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

Abigail Walsh, Teresa Lynn, Jennifer Foster

PARSEME Shared Task 1.2

- (Ramisch et al. 2020)
- Automatic identification of *unseen* verbal MWEs
- 14 languages
  - First time including **Irish** (Walsh et al. 2020)
- 9 systems: 7 *open* track & 2 *closed* track
  - 4 systems used **Pre-trained language models**

Background

- **Pre-trained language models** have seen widespread use in many NLP applications
- **Monolingual language models** have been shown to give better model performance than multilingual language models for certain tasks
- We compare results obtained using a **multilingual language model** (mBERT) (Devlin et al. 2019) with an **Irish monolingual language model** (gaBERT) (Barry et al. 2022) for the task of **automatic identification of verbal MWEs (vMWEs) in Irish**

Challenges for Irish dataset

- Systems performed the **most poorly** on Irish data
- **Many labels used** (7 compared to average of 5 across languages)
- **High ratio of *unseen* vMWEs** (69% compared to average of 33% across languages)
- **Small number** of training and tuning **examples** (226 compared to average of 3645 across languages)
- Irish vMWEs exhibit **high degree of variability**

Model Instability

- Known issue in **Transformer** architecture
- Training a model with 10 random seed values shows **variable F1 scores**

Run	Precision	Recall	$F1$
1	0.3288	0.2330	0.2727
2	0.3158	0.2330	0.2682
3	**0.0**	**0.0**	**0.0**
4	0.2870	0.1602	0.2056
5	0.3401	0.2427	0.2833
6	0.2566	0.1408	0.1818
7	**0.0**	**0.0**	**0.0**
8	0.2727	0.1602	0.2018
9	0.3008	0.1942	0.2360
10	0.2966	0.1699	0.2160

Experiment Series 1

Hyperparameter tuning

Number of layers

- Range [0–12]
- Training on **all 12 layers** gives better performance

Batch size

- Range [1–20]
- **Lower batch size** (1–4) improves performance

Number of epochs

- Range [5–40]
- Training on **more epochs** improves model performance and stability

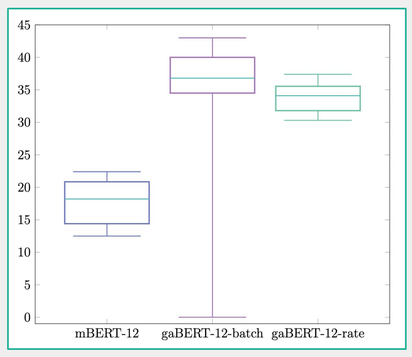
Learning rate

- Range [1e-6–0.8]
- Better model performance found with learning rate between **2e-5 and 8e-4**
- Exception was models with all 12 layers frozen: required **larger learning rate**

Random seed

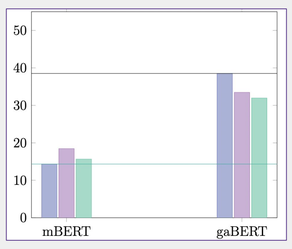
- 20 trials using best performing hyperparameters
- Random seed values selected evenly from **5-100**
- It was found that combining a **batch size of 2** and **learning rate of 2e-4** results in a model that **does not predict** any MWEs
- Instead, two random seed experiments devised for gaBERT models using **best learning rate** (gaBERT-12-rate) and **best batch size** (gaBERT-12-batch)

Parameter	mBERT-12	gaBERT-12-rate	gaBERT-12-batch
Num Epochs	30	30	30
Batch size	4	8	2
Learning rate	4e-5	2e-4	2e-5



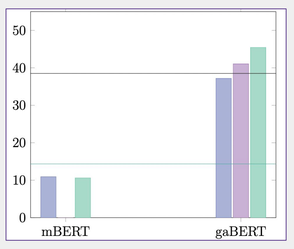
Experiment Series 2

Addressing dataset challenges



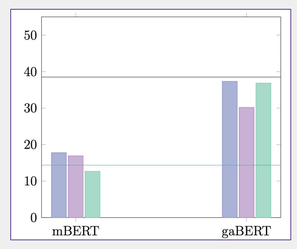
Experiment 1

- Using baseline data
- Comparing across 3 labelling schemes



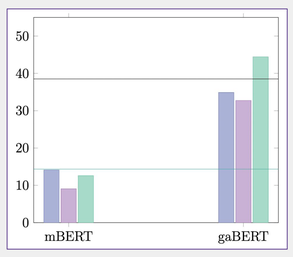
Experiment 2A

- Addressing label complexity by **merging two MWE labels**
  - `LVC.full` and `LVC.cause` → `LVC`
  - `VPC.full` and `VPC.semi` → `VPC`



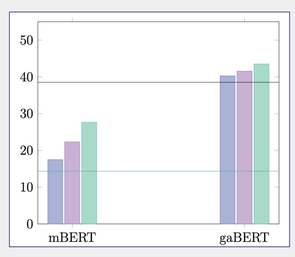
Experiment 2B

- Addressing label complexity by **merging all MWE labels**
  - Any category → `MWE`



Experiment 3

- Addressing data complexity by **removing difficult MWEs**
  - Removed controversial and infrequent `IRV`
  - Removed diverse `VID`



Experiment 4:

- Addressing data scarcity by **reshuffling dataset splits**
  - Training data: 219 (**+119**) vMWEs
  - Tuning data: 216 (**+90**) vMWEs
  - Testing data: 230 (**-212**) vMWEs

Labelling Schemes

Sentence:	**Dhein**	**sé**	**an-chuid**	**staidéir**	**agus**	**taighde**
CUPT:	**1:LVC.full; 2:LVC.full**	*	*	**1**	*	**2**
IOB2:	**B-LVC.full**	**O**	**O**	**I-LVC.full**	**O**	**I-LVC.full**
IOB2-d:	**B-LVC.full**	**O**	**O**	**I-LVC.full**	**O**	**B-LVC.full**
bigappy-uni-d:	**B-LVC.full**	**o**	**o**	**i-LVC.full**	**o**	**B-LVC.full**

Annotating sentence "He **did** a lot of **study** and **research**"

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

Comparing precision, recall and F1 scores of our **optimised gaBERT** and **mBERT-based models** with systems submitted to the PARSEME shared task 1.2 for the Irish dataset

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

10 most frequently labelled words for mBERT-optimised and gaBERT-optimised models.

Conclusions

- Results demonstrate that **monolingual pre-trained language models** can achieve surprisingly good results even on small datasets
- Instability is an issue, particularly with small datasets
  - Can be combated through **training for more epochs**, and careful selection of **learning rate**
- Addressing dataset challenges shows **inconclusive results**
- Similarly with **alternative labelling schemes**
- Possible that effects would be more noticeable using different hyperparameters and larger datasets

Fondúreacht Eolaíochta Éireann
Dá bhfuil romhainn
Science Foundation Ireland
For what's next

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Science Foundation Ireland

An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.