INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A 43 Language Multilingual Punctuation Prediction Neural Network Model

Xinxing Li, Edward Lin

Microsoft China
{xinxili, edlin}@microsoft.com

Abstract

Punctuation prediction is a critical component for speech recognition readability and speech translation segmentation. When considering multiple language support, traditional monolingual neural network models used for punctuation prediction can be costly to manage and may not produce the best accuracy. In this paper, we investigate multilingual Long Short-Term Memory (LSTM) modeling using Byte Pair Encoding (BPE) for punctuation prediction to support 43 languages[1] across 69 countries. Our findings show a single multilingual BPE-based model can achieve similar or even better performance than separate monolingual word-based models by benefiting from shared information across different languages. On an in-domain news text test set, the multilingual model achieves on average 80.2% *F1*-score while on out-of-domain speech recognition text, it achieves 73.5% *F1*-score. We also show that the shared information can help in fine-tuning for low-resource languages as well.

Index Terms: multilingual punctuation prediction, byte pair encoding, long short term memory, speech recognition

1. Introduction

Punctuation prediction provides segmentation of Speech Recognition (SR) output for human and machine consumption. This is critical for applications where text converted from speech needs to be readable such as dictation, visual voicemail, and meeting transcriptions. Similarly, downstream Natural Language Processing (NLP) tasks such as machine translation, intent recognition, and analytics utilize punctuation to provide more accurate results. In most speech recognition production systems, punctuation prediction is an independent component, added after text is produced.

Punctuation prediction is a sequence tagging problem, like Part-of-Speech Tagging and Named Entity Recognition. Traditional punctuation prediction methods rely only on lexical features [1, 2], while others both lexical and acoustic features [3, 4]. In [5] researchers adopt lexical features and build an extended language model taking punctuation as inner tokens. [6] uses a maximum entropy model that uses both lexical and acoustic features for prediction. There is also other combination method which separately models lexical and acoustic features with N-gram and Decision Tree (DT) and then tries model combination method [7]. Considering contextual information, Conditional Random Fields (CRF) further improved punctuation quality by combining the punctuation probability distribution with the text sequence [8]. Compared with only using lexical features, CRF model shows great improvement combining various features, such as LM scores, n-gram tokens, sentence length and syntactic features [9].

More recently, Recurrent Neural Network (RNN), and the variant Long Short-Term Memory (LSTM), have shown better results by being able to consider long range context [10]. Bidirectional LSTM (BLSTM) performs better than unidirectional LSTM by utilizing future context [11]. Based on this, [12] proposes BLSTM model with attention mechanism to further improve the performance by weighting the relevant parts of the context for punctuation prediction. For better modeling word relationships, encode-decoder model structure has been used [13, 14], which learns from the experience from neural machine translation. In [15] researcher applies self-attention based model to predict punctuation using word and speech embedding features. There is also attempt to multilingual punctuation generation, which proposes a transition-based LSTM method on 5 languages [16].

The most common punctuation symbols are generally universal across all languages—period, comma, question mark, but because each language has a unique set of words in the vocabulary, to support punctuation in many languages, requires a punctuation model for each language or a single model that must support an extremely large vocabulary set. If regional linguistic similarities are considered, such as the Romance languages (French, Portuguese, Spanish, and Italian), words may be shared across languages. Fortunately, research into subword representations [14, 17] have shown to yield promising results in tasks such as multilingual machine translation [18, 19]. Moreover, inspired by utilizing the shared linguistic information across languages, many NLP work have tried multilingual method [20, 21]. On Name Entity Recognition, [22] shows surprisingly good results at cross-lingual model transfer for zero-shot languages with a pre-trained multilingual model.

In this paper, we investigate applying multilingual techniques to the punctuation prediction task to support 43 languages across 69 countries with a single model. Instead of relying on spoken transcription punctuation labels which would be cost prohibitive, we rely exclusively on text from news assets worldwide. We utilize Byte Pair Encoding (BPE) [23] to model sub-word units across languages, and use an LSTM to predict punctuation. Our findings show that a single multilingual model can get comparable or even better results than the monolingual word-based, by adding more languages and deeper models. We also show that fine-tuning the multilingual model brings improvement on low-resource languages.

The rest of the paper is organized as follows. Section 2 introduces the multilingual punctuation prediction architecture. Section 3 describes the experimental setup. Section 4 gives the experiment results and analysis. Section 5 is the conclusions drawn from the experiment.

