gaBERT — an Irish Language Model

Anonymous ACL submission

Abstract

The BERT family of neural language models have become highly popular due to their ability to provide sequences of text with rich context-sensitive token encodings which are able to generalise well to many NLP tasks. We introduce, gaBERT, a monolingual BERT model for the Irish language. We compare our gaBERT model to multilingual BERT and monolingual WikiBERT, and we show that gaBERT provides better representations for a downstream parsing task. We also show how different filtering criteria, vocabulary size and the choice of subword tokenisation model affect downstream performance. We release gaBERT and related code to the community.

1 Introduction

The technique of fine-tuning a self-supervised language model has become ubiquitous in Natural Language Processing (NLP) because models trained in this way have advanced evaluation scores on many tasks (Radford et al., 2018; Peters et al., 2018; Devlin et al., 2019). Arguably the most popular architecture is BERT (Devlin et al., 2019) which uses stacks of transformers to predict the identity of a masked token and to predict whether two sequences are contiguous. It has spawned many variants (Liu et al., 2019; Lan et al., 2019) and much analysis (Jawahar et al., 2019; Chi et al., 2020; Rogers et al., 2020). In this paper, we introduce gaBERT, a monolingual model of Irish.

Although Irish is the first official language of the Republic of Ireland, a minority, 1.5% of the population (CSO, 2016), use it in their everyday lives outside of the education system. As the less dominant language in a bilingual community, the availability of Irish language technology is important since it makes it easier for Irish speakers and learners to use the language in their daily lives. Building on recent progress in data-driven Irish NLP (Lynn et al., 2012, 2015; Walsh et al., 2019), we release gaBERT with the hope that it will contribute to preserving Irish as a living language in the digital age.

While there is evidence to suggest that dedicated monolingual models can be superior to a multilingual model for within-language downstream tasks (de Vries et al., 2019; Virtanen et al., 2019; Farahani et al., 2020), other studies suggest that a multilingual model such as mBERT is a good choice for low-resourced languages (Wu and Dredze, 2020; Rust et al., 2020; Chau et al., 2020). We compare gaBERT to mBERT and to the monolingual Irish WikiBERT, both using Wikipedia as source of training data. We base our comparison on the downstream task of universal dependency (UD) parsing, since we have labelled Irish data in the form of the Irish UD Treebank (Lynn and Foster, 2016; McGuinness et al., 2020). We find that parsing accuracy improves when using gaBERT – by 3.7 and 3.6 LAS points over mBERT and WikiBERT respectively. Continued pretraining of mBERT using the gaBERT training data results in a recovery of 2 LAS points over the off-the-shelf version. The benefit of the gaBERT training data is also shown in a manual analysis which compares the models on their ability to predict a masked token.

We detail our hyperparameter search for our final model, where we consider the type of text filtering to apply, the vocabulary size and tokenisation model. We release our experiment code through GitHub[1] and our models through the Hugging Face (Wolf et al., 2020) model repository.[2]

2 Data

We use the following to train gaBERT:

CoNLL17: The Irish data from the CoNLL'17 raw text collection (Ginter et al., 2017) released as part of the 2017 CoNLL Shared Task on UD Parsing (Zeman et al., 2017).

IMT: A collection of Irish texts used in Irish machine translation research (Dowling et al., 2018, 2020), including legal text, general administration and data crawled from public body websites.

NCI: The New Corpus for Ireland (Kilgarriff et al., 2006), which contains a wide range of texts in Irish, including fiction, news reports, informative texts and official documents.

OSCAR: The unshuffled Irish portion of the 2019 OSCAR corpus (Ortiz Suárez et al., 2019), a subset of CommonCrawl.

Paracrawl: The Irish side of the `ga-en` bitext pair of ParaCrawl v7 (Bañón et al., 2020), which is a collection of parallel corpora crawled from multi-lingual websites.

Wikipedia: Text from Irish Wikipedia, an online encyclopedia.[3]

The sentence counts in each corpus are listed in Table 1 after tokenisation and segmentation but before filtering described below. See Appendix A for more information on the content of these corpora, including license information. We apply corpus-specific pre-processing, sentence-segmentation and tokenisation described in Appendix B.

Corpus	Num. Sents	Size (MB)
CoNLL17	1.7M	138
IMT	1.4M	124
NCI	1.6M	174
OSCAR	0.8M	89
ParaCrawl	3.1M	380
Wikipedia	0.7M	38
Overall	9.3M	943

Table 1: Sentence counts and plain text file size in megabytes for each corpus after tokenisation and segmentation but before applying sentence filtering.

