenarios. Even the results presented by Dong et al. (2022) where they sample 1000 pronunciations to simulate a low-resourced scenario is not representative of an extremely low-resourced language like Manx where very few data points are available to train the model (for details see Section 4).

In this paper, we study the impact of two different data augmentation strategies which we call DA1 and DA2 (for details see Section 3) as well as that of monolingual and multilingual training on the G2P conversion task. Specifically, we empirically analyze what is the optimal choice of data augmentation technique, training strategy and choice of model for G2P conversion of Manx, an extremely low-resourced language. We are particularly interested in how data from related languages can improve the performance in the multilingual training regime.

2 Related Works

G2P conversion has been an active area of research with a wide variety of methods being employed to tackle this problem (Taylor, 2005; Bisani and Ney, 2008; Rao et al., 2015; Chen, 2003; Novak et al., 2012; Dong et al., 2022). Braga et al. (2006) propose a rule-based system for G2P conversion of European Portuguese. The proposed system is intended as an unit of a larger TTS system. Their paper illustrates the G2P rules in European Portuguese and reports a very high phoneme accuracy rate of 98.80% achieved by the system. Deep learning based methods have achieved good performance on the G2P conversion task with LSTMs (Hochreiter and Schmidhuber, 1996) and Transformers (Vaswani et al., 2017) at the forefront of deep learning research in this area. Yolchuyeva et al. (2019) propose the use of the Transformer architecture for building a G2P conversion system for English. They train and evaluate the proposed model on the CMUDict and NetTalk datasets and report low ($\sim$ 5%) Phoneme Error Rate (PER). Juzová et al. (2019) propose an encoder-decoder architecture composed of bi-LSTMs to tackle the G2P problem for English, Czech and Russian. They report high phoneme accuracy rates for all of the three languages. Dong et al. (2022) propose GBERT, a multi-layer Transformer encoder inspired by the BERT architecture (Kenton and Toutanova, 2019). Monolingual word lists with randomly masked graphemes (letters) are used to pre-train the GBERT encoder with the masked grapheme objective. The GBERT encoder is then trained/fine-tuned on the G2P conversion task with a Transformer decoder. Experiments have been carried out in the low and medium resourced scenarios and the results indicate the better performance achieved by masked grapheme pre-training.

The DA1 augmentation scheme proposed in this paper is closely related to the work done by Hammond (2021). They propose the use of LSTM (Hochreiter and Schmidhuber, 1996) to tackle G2P conversion for 10 low-resourced languages. Each of these languages has 800 word-pronunciation pairs available for training; in order to augment the training sets splitting of words based on unambiguous mapping of peripheral grapheme sequences to phoneme sequences is proposed. Multilingual training for G2P conversion of Manx in this paper was inspired by the work carried out by Vesik et al. (2020) where they propose the use of multilingual training of Transformers (Vaswani et al., 2017) on the G2P conversion task. They carry out experiments on 15 languages with relatively larger datasets of 4050 samples. The system was trained in a multilingual setting where each source grapheme sequence was prepended with the corresponding language identifier to allow the model to learn meaningful representations from the combined dataset while having the ability to discriminate amongst the languages during inference. The results show an improvement of over 50% in the phoneme and word error rates (PER and WER). We have also carried out experiments to empirically analyze the method proposed by Prabhu and Kann (2020) where they train a Transformer model jointly on grapheme-to-phoneme as well as phoneme-to-grapheme tasks i.e both the forward and the backward directions at each time step of the training. Their results indicate marginal improvement in performance on joint training. Novak et al.

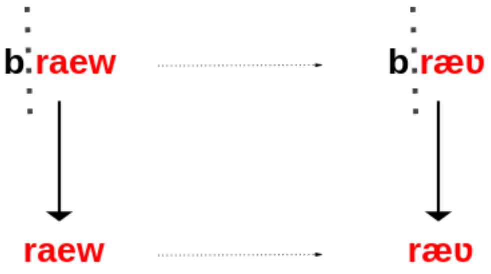

Figure 1: DA1 applied to *braew* such that it is split into two grapheme sequences *b* and *raew*. The mapping of *raew* to ræ*υ* is independent of *b* and therefore is treated as a separate datapoint in addition to the original word i.e. *braew*. This split point is not based on linguistic rules but an observation of the grapheme and the phoneme sequences which shows that there is a direct correspondence between the phoneme *b* and the grapheme *b* and thus the split point at *b*.

