A BERT's Eye View: Identification of Irish Multiword Expressions Using Pre-trained Language Models

Abigail Walsh[1], Teresa Lynn[1], Jennifer Foster[2]
ADAPT Centre
Dublin City University
[1] {firstname.lastname}@adaptcentre.ie
[2] {firstname.lastname}@dcu.ie

Abstract

This paper reports on the investigation of using pre-trained language models for the identification of Irish verbal multiword expressions (vMWEs), comparing the results with the systems submitted for the PARSEME shared task edition 1.2. We compare the use of a monolingual BERT model for Irish (gaBERT) with multilingual BERT (mBERT), fine-tuned to perform MWE identification, presenting a series of experiments to explore the impact of hyperparameter tuning and dataset optimisation steps on these models. We compare the results of our optimised systems to those achieved by other systems submitted to the shared task, and present some best practices for minority languages addressing this task.

Keywords: Irish, BERT, multiword expressions, identification, pre-trained language models, hyperparameter-tuning, supervised learning, low-resource language NLP

1. Introduction

The automatic identification of multiword expressions (MWEs) has been highlighted as one of the two main subtasks of MWE processing (Constant et al., 2017), with their successful identification assisting a number of NLP tasks, such as parsing, machine translation and information retrieval. The PARSEME shared task on the automatic identification of verbal MWEs (vMWEs) (Savary et al., 2017), now in its third iteration, has recognised vMWEs as being of particular interest in this task, due to challenging properties that they can present, such as variability, ambiguity, and discontiguity. Its most recent edition (1.2), further highlighted the challenges inherent to identifying *unseen* vMWEs, that is, vMWEs that did not occur in either the training or development stage of model learning (Ramisch et al., 2020).

In this paper, we present a system for the identification of vMWEs in Irish and compare our results to other systems submitted to the PARSEME shared task. We use multilingual and monolingual language models, and demonstrate that monolingual models can lead to superior results, even compensating for small amounts of data. We also explore some of the optimisation steps that allow for lower-resourced languages, such as Irish, to fully exploit such resources, and report on patterns we find in these optimisation experiments.

2. Background

The Irish language is a minority language of the Celtic family of languages. Despite its status as the official language of Ireland, and an official working language of the European Union, it is recognised as a low resource language, particularly in the field of NLP (Judge et al., 2012; Lynn, 2022). Many NLP tasks lack the necessary resources for research in Irish, and the development of these resources has been an ongoing initiative for the past several years. Research into MWEs is one of those areas.

The PARSEME shared task on the identification of verbal MWEs came about as demand for a multilingual framework for the treatment of MWEs in NLP increased. Verbal MWEs, or vMWEs, are MWEs with a head verbal component, and include Light Verb Constructions ('LVCs') such as '***make** a **decision***', and Verbal Idioms 'VIDs' such as '***a little birdie told me***'. The latest edition (1.2) saw 14 languages included, as systems attempted to tackle the problem of *unseen* vMWEs, which has been recognised as a significant challenge in the task of MWE identification to date. Irish was one of the languages included, with the creation of the PARSEME annotated corpus of verbal MWEs for Irish (Walsh et al., 2020).

Of the nine systems participating in the shared task, five systems made use of neural networks: MultiVitaminBooster (Gombert and Bartsch, 2020), TRAVIS-mono and TRAVIS-multi (Kurfalı, 2020), MTLB-STRUCT (Taslimipoor et al., 2020) and ERMI (Yirmibeşoğlu and Güngör, 2020). Three used methods based on filtering using association measures: HMSid (Colson, 2020), Seen2Seen (Pasquer et al., 2020b) and Seen2Unseen (Pasquer et al., 2020a), while one system used a rule-based joint parsing and MWE identification system: FipsCo. Of the systems using neural networks, four of them included the use of pre-trained language models, those being multilingual BERT, monolingual BERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020).

Pre-trained language models have become the defacto standard language resource for many NLP tasks, with

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 89–99
Marseille, 25 June 2022.

a track record of beating previous SOTA results (Min et al., 2021). The MTLB-STRUCT system, which uses multilingual BERT fine-tuned for joint parsing and identification, achieved the best results for the open track in both the tasks of the identification of vMWES, and the subtask of identifying *unseen* vMWEs, when averaged across all languages. For individual languages, the only system in the open track to outperform the MTLB-STRUCT system was the TRAVIS-mono system, which uses a monolingual BERT model with a classification layer for MWE identification, where that language had a monolingual BERT model.

2.1. BERT and gaBERT

Bidirectional Encoder Representations from Transformers (BERT) is the transformer-based pre-trained language model that has seen applications in a wide variety of NLP tasks (Devlin et al., 2019). It is trained on two tasks: (1) a masked language modelling task, where words are masked and then predicted from their context, and (2) next sentence prediction, where the task is to determine if the second sentence in a pair follows the first one. These two tasks have proven to be sufficiently general that the resulting language model can be fine-tuned on a large number of NLP tasks, through the addition of a classification layer, and the adjustment of model parameters.[1] Two English language BERT models (BERT-base and BERT-large) were released, along with a multilingual BERT model (mBERT), which had been trained on a concatenation of Wikipedia data for 104 languages. Since the release of BERT, monolingual models have been built for many other languages, including Irish.

gaBERT (Barry et al., 2022) is a monolingual language model for Irish trained on approximately 7.9 million sentences in Irish. The training process and hyperparameters were largely kept the same as that of BERT, with the distinction of a smaller batch size to accommodate memory size limitations. gaBERT was evaluated on dependency parsing and a cloze test, and the results were compared with mBERT, showing that gaBERT was more effective than mBERT for both these tasks.