[1] 43 languages: Afrikaans, Armenian, Catalan, Chinese (Simplified, Traditional), Croatian, Czech, Danish, Dutch, English, Estonian, Farsi, Finnish, French, Galician, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Malay, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovenian, Slovak, Spanish, Swedish, Turkish, Tamil, Ukrainian, Urdu, Vietnamese

http://dx.doi.org/10.21437/Interspeech.2020-2052

2. Multilingual Punctuation Architecture

In this section we describe the 2 main components of our architecture, BPE and LSTM, and an overview of our multilingual model.

2.1. Multilingual BPE

BPE is a data compression technique to represent words in smaller units to reduce overall vocabulary size and word sparsity by sharing the same sub-word units between different words. In BPE model training, characters are paired according to the highest frequency, treated as a new char and merging continuous until the targeted vocabulary size is achieved. In our multilingual work, we adopt the public multilingual BPE model, MultiBPEmb[2][24], which is trained with 275 languages wikipedia data, which covers all the languages in our punctuation model.

To evaluate the efficacy of the BPE model we analyzed a set of similar languages, French, Italian and Spanish, to understand how much overlap exists between them. After encoding the 3 languages' training data with MultiBPEmb, French has 23.6k unique tokens, Italian has 24.3k and Spanish has 22.9k. We calculate the overlap percent of the 3 languages, here is the overlap percent,

$$Spanish \cap (French \cup Italian) = 93.2\%$$
$$French \cap (Italian \cup Spanish) = 95.4\%$$
$$Italian \cap (French \cup Spanish) = 95.8\%$$

The overlap percent shows that each of these languages are closely related. From the overlap percent we can believe that MultiBPEmb is able to help us explore the shared sub-words across languages.

2.2. LSTM with shift

We adopt LSTM to perform the sequence labeling task, because it combines the benefit of lower resource utilization when compared to BLSTM or encoder-decoder models but still provides high accuracy.

To improve accuracy, we want the benefit of looking ahead a few BPE tokens in order to utilize the future context to help prediction as BLSTM is able to, so we added a token shift with LSTM model. Suppose *shift=2*, we will append 2 special tokens ⟨PAD⟩ at the end of text sequence and add 2 ⟨PAD⟩ before the punctuation sequence, then we re-align the text and punctuation sequence and feed into the LSTM model. In this way, the model can lookahead some future tokens.

In [11], punctuation has shown improvements with *shift=1*. In our Mandarin Chinese monolingual experiment, LSTM with *shift=3* has around 2% *F1-score* gap with BLSTM model, without *shift* the gap increases to 15%.

2.3. Model structure

The complete illustration of our multilingual model is shown in Figure 1. Our prediction labels include comma, period, question mark and ∅ representing no punctuation.

Our method is language independent, so we don't pass any language information into either BPE component or LSTM model. After multilingual BPE, words are split into BPE tokens. We remove the word boundary tag attached by MultiBPEmb to further decrease the token size. The punctuation is only placed after the last BPE token of the word, other inner tokens are labelled with ∅. We don't adopt pre-trained embedding vectors for the BPE tokens and the embedding layer is trained together with the LSTM layer.

[2]https://github.com/bheinzerling/bpemb

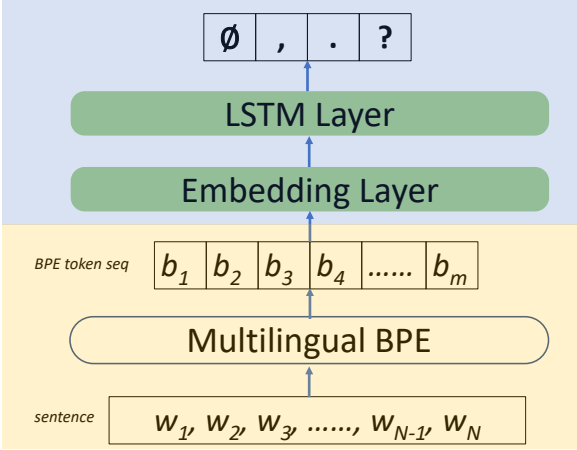

Figure 1: *Illustration of multilingual model*

3. Experimental Setup

3.1. Data

We utilize 43 languages news text across 69 countries as training data. Among these languages, only Chinese (Traditional, Simplified) and Japanese are non-spacing language. For the non-spacing languages, BPE also plays the role of word breaking. We do some pre-process on the data, like replacing semicolon with comma, replacing exclamation mark, ellipsis with period, converting all digits to "0".