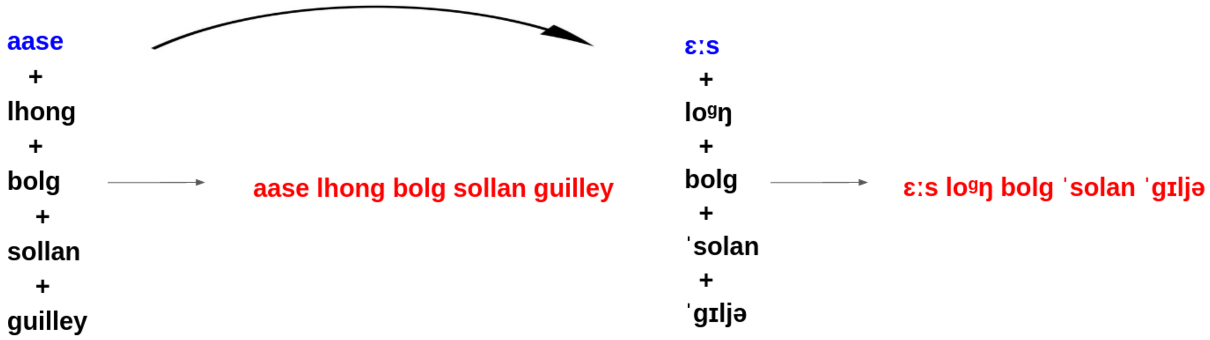

Figure 2: For *aase* we randomly sample 4 words *lhong*, *bolg*, *sollan* and *guilley* and concatenate them together to form the string *aase lhong bolg sollan guilley*, which is a new data point. The corresponding phonemic representations are also concatenated as illustrated in the figure.

(2016) introduced Phonetisaurus a joint n-gram based grapheme-to-phoneme toolkit built upon OpenFST framework[1]. El-Hadi and Mhania (2017) carry out experiments on letter-to-sound mapping using Phonetisaurus and demonstrate good results thereby demonstrating its applicability to this task.

3 Data Augmentation

We introduce two data augmentation techniques namely, DA1 and DA2. The idea behind the DA1 augmentation scheme is that certain grapheme segments which are substrings of the original word can be mapped unambiguously to phoneme segments given that appropriate splitting points are found in the original word (see Figure 1 for details). There can be multiple such splitting points in a word leading to the creation of multiple such data points from one word-phoneme pair. The hypothesis is that creation of such subword level pronunciation pairs improves the learnability of the model with regards to the fine-grained grapheme-to-phoneme rules.

In DA2 augmentation scheme for every word in the pronunciation list we randomly sample 4 other words from the word list and concatenate all the 5 words and correspondingly their 5 pronunciations (see Figure 2 for details). The resultant sequence-phoneme pair is now treated as a new datapoint and used in training. The hypothesis is that longer and more diverse sequences would help improve the performance of the model.

4 Dataset

The problem statement has been framed as a supervised learning problem and therefore a parallel word list comprising of words and their corresponding phonemic representations (pronunciations) is needed to train the model. In the multilingual training regime the idea is to leverage the phonetic and orthographic similarity of related languages to augment the Manx data available for training. Irish and Scottish Gaelic belong to the the same Goidelic language family as Manx and have a similar phonology (Paul, 2014), Welsh and English

[1] https://www.openfst.org

Language	Train	Valid	Test
English	1,264	___	___
Irish	1,032	___	___
Welsh	512	___	___
Manx	77	34	28
Scottish Gaelic	86	___	___

Table 1: Split Statistics after Data augmentation

have an orthography similar to that of Manx (Gelb, 1968). Therefore, we collect pronunciation lists for English, Welsh, Scottish Gaelic and Irish. In order to collect the data required for the experiments, we use the Wikipron library (Lee et al., 2020) which allows the extraction of pronunciations from Wiktionary[2]. It must be noted that during data collection we collect all available data points for Manx, Welsh, Irish and Scottish Gaelic. However, we limit the number of English samples to 1300 words. The reason behind doing so is to simulate situations where the main language (Manx in this case) as well as all related languages are low-resourced. Furthermore, we observe the presence of repeated entries in the English dataset. On removing these repeated entries we are left with 1264 words.

