Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese

Kurt Micallef[1]
kurt.micallef@um.edu.mt

Albert Gatt[2,3]
a.gatt@uu.nl

Marc Tanti[3]
marc.tanti@um.edu.mt

Lonneke van der Plas[4,3]
lonneke.vanderplas@idiap.ch

Claudia Borg[1]
claudia.borg@um.edu.mt

[1]Department of Artificial Intelligence, University of Malta
[2]Information and Computing Sciences, Utrecht University
[3]Institute of Linguistics and Language Technology, University of Malta
[4]Idiap Research Institute

Abstract

Multilingual language models such as mBERT have seen impressive cross-lingual transfer to a variety of languages, but many languages remain excluded from these models. In this paper, we analyse the effect of pre-training with monolingual data for a low-resource language that is not included in mBERT – Maltese – with a range of pre-training set ups. We conduct evaluations with the newly pre-trained models on three morphosyntactic tasks – dependency parsing, part-of-speech tagging, and named-entity recognition – and one semantic classification task – sentiment analysis. We also present a newly created corpus for Maltese, and determine the effect that the pre-training data size and domain have on the downstream performance. Our results show that using a mixture of pre-training domains is often superior to using Wikipedia text only. We also find that a fraction of this corpus is enough to make significant leaps in performance over Wikipedia-trained models. We pre-train and compare two models on the new corpus: a monolingual BERT model trained from scratch (BERTu), and a further pre-trained multilingual BERT (mBERTu). The models achieve state-of-the-art performance on these tasks, despite the new corpus being considerably smaller than typically used corpora for high-resourced languages. On average, BERTu outperforms or performs competitively with mBERTu, and the largest gains are observed for higher-level tasks.

1 Introduction

Language Models have become a core component in many Natural Language Processing (NLP) tasks. These models are typically pre-trained on unlabelled texts, and then further fine-tuned using labelled data relevant to the target task. Transformer-based (Vaswani et al., 2017) contextual models such as BERT (Devlin et al., 2019) have gained success since the fine-tuning step is relatively inexpensive, while attaining state-of-the-art results in various syntactic and semantic tasks.

While the bulk of work with the BERT family of models focuses on English, there have been some monolingual models developed for other languages as well (Martin et al., 2020; Polignano et al., 2019; Antoun et al., 2020; de Vries et al., 2019; Virtanen et al., 2019; Agerri et al., 2020; inter alia). These monolingual models have been trained on large volumes of data, typically amounting to billions of tokens. In contrast, it is challenging to find publicly available corpora of this size for low-resource languages. The evaluation benchmarks for downstream tasks on these languages are also limited, and tend to be dominated by low-level structural tagging tasks.

To counteract the lack of large volumes of monolingual corpora for low-resource languages, a number of multilingual models have been released, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). These multilingual models were pre-trained on more than one language at a time by combining corpora from different languages, usually sourced from Wikipedia. Several works have demonstrated the efficacy of these multilingual models, especially for languages without a language-specific model (Kondratyuk and Straka, 2019; Wu and Dredze, 2019). Benchmark results have improved for many languages by leveraging cross-linguistic features learnt by these multilingual models (Conneau et al., 2020).

However, the gains with multilingual models may vary depending on the language being considered. The "curse of multilinguality" limits the language-specific features that these models can learn, since the limited model capacity has to be

shared between multiple languages (Conneau et al., 2020). Models such as mBERT use WordPiece tokenisation (Johnson et al., 2017), which splits words into various sub-tokens, thereby reducing the number of unknown tokens. However, the vocabulary representations for multilingual models tend to be sub-optimal for specific languages, because words tend to be split into a higher number of sub-tokens (Rust et al., 2021). Moreover, these models may still be biased in favour of over-representing sub-tokens common to a certain subset of languages over others. Due to the data imbalance across languages, lower-resourced languages tend to be disadvantaged, as there is relatively less pre-training data available compared to the other languages considered in the multilingual model (Wu and Dredze, 2020).

Apart from the tension between languages in a multilingual model, other factors are at play as well. Most prominently, many languages are never seen by these multilingual models (Muller et al., 2021), since these are typically trained on the largest-available corpora (e.g. mBERT was pre-trained on the 104 languages with the greatest Wikipedia presence). Such criteria exclude many of the world's languages, including Maltese, the focus of this paper. This issue is exacerbated even further when the language uses a script which is either different to its closely related languages (Muller et al., 2021), or which is never seen during pre-training, thereby encoding most of the input with out-of-vocabulary tokens (Pfeiffer et al., 2021). In fact, Muller et al. (2021) show that the language transfer capability of a multilingual model to an unseen language is dependent on the degree to which the target language is related to languages already included in the multilingual model.