2.2. Irish in the PARSEME Shared Task

Until recently, Irish research on MWEs has been mostly limited to the field of theoretical linguistics or corpus linguistics. Developments on this topic for NLP include the publication of the Peadar Ó Laoghaire collection of idioms (Ní Loingsigh and Ó Raghallaigh, 2016), and the creation of a lexicon of Irish MWEs for research purposes (Walsh et al., 2019). The Irish UD treebank (Lynn and Foster, 2016) recently saw a unified treatment of MWEs applied to the data (McGuinness et al., 2020). The release of the PARSEME annotated corpus of verbal MWEs for Irish was the first corpus to be manually annotated for these types of verbal MWEs in Irish (Walsh et al., 2020). The corpus[2] consists of 1700 sentences originally from the Irish UD Treebank[3], which includes gold-standard POS-information, morphological features, and dependency relations. These sentences are manually annotated with seven categories of verbal MWEs: Light verb constructions ('LVC.full' and 'LVC.cause'), Inherently Adpositional Verbs ('IAV'), Verbal Idioms ('VID'), Verb-Particle constructions ('VPC.full' and 'VPC.semi'), and Inherently Reflexive Verbs ('IRV').

'LVCs' are the most numerous label in the Irish corpus, including constructions such as the 'LCV.full' *déan iarracht* 'make an attempt/try', or the 'LVC.cause' *cuir tús* 'put a start/start'. 'IAVs' are also frequent in Irish, such as *buail le* (lit. hit with) 'meet' or *éirigh le* (lit. rise with) 'succeed'.

The corpus was split according to the specifications of the PARSEME shared task (Ramisch et al., 2020), with a training dataset size of 257 sentences (100 vMWEs) and a development dataset size of 322 sentences (126 vMWEs), with the rest of the data in the test set (1120 sentences, and 442 vMWEs). Compared to the other languages in the shared task, the Irish corpus is small, with only Hindi (1684 sentences) being smaller. The number of vMWEs annotated in the corpus was also low, with only 662 vMWEs in total, compared to 1034 for Hindi. This, combined with the high ratio of *unseen* vMWEs present (69% of the vMWEs occurring in the test set were not present in either the training data or development data), as well as the relatively high numbers of categorisation labels used (7 labels, compared to a language average of 5), makes the task of vMWE identification in Irish particularly challenging.

3. Experiment Design

Approaching the task of vMWE identification as a sequence labelling task, we follow the example of the TRAVIS system and fine-tune both a multilingual BERT model (mBERT) and a monolingual BERT model (gaBERT) with a classification layer on this task, and compare the results. The classification layer is a linear layer connected to the language models' hidden states to perform token-level classification. The HuggingFace Transformers library (Wolf et al., 2020) provides both the mBERT (Devlin, J. et. al., 2018) and gaBERT (Barry, J. et. al., 2021) models, which can be integrated with their tokenising library to easily fine-tune language models.

The data we use is in `cupt` format, which is a combination of `CoNLL-U` format and `parseme-tsv` format. For this sequence labelling task, we only required

[1] While the precise reason for this ability for language models to generalise across many tasks is not well understood due to the black box nature of the pre-training, Zhang and Hashimoto (2021) suggests that the MLM task encourages the LM to capture statistical dependencies, which corresponds to general syntactic information.

[2] (Walsh, A. et. al., 2020)

[3] (Lynn, T. et. al., 2015)

Hyperparameter	Default	Tuning range
Number of epochs	20	5, 10, 15, 20, 25, 30, 35, 40
Batch size	8	1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20
Learning rate	2e-5	1e-6, 2e-6, 1e-5, 2e-5, 1e-4, 2e-4, 1e-3, 2e-3, 1e-2, 2e-2, 0.1, 0.2
Random seed	10	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100

Table 1: Default values used and range for tuning each hyperparameter.

the tokens and the MWE labels, so the data was processed into a json format containing this information. For labelling the vMWES, we used a modified *IOB2* scheme, as described in Section 3.2.1.

3.1. Series 1: Hyperparameter Optimisation

Fine-tuning hyperparameters is an important step in optimising a model's performance on a task, and even small adjustments to the hyperparameters can have a huge impact on model performance. There are many options to consider when tuning hyperparameters, from the selection of which hyperparameters to adjust, to the range of values being evaluated, to the method of hyperparameter optimisation being applied.

In this first series of experiments, we explore the impact that adjusting certain key hyperparameters has on our dataset, which is notably small. To best analyse the impact of this tuning, we opted to tune the hyperparameters **manually** and **individually**, selecting a combination of the best performing hyperparameters to fine-tune an optimised system for both the mBERT and gaBERT models.

We selected *learning rate*, *batch size*, *number of training epochs*, *the number of layers being fine-tuned*, and the *random seed variable* as our tunable parameters, defaulting to the values used by Devlin et al. (2019) for all other hyperparameters. We train both the mBERT and gaBERT models on three settings of layers: (i) fine-tuning all layers of the model, (ii) freezing layers 1-8 and only fine-tuning on the top-most 4 layers, (iii) freezing all layers and fine-tuning only the classification layer. The default values and range of values for tuning are represented in Table 1. To avoid over-fitting with the test data, we evaluated the model performance on the development set, and selected the best performing parameters based on these results.

