

gaBERT — an Irish Language Model

James Barry¹, Joachim Wagner², Lauren Cassidy¹

Alan Cowap³, Teresa Lynn¹, Abigail Walsh¹

Mícheál J. Ó Meachair⁴, Jennifer Foster²

^{1,2,3}School of Computing, Dublin City University, ¹ADAPT Centre

³SFI Centre for Research Training in Machine Learning at Dublin City University

⁴Fiontar & Scoil na Gaeilge

¹ {firstname.lastname}@adaptcentre.ie

² {firstname.lastname}@dcu.ie

³ alan.cowap2@mail.dcu.ie

⁴ micheal.omeachair@dcu.ie

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 the monolingual Irish 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 compare the results of fine-tuning a gaBERT model with an mBERT model for the task of identifying verbal multiword expressions, and show that the fine-tuned gaBERT model also performs better at this task. We release gaBERT and related code to the community.

Keywords: BERT, Irish

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 transformer blocks 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, only 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 facilitates Irish speakers and learners to continue to use the language in their increasingly digital daily lives. In terms of technological support however, Irish is a low-resourced language and significantly lacking in speech and language tools and resources (Lynn, 2022).

From a linguistic perspective, the Irish language is an inflected language, sharing linguistic features with other Celtic languages such as verb-subject-object (VSO) word order, initial mutation (lenition and eclipsis) and inflected prepositions. Inflection is common through

suffixation, marking tense, number and person, while nouns are inflected for number and case. Nouns are either masculine or feminine in grammatical gender, which in turn influences declension-dependent inflections. Its inflected nature has already been shown to impact data-driven NLP tools due to data sparsity (Lynn et al., 2013), as has the frequent use of clefting (fronting), two forms of the verb ‘to be’ and prevalence of variable and discontinuous multiword expressions.

Building upon recent progress in data-driven Irish NLP (Lynn et al., 2012; Lynn et al., 2015; Walsh et al., 2019; Cassidy et al., 2022), 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 the 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, as well as a Multiword Expression (MWE) identification task, where a token classification layer is trained to locate and classify verbal MWEs in text.

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 GitHub¹ and our models through the HuggingFace (Wolf et al., 2020) model repository.^{2,3}

2. Data

We use the following to train gaBERT:

- **CoNLL17**: The Irish data from the CoNLL17 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; Dowling et al., 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.⁴

The sentence and word 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.

¹<https://github.com/jbrry/Irish-BERT>

²<https://huggingface.co/DCU-NLP>

³We also release gaELECTRA described in Appendix D.

⁴We use the articles from <https://dumps.wikimedia.org/gawiki/20210520/>

Corpus	Num. Sents	Num. Tokens	Size (MB)
CoNLL17	1.7M	24.7M	138
IMT	1.4M	22.6M	124
NCI	1.6M	33.5M	174
OSCAR	0.8M	16.2M	89
ParaCrawl	3.1M	67.5M	380
Wikipedia	0.7M	6.8M	38
Overall	9.3M	171.3M	943

Table 1: Sentence and word 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 OpusFilter (Aulamo et al., 2020) with 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.

OpusFilter-basic-char-lang The same filters are used as **OpusFilter-basic** but with additional character script and language ID filters:

- **CharacterScoreFilter**: All alphabetic characters in a sentence must be in Latin script.

- `LanguageIDFilter`: The confidence scores from the language ID tools must be > 0.8 . We use two language identification tools: `langid.py` (Lui and Baldwin, 2012) and `CLD2`.⁵

3.2. Vocabulary Creation

To create a model vocabulary, we experiment with the SentencePiece (Kudo and Richardson, 2018) and WordPiece tokenisers.⁶ 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 and the default hyperparameters and model architecture of BERT_{BASE} (Devlin et al., 2019).⁷ Training for development runs of gaBERT took just under 48 hours on GPU. While a seed for data preparation can be set (we do not change the default 12345), the BERT implementation does not provide an option to set a seed for model initialisation and we did not find code that sets a seed for pretraining internally, suggesting initialisation is non-deterministic. 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 Engine⁸ and use a batch size of 128. Training gaBERT on TPU for 1M steps took around 37.5 hours.

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. Model hyperparameters are given

⁵<https://github.com/CLD2Owners/cld2>

⁶As BERT expects WordPiece tokenisation, a heuristic tool is used to map the SentencePiece vocabulary to WordPiece (<https://github.com/spyysalo/sent2wordpiece>).