3 Experimental Setup

After initial corpus pre-processing, all corpora are merged and we use the WikiBERT pipeline (Pyysalo et al., 2020) to create pretraining data. We experiment with four corpus filtering settings, five vocabulary sizes and three tokenisation models.

3.1 Corpus Filtering

The WikiBERT pipeline contains a number of filters which dictate whether a document should be kept. As we are working with data sources where there may not be clear document boundaries, or where there are no line breaks over a large number of sentences, document-level filtering may be inadequate for such texts. Consequently, we also experiment with using OpusFilter (Aulamo et al., 2020), which filters individual sentences, thereby giving us the flexibility of filtering noisy sentences while not discarding full documents.

For each filter setting below, we train a BERT model on the data which remains after filtering:

- **No-filter**: All collected texts are included in the pre-training data.
- **Document-filter**: The default document-level filtering used in the WikiBERT pipeline.
- **OpusFilter-basic**: We use OpusFilter with basic filtering described in Appendix B.4.
- **OpusFilter-basic-char-lang** We use OpusFilter with basic filtering as well as character-script and language filters described in Appendix B.4.

3.2 Vocabulary Creation

To create a model vocabulary, we experiment with the SentencePiece (Kudo and Richardson, 2018) and WordPiece tokenisers. Using the model with highest median LAS from the filtering experiments, we try vocabulary sizes of 15K, 20K, 30K, 40K and 50K. We then train a WordPiece tokeniser, keeping the vocabulary size that works best for the SentencePiece tokeniser. We also train a BERT model using the union of the two vocabularies.

3.3 BERT Pretraining Parameters

We use the original BERT implementation of Devlin et al. (2019). For the development experiments, we train our BERT model for 500K steps with a sequence length of 128. We use whole word masking

[1] GitHub URL will appear here in the published paper.
[2] Hugging Face URL will appear here in the final paper.
[3] We use the articles from `https://dumps.wikimedia.org/gawiki/20210520/`

and the default hyperparameters and model architecture of $\mathrm{BERT}_{\mathrm{BASE}}$ (Devlin et al., 2019).[4]

For the final gaBERT model, we train for 900k steps with sequence length 128 and a further 100k steps with sequence length 512. We train on a TPU-v2-8 with 128GB of memory on Google Compute Engine[5] and use a batch size of 128.

4 Evaluation Measures

Dependency Parsing The evaluation measure we use to make development decisions is dependency parsing labelled attachment score (LAS). To obtain this measure, we fine-tune a given BERT model in the task of dependency parsing and measure LAS on the development set of the Irish-IDT treebank in version 2.8 of UD. We report the median of five fine-tuning runs with different random initialisation. For the dependency parser, we use a multitask model which uses a graph-based parser with biaffine attention (Dozat and Manning, 2016) as well as additional classifiers for predicting POS tags and morphological features. We use the AllenNLP (Gardner et al., 2018) library to develop our multitask model.

Cloze Test To compile a cloze task test set, 100 strings of Irish text (4–77 words each) containing the pronouns ‘é’ (‘him/it’), ‘í’ (‘her/it’) or ‘iad’ (‘them’) are selected from Irish corpora and online publications. One of these pronouns is masked in each string for the cloze test.[6]

Following Rönnqvist et al. (2019), the models are evaluated on their ability to generate the original masked token, and a manual evaluation of the models is performed wherein predictions are classified into the following exclusive categories:

- **Match**: The predicted token fits the context grammatically and semantically. This may occur when the model predicts the original token or another token which also fits the context.
- **Mismatch** The predicted token is a valid Irish word but is unsuitable given the context.
- **Copy** The predicted token is an implausible repetition of another token in the context.
- **Gibberish** The predicted token is not a valid Irish word.

[4]We use a lower batch size of 32 in order to train on NVIDIA RTX 6000 GPUs with 24 GB RAM.

[5]TPU access was kindly provided to us through the Google Research TPU Research Cloud.

[6]All the masked tokens exist in the vocabularies of the candidate BERT models and are therefore possible predictions.

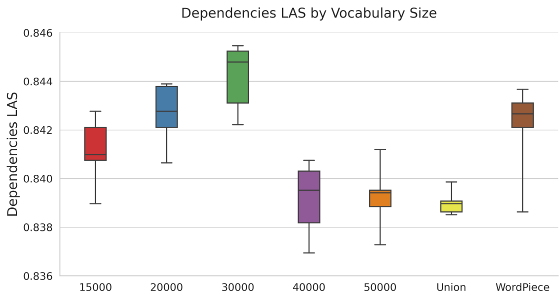

Figure 1: Dependency parsing LAS for each filter type and vocabulary type (five runs each).