3.1.1. Transformer Instability

A known issue in the training of transformer models is the tendency for instability of those models (Dodge et al., 2020; Bouscarrat et al., 2021; Mosbach et al., 2021), where the selection of a random seed value can have a significant impact on model performance. This effect appears to be magnified when training on small datasets, although, as Mosbach et al. (2021) find, this may be in fact due to the reduced number of iterations that training on smaller datasets may produce. To demonstrate the effects of this factor, we trained 10 mBERT-based models with a different random seed each time, which resulted in 2 models that failed to predict any MWEs at all.

Accounting for this instability, we perform the hyperparameter tuning first on the *number of epochs*, *batch size* and *learning rate*, before selecting the best performing hyperparameters and tuning this model on the *random seed value*, selecting the best performing random seed as our optimised hyperparameter. While the random seed value does not provide information on the model's structure as with the other hyperparameters, the selection of this value can drastically affect the performance of the model. The goal of our first series of experiments is to select an optimised model for comparison with the systems submitted at the PARSEME shared task so we have elected to tune this variable also.

3.2. Series 2: Data and Labelling

The second series of experiments focuses on some of the challenges for this task that are related to the data, such as data scarcity, relatively high number of labels, and the labelling scheme applied. We attempt to address these issues through optimising the data and comparing the baseline results to the models trained on these adjusted datasets.

3.2.1. Labelling Scheme

The labelling scheme used for the first series of experiments was a modified *IOB2* scheme, which is a version of the Inside-Outside-Beginning (*IOB*) format designed for chunking tasks such as NER detection (Ramshaw and Marcus, 1995). This required a conversion from the labels used in cupt, where vMWES are tagged with a number corresponding to the order the vMWE occurs within the sentence, to which a category label was appended for the first token of the vMWE. Our initial method was simply to select the first MWE label attached to each token, and discard any subsequent labels, a solution which was not always adequate. For instance, in converting from the *cupt* (Example 1) to the *IOB2* labelling (Example 2), the two LVCs *dhein staidéir* 'did study' and *dhein taighde* 'did research' have been incorrectly reduced to what appears to be a single LVC *dhein staidéir taighde* 'did study research'.

(1)

dhein	sé	an-chuid	**staidéir**	agus	**taighde**
1:LVC.full;2:LVC.full	*	*	1	*	2

(2)

dhein	sé	an-chuid	**staidéir**	agus	**taighde**
B-LVC.full	O	O	I-LVC.full	O	I-LVC.full

'he did a lot of study and research'

To address this, we propose a modified scheme *IOB2-double*, which uses the same labels as *IOB2*, but attempts to represent these 'doubly-annotated' tokens by adjusting the usage of the 'B-' labels: when encountering a token which has more than one vMWE label, the 'B-' prefix can be applied to both the initial token (vMWE #1), and the first subsequent token in the second vMWE (vMWE #2), as in Example 3. Using this scheme, the two LVCs are represented as 'did study' and 'research', which still does not capture the full picture, but prevents the loss of vMWEs through merging labels.

(3)

dhein	sé	an-chuid	**staidéir**	agus	**taighde**
B-LVC.full	O	O	I-LVC.full	O	B-LVC.full

This labelling scheme does not address the discontiguity of *dhein*, *staidéir* and *taighde*, which are interleaved with non-lexicalised components. Berk et al. (2019) discuss this issue, and propose an alternative labelling scheme, *bigappy-unicrossy*, which uses lower case labels and label prefixes ('b-', 'i-', 'o') to allow for one level of nested MWEs, two levels of discontinuity of MWEs (including nested discontinuous MWEs), and one level of crossing MWEs. Their scheme does not address the issue of double-tagged tokens or overlapping vMWEs, so we apply our adjusted 'B-' criteria. In this scheme, the previous text is annotated as in Example 4. The lower case labels indicate that the vMWE *dhein staidéir* 'do study' is partially nested, as elements of it come between construction *dhein taighde* 'do research'.

(4)

dhein	sé	an-chuid	**staidéir**	agus	**taighde**
B-LVC.full	o	o	i-LVC.full	o	B-LVC.full

3.2.2. Data Optimisation

The data-optimisation experiments address potential challenges that the Irish dataset presents over other languages: (i) the number of tags in the tagset, (ii) the complexity of the data, and (iii) the small size of the training and development datasets. To address these challenges, Exp 2A reduces the number of tags through first merging the two fine-grained labels ('LVC.full' and 'LVC.cause' → 'LVC'; 'VPC.full' and 'VPC.semi' → 'VPC'), and Exp 2B merges all tags into a single 'MWE' tag. Exp 3 reduces the complexity of the data through removing two challenging vMWE labels ('IRV' and 'VID'), while Exp 4 increases the size of the training and development datasets through re-splitting of the data, with 219 vMWEs annotated in the training data (+119 vMWEs), 216 vMWEs annotated in development data (+90 vMWEs) and 230 vMWEs in the test data (-212 vMWEs).

Of note, one of the so-called challenging vMWE labels, the 'IRV' label (e.g. ***iompair*** *mé* ***mé féin*** 'I **behaved myself**'), was identified previously (Walsh et al., 2020) as a label potentially worth removing due to the scarcity of this label occurring in the data (only 6 instances of this label were annotated) and the controversial nature of the label. The 'VID' label (e.g. ***cuir isteach sa chomhrá*** (lit. put into the conversation) 'intervene', ***dar le*** 'according to') presents the most syntactically and semantically diverse of the vMWE categories, given the highly variable nature of verbal idioms, whose lexicalised components can differ by part-of-speech, number, open-slots, etc.