In this work we focus on the Maltese language, an official EU language spoken primarily in Malta and in some small communities around the world (Brincat, 2011). It is the only Semitic language written exclusively with a Latin script, containing a few additional characters with diacritic marks (ċ, ġ, ħ, ż). The language also has strong influences from Romance languages such as Italian, as well as English. The Semitic influence is largely exhibited in the grammatical structure through complex morphological characteristics, whilst the non-Semitic aspect is predominantly observed in its vocabulary, with extensive lexical borrowing from Italian and English.

In the context of NLP, Maltese is a low-resource language (Rosner and Borg, 2022) and is not part of the languages covered by either mBERT or XLM-R. Muller et al. (2021) find that mBERT underperforms non-contextual baselines on Maltese, but benefits when pre-trained further on raw Maltese data. Similarly, Chau et al. (2020) further pre-train mBERT but impute the 99 unused tokens present in the model with language specific tokens, yielding better results. This confirms previous findings by Wang et al. (2020), who also extend mBERT's vocabulary to accommodate unseen languages, but do so by extending the vocabulary and model dimensionality, hence increasing its footprint.

Motivated by the limitations of existing multilingual models and the deficiency of publicly available corpora for Maltese, we set out to pre-train a new monolingual language model for Maltese and compare it to the alternative strategy of further pre-training an existing multilingual model. We study, in particular, the impact that the pre-training data size and domain has on the performance in downstream tasks. The main contributions of this work are as follows:

1. We develop a new corpus of Maltese text.

2. Using this new data, language models for Maltese are pre-trained.

3. We compare the newly pre-trained models and find that both models improve the state-of-the-art on three structural tagging tasks – dependency parsing, part-of-speech tagging, and named-entity recognition – and one semantic classification task – sentiment analysis.

4. We demonstrate that in a low-resource setting, pre-training using text from varied domains is often superior to solely using Wikipedia, and that matching the domain to target task is beneficial when this is available.

5. We also provide an analysis on the effects of the pre-training size, shedding new light on how much pre-training data is needed to attain significant improvements in performance.

We make this new corpus, the newly pre-trained language models, and the code publicly available[1].

---

2 Corpus

In this work, we build a new unlabelled text corpus, which we call the **Korpus Malti v4.0 (KM)**. This builds on and extends an existing corpus, Korpus Malti v3.0[2], which is approximately half the size.

Rather than scraping the web randomly for Maltese text, we collect text data from specific sources, including both online and offline. Although this does incur additional effort in data collection, and results in a smaller dataset compared to large-scale web-scraping initiatives, it has the benefit of resulting in a less noisy dataset, while offering greater control over sources. For comparison, the Maltese portion of the OSCAR data (Ortiz Suárez et al., 2019), which is sourced entirely from the web, contains texts which, to a native speaker, suggest that they are automatically generated through the use of a low-quality machine translation system, a common pitfall of web-scraping for low-resource languages (Kreutzer et al., 2022). We also expect to find a small proportion of code-switched texts, as this is a pre-dominant phenomenon for Maltese in domains such as social media or transcribed speech. In addition, the data is separated into different domains, and the source for each document is available as part of the metadata. This allows data users to select data subsets which are more appropriate for their particular use-case, such as domain-adaptive pre-training (Lee et al., 2019; Gururangan et al., 2020; inter alia), whilst enabling tracing back to the original source, or omission in case unforeseen ethical or privacy issues come to light. In short, the goal was to build a good quality training dataset, while avoiding at least some of the pitfalls identified with opportunistic, web-scale data initiatives (Bender et al., 2021; Rogers, 2021).

Data is collected from a variety of sources, including online news sources, legal texts, transcripts of speeches and debates, blogs, Wikipedia, etc. Before texts are included in the corpus, we filter non-Maltese sentences using language identification using LiD (Lui and Baldwin, 2014), and perform de-duplication using Onion (Pomikálek, 2011).

The resulting data, split into 19 different domains, is summarised by Table 1.

To the best of our knowledge, there is no corpus of this size available for Maltese. We also note that this data is a significant increase over Wikipedia data, which is what is usually available and used in low-resource scenarios. The Wikipedia data makes up less than 1% of the entire corpus in terms of both tokens and sentences.