⁷We use a lower batch size of 32 in order to train on NVIDIA RTX 6000 GPUs with 24 GB RAM.

⁸TPU access was kindly provided to us through the Google Research TPU Research Cloud.

in Appendix C.1. 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.⁹ 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 also 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. This might occur in the form of a subword or sequence of punctuation not forming a meaningful word.

MWE Identification task Multiword expressions (MWEs) pose a challenge in many tasks in NLP, including parsing. Treatment of MWEs can range between considering them as syntactically fixed words-with-spaces (ex: ‘out of’, ‘Every cloud has a silver lining’, ‘sooner or later’), to syntactically flexible constructions that display idiosyncratic behaviours (ex: ‘touch up’, ‘life hack’, ‘get something out of your system’). In addition to varying syntactic structures, MWEs also present issues of discontinuity, disambiguation, productivity, and can be more or less semantically opaque (Sag et al., 2002). The task of automatically identifying multiword expressions (MWEs) has been explored in the series of shared tasks organised by the PARSEME network (Savary et al., 2017), focusing on verbal MWEs, i.e. MWEs headed by a verb, as they pose a particular challenge in terms of automatic identification.

We design an experiment to compare the results of fine-tuning both a gaBERT model and an off-the-shelf mBERT model for the task of identifying MWEs in Irish text. We used the Irish portion of the PARSEME 1.2 shared task data (Walsh et al., 2020), which has been manually annotated with six types of verbal MWEs. The annotations were converted to a modified version of BIO tagging, based on the work of (Schneider et al., 2014), and a linear layer for token classification was added for the task of identifying the correct label for each word.

⁹All the masked tokens exist in the vocabularies of the candidate BERT models and are therefore possible predictions.

Filter	Sentences	Tokens
No-filter	9.2M	171.3M
Document-filter	7.9M	161.0M
OpusFilter-basic	9.0M	170.8M
OpusFilter-basic-char-lang	7.7M	161.2M

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

5. Results

5.1. Development Results

Filter Settings The overall number of sentences and words 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 Fig. 1. `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`, the additional character script and language ID filters did not lead to a noticeable change in LAS. 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 Fig. 2. 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. A manual inspection of the two vocabularies showed that the `WordPiece` tokeniser created more entries consisting of foreign characters and emojis at the expense of Irish words/word-pieces, which may account for the lower performance of settings using this tokeniser.

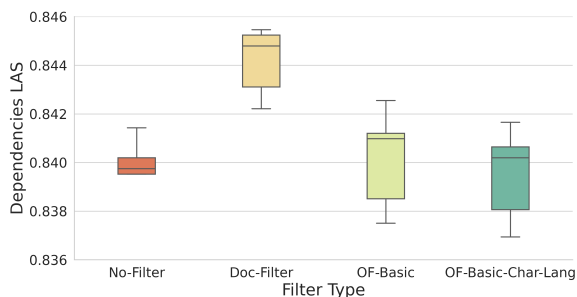


Figure 1: Dependency parsing LAS for each filter type.

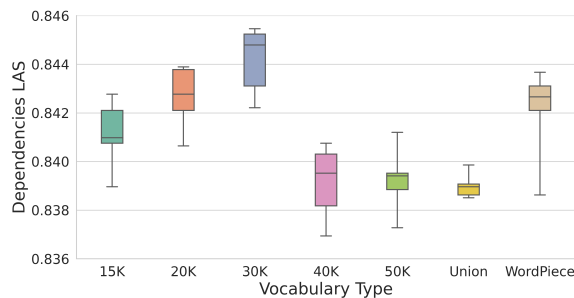


Figure 2: Dependency parsing LAS for each vocabulary type.

5.2. Model Comparison

We compare our final gaBERT model with off-the-shelf mBERT and the monolingual Irish WikiBERT-ga model, as well as an mBERT model obtained with continued pre-training on our corpora (mBERT-cp).¹⁰

Dependency Parsing Table 3 shows the results for dependency parsing. The first row (No BERT) is a baseline which does not use a pretrained BERT model but uses a BiLSTM encoder operating over token and character-level features instead. Using mBERT off-the-shelf results in a test set LAS of 80.3, an absolute improvement of 8.9 points over the baseline. 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. The results are lower for both settings due to the fewer amount of trees in v2.5 of the treebank¹¹ and a manual effort to clean up some inconsistent annotations (McGuinness et al., 2020). Our model outperforms this approach, likely due to our inclusion of a wider variety of corpora as well as our dedicated Irish vocabulary.