The training set contains 545.7M sentences. The training set is unbalanced, we get much data of English, Spanish, French, German, and so on, while for Afrikaans, Galician, Icelandic, Tamil and Urdu, we don't get much data.

Table 1: *Avg statistics in test set*

Token	News	SR text
BPE tokens	84.8	21.4
comma	3.47	1.3
period	2.46	1
question	0.22	0.31

We use two kinds of data to evaluate our model, in-domain news test set to test our model in the same condition in which the model is trained, and another out-of-domain SR text set to test how the model generalizes to speech data. The news test set covers all the 43 languages, but the SR test set only contains data from 10 of the languages[3]. We collect SR text set from the recognition results of TV show, simulated voice-mail and simulated dictation audios with low word error rate. We label the text with punctuation while providing audio to listen to. Each sentence is annotated by 3 annotators, the inter-annotator agreement is 2. The news test set contains 1.678M sentences while

[3]SR test set: Chinese(Simplified), English, German, French, Hindi, Italian, Japanese, Portuguese, Russian, Spanish

SR test set contains 29.3K sentences. Table 1 shows the comparison of the average count in each sentence between news and SR test set, from which we can see news data is much longer and has more comma and period, while less question mark. Training data has the same distribution as the news test text. Besides this, SR text also has a spoken expression style unique to news, like filler words, or spoken repetition which also can influence the punctuation prediction.

3.2. Model

After BPE, we totally have 133K BPE tokens. The embedding layer is 256 dims, we trained 1, 2 and 4 layers LSTM model with 1024 nodes each layer separately. We also trained monolingual word-based model for the 10 languages as baseline, which had 1 layer with 1024 nodes. We set *shift=4* for word model and *shift=8* for BPE model empirically, as BPE token is smaller unit.The model is trained with PyTorch[4].

3.3. Metrics

We use Precision, Recall and *F1*-score to evaluate the model performance, and we distinguish the period and question mark whether they occur in the middle of sentence or at the end.

4. Experimental Results

4.1. Romance languages

Our first multilingual experiment examines word and BPE-based on 4 Romance languages: French, Italian, Spanish and Portuguese. We trained a 3 language multilingual model with the first 3 languages, and we take Portuguese as a low-resource language and fine-tune with the multilingual model. The training set contains 25M sentences, and we take 10K sentences of each language as test set.

Table 2: *F1-score of the Romance language*

Language	word-based	MonoBPE	MultiBPE
French	76.4%	73.8%	76.1%
Italian	66.8%	63.5%	67.1%
Spanish	81.9%	81.0%	81.7%

Table 2 shows the word-based, monolingual BPE and multilingual BPE model *F1*-score of the 3 languages. The multilingual BPE model *F1*-score is similar to the word-based model, while monolingual BPE model has 2.3%, 3.6% and 0.7% gap on the 3 languages compared with multilingual BPE model. The results show that while BPE performs worse than word, adding additional languages to the model makes up for the difference because of the shared knowledge the model learns.

Next we adapted the 3 language multilingual BPE model with Portuguese data, treating it like a low resource language. What needs to be pointed out is that the overlap percent of Portuguese BPE token with the union of the 3 languages is 97.6%. We increase the training data size, observing the *F1*-score and comparing it with a model trained with only Portuguese data. As shown in Figure 2, the most obvious thing is that *F1*-score gets greatly improved along with the increase of data size, both for adapted model and that from zero. Secondly, the model from zero always has around 2% *F1*-score gap compared with the adapted multilingual model, which highlights that Portuguese is able to benefit from the relationship to the other 3 languages.

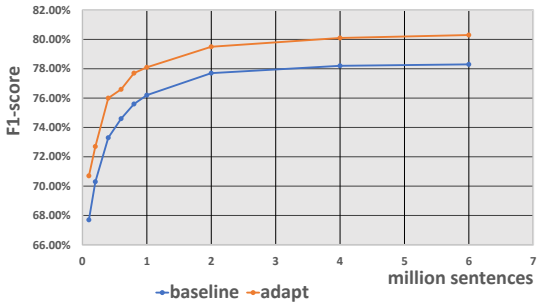

Figure 2: *Trend of Portuguese F1-score with data size*

4.2. 43 Language experiment result

After the experiment with the Romance languages, we expand our experimentations to 43 languages and varied the number of layers to capture the improvement from deeper models.