Despite this substantial increase in data, we emphasise that a corpus of under 500M tokens is still substantially smaller than is typically used for higher-resourced languages. For example, Devlin et al. (2019) pre-train BERT using a combined corpus of 3.3B words for English (approximately 16GB). Larger models have since exceeded these pre-training sizes by a wide margin – for example, RoBERTa is pre-trained on 161GB of text (Liu et al., 2019). Monolingual models for languages other than English, typically use smaller corpora than English models, but their size is still significantly larger than ours – for example AraBERT was pre-trained on a corpus of 24GB (Antoun et al., 2020) and BERTje was pre-trained on a corpus of 12GB (de Vries et al., 2019).

3 Language Models

Using this new corpus, two new language models are pre-trained for Maltese: a monolingual model (**BERTu**) and a multilingual model (**mBERTu**). In both cases, pre-training is performed using the Masked Language Modelling Objective (MLM) only, since the Next Sentence Prediction (NSP) objective was found to be detrimental to downstream performance (Joshi et al., 2020; Liu et al., 2019). Other than that, pre-training largely follows the pre-training setup of BERT (Devlin et al., 2019). This allows for a better comparison with already available models. The pre-training data from all domains is combined, shuffled, and split into 85% and 15% for training and validation sets respectively.

BERTu We pre-train a monolingual BERT model from scratch on the new unlabelled data, using the BERT$_{\text{BASE}}$ architecture with 12 transformer layers, a hidden size of 768, and 12 attention heads. The vocabulary is initialised with 52K tokens. Pre-training is done across 1M steps, with a sequence length of 128 for the first 90% of the steps and a sequence length of 512 for the remaining 10% steps. A batch size of 512 is used, which amounts to approximately 30 epochs in total, and a warmup of 1% of the total number of steps. We use mixed-precision training to ease memory requirements. Training was performed on 8 A100 GPUs for the first 90% steps and 16 A100 GPUs for the remaining 10% steps, taking approximately 53 hours.

---

[2]See: https://mlrs.research.um.edu.mt

data subset	documents	sentences	tokens	size
belles_lettres	195	299 762	4 454 906	21.82MB
blogs	25 436	807 628	14 562 039	74.45MB
comics	62	2 413	44 768	233.22KB
court	2 663	694 227	11 881 638	61.91MB
eu_docs	2 974	5 099 564	135 811 945	773.25MB
government_gazette	2 974	1 881 034	39 771 556	203.61MB
gov_docs	272	120 209	1 900 842	10.79MB
law_eu	71	4 433 235	98 582 031	541.13MB
law_mt	2 596	401 118	7 631 651	38.84MB
legal	3	4 784	83 581	490.67MB
nonfiction	2 177	208 763	3 902 436	20.01MB
parliament	6 198	3 935 906	82 294 520	433.09MB
press_eu	5 483	413 317	9 774 919	55.73MB
press_mt	46 782	713 886	17 679 904	93.15MB
speeches	62	2 067	51 259	286.63MB
theses	19	11 545	310 243	1.63MB
umlib_oar	11 688	963 606	21 235 949	106.11MB
web_general	2	685 873	14 741 525	75.22MB
wiki	3 469	79 134	1 885 661	9.73MB
all	131 429	20 758 071	466 601 373	2.52GB

Table 1: Korpus Malti v4.0 corpus distribution. *belles_lettres* is largely composed of literary works; the *government_gazzette* consists of text from the official newsletter of the Maltese government; *umlib_oar* is a miscellaneous collection of previously published non-fiction texts, available in the public domain via the University of Malta Library Open Access Repository.

mBERTu Similar to Chau et al. (2020) and Muller et al. (2021) we also pre-train mBERT further on Maltese. Since the embedding weights are not randomly initialised, as is the case for the monolingual model, we follow Rust et al. (2021) and pre-train for 250K steps. A sequence length of 512 is used throughout, keeping the rest of the hyper-parameters the same as the monolingual pre-training. To better fit the Maltese language, the mBERT vocabulary is augmented with Maltese tokens following the procedure from Chau et al. (2020), by replacing the unused tokens reserved in the original vocabulary. Specifically, we train a tokeniser with a vocabulary size of 5 000 tokens on the data and choose the set of 99 tokens which reduce the number of [UNK] tokens the most in the target data. Training was performed on 32 A100 GPUs, and took around 46 hours to complete.