Filter	Sentences
No-filter	9.3M
Document-filter	7.9M
OpusFilter-basic	9.0M
OpusFilter-basic-char-lang	7.7M

Table 2: The number of sentences which remain after applying the specific filter.

5 Results

5.1 Development Results

Filter Settings The overall number of sentences which remain after applying each filter are shown in Table 2. The results of training a dependency parser with the gaBERT model produced by each setting are shown in the top half of Fig. 2. `Document-Filter` has the highest LAS score. As the BERT model requires contiguous text for its next-sentence-prediction task, filtering out full documents may be more appropriate than filtering individual sentences. The two `OpusFilter` configurations perform marginally worse than the `Document-Filter`. In the case of `OpusFilter-basic-char-lang`, perhaps the lower number of sentences overall translates to lower LAS scores. Finally, `No-Filter` performs in the same range as the two `OpusFilter` configurations but has the lowest median score, suggesting that some level of filtering is beneficial.

Vocabulary Settings The results of the five runs testing different vocabulary sizes are shown in the bottom half of Fig. 1. A vocabulary size of 30K performs best for the SentencePiece tokeniser, which outperforms the WordPiece tokeniser with the same vocabulary size. The union of the two vocabularies results in 32,314 entries, and does not perform as well as the two vocabularies on their own.

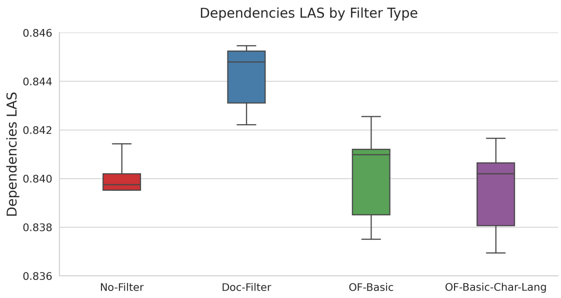

Figure 2: Dependency parsing LAS for each filter type.

		LAS	
Model	UD	Dev	Test
mBERT	2.8	81.8	80.3
WikiBERT	2.8	81.9	80.4
mBERT-cp	2.8	84.3	82.3
gaBERT	2.8	**85.6**	**84.0**
Chau et al. (2020)	2.5	-	76.2
gaBERT	2.5	-	77.5

Table 3: LAS in dependency parsing (UD v2.8) for selected models. Median of five fine-tuning runs. Scores are calculated using the official UD evaluation script (*conll18_ud_eval.py*).

5.2 Model Comparison

We compare our final gaBERT model with off-the-shelf mBERT and WikiBERT-ga, as well as an mBERT model obtained with continued pre-training on our corpora.

Dependency Parsing Table 3 shows the results for dependency parsing.[7] Using mBERT off-the-shelf results in a test set LAS of 80.3. The WikiBERT-ga model performs slightly better than mBERT. By training mBERT for more steps on our corpora, LAS can be improved by 2 points. Our gaBERT model has the highest LAS of 84. The last two rows compare gaBERT, on v2.5 of the treebank, with the system of Chau et al. (2020) who augment the mBERT vocabulary with the 99 most frequent Irish tokens and fine-tune on Irish Wikipedia. Our model outperforms this approach.

Cloze Test Table 4 shows the accuracy of each model with regard to predicting the original masked token. mBERT-cp is the most accurate and gaBERT is close behind. Table 5 shows the manual evaluation of the tokens generated by each model, accounting for plausible answers deviating from the original token and separately reporting copying of content and production of gibberish. These results echo those of the original masked token prediction evaluation in so far as they rank the models in the same order. Further detail, examples and analysis of the cloze test can be found in Appendix C.

[7] A competitive parser, UDPipe, trained on the Irish-IDT Treebank 2.8 without external embeddings achieves 72.59 LAS on the test set.

Model	Original Token Prediction
mBERT	16
WikiBERT	53
mBERT-cp	**78**
gaBERT	75

Table 4: The number of times the original masked token was predicted (100 test items).

Model	Match	Mism.	Copy	Gib
mBERT	41	42	4	13
WikiBERT	62	31	1	6
mBERT-cp	**85**	12	1	2
gaBERT	83	14	2	1

Table 5: The number of matches, mismatches, copies and gibberish predicted by each model (100 test items).