Initially, 106 Manx samples are collected for Manx using the Wikipron API. We then manually apply DA1 to these 106 words and observe that 33 word-pronunciation pairs can be split into two as illustrated in Figure 1 leading to the creation of 33 additional datapoints. Thus, a total of 139 grapheme-phoneme pairs are obtained after applying DA1. In order to compare DA1 and DA2 we then choose the same 33 words from the original pronunciation list and apply DA2 to each of these 33 word pronunciation pairs i.e for each of these 33 words we randomly choose 5 more words and concatenate them to the originally chosen word; the corresponding pronunciations are also concatenated. Thus, 139 samples are generated by applying the DA2 augmentation scheme. The Manx dataset obtained after the data augmentation has 139 samples and is split in the ratio of 80:20 train-test split. The train dataset is further split in the ratio of 70:30 train-validation split. The resultant dataset statistics are illustrated in Table 1. It illustrates the extremely low-resourced nature of Manx and reinforces the previously mentioned challenges associated with building deep learning systems that are capable of mapping graphemes to phonemes with such few datapoints.

5 Background

5.1 IBM Model 2

IBM Model 2 is a translation model that was introduced by Brown et al. (1993) and is based on the noisy-channel model of parameter estimation (Weaver, 1949). It is important to note here that in this case the words are the source sequences and the corresponding pronunciations are the target sequences. The source sequences are translated into the target sequences according to a translation table and an alignment function which are learned from the data. For more details on IBM Model 2 we refer the reader to Brown et al. (1993).

5.2 LSTM

Recurrent Neural Networks (RNNs) are a class of neural networks that are capable of modelling time-distributed data sequences (Rumelhart, 1986). However, they suffer from the problem of vanishing gradients over a larger number of time steps (Basodi et al., 2020). Long Short-term Memory network (LSTM) first introduced by Hochreiter and Schmidhuber (1997) mitigate this problem by selectively retaining information over a larger number of time steps. LSTMs have achieved good performance across a wide variety of NLP tasks such as language modelling (Sundermeyer et al., 2012), sentiment classification (Wang et al., 2016), speech recognition (Graves et al., 2013) and named entity recognition (Jin et al., 2019). For further details on the gated architecture of a LSTM cell we refer the reader to Hochreiter and Schmidhuber (1997).

5.3 Phonetisaurus

Phonetisaurus is an open-source grapheme-to-phoneme converter based on the OpenFST frame-

[2] https://en.wiktionary.org/w/index.php?title=Category:Terms_with_IPA_pronunciation_by_language&from=W

work first introduced by Novak et al. (2016). It uses joint n-gram models to learn a mapping from graphemes to phonemes. The first step in the Phonetisaurus pipeline is the alignment of the source and the target sequences based on a modified form of the algorithm proposed by Jiampojamarn et al. (2007). The next step involves training a n-gram language model which is then used to construct a Weighted Finite State Transducer (WFST) (Novak et al., 2012). The final step involves decoding using the WFST constructed in the previous step, the decoder finds the optimal phoneme sequence for a given input sequence of graphemes. For more details on the Phonetisaurus pipeline we refer the reader to Novak et al. (2016).

5.4 Transformer

The Transformer architecture first proposed by Vaswani et al. (2017) was introduced with the objective of mitigating the challenges associated with the recursive structure of sequence modelling neural architectures such as RNN and LSTM. The Transformer architecture is an encoder-decoder architecture with both the encoder and the decoder composed entirely of attention (Bahdanau et al., 2015) blocks. Transformer and modifications to its architecture such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020) have achieved state-of-the-art results on various natural language processing tasks (Patil et al., 2022; Do and Phan, 2022; Yang et al., 2022). For further details on the Transformer architecture we refer the reader to Vaswani et al. (2017).