4   Evaluation

An evaluation for the language models described in Section 3 is presented here. **mBERT** without any additional pre-training is used as one of the baselines. In addition, we pre-train two language models on the Maltese Wikipedia data as additional baselines. This allows us to analyse the limitations that could be faced when following the common practice of using Wikipedia data, for the specific case of low-resource languages with a comparatively small Wikipedia footprint.

Following the same setup of the main models, a monolingual model (**BERTu Wiki**) and a multilingual model (**mBERTu Wiki**) are pre-trained. The same hyper-parameters described in Section 3 are used, but the batch size and number of steps are decreased to prevent overfitting due to the smaller data size. To this end, the batch size is set to 64 and the total number of steps set to 30 500 and 7 600 steps for the monolingual and multilingual models, respectively. This was deemed appropriate since it would amount to the same number of epochs as the models pre-trained on the entire corpus.

4.1   Tasks

The language models are fine-tuned on the following downstream tasks. A summary of the datasets and fine-tuning architectures used is given below.

Dependency Parsing (DP) The Maltese Universal Dependencies Treebank (MUDT) (Čéplö, 2018) is used for this task using the provided training, validation, and testing splits. The data is composed of 2 074 human-annotated sentences from 4 different high-level domains. Similar to Chau et al. (2020), Muller et al. (2021), and Chau and Smith (2021), we use a Biaffine graph-based prediction layer (Dozat and Manning, 2017) and use the Labelled Attachment Score (LAS) as the main evaluation metric, but also report the Unlabelled Attachment Score (UAS).

Part-of-Speech Tagging (POS) The MLRS POS data (Gatt and Čéplö, 2013), is used for this task. This data is composed of 6 167 human-annotated sentences – 426 of which overlap with the MUDT data (Čéplö, 2018) – and are stratified into 8 domains. We combine the data from the different domains, shuffle it, and split the data into 80%, 10%, and 10% for training, validation, and testing sets, respectively. The annotations are language-specific tags (using the XPOS scheme) and we follow the tag mapping in MUDT (Čéplö, 2018) to also produce tags in the Universal Part of Speech tagset (UPOS). To evaluate tagging with these two tagsets, we use a linear layer, and use accuracy as the evaluation metric.

Named-Entity Recognition (NER) The Maltese annotations for the WikiAnn data (Pan et al., 2017) are used for this task, using the data splits from Rahimi et al. (2019). The data is made up of 300 sentences derived from Wikipedia. Following Chau and Smith (2021), a Conditional Random Field layer is used for this task, and we use F1 as the evaluation metric.

Sentiment Analysis (SA) We use the Maltese sentiment analysis dataset by Martínez-García et al. (2021), which is a collection of 815 sentences, using the provided training, validation, and testing splits. The texts in this data originate from comments on news articles and social media posts, and are a combination of two datasets from Cortis and Davis (2019) and Dingli and Sant (2016). A linear prediction layer is used, and we use the macro-averaged F1 score as the evaluation metric.

We largely use the hyper-parameters from Chau and Smith (2021), but optimise the learning rate, batch size, and dropout on the validation set of each task. Table 2 shows the chosen hyper-parameters. Fine-tuning is performed for at most 200 epochs,

Name	DP	POS	NER	SA
Learning Rate	5e-4	5e-4	5e-4	1e-4
Batch Size	128	128	64	32
Dropout	0.3	0.3	0.2	0.5

Table 2: Fine-tuning hyper-parameters

with an early stopping of 20 epochs on the validation set.

4.2 Results

The results on all tasks are summarised in Table 3. Consistent with the results reported by Muller et al. (2021) and Chau and Smith (2021), BERTu Wiki generally underperforms mBERT, and mBERTu Wiki performs better than mBERT. Whilst they show this for Dependency Parsing, Part-of-Speech tagging, and Named Entity Recognition, we demonstrate that this also holds for Sentiment Analysis.

Our baseline results diverge slightly from previous results on the Named-Entity Recognition task, where BERTu Wiki performs slightly better than mBERT. We suspect that this is due to a slightly different pre-training setup than that used by Muller et al. (2021) and Chau and Smith (2021)[3], but we analyse this further in Section 5.1. However, these results remain consistent with regards to BERTu Wiki not performing as well as mBERTu Wiki.

Both language models pre-trained with the KM data perform significantly better than all the other baselines, on all tasks except for Named-Entity Recognition, where the trend is similar, but does not reach statistical significance. This underlines the value of this new corpus. When compared to the Wikipedia language models, the most noticeable improvements can be seen between the BERTu models, across all tasks. A more detailed analysis on this is presented in Section 5.2, but intuitively this finding makes sense since mBERTu models are exposed to significantly more data, making them less specific to Maltese.