6 Friends of gaBERT

In subsequent experiments, we look at variants of BERT, including RoBERTa (Liu et al., 2019). The multilingual XLM-$\mathrm{R_{BASE}}$ (Conneau et al., 2020) clearly outperforms both variants of mBERT but underperforms gaBERT. We tried training a RoBERTa$_{\mathrm{BASE}}$ model but could only obtain LAS scores comparable to off-the-shelf mBERT and leave finding suitable hyperparameters to future work. We train an ELECTRA model (Clark et al., 2020), which performs better than both mBERT models and the WikiBERT model but slightly below gaBERT. See Appendix D-F for details.

7 Conclusions

We release gaBERT, a BERT model trained on over 7.9M Irish sentences, combining Irish language text from a variety of sources, and evaluate it in dependency parsing and in a pronoun cloze test task, showing improvements over three baselines, multilingual BERT, WikiBERT-ga and XML-$\mathrm{R_{BASE}}$.

8 Ethical Considerations

No dataset is released with this paper, however most of the corpora are publicly available as described in Appendix A. Furthermore, where an anonymised version of a dataset was available it was used. We release the gaBERT language model based on the $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) autoencoder architecture. We note that an autoregressive architecture may be susceptible to training data extraction, and that larger language models may be more susceptible (Carlini et al., 2021). However, gaBERT is an autoencoder architecture and a smaller language model which may help mitigate this potential vulnerability.

Possible harms of language model pre-trained on web-crawled text have been widely discussed (Bender et al., 2021). Since gaBERT uses CommonCrawl data, there is a risk that the gaBERT model may, for example, produce unsuitable text outputs when used to generate text. To mitigate this possibility we include the following caveat with the released code and model cards:

> We note that some data used to pre-train gaBERT was scraped from the web which potentially contains ethically problematic content (bias, hate, adult content, etc.). Consequently, downstream tasks/applications using gaBERT should be thoroughly tested with respect to ethical considerations.

We do not discuss in detail how gaBERT can be used in actual use cases as we expect the use of BERT-style models to be essential knowledge for NLP practitioners up-to-date with current research. There are many downstream tasks which can use gaBERT, including machine translation, educational applications, predictive text, search and games. The authors hope gaBERT will contribute to the ongoing effort to preserve the Irish language as a living language in the technological age. Supporting a low-resourced language like Irish in a bilingual community will make it easier for Irish speakers, and those who wish to be Irish speakers, to use the language in practice.

Each use case or downstream application may rank the available pre-trained language models differently in terms of suitability. We urge NLP practitioners to compare available models such as those tested in this paper in their application rather than relying on results for a different task.

References

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of The Eighth International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

CSO. 2016. Census of Population 2016 – Profile 10 Education, Skills and the Irish Language. Publisher: Central Statistics Office.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. ArXiv 1912.09582v1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Meghan Dowling, Sheila Castilho, Joss Moorkens, Teresa Lynn, and Andy Way. 2020. A human evaluation of English-Irish statistical and neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 431–440, Lisboa, Portugal. European Association for Machine Translation.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA. Association for Machine Translation in the Americas.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. ParsBERT: Transformer-based model for Persian language understanding. ArXiv 2005.12515v1.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation*, 40:127–152.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. ArXiv 1909.11942v6.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. ArXiv 1907.11692v1.

Teresa Lynn, Ozlem Cetinoglu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith. 2012. Irish treebanking and parsing: A preliminary evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1939–1946, Istanbul, Turkey. European Language Resources Association (ELRA).

Teresa Lynn and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Proceedings of the Second Celtic Language Technology Workshop*, July, pages 79–92, Paris.

Teresa Lynn, Kevin Scannell, and Eimear Maguire. 2015. Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 1–8, Beijing, China. Association for Computational Linguistics.

Sarah McGuinness, Jason Phelan, Abigail Walsh, and Teresa Lynn. 2020. Annotating MWEs in the Irish UD treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 126–139, Barcelona, Spain (Online). Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2020. WikiBERT models: deep transfer learning for many languages. ArXiv 2006.01538v1.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI Preprint.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. ArXiv 2012.15613v2.

Milan Straka and Jana Straková. 2017. 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*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. ArXiv 1912.07076v1.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 162–168, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phùòng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lùòng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Data Licenses

This Appendix provides specific details of the licence for each of the datasets used in the experiments.

A.1 CoNLL17

The Irish annotated CoNLL17 corpus can be found here: http://hdl.handle.net/11234/1-1989 (Ginter et al., 2017).