4.2.1. In-domain News Test Set

Table 3 presents the in-domain results on 43 languages news test set. Precision, recall and *F1*-score on all the punctuation increases with a deeper model. When the model goes from 1 layer to 2 layers, the overall *F1*-score improves from 76.1% to 78.1%, and increases from 78.1% to 80.2% when model goes from 2 layers to 4 layers. In Total we obtain 4.1% absolute gain on overall *F1*-score from 1 layer to 4 layers.

We also can find that all the 3 models generate high precision and low recall. From the perspective of user experience, it is more friendly, as incorrect sentence breaking is more harmful for reading. We observe that deeper model has more improvement on recall than precision, from 71.8% to 77.2% on recall, while precision improved from 81.0% to 83.4%.

4.2.2. Out-of-domain Speech Test Set

Table 4 shows the results on out-of-domain SR text test set, where we compare the BPE-based multilingual models with the word-based monolingual baseline model of the 10 languages. From the table we can see that all the 3 multilingual models perform better than the monolingual word-based model. 1 layer model has the same depth with the monolingual models, and it gets 1.2% gain, 4 layer model improves 2.3% *F1*-score compared with the baseline. The results show that language text sharing benefits punctuation prediction of a single language. We believe although the training and test data source differ, the added languages make the multilingual model more robust to content style variation.

When examining individual punctuation symbols, we observe that monolingual model performs better than multilingual model on end-of-text punctuation, while multilingual model performs a little better on middle-text punctuation.

Another finding is that unlike the in-domain news test, improvement is not obvious on the out-of-domain test set as stacking more layers does not get much more improvement from 1 layer to 2 layers. 4 layer model has 1.1% gain over 1 layer model.

[4]https://github.com/pytorch/pytorch

Table 3: *Punctuation results on 43 language news test set*

Punctuation	**1-LSTM**			**2-LSTM**			**4-LSTM**		
	Rec.	*Pre.*	F_1	*Rec.*	*Pre.*	F_1	*Rec.*	*Pre.*	F_1
comma	67.0%	77.9%	72.1%	70.3%	78.7%	74.3%	72.7%	80.2%	76.3%
mid-period	70.3%	77.8%	73.9%	73.8%	79.9%	76.8%	77.1%	81.9%	79.5%
mid-question	53.9%	70.0%	60.9%	59.4%	72.5%	65.3%	64.1%	75.8%	69.5%
end-period	94.4%	97.1%	95.7%	93.4%	97.4%	95.4%	95.7%	97.8%	96.7%
end-question	75.7%	90.1%	82.3%	78.2%	90.9%	84.1%	81.0%	91.4%	85.9%
Overall	71.8%	81.0%	**76.1%**	0.745	82.0%	**78.1%**	77.2%	83.4%	**80.2%**

Table 4: *Punctuation results on 10 language SR text test set*

Punctuation	word model			**1-LSTM**			**2-LSTM**			**4-LSTM**		
	Rec.	Pre.	F1	*Rec.*	*Pre.*	*F1*	*Rec.*	*Pre.*	*F1*	*Rec.*	*Pre.*	*F1*
comma	61.2%	64.9%	63.0%	65.9%	65.2%	65.5%	67.3%	64.9%	66.1%	68.6%	65.6%	67.1%
mid-period	38.1%	39.1%	38.6%	32.3%	52.5%	40.0%	34.4%	52.3%	41.5%	36.1%	52.9%	42.9%
mid-question	60.4%	48.0%	53.4%	59.3%	53.3%	56.2%	62.8%	52.8%	57.4%	62.8%	52.6%	57.2%
end-period	93.8%	93.9%	93.9%	92.9%	92.3%	92.6%	92.6%	92.5%	92.5%	92.9%	93.0%	93.0%
end-question	87.0%	81.9%	84.4%	81.9%	78.6%	80.2%	83.0%	78.1%	80.5%	84.5%	79.1%	81.7%
Overall	70.5%	71.8%	**71.2%**	71.5%	73.2%	**72.4%**	72.5%	72.9%	**72.7%**	73.6%	73.4%	**73.5%**

Compared with the in-domain results in Table 3, there is an expected accuracy gap of the same model between in-domain and out-of-domain test set. 1 layer model gap is 3.7%, 2 layer model is 5.4% and 4 layer is 6.7% between in-domain and out-of-domain. The gap is reasonable considering the difference between news article and spoken text style. News article is more formal, while spoken text is more casual. Another difference is that it's more natural to use punctuation while writing a news article, so the punctuation usage is more accurate, comparing in SR text the punctuation can be ambiguous and subjective.