6 Experiments

As mentioned previously the development of rule-based systems for low-resourced languages such as Manx is challenging due to the absence of linguistic expertise. Concretely, there are three primary challenges:

- The curation of G2P rules for Manx often depends on the number of syllables in a word and whether the consonants are broad or slender (Pickeral III, 1990). Ascertaining these for a particular word requires specialist linguistic knowledge of Manx.

- The quality of a vowel depends on factors such as height of the tongue with relation to the jaw and horizontal position of the tongue in the mouth. Such variation in the quality of a vowel leads to difference in pronunciation in different contexts (Pickeral III, 1990). As a result vowel letters often have one-to-many mappings with phonemes and thus the curation of rules mapping vowels to their corresponding phonemes is a linguistically involved task.

- Manx exhibits initial consonant mutation. The pronunciation of the initial consonant of a word alters depending on the morpho-syntactic context (Hannahs, 2013). Such alterations further complicate the curation of grapheme-to-phoneme rules for the language.

We carry out experiments with deep learning based methods and WFST based Phonetisaurus to empirically study their suitability for building G2P systems for Manx. The optimal hyperparameters are found by training on the train data and manual tuning on the validation set. 5 trials were conducted for hyperparameter search on the LSTM model using only Manx data during training, whereas the optimal hyperparameters for the Transformer model were found in 9 search trials using only Manx data. The test results have been reported in the form of mean and standard deviation of 5 evaluations on the test set using the optimal hyperparameters.

Data Augmentation	**PER**
No Data Augmentation	90.75 ±1.23
DA1	**87.52 ± 0.75**
DA2	280.94 ±1.65

Table 2: Preliminary Results

6.1 Preliminary Experiments

We carry out preliminary experiments to study the impact of the two proposed data augmentation schemes on performance. Both DA1 and DA2 are applied to the original dataset independently and resultant datasets are used to train LSTM based sequence-to-sequence models for Manx G2P conversion. Furthermore, the unaugmented dataset is also used to train a model on the same task to establish a baseline. Phoneme error rate (PER) is used as the evaluation metric. It is a measure of the percentage of phonemes incorrectly generated by the model for each word. The results illustrated in Table 2 show that the performance significantly deteriorates with DA2 and marginal improvement

Model	LangID	gv	gv+ga+gd	gv+ga	gv+gd	gv+cy	gv+en	gv+cy+en
IBM 2	No	73.58±1.45	73.48±4.87	73.46±3.89	73.89±0.64	75.01±3.21	79.05±3.99	81.79±0.01
	Yes	73.58±1.45	73.48±4.87	73.46±3.89	73.89±0.64	75.01±3.21	79.05±3.99	81.79±0.01
LSTM	No	86.98±3.99	96.58±5.32	98.52±1.23	86.23±0.23	116.10±1.68	84.98±2.99	139.47±1.00
	Yes	70.89±2.09	62.00±1.99	64.96±2.43	70.89±1.78	112.98±3.56	65.82±4.56	73.39±5.32
Transformer	No	96.35±1.89	58.71±3.48	64.96±2.79	73.67±3.45	61.42±1.65	61.99±2.45	55.39±4.87
	Yes	73.89±1.00	59.14±2.67	64.01±0.24	70.49±0.98	58.86±6.25	62.13±1.12	**49.53±0.01**
Phonetisaurus	No	57.24±0.56	103.49±0.05	104.91±1.26	69.81±0.09	74.71±1.19	72.00±0.85	68.56±0.02

Table 3: PER without Language Identifiers

over the baseline is observed with DA1, thereby indicating the better performance of DA1 scheme on the G2P task. Thus, going forward all experiments are carried out with the DA1 augmented dataset.

6.2 Multilingual Training

The hypothesis is that training the models on the combined datasets would allow them to learn meaningful representations by leveraging the additional training data from related languages. However, this raises a question on the models' ability to discriminate amongst languages during inference. The same grapheme might have same or different phoneme mappings across languages. To mitigate this problem, we prepend language specific identifiers to words and their phonemic representations. We hypothesize that adding these identifiers would facilitate the learning of language specific representations which in turn would allow the model to meaningfully utilize data from related languages to learn grapheme-to-phoneme rules while also enabling distinction amongst the languages during inference.