The automatically generated annotations on the raw text data are available under the CC BY-SA-NC 4.0 licence. Wikipedia texts are available under the CC BY-SA 3.0 licence. Texts from Common Crawl are subject to Common Crawl Terms of Use, the full details of which can be found here: https://commoncrawl.org/terms-of-use/full/.

A.2 IMT

The Irish Machine Translation datasets contains text from the following sources:

- Text crawled from the Citizen's Information website, contains Irish Public Sector Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence: https://www.citizensinformation.ie/ga/.
- Text crawled from Comhairle na Gaelscolaíochta website: https://www.comhairle.org/gaeilge/.
- Text crawled from the FÁS website (http://www.fas.ie/), accessed in 2017. The website has since been dissolved.
- Text crawled from the Galway County Council website: http://www.galway.ie/ga/.
- Text crawled from https://www.gov.ie/ga/, the central portal for government services and information.
- Text crawled from articles on the Irish Times website.
- Text crawled from the Kerry County Council website: https://ciarrai.ie/.
- Text crawled from the Oideas Gael website: http://www.oideas-gael.com/ga/.
- Text crawled from articles generated by Teagasc, available under PSI licence.
- Text generated by Conradh na Gaeilge, shared with us for research purposes.
- The Irish text from a parallel English–Irish corpus of legal texts from the Department of Justice. This dataset is available for reuse on the ELRC-SHARE repository under a PSI license: https://elrc-share.eu
- Text from the Directorate-General for Translation (DGT), available for download from the European Commission website. Reuse of the texts are subject to Terms of Use, found on the website: https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory.
- Text reports and notices generated by Dublin City Council, shared with us for research purposes.
- Text uploaded to ELRC-share via the National Relay Station, shared with us for research purposes.
- Text reports and reference files generated by the Language Commissioner, available on ELRC-share under PSI license: https://elrc-share.eu/.
- Text generated by the magazine Nós, shared with us for research purposes.
- Irish texts available for download on OPUS, under various licenses: https://opus.nlpl.eu/
- Text generated from in-house translation provided by the then titled Department of Culture, Heritage and Gaeltacht (DCHG), provided for research purposes. The anonymised dataset is available on ELRC-share, under a CC-BY 4.0 license: https://elrc-share.eu/.
- Text reports created by Údarás na Gaeilge, uploaded to ELRC-share available under PSI license: https://elrc-share.eu/.
- Text generated by the University Times, shared with us for research purposes.

A.3 NCI

The corpus is compiled and owned by Foras na Gaeilge and is provided to us for research purposes.

A.4 OSCAR

The unshuffled version of the Irish part of the OSCAR corpus was provided to us by the authors for research purposes.

A.5 ParaCrawl

Text from ParaCrawl v7, available here: `https://www.paracrawl.eu/v7`. The texts themselves are not owned by ParaCrawl, the actual packaging of these parallel data are under the Creative Commons CC0 licence ("no rights reserved").

A.6 Wikipedia

The texts used are available under a CC BY-SA 3.0 licence and/or a GNU Free Documentation License.

B Corpus Pre-processing

This appendix provides specific details on corpus pre-processing, and the OpusFilter filters used.

CoNLL17 The CoNLL17 corpus is already tokenised, as it is provided in CoNLL-U format, which we convert to one-sentence-per-line tokenised plain text.

IMT, OSCAR and ParaCrawl The text files from the IMT, OSCAR and ParaCrawl contain raw sentences requiring tokenisation. We describe the tokenisation process for these corpora in Appendix B.1.

Wikipedia For the Wikipedia articles, the Irish Wikipedia dump is downloaded and the WikiExtractor tool[8] is then used to extract plain text. Article headers are included in the extracted text files. Once the articles have been converted to plain text, they are tokenised using the tokeniser described in Appendix B.1.

NCI As many of the NCI segments marked up with ⟨s⟩ tags contain multiple sentences, we further split these segments with heuristics described in Appendix B.3.

B.1 Tokenisation and Segmentation

Raw texts from the IMT, OSCAR, ParaCrawl and Wikipedia corpora are tokenised and segmented with UDPipe (Straka and Straková, 2017) trained on a combination of the Irish-IDT and English-EWT corpora from version 2.7 of the Universal Dependencies (UD) treebanks (Zeman et al., 2020). We include the English-EWT treebank in the training data to expose the tokeniser to more incidences of punctuation symbols which are prevalent in our pre-training data. This also comes with the benefit of supporting the tokenisation of code-mixed data. We upsample the Irish-IDT treebank by ten times to offset the larger English-EWT treebank size. This tokeniser is applied to all corpora apart from the NCI, which is already tokenised by Kilgarriff et al. (2006), and the CoNLL17 corpus as this corpus is already tokenised in CoNLL-U format.