4. Results and Analysis

4.1. Evaluation Metrics

We use both the evaluation library provided by seqeval (Nakayama, 2018), as well as the evaluation algorithm used in the PARSEME Shared Task (Ramisch et al., 2018) to evaluate our models, reporting *precision*, *recall* and *F1* scores. Two important differences between these algorithms are noted: (i) discontinuous MWE chunks are counted as separate MWEs by the seqeval calculations, and (ii) the PARSEME shared task evaluation metrics allow for partial matches of predicted vMWEs that share tokens with the gold annotated vMWEs ('Token-based' measures). When comparing our systems with those submitted for the PARSEME shared task, we limit the evaluation to the metrics calculated by the evaluation script provided for that task.

4.1.1. Analysis of Series 1

We trained each language model on the three layer settings mentioned in Section 3.1, resulting in six models for each hyperparameter tuning step: `mBERT-0` and `gaBERT-0` (layers 1-12 frozen, fine-tuned on 0 layers of language model), `mBERT-4` and `gaBERT-4` (layers 1-8 frozen, fine-tuned on final 4 layers of language model), and `mBERT-12` and `gaBERT-12` (no layers frozen, fine-tuned on all 12 layers).

`mBERT-12` and `gaBERT-12` models generally performed the best across our experiments, while `mBERT-0` and `gaBERT-0` generally performed the worst. From our experiments, we found training the models for more epochs improved performance, while batch size was inversely correlated with performance. The range of values containing the optimal learning rate varies depending on the layer settings, with `mBERT-0` and `gaBERT-0` requiring a larger learning rate. These trends are explained in more detail.

Number of Epochs: Training `mBERT-4` and `mBERT-12` for less than 5 epochs almost always produced a model that failed to predict any vMWE labels at all, with the same applying to `gaBERT-4` and `gaBERT-12`. This tendency to not predict labels decreased significantly as the number of epochs approached 15, while the $F1$ score for the models increased. This increase in $F1$ score continued an upwards trend to our upper bound of 40 epochs, though improvement slowed after 20 epochs.

Batch Size: The $F1$ score followed an inverse trend for batch size, with the peak $F1$ score achieved when batch size was between 1-4 for models `mBERT-4`, `mBERT-12`, `gaBERT-4` and `gaBERT-12`. Of note, when training with batch size of 20 for `mBERT-12`, the training halted due to memory limitations, highlighting the impact of hardware limitations on such experiments.

Learning Rate: Initially, the learning rates tuned were those described in Table 1. Noting the range of values that yielded the best performing models, we conducted a secondary tuning experiment using these optimised learning rates as anchor values for each of the layer settings, and training on a range of values on either side of these initial values. For `mBERT-4`, `mBERT-12`, `gaBERT-4` and `gaBERT-12`, when combined with the other default parameters, learning rates needed to be small; if the learning rate was larger than $8e$-4 it invariably produced a model that failed to predict any MWE labels. The best performing models used a learning rate of $4e$-5 for `mBERT-4` and `mBERT-12`, and $2e$-4 for `gaBERT-4` and `gaBERT-12`.

For `mBERT-0` and `gaBERT-0`, a larger learning rate was necessary to train a model that predicted MWE labels (greater than $2e$-4), and even learning rates as large as 0.8 will result in an $F1$ score of 10.1 (`mBERT-0`) and 23.2 (`gaBERT-0`). Combining this larger learning rate with the other hyperparameter tuning steps may result in even better performance.

Following these investigative experiments, we selected the best performing hyperparameter values from each trial and performed a series of experiments tuning the random seed value. As the best results for both models was consistently achieved for `mBERT-12` and `gaBERT-12`, we limited tuning to these layer settings. When using a combination of the best learning rate and batch size for `gaBERT-12`, we found that none of the models across any of the seed values succeeded in predicting any MWE labels, indicating that this particular combination of hyperparameters was not useful for our task. To find an optimised mode, we trained one series of models using the optimised learning rate parameter (`gaBERT-12-rate`) and one series of models using the optimised batch size (`gaBERT-12-batch`), with the values for the other hyperparameters taken from the default values in Table 2.

Random Seed: The box plot average of $F1$ scores from random seed tuning experiments are shown in Figure 1. We can see from the diagram that the `gaBERT-12-batch` model was more sensitive to instability than the `gaBERT-12-rate`, with the highest performing model achieving an $F1$ score of 43.0, but several seed values yielded a model that gave an $F1$ score of 0.0. The optimised gaBERT model was found with `gaBERT-12-rate` trained on random seed 10, while the optimised mBERT model (`mBERT-12`) was found on random seed 75.

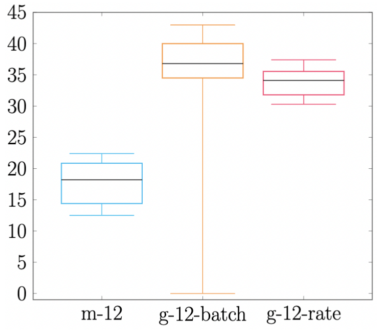

Figure 1: Box plot of $F1$ scores generated by mBERT-12, gaBERT-12-batch and gaBERT-12-rate models trained across 20 random seed values.

4.1.2. Analysis of Series 2

In our data optimisation experiments, we compare the results of models trained with the optimised hyperparameters of Series 1 on the baseline dataset (Exp 1), and datasets modified to address the challenges outlined in Section 4.1.2 (Exps 2A, 2B, 3 and 4). For each experiment, we apply the three labelling schemes discussed in Section 3.2.1: *IOB2*, *IOB2-double*, and *bigappy-unicrossy*. The $F1$ scores for each of these three datasets are displayed in Figure 2 (Exp 1), Figures 3 and 4 (Exp 2A and Exp 2B), Figure 5 (Exp 3) and Figure 6 (Exp 4). The precision, recall and $F1$ scores for each of these experiments are displayed in full in Table 3.