The gap in performance between the BERTu and mBERTu models is much less for the KM pre-trained models than it is for the Wikipedia pre-trained models. In fact, on average, the BERTu KM model performs better than the mBERTu KM model on all tasks except Part-of-Speech tagging.

---

[3]Muller et al. (2021) pre-train for at most 10 epochs whilst Chau and Smith (2021) pre-train for at most 20 epochs (choosing the best performing model on based on the validation set). Both use a smaller-sized BERT architecture with 6 layers and pre-train with a maximum sequence length of 128.

Data	Model	UAS	LAS
Wiki	BERTu	80.95 ± 0.25	74.16 ± 0.20
	mBERTu	88.74 ± 0.11	82.59 ± 0.19
N/A	mBERT	84.83 ± 0.31	77.22 ± 0.34
KM	BERTu	**92.31 ± 0.15**	**88.14 ± 0.21**
	mBERTu	*92.10 ± 0.14	*87.87 ± 0.18

(a) Dependency Parsing

Data	Model	UPOS	XPOS
Wiki	BERTu	97.27 ± 0.11	97.01 ± 0.07
	mBERTu	97.95 ± 0.13	97.83 ± 0.08
N/A	mBERT	97.26 ± 0.15	97.20 ± 0.14
KM	BERTu	98.58 ± 0.02	*98.54 ± 0.03
	mBERTu	**98.66 ± 0.03**	**98.58 ± 0.04**

(c) Part-of-Speech tagging

Data	Model	span F1
Wiki	BERTu	67.96 ± 2.20
	mBERTu	*85.01 ± 2.92
N/A	mBERT	65.41 ± 2.06
KM	BERTu	**86.77 ± 3.55**
	mBERTu	*86.60 ± 2.49

(b) Named Entity Recognition

Data	Model	macro-F1
Wiki	BERTu	53.95 ± 2.70
	mBERTu	56.05 ± 3.24
N/A	mBERT	55.99 ± 3.63
KM	BERTu	**78.96 ± 1.95**
	mBERTu	*76.79 ± 1.79

(d) Sentiment Analysis

Table 3: Experimental results, grouped by the underlying language model and additional pre-training data used. All figures shown are the mean and standard deviations over 5 runs with different random seeds. The best performing models for each metric are **bolded**. Values marked with * are not found to be significantly worse than the best model (using a 1-tailed t-test with a p-value = 0.05 with Bonferroni correction).

For Part-of-Speech tagging we also note that the baseline results are already quite high, probably due to the relatively larger labelled data, which may partially mask the effects of the KM models.

Overall, Sentiment Analysis is the task where the most gains are made with respect to the baselines. The KM-trained models are over 20 F1 points higher than the best performing baseline. This finding provides evidence that, unlike syntactic tasks, where structural information could potentially be shared across related clusters of languages, semantic tasks such as Sentiment Analysis will benefit much more from language-specific embeddings.

5 Analysis

In this section we build on the results presented in Section 4.2, and analyse the effect that the pre-training data has on performance on the downstream tasks.

5.1 Data Domain

In this subsection we take advantage of the fact that the data is stratified by domains. Here, we analyse the impact of pre-training using data from different domains, compared to a single domain, namely Wikipedia, which is commonly used in multilingual models and low-resource settings. For this purpose, we consider the BERTu Wiki and mBERTu Wiki baselines from Section 4 as single-domain models. We compare them to language models pre-trained with the same amount of data

but from different domains, which are referred to as "Mixed" in this discussion.

We determine the size using the number of sentences, since this would directly effect the number of epochs and allows us to keep an identical pre-training setup as the Wikipedia-trained models. Since the Maltese Wikipedia data is composed of 79 134 sentences, the Mixed language models are also pre-trained with the same amount of sentences, split into training and validation sets as the Wikipedia models. A comparison of the downstream task performance for these language models is plotted in Figure 1.

At this small scale of data, both Wiki and Mixed mBERTu models consistently perform better than the mBERT models, owing to the multilingual representation power of these models. The Mixed models perform better than their Wikipedia counterparts on Part-of-Speech tagging and Sentiment Analysis. On Dependency Parsing, there is a slight improvement on the mBERTu model but a slight degradation on the BERTu model.

