mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model

Rasmus Kær Jørgensen[1,2] and **Mareike Hartmann**[3*] and **Xiang Dai**[1] and **Desmond Elliott**[1]
[1]University of Copenhagen, Denmark
[2]PricewaterhouseCoopers (PwC), Denmark
[3]German Research Center for Artificial Intelligence (DFKI), Germany
rasmuskj,xiang.dai,de@di.ku.dk
mareike.hartmann@dfki.de

Abstract

Domain adaptive pretraining, i.e. the continued unsupervised pretraining of a language model on domain-specific text, improves the modelling of text for downstream tasks within the domain. Numerous real-world applications are based on domain-specific text, e.g. working with financial or biomedical documents, and these applications often need to support multiple languages. However, large-scale domain-specific multilingual pretraining data for such scenarios can be difficult to obtain, due to regulations, legislation, or simply a lack of language- and domain-specific text. One solution is to train a single multilingual model, taking advantage of the data available in as many languages as possible. In this work, we explore the benefits of domain adaptive pretraining with a focus on adapting to multiple languages within a specific domain. We propose different techniques to compose pretraining corpora that enable a language model to both become domain-specific and multilingual. Evaluation on nine domain-specific datasets—for biomedical named entity recognition and financial sentence classification—covering seven different languages show that a single multilingual domain-specific model can outperform the general multilingual model, and performs close to its monolingual counterpart. This finding holds across two different pretraining methods, adapter-based pretraining and full model pretraining.

1 Introduction

The unsupervised pretraining of language models on unlabelled text has proven useful to many natural language processing tasks. The success of this approach is a combination of deep neural networks (Vaswani et al., 2017), the masked language modeling objective (Devlin et al., 2019), and large-scale corpora (Zhu et al., 2015). In fact, unlabelled data is so important that better downstream task performance can be realized by pretraining models on more unique tokens, without repeating any examples, instead of iterating over smaller datasets (Raffel et al., 2020). When it is not possible to find vast amounts of unlabelled text, a better option is to continue pretraining a model on domain-specific unlabelled text (Han and Eisenstein, 2019; Dai et al., 2020), referred to as domain adaptive pretraining (Gururangan et al., 2020). This results in a better initialization for consequent fine-tuning for a downstream task in the specific domain, either on target domain data directly (Gururangan et al., 2020), or if unavailable on source domain data (Han and Eisenstein, 2019).

The majority of domain-adapted models are trained on English domain-specific text, given the availability of English language data. However, many real-world applications, such as working with financial documents (Araci, 2019), biomedical text (Lee et al., 2019), and legal opinions and rulings (Chalkidis et al., 2020), should be expected to work in multiple languages. For such applications, annotated target task datasets might be available, but we lack a good pretrained model that we can fine-tune on these datasets. In this paper, we propose a method for domain adaptive pretraining of a single domain-specific multilingual language model that can be fine-tuned for tasks within that domain in multiple languages. There are several reasons for wanting to train a single model: (i) Data availability: we cannot always find domain-specific text in multiple languages so we should exploit the available resources for effective transfer learning (Zhang et al., 2020). (ii) Compute intensity: it is environmentally unfriendly to domain-adaptive pretrain one model per language (Strubell et al., 2019), and BioBERT was domain adaptive pretrained for 23 days on 8×Nvidia V100 GPUs. (iii) Ease of use: a single multilingual model eases deployment when an organization needs to work with multiple

[*] The research was carried out while the author was employed at the University of Copenhagen.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3404–3418
November 7–11, 2021. ©2021 Association for Computational Linguistics

languages on a regular basis (Johnson et al., 2017).

Our method, multilingual domain adaptive pretraining (MDAPT), extends domain adaptive pretraining to a multilingual scenario, with the goal of training a single multilingual model that performs, as close as possible, to N language-specific models. MDAPT starts with a base model, i.e. a pretrained multilingual language model, such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020). As monolingual models have the advantage of language-specificity over multilingual models (Rust et al., 2020; Rönnqvist et al., 2019), we consider monolingual models as upper baseline to our approach. We assume the availability of English-language domain-specific unlabelled text, and, where possible, multilingual domain-specific text. However, given that multilingual domain-specific text can be a limited resource, we look to Wikipedia for general-domain multilingual text (Conneau and Lample, 2019). The base model is domain adaptive pretrained on the combination of the domain-specific text, and general-domain multilingual text. Combining these data sources should prevent the base model from forgetting how to represent multiple languages while it adapts to the target domain.

Experiments in the domains of financial text and biomedical text, across seven languages: French, German, Spanish, Romanian, Portuguese, Danish, and English, and on two downstream tasks: named entity recognition, and sentence classification, show the effectiveness of multilingual domain adaptive pretraining. Further analysis in a cross-lingual biomedical sentence retrieval task indicates that MDAPT enables models to learn better domain-specific representations, and that these representations transfer across languages. Finally, we show that the difference in tokenizer quality between mono- and multilingual models is more pronounced in domain-specific text, indicating a direction for future improvement.