No clear discernible pattern emerges as to which labelling scheme produces the best results. In Figure 6 we see the results of models trained on reshuffled data (Exp 4) appears to show the *IOB2-double* labelling scheme out-performing *IOB2* labelling, with *bigappy-unicrossy* labelling giving the best results, however this trend was reversed for the mBERT model in Exp 2B, and for gaBERT in Exp 1.

The results of experiments 2A, 2B and 3 show that while modifying the dataset impacts the results of the model, it is difficult to predict whether this impact will be positive or negative. The results for Exp 2A demonstrate that the mBERT model trained on *IOB2-double* data failed to predict any MWE labels, again highlighting the model's susceptibility to instability. The experiments indicate that the language models' sensitivity to changes in dataset make it difficult to draw conclusions regarding the impact of the dataset optimisation, without further investigation into hyperparameter tuning.

4.2. Manual Inspection of Data

After inspecting the predicted labels, a large number of single-token predicted vMWEs were found. While single-token vMWEs did occur in the data as a result of converting from doubly-annotated tokens (see Section 3.2.1), these are relatively rare occurrences, and will only ever occur in combination with a multi-token vMWE. In contrast, the predicted single-token vMWEs would often occur with no other vMWE in context.

Parameter	mBERT-12	gaBERT-12-rate	gaBERT-12-batch
Number of epochs	30	30	30
Batch size	4	8	2
Learning rate	$4e$-5	$2e$-4	$2e$-5

Table 2: Hyperparameter settings for random seed tuning experiments.

Experiment	Model	Labelling	Precision	Recall	$F1$
Exp 1: Baseline dataset	mBERT-op	IOB2	16.09	12.93	14.34
		IOB2-d	20.05	17.09	14.34
		bi-uni	17.96	13.86	15.65
	gaBERT-op	IOB2	41.67	35.80	38.51
		IOB2-d	39.37	29.10	33.47
		bi-uni	39.59	26.79	31.96
Exp 2A: Fine-grained MWE labels merged	mBERT-op	IOB2	12.85	9.51	10.93
		IOB2-d	0.00	0.00	0.00
		bi-uni	12.83	9.05	10.61
	gaBERT-op	IOB2	46.21	31.09	37.17
		IOB2-d	45.25	37.59	41.06
		bi-uni	48.55	42.69	45.43
Exp 2B: All MWE labels merged	mBERT-op	IOB2	22.83	14.55	17.77
		IOB2-d	20.19	14.55	16.91
		bi-uni	16.86	10.16	12.68
	gaBERT-op	IOB2	41.83	33.72	37.34
		IOB2-d	36.75	25.64	30.20
		bi-uni	41.69	33.03	36.86
Exp 3: VID and IRV removed	mBERT-op	IOB2	15.69	12.83	14.12
		IOB2-d	9.35	8.82	9.08
		bi-uni	14.33	11.23	12.59
	gaBERT-op	IOB2	43.43	29.15	34.88
		IOB2-d	41.56	27.01	32.74
		bi-uni	48.28	41.18	44.44
Exp 4: Data resplit	mBERT-op	IOB2	18.06	16.96	17.49
		IOB2-d	22.47	22.17	22.32
		bi-uni	32.00	24.35	27.65
	gaBERT-op	IOB2	42.51	38.26	40.27
		IOB2-d	46.03	37.83	41.53
		bi-uni	46.53	40.87	43.52

Table 3: Precision, recall and $F1$ scores for the mBERT and gaBERT models trained on experiment data from Experiments 1–4, using optimised hyperparameters found in Series 1. Results obtained using the PARSEME ST evaluation script for global MWE-based evaluation, before the post-processing script was applied.

A post-processing script was added to each system where these single-token vMWEs were removed from the data, and this resulted in improved MWE-based precision and $F1$ scores for both models, an increase of 5.59 and 7.15 for global MWE-based $F1$ scores for mBERT- and gaBERT-optimised models respectively.

Between the models, this tendency to predict single-token vMWEs is more prevalent with the mBERT-based models than with gaBERT-based models, with the rate of single-token to multi-token MWE predictions almost double for the mBERT models, across all labelling schemes. Additionally, generating a bag-of-words of the predicted tokens of both models shows gaBERT-based models predict labels attached to a wider variety of tokens than mBERT-based models, particularly for 'LVC' type vMWEs.

Certain patterns in predictions were consistent across all experiments. Most of the 'VPC' label predictions were assigned to the tokens *bain* + *amach* (extract out) 'get', or some variation of these tokens, which make up the majority of the 'VPC' annotations in the training and development data. Verbs such as *cuir* 'put' were highly associated with 'LVC.cause' labels, reflecting the use of this verb in causative constructions, e.g. *cuir fearg (ar)* (put anger (on)) 'anger', while *déan* and *tabhair* ('make/do' and 'give') are highly associated with 'LVC.full', e.g. *déan iarratas* 'make an application'.

On examining individual categories of vMWES, it ap-

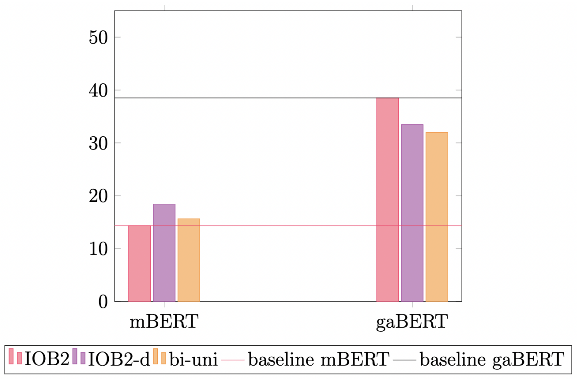

Figure 2: $F1$ scores for mBERT and gaBERT models for Exp 1: Using baseline data and comparing performance of labelling schemes (*IOB2*, *IOB2-double* and *bigappy-unicrossy*).