For Named-Entity Recognition, the Wikipedia models perform better than the Mixed ones. Since the dataset for this task originates from Wikipedia as well, this indicates that matching the pre-training data domain to the target domain boosts performance, supporting the findings by Gururangan et al. (2020). In fact, mBERTu Wiki surpasses mBERT, and proves to be a competitive baseline, as shown in Table 3b. On the other hand, mBERTu Mixed
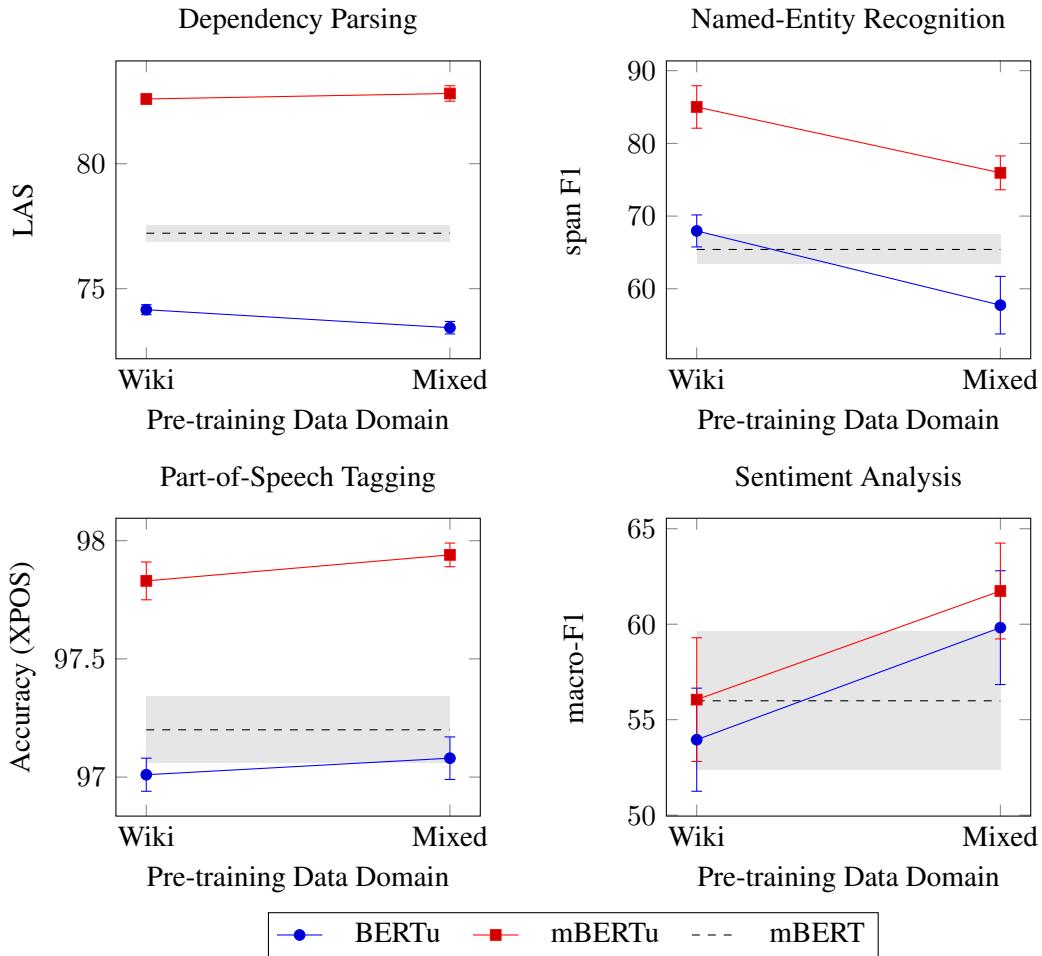
Figure 1: Downstream task performance with different pre-training data domains. All values are the mean over 5 runs with different random seeds. The standard deviation is represented by the corresponding error bars and shaded area.

performs worse than mBERT.

The opposite is true for Sentiment Analysis, as BERTu Mixed turns out to be a more competitive baseline than mBERT. The improvement is so pronounced for this task that the BERTu Mixed model not only performs better than the BERTu Wiki counterpart, but also better than mBERTu Wiki. Even though this dataset contains texts exhibiting stylistic features expected in social media text, a mixture of domains is helpful, probably since Wikipedia texts tend to be quite structured and neutral in terms of the writing style and tone. The results on sentiment analysis suggest that pre-training on a diversity of domains contribute to more effective learning of features relevant to discourse semantic tasks, compared to tasks involving morpho-syntactic tagging. We leave further investigation of this, on a broader range of semantically-oriented tasks, for future work.