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

¹⁰Since training the gaBERT model, other BERT models supporting Irish we found are BERTreach (<https://huggingface.co/jimregan/BERTreach>) and LaBSE (<https://huggingface.co/setu4993/LaBSE>). BERTreach is a monolingual model trained on 47 million tokens. LaBSE is a multilingual model trained to encode the meaning of sentences and covers 109 languages including Irish.

¹¹v2.5 has only 858 trees compared to the 4,005 in v2.8.

Model	LAS		
	UD	Dev	Test
No BERT	2.8	73.4	71.4
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*).

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).

token prediction evaluation in so far as they rank the models in the same order.

Table 6 provides one example per classification category of masked token predictions generated by the language models during our cloze test evaluation. In the *match* 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 prediction of mBERT-cp. However, the latter construction is grammatically and semantically acceptable. 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

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).

Table 6, the predicted token does not form a valid Irish word and the resulting sentence is ungrammatical.

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.

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

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

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

- (1) *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 Example 1 to be easy for the task of token prediction due to the following grammatical markers:

- ‘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 grammatical markers indicate that the missing pronoun will be feminine and singular.

- (2) *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 Example 2. This example is more challenging as the only grammatical marker is the feminine singular noun ‘béile’ which is 11 tokens in distance from the masked token.

MWE Identification MWE identification is a difficult task, and according to the system results of the most recent edition of the PARSEME shared task,¹² it appears to be particularly challenging for Irish, with the

¹²Full results: <http://multiword.sourceforge.net/sharedtaskresults2020/>

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

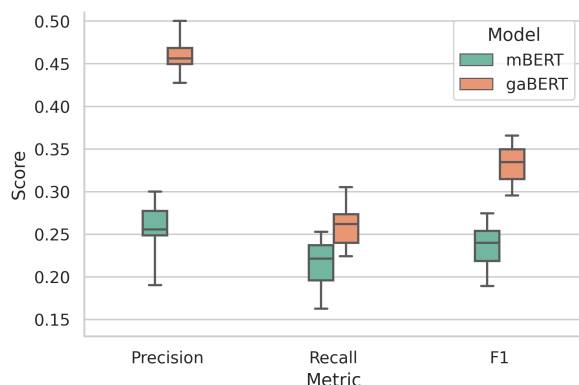


Figure 3: Verbal MWE Identification: Precision, Recall and F1 scores for each model across 20 random seed values

majority of systems performing most poorly on the Irish dataset. This may be due to the smaller size of the data, coupled with the relatively high number of MWE labels to classify (Walsh et al., 2020). We attempt a series of fine-tuning experiments varying the learning rate, batch size and initial random seed, and found model performance is sensitive to changes in hyperparameters. Figure 3 shows the results of training twenty models with different random seed values. It is evident that gaBERT outperforms mBERT in precision, recall and F1 scores on the test set.

Table 8 records the Precision (P), Recall (R), and F1

Model	P	R	F1
mBERT	0.342	0.245	0.285
gaBERT	0.523	0.361	0.427

Table 8: Verbal MWE Identification: (P)recision, (R)ecall and F1 scores of the best performing gaBERT and mBERT model

scores for the best performing gaBERT and mBERT model found during the manual tuning of hyperparameters (see Appendix C.2 for details). gaBERT performs better using these optimised parameters, particularly for precision scores, indicating that the gaBERT model tends to be correct more often when classifying MWEs than the mBERT model.

In comparison to other systems submitted to the PARSEME shared task on the Irish data, both models perform well. The best performing model for MWE identification had an F1 score of 0.306, which the gaBERT model exceeds by 0.121.¹³ On a multilingual level, the averaged F1 score for overall MWE identification of the highest ranking system was 0.701, and even with the improved F1 score of the best performing gaBERT model, results for Irish are still below the best system for Hebrew (0.483), which was the language where systems had the second-lowest performance.

¹³The results are not directly comparable, due to minor differences in calculating F1, so comparisons between our model and those systems submitted to the PARSEME shared task may be subject to slight variation when the same F1 calculation is used for both systems. Furthermore, emphasis in the most recent edition of the PARSEME shared task was on the identification of MWEs that had not been seen previously during the training phase. The highest ranking system for Irish actually had the second-highest F1 score for the task of global (seen and unseen) MWE identification, so we compare to the system with the highest F1 score for global MWE identification.