B.2 NCI

Foras na Gaeilge provided us with a `.vert` file[9] containing 33,088,532 tokens in 3,485 documents. We extract the raw text from the first tab-separated column and carry out the following conversions (number of events):

- Replace `"` with a neutral double quote (4408).
- Replace the standard xml/html entities quot, lt, gt and amp tokenised into three tokens, e.g. `&↵quot↵;`, with the appropriate characters (128).
- Replace the numeric html entities 38, 60, 147, 148, 205, 218, 225, 233, 237, 243 and 250, again spanning three tokens, e. g. `&↵#38↵;`, with the appropriate Unicode characters (3679).
- Repeat from the start until the text does not change.

We do not modify the seven occurrences of `\x\x13` as it is not clear from their contexts how they should be replaced. After pre-processing and treating all whitespace as token separators, e.g. in the NCI token “go leor”, we obtain 33,472,496 tokens from the NCI.

B.3 Sentence Boundary Detection

Many of the NCI segments marked up with ⟨s⟩ tags contain multiple sentences. We treat each segment boundary as a sentence boundary and further split segments into sentences recursively, finding the best split point according to the following heuristics, splitting the segment into two halves and applying the same procedure to each half until no suitable split point is found.

[8] `https://github.com/attardi/wikiextractor`

[9] MD5 7be5c0e9bc473fb83af13541b1cd8d20

- Reject if the left half contains no letters and is short. This covers cases where the left half is only a decimal number.
- Reject if the right half has no letters and is short or is an ellipsis.
- Reject if the right half's first letter, skipping enumerations, is lowercase.
- Reject if the left half only contains a Roman number (in addition to the full-stop).
- Reject if inside round, square, curly or angle brackets and the brackets not far away from the candidate split point.
- If sentence-ending punctuation is followed by two quote tokens we also consider splitting between the quotes and prefer this split point if not rejected by above rules.
- If sentence-ending punctuation is followed by a closing bracket we also consider splitting after the closing bracket and prefer this split point if not rejected by above rules.
- If a question mark is followed by more question marks we also consider splitting after the end of the sequence of question marks and prefer this split point if not rejected by above rules.
- If a exclamation mark is followed by more exclamation marks we also consider splitting after the end of the sequence of exclamations marks and prefer this split point if not rejected by above rules.
- If a full-stop is the first full-stop in the overall segment, the preceding token is "1", there are more tokens before this "1" and the token directly before "1" is not a comma or semicolon we assume that this is an enumeration following a heading and prefer splitting before the "1".
- We do not insert new sentence boundaries at a full-stop after "DR", "Prof" and "nDr", and, if followed by a decimal number, after "No", "Vol" and "Iml".
- Splitting after a full-stop following decimal numbers in all other cases is dispreferred, giving the largest penalty to small numbers as these are most likely to be part of enumerations. An exception is "Airteagal" followed by a token ending with a full-stop, a number, a full-stop, another number and another full-stop. Here, we implemented a preference for splitting after the first separated full-stop, assuming the last number is part of an enumeration.
- Prefer a split point balancing the lengths of the halves in characters.

B.4 OpusFilter Filters

For **OpusFilter-basic**, we include the following filters:

- `LengthFilter`: Filter sentences containing more than 512 words.
- `LongWordFilter`: Filter sentences containing words longer than 40 characters.
- `HTMLTagFilter`: Filter sentences containing HTML tags.
- `PunctuationFilter`: Filter sentences which are over 60% punctuation.
- `DigitsFilter`: Filter sentences which are over 60% numeric symbols.

For **OpusFilter-basic-char-lang**, we use the same filters as in **OpusFilter-basic** but include the following character script and language ID filters:

- `CharacterScoreFilter`: Filter sentences which are below a ratio r of Latin characters, where $r \in \{1.0\}$.
- `LanguageIDFilter`: Filter sentences where the language ID tools have a lower confidence score than c, where $c \in \{0.8\}$.

C Cloze Test Examples

C.1 Prediction Classification

Table 6 provides one example per classification category of masked token predictions generated by the language models during our cloze test evaluation.