Overall, these results emphasise the importance

of having pre-training data from sources close to the target data, even for low-resource settings.

5.2 Data Size

From Table 3, it is clear that the KM corpus translates to better performance on downstream tasks, regardless of whether a monolingual or a multilingual model is used. To better understand the relationship between the data size and performance, we pre-train several language models with varying data sizes.

We do this by fixing the desired data proportion and scaling the pre-training data to satisfy this proportion, keeping the original training and validation split. In tandem, the original 1M and 250K steps and batch size used in Section 3 are scaled down with the data size to pre-train for the same number of epochs as the models with the entire data. Language models at 10% intervals are pre-trained, with 100% being the original models from
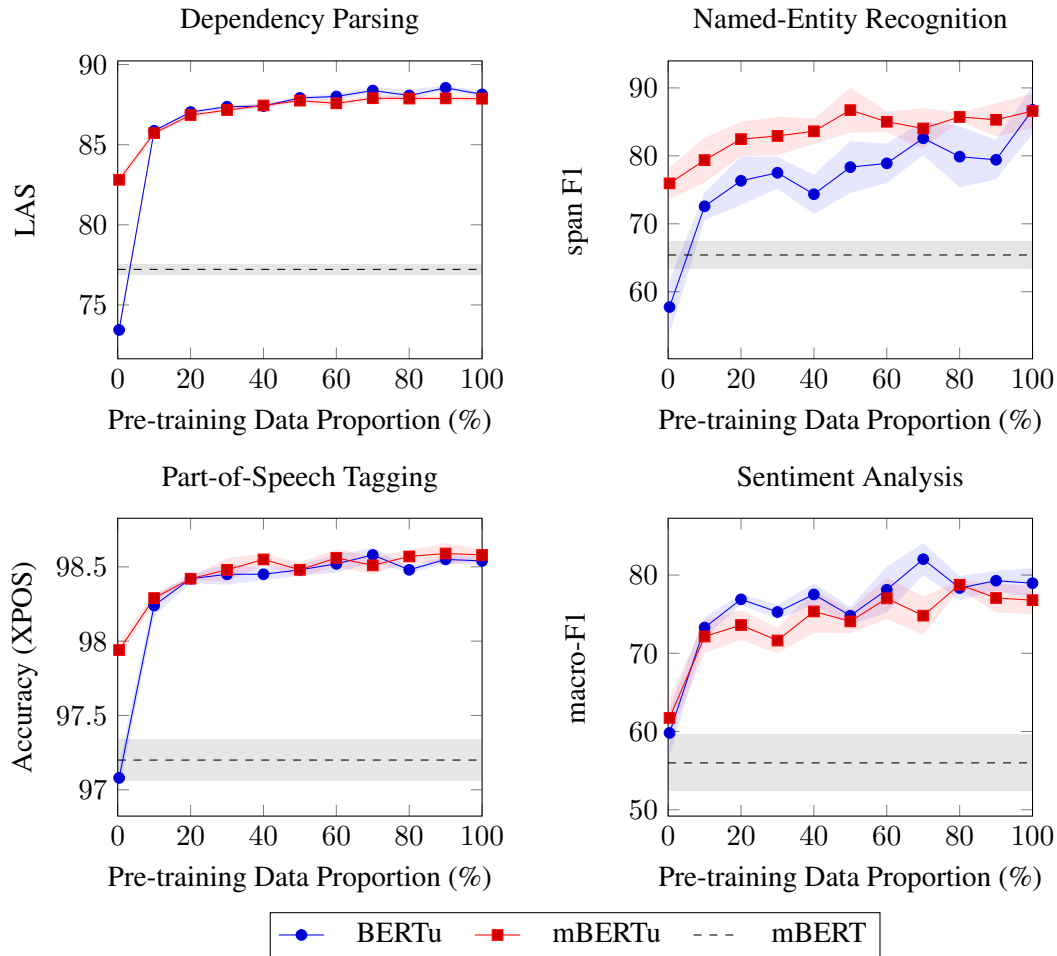
Figure 2: Downstream task performance as the pre-training data size grows. All values are the mean over 5 runs with different random seeds. The standard deviation being represented by the corresponding shaded area.

Section 3. In this analysis, we also include the BERTu Mixed and mBERTu Mixed models from Section 5.1, which use 0.38% of the data, estimated as a proportion of sentences.