6. Friends of gaBERT

In subsequent experiments (see Appendix D-F for details), we look at variants of BERT, including RoBERTa (Liu et al., 2019). The multilingual XLM-R_{BASE} (Conneau et al., 2020) clearly outperforms both variants of mBERT but underperforms gaBERT. We expect that the more diverse crawled data found in the XLM-R pretraining data makes it more competitive than mBERT. We tried training a RoBERTa_{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 slightly below gaBERT but better than both mBERT models and the WikiBERT model. As with gaBERT, this is likely due to the use of a dedicated Irish vocabulary which is absent in the multilingual models, and being exposed to more diverse data than Irish Wikipedia in the case of WikiBERT.

7. Conclusions

We release gaBERT, a BERT model trained on over 7.9M Irish sentences (containing approximately 161M words), combining Irish language text from a variety of sources, and evaluate it in dependency parsing, a pronoun cloze test task, and a MWE identification task, showing improvements over three baselines, multilingual BERT, WikiBERT-ga and XLM-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 BERT_{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 pretrain 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.

Acknowledgements

This research is supported by Science Foundation Ireland (SFI) through the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This research is also supported through the SFI Frontiers for the Future programme (19/FFP/6942) and SFI Centre for Research Training in Machine Learning (18/CRT/6183), as well as by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project. We would like to thank Chris Larkin from the TPU Research Cloud (TRC) for generously providing TPU access and the anonymous reviewers for their helpful feedback and suggestions. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- Aulamo, M., Virpioja, S., and Tiedemann, J. (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, July. Association for Computational Linguistics.
- Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Semper, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (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, July. Association for Computational Linguistics.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (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, March.
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models.
- Cassidy, L., Lynn, T., Barry, J., and Foster, J. (2022). TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May. Association for Computational Linguistics.
- Chau, E. C., Lin, L. H., and Smith, N. A. (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, November. Association for Computational Linguistics.
- Chi, E. A., Hewitt, J., and Manning, C. D. (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, July. Association for Computational Linguistics.
- Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of The Eighth International Conference on Learning Representations (ICLR)*.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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, July. Association for Computational Linguistics.
- CSO. (2016). Census of Population 2016 – Profile 10 Education, Skills and the Irish Language. Publisher: Central Statistics Office.
- de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., Noord, G. v., and Nissim, M. (2019). BERTje: A Dutch BERT model, December. arXiv 1912.09582v1.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.
- Dowling, M., Lynn, T., Poncelas, A., and Way, A. (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, March. Association for Machine Translation in the Americas.
- Dowling, M., Castilho, S., Moorkens, J., Lynn, T., and Way, A. (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, November. European Association for Machine Translation.
- Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.
- Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2020). ParsBERT: Transformer-based model for Persian language understanding. arXiv 2005.12515v1.
- Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (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, July. Association for Computational Linguistics.
- Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (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.
- Jawahar, G., Sagot, B., and Seddah, D. (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, July. Association for Computational Linguistics.
- Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E. (2006). Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation*, 40:127–152.
- Kudo, T. and Richardson, J. (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, November. Association for Computational Linguistics.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. arXiv 1909.11942v6.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. arXiv 1907.11692v1.
- Lui, M. and Baldwin, T. (2012). langid.py: An off-the-

- shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.
- Lynn, T. and Foster, J. (2016). Universal Dependencies for Irish. In *Proceedings of the Second Celtic Language Technology Workshop (CLTW 2016)*, pages 79–92, Paris, France, July.
- Lynn, T., Cetinoglu, O., Foster, J., Dhonnchadha, E. U., Dras, M., and van Genabith, J. (2012). Irish tree-banking and parsing: A preliminary evaluation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1939–1946, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- Lynn, T., Foster, J., and Dras, M. (2013). Working with a small dataset - semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Lynn, T., Scannell, K., and Maguire, E. (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, July. Association for Computational Linguistics.
- Lynn, T. (2022). Report on the Irish language. <https://european-language-equality.eu/deliverables/>. Technical Report D1.20, European Language Equality Project.
- McGuinness, S., Phelan, J., Walsh, A., and Lynn, T. (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), December. Association for Computational Linguistics.
- Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, et al., editors, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, United Kingdom, July. Leibniz-Institut für Deutsche Sprache.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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, June. Association for Computational Linguistics.
- Pyysalo, S., Kanerva, J., Virtanen, A., and Ginter, F. (2020). WikiBERT models: deep transfer learning for many languages. arXiv 2006.01538v1.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Preprint.
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Rönnqvist, S., Kanerva, J., Salakoski, T., and Ginter, F. (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, September. Linköping University Electronic Press.
- Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2020). How good is your tokenizer? on the monolingual performance of multilingual language models. arXiv 2012.15613v2.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing, Third International Conference*, pages 1–15, Mexico City, Mexico, 02.
- Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain, April. Association for Computational Linguistics.
- Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.
- Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UD-Pipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. arXiv 1912.07076v1.
- Walsh, A., Lynn, T., and Foster, J. (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, August. Association for Computational Linguistics.
- Walsh, A., Lynn, T., and Foster, J. (2020). Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online, December. Association for Computational Linguistics.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz,