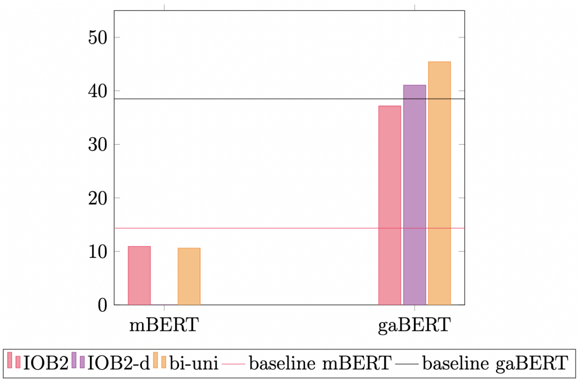

Figure 3: $F1$ scores for mBERT and gaBERT models for Exp 2A: Simplifying tagset by merging ‘LVC’ and ‘VPC’ sub-tags. Data labelled using *IOB2*, *IOB2-double* and *bigappy-unicrossy*.

mBERT	Freq	gaBERT	Freq
le	35	*le*	39
cuir	25	*cuir*	23
déan	23	*ar*	18
déanamh	16	*déan*	18
ar	14	*déanamh*	15
bain	12	*cur*	14
éirigh	11	*bain*	13
amach	10	*tabhair*	11
as	9	*éirigh*	11
tabhair	8	*i*	10

Table 4: Table showing 10 most frequently labelled words for mBERT-optimised and gaBERT-optimised models.

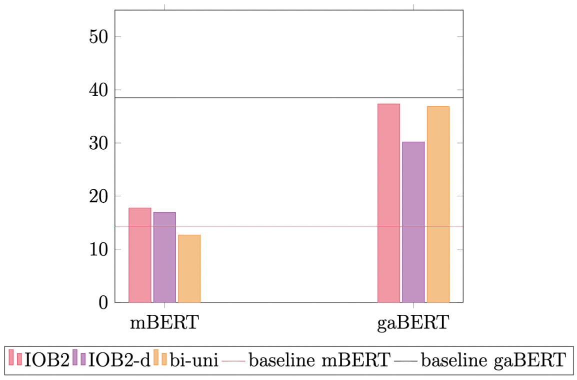

Figure 4: $F1$ scores for mBERT and gaBERT models for Exp 2B: Simplifying tagset by merging all vMWE labels. Data labelled using *IOB2*, *IOB2-double* and *bigappy-unicrossy*.

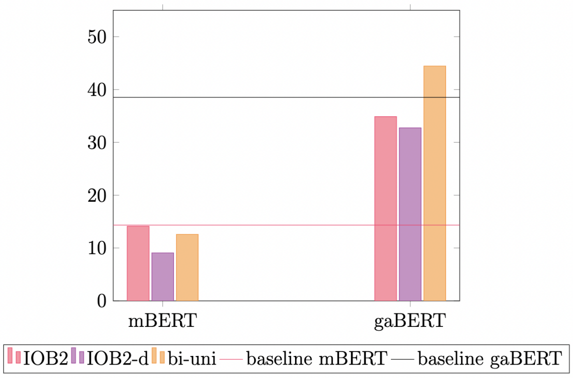

Figure 5: $F1$ scores for mBERT and gaBERT models for Exp 3: Simplifying dataset by removing challenging vMWEs ‘IRV’ and ‘VID’. Data labelled using *IOB2*, *IOB2-double* and *bigappy-unicrossy*.

pears some labels were easier to predict than others. Both gaBERT and mBERT appear to achieve high precision but low recall for ‘VPC.full’ MWEs, reflecting the scarcity of this label in the training data. gaBERT-based models appear to perform better on predicting both ‘LVC.full’ and ‘LVC.cause’ MWEs than mBERT-based models, with the baseline results showing a difference of 27.86 and 41.71 in the MWE-based $F1$ scores, respectively. ‘VID’ vMWEs proved challenging for both models to predict, with mBERT-based models outperforming gaBERT-based models, with an MWE-based $F1$ score of 12.35 vs 10.64.[4] These scores decreased further with the reshuffled dataset, with the mBERT-based model achieving an $F1$ score of 5.56 and the gaBERT-based model scoring 4.48.[4]

4.3. Optimised Model

When comparing the results of Exp 4 with the results of our optimised baseline model, we noted that the mBERT-based model sees a significant improvement for each of the evaluation metrics with the additional data, however, the gaBERT-based model actually saw a slight decline in the token-based and unseen MWE-based scores, particularly in precision scores. This result may be due to the addition of a larger variety of

[4] MWE-based $F1$ scores after removing single-token predictions.

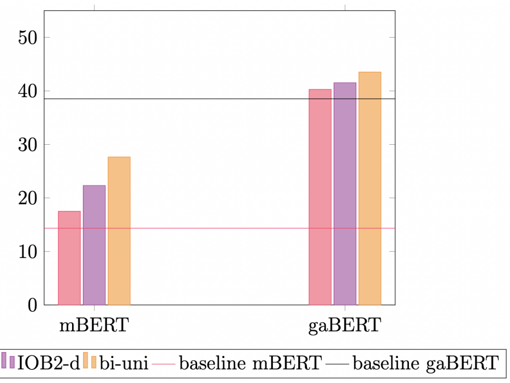

Figure 6: $F1$ scores for mBERT and gaBERT models for Exp 4: Increasing training and development data by reshuffling dataset splits. Data labelled using *IOB2*, *IOB2-double* and *bigappy-unicrossy*.

certain vMWEs such as 'VPCs', which in turn may prompt the model to attempt to predict these vMWEs attaching to a wider variety of tokens, making some incorrect predictions.