After pre-training, each language model is fine-tuned on each of the downstream tasks in the same setup considered in Section 4. These results are visualised in Figure 2.

As expected, the performance generally improves with more pre-training data. Surprisingly, the performance gap between the monolingual and multilingual models is drastically reduced with just 10% of the data. With this little data all configurations outperform mBERT. For Named-Entity Recognition this is also the case but it takes around 70% of the data for BERTu and mBERTu to start achieving very close performance.

It is also noticeable that the gradual increase is not monotonic, although it is more stable for Dependency Parsing and Part-of-Speech tagging. Suprisingly, BERTu with 70% of the data performs

better than with 100% of the data on Sentiment Analysis. Similarly, mBERTu with 50% of the data performs better than with 100% of the data on Named-Entity Recognition. One possible explanation may be due to the relationship between the number of steps and batch size chosen, but further investigation is warranted.

On Sentiment Analysis, BERTu is consistently better than mBERTu with 10% or more of the data, and is at times significantly better. This finding gives some evidence that monolingual representations seem better suited for fine-tuning on semantic tasks in a specific language.

6 Conclusion

In this work we analyse the impact of pre-training data on downstream task performance in a low-resource setting, specifically focusing on Maltese. We present a newly developed corpus of around 500M tokens, which allows us to study how the pre-

training data size and domain translates in downstream performance differences. Using BERT as our architecture, we compare a monolingual language model, pre-trained from scratch, to a further pre-trained multilingual model, in a number of pre-training configurations. We conduct an evaluation on a both syntactic and semantic tasks.

In line with previous findings on domain pre-training (Gururangan et al., 2020; inter alia), we find that matching the pre-training domain to the target task domain, results in improvements. Moreover, we demonstrate that pre-training language models with varied domains is often beneficial over pre-training solely with Wikipedia. These adjustments were in certain cases enough to surpass mBERT, underlining the importance of having pre-training data more suited to the target task, even at a small scale.

Whilst we show that further pre-training data does improve downstream performance, the gains are linear with exponential increases in data. In fact, substantial improvements are observed with a small proportion of the pre-training data, over language models trained with Wikipedia-sized data. This echoes the findings made by Martin et al. (2020) with a small pre-training subset, although our reduced data setup is considerably smaller.

Using the whole corpus, we also pre-train two new language models: BERTu, a monolingual BERT model, trained from scratch, and mBERTu, which is the result of further pre-training mBERT. These models demonstrate state-of-the-art results in Dependency Parsing, Part-of-Speech Tagging, Named-Entity Recognition, and Sentiment Analysis. Moreover, we show that in general, BERTu performs better than mBERTu, as well as other baselines. Through this, we also demonstrate that language-specific pre-training is most beneficial for higher-level tasks.

Despite these considerable improvements, the pre-training setups used in this work are as close as possible to the baselines, to allow for a more controlled comparison. Hence, in the future, we plan to experiment with more language-specific tuning to push the state-of-the-art even further.

Even though this new corpus will undoubtedly improve the state of resources available for Maltese, the language is by no means a highly-resourced one. The corpus we use is significantly smaller than typically used corpora for higher-resourced languages. We also remark that the quantity of labelled data

is still scarce, and at times non-existent for certain tasks. Although we include a semantically-oriented task in our evaluation, future work should investigate the efficacy of these models in more complex Natural Language Understanding scenarios.

We make this corpus and the models publicly available to foster further work and improvements for various NLP applications for Maltese. We also hope that this work inspires work in other low-resource languages, since we show that the amount of data needed to achieve considerable improvements, does not need to be overly ambitious.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the fourth ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*, Online. Association for Computing Machinery.

Joseph Brincat. 2011. *Maltese and other languages: A linguistic history of Malta*. Midsea Books, Malta.

Slavomír Čéplö. 2018. *Constituent order in Maltese: A quantitative analysis*. Ph.D. thesis, Charles University, Prague.

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

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

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

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

Alexei Dingli and Nicole Sant. 2016. Sentiment analysis on Maltese using machine learning. In *The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Albert Gatt and Slavomír Čéplö. 2013. Digital Corpora and Other Electronic Resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*, pages 96–97. UCREL, Lancaster, UK.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian bert language understanding model for NLP challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Mike Rosner and Claudia Borg. 2022. *Report on the Maltese Language*. Language Technology Support of Europe's Languages in 2020/2021. Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne (Series Editors). Available online at https://european-language-equality.eu/deliverables/.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Llion Uszkoreit, Jakob an Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

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