- M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (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, October. Association for Computational Linguistics.
- Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130. Online, July. Association for Computational Linguistics.
- Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Drozanova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (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, August. Association for Computational Linguistics.
- Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranzabe, M. J., Arnardóttir, H., Arutie, G., Arwidarasti, J. N., Asahara, M., Ateyah, L., Atmaca, F., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Balasubramani, K., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Batchelor, C., Bauer, J., Bedir, S. T., Bengoetxea, K., Berk, G., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Bjarnadóttir, K., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Çetinoğlu, Ö., Chalub, F., Chi, E., Cho, Y., Choi, J., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Derin, M. O., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dinakaramani, A., Dione, B., Dirix, P., Dobrovolsky, K., Dozat, T., Drozanova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Facundes, S., Farkas, R., Fernanda, M., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerardi, F. F., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grobol, L., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Güngör, T., Habash, N., Hafsteinsson, H., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Hanifmuti, M. Y., Hardwick, S., Harris, K., Haug, D., Heinecke, J., Hellwig, O., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Huber, E., Hwang, J., Ikeda, T., Ingason, A. K., Ion, R., Irimia, E., Ishola, O., Jelínek, T., Johannsen, A., Jónsdóttir, H., Jørgensen, F., Juutinen, M., K, S., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Köksal, A., Kopacewicz, K., Korkiakangas, T., Kotsyba, N., Kovalevskaitė, J., Krek, S., Krishnamurthy, P., Kwak, S., Laippala, V., Lam, L., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Levina, M., Li, C. Y., Li, J., Li, K., Li, Y., Lim, K., Lindén, K., Ljubešić, N., Logonova, O., Luthfi, A., Luukko, M., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsuda, H., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Mischenkova, K., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Mojiri Foroushani, A., Moloodi, A., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Mori, S., Morioka, T., Moro, S., Mortensen, B., Moskalevskiy, B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Nakhle, M., Navarro Horniáček, J. I., Nedoluzhko, A., Nešpore-Běrzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nourian, A., Nurmi, H., Ojala, S., Ojha, A. K., Olúokun, A., Omura, M., Onwuegbuzia, E., Osenova, P., Östling, R., Øvrelid, L., Özateş, Ş. B., Özgür, A., Öztürk Başaran, B., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perkova, N., Perrier, G., Petrov, S., Petrova, D., Phelan, J., Pitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rade-maker, A., Rama, T., Ramasamy, L., Ramisch, C., Rashel, F., Rasooli, M. S., Ravishankar, V., Real, L., Rebeja, P., Reddy, S., Rehm, G., Riabov, I., Riebler,

M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Rögnvaldsson, E., Romanenko, M., Rosa, R., Roşca, V., Rovati, D., Rudina, O., Rueter, J., Rúnarsson, K., Sadde, S., Safari, P., Sagot, B., Sahala, A., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulite, B., Sawanakunanon, Y., Scannell, K., Scarlata, S., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, Dmitry Simionescu, R., Simkó, K., Šimková, M., Simov, K., Skachodubova, M., Smith, A., Soares-Bastos, I., Spadine, C., Steingrímsson, S., Stella, A., Straka, M., Strickland, E., Strnadová, J., Suhr, A., Sulestio, Y. L., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tan, M. A. C., Tanaka, T., Tella, S., Tellier, I., Thomas, G., Torga, L., Toska, M., Trosterud, T., Trukhina, A., Tsarfaty, R., Türk, U., Tyers, F., Uematsu, S., Untilov, R., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wakasa, A., Wallenberg, J. C., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Widmer, P., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamashita, K., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zahra, S., Zeldes, A., Zhu, H., and Zhuravleva, A. (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.