4.3.1. Comparison with Systems Submitted to the PARSEME Shared Task

Table 5 displays the results of systems submitted to the open track of the PARSEME shared task 1.2 for the Irish language. We see that our fine-tuned mBERT model from Series 1 compares favourably with the systems submitted for this task in Irish. Our mBERT-based system, if hypothetically submitted to the open track for Irish, would rank 3rd for unseen MWE identification, as well as MWE-based and token-based rankings. Our gaBERT-based system outperforms all other systems in this track, ranking 1st across all metrics, beating the MTLB-STRUCT system's MWE-based $F1$ score by 20.79 for unseen vMWE identification.

On the multilingual level, MTLB-STRUCT, the overall highest-performing system, achieved an MWE-based $F1$ score of 38.53 on unseen MWEs, a global MWE-based $F1$ score of 70.14, and a Token-based $F1$ score of 74.14, when averaged across all 14 languages. Even with the improvement in scores generated by the gaBERT-based model, Irish is still the language with the lowest performance score for global MWE-based and Token-based scores. However, the unseen MWE-based $F1$ score given by gaBERT is actually higher than the language average, and gaBERT outperforms the best system for several other languages (Basque, Hebrew, Italian, Portuguese and Romanian). This could be due to many Irish vMWE constructions consisting of common verbs (*bain* 'extract', *cuir* 'put', *tabhair* 'give', *faigh* 'get') and the language's proclivity for 'LVC' and 'IAV' constructions, which follow regular syntactic patterns.

4.4. Lessons Learned for Low-Resource MWE Identification

Following these experiments, we draw some conclusions from our method, and hope these learnings will be applicable to other lower-resourced languages tackling this task.

The results demonstrate the value of **monolingual language models** in such tasks. Our gaBERT-based models outperformed the mBERT-based models in almost all experiments conducted, barring some models which failed to predict any MWEs at all. This significant increase in performance is particularly reflected in the case of unseen VMWEs, which by their nature, present a great challenge to low-resource languages, as they are likely to be more prevalent where there is a scarcity of data/resources. Our experiments show how even a very small dataset can yield results similar to languages with much larger datasets (e.g. Portuguese, which had 6437 annotated vMWEs, almost 10 times the number annotated in the Irish dataset).

Clearly, such monolingual language models are expensive to train, both in language resources and in hardware required, and may be a challenge for lower-resource languages to build. However, our experiments show that multilingual models such as mBERT show promising capabilities to capture even unseen vMWES, and even small additions to the data can dramatically improve these results. These experiments also highlighted the importance of careful **hyperparameter tuning**, as the manual explorations of the hyperparameter space resulted in an improvement of 4.73 (8.86 after single-tokens were removed) in the unseen MWE-based $F1$ score compared to the mBERT-based system submitted by TRAVIS-multi.

Our experiments confirm the susceptibility of transformer-based models to **instability**, where even small variations in the data or in the hyperparameters selected (particularly the varying of the random seed variable) can result in a model that fails to predict any labels whatsoever. This problem seems to be exacerbated by the small size of the training data. However, our experiments indicate that the issue can be combatted through increasing the number of epochs trained for, and by varying the learning rate. This finding of ours parallels the work of Mosbach et al. (2021) who, upon investigating the topic of instability in fine-tuning BERT, recommend using small learning rates with bias correction to avoid vanishing gradients early in training, and increasing the number of iterations considerably and training to near zero training loss. However, as discussed in Section 4.1.1, some combinations of hyperparameters may result in unexpected model behaviour during training. As such, a random search hyperparameter tuning approach may be the most effective, as there is little guarantee that a well-performing hyperparameter setting will still perform well when combined with a different well-performing hyperparameter.

Category	Model	Precision	Recall	$F1$
Unseen MWE-based	gaBERT-optimised	53.30	32.44	**40.33**
	MTLB-STRUCT	23.08	16.94	19.54
	Seen2Unseen	21.74	9.97	13.67
	mBERT-optimised	25.88	07.36	11.46
	Travis-multi	3.75	1.99	2.6
	MultiVitaminBooster	0.0	0.0	0.0
Global MWE-based	gaBERT-optimised	63.01	35.80	**45.66**
	MTLB-STRUCT	37.72	25	30.07
	Seen2Unseen	44.16	23.39	30.58
	mBERT-optimised	43.41	12.93	19.93
	Travis-multi	12.36	5.05	7.17
	MultiVitaminBooster	0.0	0.0	0.0
Global Token-based	gaBERT-optimised	74.31	42.89	**54.38**
	MTLB-STRUCT	65.02	33.79	44.47
	Seen2Unseen	50.41	24.11	32.62
	mBERT-optimised	65.76	19.30	29.85
	Travis-multi	65.48	16.3	26.11
	MultiVitaminBooster	0.0	0.0	0.0

Table 5: Precision, recall and $F1$ scores for unseen MWE-based, global MWE-based and global Token-based metrics for open-track systems submitted to the PARSEME shared task 1.2 for the Irish annotated corpus, with our optimised gaBERT and mBERT-based models included for comparison.