In the *matc*h example in Table 6 , the original meaning ('What are those radical roots?') differs to the meaning of the resulting string ('What about those radical roots?') in which the masked token is replaced by the predicted by mBERT-cp. However, the latter construction is grammatically and semantically acceptable.

Context Cue	Masked Word	Model	Prediction	Classification
Céard [MASK] na préamhacha raidiciúla sin? ('What [MASK] those radical roots?')	*iad* ('them')	mBERT-cp	*faoi* ('about')	match
Agus seo [MASK] an fhadhb mhór leis an bhfógra seo. ('And this [MASK] the big problem with this advert.')	*í* ('it')	WikiBERT	*thaitin* ('liked')	mismatch
Cheannaigh Seán leabhar agus léigh sé [MASK]. ('Seán bought a book and he read [MASK].')	*é* ('it')	gaBERT	*leabhar* ('a book')	copy
Ní h[MASK] sin aidhm an chláir. ('[MASK] is not the aim of the programme.')	*##é* ('it')	mBERT	- (minus sign)	gibberish

Table 6: Examples of cloze test predictions and classifications.

Model	Short	Medium	Long
mBERT	20.69%	55.56%	41.67%
wikibert	51.72%	58.33%	74.29%
mBERT-cp	75.86%	**83.33%**	**94.29%**
gaBERT	**79.31%**	**83.33%**	85.71%
gaELECTRA	**79.31%**	77.78%	88.57%

Table 7: Accuracy of language models segmented by length of context cue where short: 4–10 tokens, medium: 11–20 tokens, and long: 21–77 tokens.

In the *mismatch* example in Table 6, the predicted token is a valid Irish word, however the resulting generated text is nonsensical.

Though technically grammatical, the predicted token in the *copy* example in Table 6 results in a string with an unnatural repetition of a noun phrase where a pronoun would be highly preferable ('Seán bought a book and he read a book.').

In the *gibberish* example in Table 6, the predicted token does not form a valid Irish word and the resulting sentence is ungrammatical and meaningless.

C.2 Effect of Length of Context on Accuracy of Prediction

In order to observe the effect that the amount of context provided has on the accuracy of the model, Table 7 shows the proportion of matches achieved by each language model when the results are segmented by the length of the context cues.

All the models tested are least accurate when tested on the group of short context cues. All except mBERT achieved the highest accuracy on the group of long sentences.

C.3 Easy and Difficult Context Cues

A context cue may be considered easy or difficult based on:

- Whether the tokens occur frequently in the training data
- The number of context clues
- The distance of the context clues from the masked token

Two Irish language context cues which vary in terms of difficulty are exemplified below.

***Bean**, agus **í** cromtha thar thralaí bia agus [MASK] ag ithe a **sáithe**.*
'A woman, bent over a food trolley while eating her fill.'

We can consider the above sentence to be easy for the task of token prediction due to the following context clues:

- 'Bean' is a frequent feminine singular noun.
- 'í' is a repetition of the feminine singular pronoun to be predicted.
- The lack of lenition on 'sáithe' further indicates that the noun it refers to may not be masculine.

These clues indicate that the missing pronoun will be feminine and singular.

*Seo **béile** aoibhinn fuirist nach dtógann ach timpeall leathuair a chloig chun [MASK] a ullmhú.*
'This is an easy, delicious meal that only takes about half an hour to prepare.'

None of the language models tested predicted a plausible token for the above sentence. This example is more challenging as the only context clue is the feminine singular noun ‘béile’ which is 11 tokens in distance from the masked token.

D gaELECTRA Model

In addition to the gaBERT model of the main paper, we release gaELECTRA, an ELECTRA model (Clark et al., 2020) trained on the same data as gaBERT. ELECTRA replaces the MLM pre-training objective of BERT with a binary classification task discriminating between authentic tokens and alternative tokens generated by a smaller model for higher training efficiency. We use the default settings of the “Base” configuration of the official implementation[10] and train on a TPU-v3-8. As with BERT, we train for 1M steps and evaluate every 100k steps. However, we train on more data per step as the batch size is increased from 128 to 256 and a sequence length of 512 is used throughout.

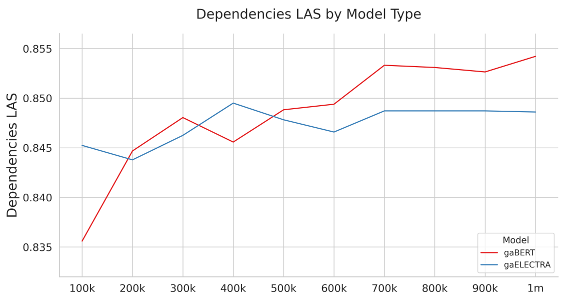

Figure 3: Dependency parsing LAS for each model type. Every 100k steps, we show the median of five LAS scores obtained from fine-tuning the respective model five times with different initialisation.