4.2.3. *Error analysis*

To better understand the multilingual model performance, we analyze the out-of-domain confusion matrix. Figure 3 presents the result of 4 layers' model on the SR text test set. We see that actual comma has 0.25 that is predicted as null, and 0.43 middle period is predicted as comma. Considering from user perception, these 2 kinds of errors are less harmful for reading and understanding: 1) comma and period both play the role of sentence breaking, and 2) comma usually represents a pause, sometimes it will be ignored in spoken text.

The errors that do have a larger effect are null and period, as period usage is less ambiguous and can change the meaning of a sentence if incorrectly added or missed. If we add a period where no punctuation should exist or label a null when period should exist, it is more harmful for reading and understanding.

Another shortcoming is the prediction of question mark. In oral expression, interrogative sentences are usually accompanied by changes in tone, especially for the short phrases. It is hard to process by the model for lack of acoustic features.

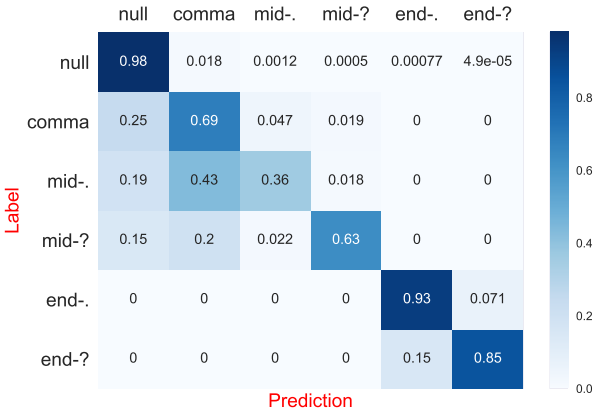

Figure 3: *Confusion matrix of 4 layers' model on SR test set*

5. Conclusion

This paper presents a 43 language multilingual method for punctuation prediction, which utilizes multilingual BPE to exploit the shared sub-words and uses a shifted LSTM model to model the sequential information across languages to predict punctuation. The multilingual model shows better performance compared with several monolingual models by sharing linguistic knowledge across languages. Moreover, fine-tuning the multilingual model also shows expected benefit for low-resource languages.

For future work, we will continue to expand to more locales. We intend on building out a pre-trained multilingual language model introducing self-unsupervised training, then fine-tune the model on the punctuation prediction task.

6. Acknowledgements

The work gets lots of help from other people. Thanks our colleagues Shawn Chang, Piyush Behre and William Gale for instruction and providing tools, also thanks our partner Xiang Li for the help on experiment.

7. References

[1] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4741–4744.

[2] D. Zhang, S. Wu, N. Yang, and M. Li, "Punctuation prediction with transition-based parsing," in *Proceedings of the 51st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 752–760.

[3] J.-H. Kim and P. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," 01 2001, pp. 2757–2760.

[4] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *ISCA tutorial and research workshop (ITRW) on prosody in speech recognition and understanding*, 2001.

[5] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 2, 1998, pp. 689–692 vol.2.

[6] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Seventh International Conference on Spoken Language Processing*, 2002.

[7] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *Fifth International Conference on Spoken Language Processing*, 1998.

[8] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 177–186.

[9] N. Ueffing, M. Bisani, and P. Vozila, "Improved models for automatic punctuation prediction for spoken and written text." in *Interspeech*, 2013, pp. 3097–3101.

[10] O. Tilk and T. Alumäe, "Lstm for punctuation restoration in speech transcripts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[11] K. Xu, L. Xie, and K. Yao, "Investigating lstm for punctuation prediction," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[12] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." in *Interspeech*, 2016, pp. 3047–3051.

[13] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5700–5704.

[14] H. Schwenk and M. Douze, "Learning joint multilingual sentence representations with neural machine translation," *arXiv preprint arXiv:1704.04154*, 2017.

[15] J. Yi and J. Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7270–7274.

[16] M. Ballesteros and L. Wanner, "A neural network architecture for multilingual punctuation generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1048–1053. [Online]. Available: https://www.aclweb.org/anthology/D16-1111

[17] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[18] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[19] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," *arXiv preprint arXiv:1601.01073*, 2016.

[20] S. Mille, M. Ballesteros, A. Burga, G. Casamayor, and L. Wanner, "Multilingual natural language generation within abstractive summarization." in *MMDA@ ECAI*, 2016, pp. 33–38.

[21] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," *Artificial Intelligence Review*, vol. 48, no. 4, pp. 499–527, 2017.

[22] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[23] P. Gage, "A new algorithm for data compression," *C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[24] B. Heinzerling and M. Strube, "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 7-12, 2018 2018.