In order to study the validity of our hypotheses related to multilingual training and language identifiers we carry out experiments with IBM model 2, LSTM and the Transformer architecture. Multilingual models are trained on a Nvidia RTX2060 GPU using various subsets of the related languages both with and without language identifiers. These models are then evaluated on the Manx test data.

The results are illustrated in Table 3 and show that performance of the LSTM and the Transformer models trained on data with language identifiers is better than those trained without these identifiers. For the purpose of brevity these languages have been referred to by the following ISO 693-1 language codes in Tables 3: Manx (*gv*), Irish (*ga*), Scottish Gaelic (*gd*), Welsh (*cy*) and English (*en*). No improvement in performance is observed with the addition of language identifiers in case IBM model 2. Furthermore, the Transformer model trained multilingually on English, Welsh and Manx data with language identifiers attains a PER of 49.53% and outperforms all other monolingual and multilingual models. The training configuration of this model is given in Table 4. It improves upon the PER (74.24%) of the baseline monolingual Transformer trained only on Manx data by a significant 24.71%.

Hyperparameter	Value
Number of Encoder & Decoder Blocks	2
Number of Attention Heads	2
Number of Training Epochs	200
Batch Size	16
Embedding Dimension	256
Maximum Sequence Length	256

Table 4: Training configuration of the best model (en+cy+gv)

6.3 Joint Training

The mappings from graphemes to phonemes (G2P) and from phonemes to graphemes (P2G) are monotonic relationships that proceed from left to right. We hypothesize that joint training of the model on both G2P and P2G tasks would facilitate the learning of the monotonic nature of these mappings. Furthermore, given that phonemes and graphemes have a bidirectional mapping between them, that is any given phoneme can be mapped to one or many graphemes and the vice-versa, we hypothesize that training the model to map a phoneme to a specific set of graphemes should introduce signals that drive the model towards optimal performance on the G2P task.

$$\ell(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log P(T^i|S^i;\theta) + \sum_{j=1}^{N}\log P(S^i|T^i;\theta) \tag{1}$$

Thus, at each training step the model loss given by Eqn.1 is used to optimize the model parameters where S is the set of words and T is the set of corresponding phonemic sequences. As shown in Section 7, the model trained on the downsized English-Welsh dataset alongwith the Manx data has the best performance on the G2P task. In order to validate our hypothesis on joint learning, we use performance as a baseline and train a Transformer model jointly on the G2P and P2G tasks using the downsized English-Welsh data. The hyperparameters used during training are listed in Table 4. During evaluation we observe a PER of 71.45%. This result invalidates our hypothesis related to improvement of performance by introduction of the auxiliary P2G task during training.

6.4 Phonetisaurus

We carry out experiments with Phonetisaurus to assess its suitability for extremely low-resourced languages like Manx. We train the model on subsets of related languages along with the DA1 augmented Manx dataset and the results are presented in Table 3. The results indicate that the performance of Phonetisaurus in general is worse than the best performing model described in Section 7. This result further reinforces the optimality of multilingual training of Transformer to tackle G2P conversion in extremely low-resourced scenarios.

7 Ablation Study

As shown in Section 6.2, the best result is achieved by using data from English and Welsh alongside Manx. English and Welsh are orthographically similar to Manx and the size of the dataset (1,776 samples) is greater than that of the combined Irish and Scottish Gaelic dataset (1,118 samples). To ascertain the impact of orthographic similarity and size of the dataset on the performance we randomly sample 1,118 datapoints from the English-Welsh dataset. The hypothesis is that if orthographic similarity amongst the related languages and Manx is the dominant factor then the performance achieved by the model trained on the downsized English-Welsh dataset should be better than that achieved by training on the phonetically similar Irish-Scottish Gaelic dataset of the same size. In order to validate our hypothesis we train a Transformer model on the downsized English-Welsh dataset with language identifiers using the training configuration demonstrated in Table 4. Then we evaluate the trained model on Manx test data and observe a PER of **47.94%**. Thus, the model trained on downsized English-Welsh data outperforms the Transformer model trained on the Irish-Scottish Gaelic (PER - 59.14%) dataset by 11.2% validating our initial hypothesis about the impact of orthographic similarity on performance of the system. Furthermore, it also marginally improves upon the performance of the model trained on the full English-Welsh dataset by 1.59%.