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 2019 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 tool¹⁴ 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 $\langle s \rangle$ tags contain multiple sentences, we further split these segments with heuristics described in Appendix B.3.

¹⁴<https://github.com/attardi/wikiextractor>

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` file¹⁵ 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, 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, with the appropriate Unicode characters (3679).
- Repeat from the start until the text does not change.

We do not modify the seven occurrences of `\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 $\langle s \rangle$ 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 (among candidate split points after “.”, “?” and “!” tokens) 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.

- Reject if the left half contains no letters and is short. This includes cases where the left half is only a decimal number such as in enumerations.
- Reject if the right half has no letters and is short or is an ellipsis.

¹⁵MD5 7be5c0e9bc473fb83af13541b1cd8d20

- Reject if the right half’s first letter, skipping alphabetic and Roman enumerations in round brackets, 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 are not far away from the candidate split point.

For full-stop only:

- Reject after “DR”, “Prof” and “nDr”.
- Reject after “No”, “Vol” and “Iml” if followed by a decimal number.

Additional candidate split points are added with the following heuristics. Furthermore, when we need to choose between multiple candidate split points that pass the above tests, we try to keep the lengths of the halves (in characters) similar but also factor in the preferences in the heuristics below.

- 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 an 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 semi-colon we assume that this is an enumeration following a heading and prefer splitting before the “1”.
- 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.

C. Hyperparameters used in the Multitask Parser and MWE Identification Task

This appendix provides specific details and hyperparameters for the multitask parser and MWE identification model.

C.1. Multitask Parser

The hyperparameters of the multitask parser are given in Table 9. For the tagging tasks, the output of the Transformer is first projected through a task-specific Feedforward network and then passed to a classification layer. For dependency parsing, the projected representations from the tagging modules are concatenated to the output of the Transformer before being passed to the parsing module.

Multitask Parser Details	
Encoder	
Word-piece embedding size	768
Word-piece type	average
Tagger (UPOS/XPOS/Feats)	
MLP size	200
Dropout MLP	0.33
Nonlinear act. (MLP)	ELU
Parser	
Arc MLP size	500
Label MLP size	100
Dropout LSTMs	0.33
Dropout MLP	0.33
Dropout embeddings	0.33
Nonlinear act. (MLP)	ELU
Optimiser and Training Details	
Optimizer	AdamW
Learning rate	3e-4
beta1	0.9
beta2	0.999
Num. epochs	50
Patience	10
Batch size	16

Table 9: Chosen hyperparameters for the multitask parser and tagger.

C.2. MWE Identification

For the task of automatically identifying MWEs, the best performing models were found using a learning rate of $2e-5$, and a random seed of 10. We trained the models for 20 epochs each. Using a batch size of 5, we found the best performing mBERT model, while the best performing gaBERT model used a batch size of 1. We fine-tuned each model on all layers.

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 implementation¹⁶ 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.

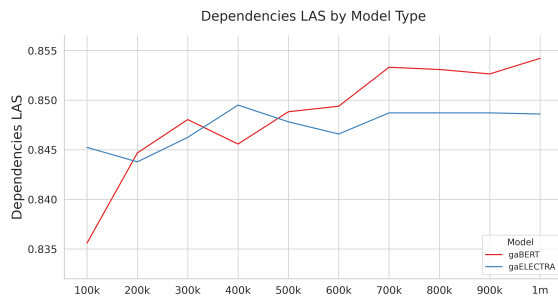


Figure 4: Dependency parsing LAS for each model type. Every 100k steps, we show the median of five LAS scores obtained from fine-tuning the respective model five times with different initialisation.

Figure 4 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 10 and 11. 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.

¹⁶<https://github.com/google-research/electra>

E. XLM-R Baseline

We add another off-the-shelf baseline by fine-tuning XLM-R_{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_{BASE} and gaELECTRA.¹⁷ Tables 10 and 11 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_{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_{BASE}. The continued pretraining of mBERT on our data enables us to close the gap between mBERT and XLM-R_{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_{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.

¹⁷We tried training a RoBERTa_{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
galextra	2.8	97.3	96.1	92.8	89.1	85.3

Table 10: 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
galextra	2.8	96.9	95.5	91.5	87.6	83.1

Table 11: Full model results on test data (os = fine-tuned off-the-shelf model, cp = continued pre-training before fine-tuning).