We also investigated the potential for **alternative sequence labelling schemes** that more accurately capture the vMWE labels. Our experiments on this topic are inconclusive, as there is no guarantee that the results we found are consistent when applied to a model trained on different hyperparameter settings. However, these alternative labelling schemes do allow for capturing doubly-annotated tokens, which previously would have been lost when using a traditional *IOB2* labelling scheme.

5. Conclusion & Future Work

In this paper we report on an exploration of the application of pre-trained language models (both multilingual and monolingual) for the task of vMWE identification in Irish. Following the example of the TRAVIS systems submitted to the PARSEME shared task 1.2, we fine-tune language models to perform sequence labelling classification of the tokens, describing two series of experiments, exploring hyperparameter tuning, and data modifications addressing potentially challenging issues. We briefly discuss the labelling scheme used, focusing on the issue of labelling doubly-annotated (overlapping) tokens.

Our results reveal patterns in hyperparameter tuning, and these insights lead us to developing an optimised mBERT and gaBERT-based model. Five experiments exploring data modification and labelling of the data show inconclusive patterns with $F1$ scores achieved. A manual inspection of the data reveals some patterns in predicted MWEs by model and category. A comparison of our optimised systems for both mBERT and gaBERT with the PARSEME shared task results demonstrate the importance of careful hyperparameter tuning.

These experiments particularly highlight the value of monolingual language models in this task, as the gaBERT-based model achieved unseen MWE-based $F1$ scores that outperformed other systems submitted for the Irish corpus, and even outperformed systems submitted for other, higher-resourced languages, indicating that high-quality language-specific resources can compensate for a lack of language data in certain NLP tasks.

Future work includes continuing hyperparameter optimisation following the data optimisation strategies explored in this work and application of alternative labelling schemes, to investigate the full impact of these changes to a potentially optimised MWE identification model. We would also consider experiments in joint-learning tasks, such as the joint parsing and MWE identification systems trained by MTLB-STRUCT, which showed promising results. Such experiments allow for exploitation of other linguistically rich Irish resources, such as the Irish UD Treebank.

6. Acknowledgements

This research is funded by the Irish Government Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media under the GaelTech Project. This work is also supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (`http://www.adaptcentre.ie`) at Dublin City University. The authors would like to thank the anonymous reviewers for their helpful feedback and suggestions.

7. Bibliographical References

Barry, J., Wagner, J., Cassidy, L., Cowap, A., Lynn, T., Walsh, A., Ó Meachair, M. J., and Foster, J. (2022). gaBERT – an Irish Language Model. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France, June.

Berk, G., Erden, B., and Güngör, T. (2019). Representing overlaps in sequence labeling tasks with a novel tagging scheme: bigappy-unicrossy. Computational Linguistics and Intelligent Text Processing, Springer International Publishing.

Bouscarrat, L., Bonnefoy, A., Capponi, C., and Ramisch, C. (2021). AMU-EURANOVA at CASE 2021 Task 1: Assessing the stability of multilingual BERT. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 161–170, Online, August. Association for Computational Linguistics.

Colson, J.-P. (2020). HMSid and HMSid2 at PARSEME shared task 2020: Computational corpus linguistics and unseen-in-training MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 119–123, online, December. Association for Computational Linguistics.

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.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Survey: Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892, December.

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.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. A. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.

Gombert, S. and Bartsch, S. (2020). MultiVitaminBooster at PARSEME shared task 2020: Combining window- and dependency-based features with multilingual contextualised word embeddings for VMWE detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 149–155, online, December. Association for Computational Linguistics.

Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. P., and Uí Dhonnchadha, E. (2012). *The Irish Language in the Digital Age*. Springer Publishing Company, Incorporated.

Kurfalı, M. (2020). TRAVIS at PARSEME shared task 2020: How good is (m)BERT at seeing the unseen? In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 136–141, online, December. Association for Computational Linguistics.

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

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.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2021). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, pages 847–869, Vienna, Apr.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Ní Loingsigh, K. and Ó Raghallaigh, B. (2016). Starting from scratch – the creation of an Irish-language idiom database. In George Meladze Tinatin Margalitadze, editor, *Proceedings of the 17th EURALEX International Congress*, pages 726–734, Tbilisi, Georgia, sep. Ivane Javakhishvili Tbilisi University Press.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020a). Seen2Unseen at PARSEME shared task 2020: All roads do not lead to unseen verb-noun VMWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online, December. Association for Computational Linguistics.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020b). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–

3345, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online, December. Association for Computational Linguistics.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

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.

Taslimipoor, S., Bahaadini, S., and Kochmar, E. (2020). MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online, December. Association for Computational Linguistics.

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.

Yirmibeşoğlu, Z. and Güngör, T. (2020). ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135, online, December. Association for Computational Linguistics.

Zhang, T. and Hashimoto, T. B. (2021). On the inductive bias of masked language modeling: From statistical to syntactic dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5131–5146, Online, jun. Association for Computational Linguistics.

8. Language Resource References

Barry, J. et. al. (2021). *gaBERT Irish language model*. distributed via Huggingface Library: DCU-NLP/bert-base-irish-cased-v1.

Devlin, J. et. al. (2018). *BERT multilingual language model*. distributed via Huggingface Library: bert-base-multilingual-cased.

Lynn, T. et. al. (2015). *Irish UD Treebank*. distributed via LINDAT/CLARIAH-CZ: `http://hdl.handle.net/11234/1-4611`.

Walsh, A. et. al. (2020). *PARSEME corpus for Irish*. distributed via LINDAT/CLARIAH-CZ: `http://hdl.handle.net/11234/1-3367`.