8 Computational Cost

The LSTM model used for preliminary experiments has 613,424 parameters whereas the transformer model used for multilingual training and joint training has 3,787,776 parameters. The average runtime of the LSTM model is 62ms per gradient step during training whereas for the Transformer architecture we observe an average runtime of 111 ms per gradient step during training. During inference, the transformer model took 15 ms per input instance and the LSTM had a runtime of 5ms per instance.

9 Error Analysis

We analyze the sequences generated by the best performing model described in Section 7 and observe that in 75% of the sequences, more than 50% of the errors were accounted for by the vowels. We observed that this is due to following two reasons primarily:

- The vowel sound is incorrectly classified altogether. Lhong should be transcribed to loᵍŋː, but is transcribed to løɔ̠ːŋ .
- The quality of the generated vowel is incorrect. For example the vowel e in ane should be transcribed to ɛːnˌ (Open-mid unrounded vowel), but it is transcribed to e:n (Close-mid unrounded vowel).

10 Results

The preliminary results demonstrated in Table 2 show that the PER achieved by LSTM models across the augmented and the original datasets is not very low. This is primarily because these models are trained only on extremely small Manx datasets which are not sufficient to train deep learning models. However, we empirically observe that multilingual training using related languages improves performance on the G2P task as shown by the results demonstrated in Table 3. The use of identifiers that enable the discrimination amongst languages during training have a positive impact on the performance of the model. Also, the optimality of Transformers for this task when they are trained on appropriate datasets is established. Furthermore, as observed in Section 7 orthographically similar languages have a greater impact on the performance of the model. This indicates that languages with similar writing systems when used in the multilingual training regime are more effective than phonetically similar languages. The experiments carried out using IBM model 2 show that there is no significant improvement in the performance of the model in the multilingual training regime. In order to validate our hypothesis as stated in Section 6.3 we conduct experiments by introducing an auxiliary P2G task during training. The results are significantly lower than those of the model described in Section 7 and invalidate our initial hypothesis; joint training on both tasks leads to catastrophic forgetting (Kirkpatrick et al., 2017) and therefore the performance of the model is suboptimal.

We also conduct experiments with Phonetisaurus to assess its applicability for this task. The result does not improve upon the performance of the multilingual model described in Section 7. Furthermore, as indicated by the results presented in Table 3, the performance of Phonetisaurus worsens when data from related languages is introduced during training. It must also be noted that the performance of the Phonetisaurus model trained only on the DA1 augmeneted Manx dataset is better than other monolingual models shown in Table 3. Finally, the PER of 47.94% achieved by the model trained on English-Welsh dataset is not optimally low, however the results indicate that design of better data augmentation schemes alongwith improved multilingual training mechanisms leave the scope open for development of G2P systems for Manx.

11 Conclusion

To conclude, we carry out experiments to identify the optimal training regime, model architecture and data augmentation scheme to build a G2P system for Manx, an extremely low-resourced language. We propose the use of two augmentation schemes DA1 and DA2 to counter the low-resourced nature of Manx and empirically observe an improvement in performance when DA1 is applied to the original dataset. The results indicate that multilingual training of Transformer on data from orthographically similar languages in the presence of language identifiers outperforms all other monolingual as well as multilingual models. This is an interesting result and opens up avenues for application of other multilingual training methodologies for G2P conversion, especially for low-resourced languages where not a lot of training data is available.

12 Acknowledgement

Author Shubhanker Banerjee was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University Of Galway.