Figure 3 shows the development LAS of gaELECTRA and gaBERT for each checkpoint. The best gaBERT checkpoint is reached at step 1 million, which may indicate that there are still gains to be made from training for more steps. The highest median LAS for gaELECTRA is reached at step 400k. It is worth noting that although the two models are compared at the same number of steps, the different pretraining hyperparameters mean they are not trained on the same number of tokens per step.

We also compare the results of the gaELECTRA model to the other models in Tables 8 and 9. gaELECTRA performs slightly below gaBERT but better than both mBERT models and the WikiBERT model.

In terms of the Cloze test experiments: First, for the original masked token prediction (Table 4), gaELECTRA predicted the correct token 75 times, which is the same number as gaBERT and is slightly below mBERT with continued pretraining, which has a score of 78. Second, for the manual evaluation of the tokens generated by each model (Table 5), gaELECTRA predicted 82 matches, 8 mismatches, 1 copy, and 9 gibberish tokens; compared to 83, 14, 2 and 1 predicted by gaBERT, respectively.

E XLM-R Baseline

We add another off-the-shelf baseline by fine-tuning XLM-R$_{\text{BASE}}$, which is a multilingual RoBERTa model introduced by Conneau et al. (2020), in the task of multitask dependency parsing and POS and morphological features tagging. This model performs better than both variants of mBERT as well as the WikiBERT model but underperforms our two monolingual models, gaBERT and gaELECTRA.

F Full Model Results

This section examines the results produced by each of our models in more detail and also presents the scores of the additional models we examine, namely XLM-R$_{\text{BASE}}$ and gaELECTRA. [11] Tables 8 and 9 list the accuracies for predicting universal part of speech (UPOS), treebank-specific part of speech (XPOS) and morphological features, as well as the unlabelled and labelled attachment score (UAS and LAS, respectively) for all models discussed in this paper.

For the multilingual models, mBERT performs worse than XLM-R$_{\text{BASE}}$, which is a strong multilingual baseline. The monolingual WikiBERT model performs slightly better than mBERT in terms of LAS but is worse than XLM-R$_{\text{BASE}}$. The continued pretraining of mBERT on our data enables us to close the gap between mBERT and XLM-R$_{\text{BASE}}$. gaBERT is still the strongest model for all metrics in terms of test set scores. gaELECTRA performs slightly below that of gaBERT but better than XLM-R$_{\text{BASE}}$. It should be noted that each row selects the model based on median LAS, therefore, all other metrics are those that this selected model achieved.

[10] https://github.com/google-research/electra

[11] We tried training a RoBERTa$_{\text{BASE}}$ model on our data but could not obtain satisfactory LAS scores (a fine-tuned model achieved a dev LAS of 81.8, which is comparable to mBERT) and leave finding suitable hyperparameters for this architecture to future work.

Model	**UD**	**UPOS**	**XPOS**	**FEATS**	**UAS**	**LAS**
mbert-os	2.8	95.7	94.7	89.2	86.9	81.8
xlmr-base-os	2.8	96.4	95.1	90.6	88.3	84.0
wikibert-os	2.8	95.9	94.9	89.4	86.8	81.9
mbert-cp	2.8	97.2	95.8	92.3	88.1	84.3
gabert	2.8	97.1	**96.2**	**93.1**	**89.2**	**85.6**
gaelectra	2.8	**97.3**	96.1	92.8	89.1	85.3

Table 8: Full model results on development data. For model name abbreviations, see test result table.

Model	**UD**	**UPOS**	**XPOS**	**FEATS**	**UAS**	**LAS**
mbert-os	2.8	95.4	94.3	88.6	86.2	80.3
xlmr-base-os	2.8	96.1	95.1	90.0	87.7	82.5
wikibert-os	2.8	95.7	94.4	88.3	85.9	80.4
mbert-cp	2.8	96.7	95.5	91.7	87.1	82.3
gabert	2.8	**97.0**	**95.7**	**91.8**	**88.4**	**84.0**
gaelectra	2.8	96.9	95.5	91.5	87.6	83.1

Table 9: Full model results on test data (os = fine-tuned off-the-shelf model, cp = continued pre-training before fine-tuning).