References

Simon Ager. 2008. Omniglot-writing systems and languages of the world. *Retrieved January*, 27:2008.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Sunitha Basodi, Chunyan Ji, Haiping Zhang, and Yi Pan. 2020. Gradient amplification: An efficient way to train deep neural networks. *Big Data Mining and Analytics*, 3(3):196–207.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Daniela Braga, Luís Coelho, and Fernando Gil Vianna Resende. 2006. A rule-based grapheme-to-phone converter for TTS systems in European Portuguese. In *2006 International Telecommunications Symposium*, pages 328–333. IEEE.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stanley F. Chen. 2003. Conditional and joint models for grapheme-to-phoneme conversion. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2033–2036.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Phuc Do and Truong HV Phan. 2022. Developing a BERT based triple classification model using knowledge graph embedding for question answering system. *Applied Intelligence*, 52(1):636–651.

Lu Dong, Zhi-Qiang Guo, Chao-Hong Tan, Ya-Jun Hu, Yuan Jiang, and Zhen-Hua Ling. 2022. Neural grapheme-to-phoneme conversion with pre-trained grapheme models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6202–6206. IEEE.

Cherifi El-Hadi and Guerti Mhania. 2017. Phonetisaurus-based letter-to-sound transcription for Standard Arabic. In *2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B)*, pages 1–4. IEEE.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu. 2021. Parallel tacotron: Non-autoregressive and controllable TTS. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5709–5713. IEEE.

I.J. Gelb. 1968. Orthography studies, articles on new writing systems: William A. Smalley and others. Helps for Translators, Volume VI (Published by the United Bible Societies, London, in co-operation with the North-Holland Publishing Company, Amsterdam, 1964). VII + 173 pages. *Lingua*, 20:319–323.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

Michael Hammond. 2021. Data augmentation for low-resource grapheme-to-phoneme mapping. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 126–130, Online. Association for Computational Linguistics.

SJ Hannahs. 2013. Celtic initial mutation: pattern extraction and subcategorisation. *Word Structure*, 6(1):1–20.

Sepp Hochreiter and Jürgen Schmidhuber. 1996. LSTM can solve hard long time lag problems. *Advances in neural information processing systems*, 9.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379.

Yanliang Jin, Jinfei Xie, Weisi Guo, Can Luo, Dijia Wu, and Rui Wang. 2019. LSTM-CRF neural network with gated self attention for Chinese NER. *IEEE Access*, 7:136694–136703.

Markéta Juzová, Daniel Tihelka, and Jakub Vít. 2019. Unified Language-Independent DNN-Based G2P Converter. In *INTERSPEECH*, pages 2085–2089.

Preethi Jyothi and Mark Hasegawa-Johnson. 2017. Low-resource grapheme-to-phoneme conversion using recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5030–5034. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Byeongchang Kim, Gary Geunbae Lee, and Jong-Hyeok Lee. 2002. Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):65–82.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2020. Phoneme-to-grapheme conversion based large-scale pre-training for end-to-end automatic speech recognition. In *INTERSPEECH*, pages 2822–2826.

Josef R Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49.

Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. 2016. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

Priyadarshini Patil, Chandan Rao, Gokul Reddy, Riteesh Ram, and SM Meena. 2022. Extractive Text Summarization Using BERT. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 741–747. Springer.

Russell Paul. 2014. *An Introduction to the Celtic Languages.* Longman Linguistics Library. Routledge.

John J Pickeral III. 1990. A Preliminary Phonology of Manx. *Orbis*, 35:81–97.

Nikhil Prabhu and Katharina Kann. 2020. Frustratingly easy multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 123–127, Online. Association for Computational Linguistics.

Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

De Rumelhart. 1986. Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1:318–362.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Paul Taylor. 2005. Hidden Markov Models for grapheme to phoneme conversion. In *Ninth European Conference on Speech Communication and Technology*. Citeseer.

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

Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2020. One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 146–152, Online. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Warren Weaver. 1949. The mathematics of communication. *Scientific American*, 181(1):11–15.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer based grapheme-to-phoneme conversion. *Proc. Interspeech 2019*, pages 2095–2099.

Chendong Zhao, Jianzong Wang, Xiaoyang Qu, Haoqian Wang, and Jing Xiao. 2022. r-g2p: Evaluating and Enhancing Robustness of Grapheme to Phoneme Conversion by Controlled Noise Introducing and Contextual Information Incorporation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6197–6201. IEEE.