All models trained with MDAPT and the new datasets used in downstream tasks and pretraining data[1] and our code is made available[2].

2 Problem Formulation

Pretrained language models are trained from random initialization on a large corpus $\mathcal{C}$ of unlabelled sentences. Each sentence is used to optimize the parameters of the model using a pretraining objective, for example, masked language modelling, where, for a given sentence, 15% of the tokens are masked in the input m, and the model is trained to predict those tokens $J(\theta) = -\log p_\theta(x_m \mid \mathbf{x}_{\setminus m})$ (Devlin et al., 2019). $\mathcal{C}$ is usually a corpus of no specific domain,[3] e.g. Wikipedia or crawled web text.

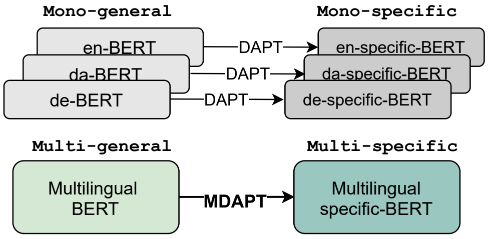

Figure 1: MDAPT extends domain adaptive pretraining to a multilingual scenario.

Domain-adaptive pretraining is the process of continuing to pretrain a language model to suit a specific domain (Gururangan et al., 2020; Han and Eisenstein, 2019). This process also uses the masked language modelling pretraining objective, but the model is trained using a domain-specific corpus $\mathcal{S}$, e.g. biomedical text if the model should be suited to the biomedical domain. Our goal is to pretrain a *single model*, which will be used for downstream tasks in multiple languages within a specific domain, as opposed to having a separate model for each language. This single multilingual domain-specific model should, ideally, perform as well as language-specific domain-specific models in a domain-specific downstream task.

In pursuit of this goal, we use different types of corpora for domain adaptive pretraining of a single multilingual model. Each considered corpus has two properties: (1) a domain property – it is a `general` or `specific` corpus; and (2) a language property – it is either `mono`linugal or `multi`lingual. These properties can be combined, for example the multilingual Wikipedia is a `multi-general` corpus, while the abstracts of English biomedical publications would be a `mono-specific` corpus. Recall that `specific` cor-

[1] https://github.com/RasmusKaer/mDAPT_supplements

[2] https://github.com/mahartmann/mdapt

[3] Text varies along different dimensions, e.g. topic or genre (Ramponi and Plank, 2020). In the context of this paper, we focus on *domain-specificity* along the topic dimension , i.e. texts are considered as *domain-specific* if they talk about a narrow set of related concepts. The domain-specific text can comprise different genres of text (e.g. financial news articles and financial tweets would both be considered as being from the financial domain).

pora are not always available in languages other than English, but they are useful for adapting to the intended domain; while `multi-general` are more readily available, and should help maintain the multilingual abilities of the adapted language model. In the remainder of this paper, we will explore the benefits of domain adaptive pretraining with `mono-specific`, `multi-specific`, *and* `multi-general` corpora. Figure 1 shows how MDAPT extends domain adaptive pretraining to a multilingual scenario.

3 Multilingual Domain Adaptive Pretraining

Recall that we assume the availability of large scale English domain-specific and multilingual general unlabelled text. In addition to these `mono-specific` and `multi-general` corpora, we collect multilingual domain-specific corpora, using two specific domains—financial and biomedical—as an example (Section 3.1). Note that although we aim to collect domain-specific data in as many languages as possible, the collected data are usually still relatively small. We thus explore different strategies to combine different data sources (Section 3.2), resulting in three different types of pretraining corpora of around 10 million sentences, that exhibit `specific` and `multi` properties to different extents: $\mathbf{E_D}$: English domain-specific data; $\mathbf{M_D}$+$\mathbf{E_D}$: Multilingual domain-specific data, augmented with English domain-specific data; and $\mathbf{M_D}$+$\mathbf{M_{WIKI}}$: Multilingual domain-specific data, augmented with multilingual general data.

We use mBERT (Devlin et al., 2019) as the multilingual base model, and employ two different continued pretraining methods (Section 3.3): adapter-based training and full model training, on these three pretraining corpora, respectively.

3.1 Domain-specific corpus

Financial domain As `specific` data for the financial domain, we use Reuters Corpora (RCV1, RCV2, TRC2),[4] SEC filings (Desola et al., 2019),[5] and FinMultiCorpus, which is an in-house collected corpus. The FinMultiCorpus consists of articles in multiple languages published on PwC website. The resulting corpus contains the following languages: *zh*, *da*, *nl*, *fr*, *de*, *it*, *ja*, *no*, *pt*, *ru*, *es*, *sv*, *en*, *tr*. Statistics on the presented languages can be found in Table 9 in the Appendix. Information about preprocessing are detailed in Appendix C.

Domain	Data	# Lang.	# Sent.	# Tokens
Fin	$\mathrm{M_D}$	14	4.9M	34.4M
	$\mathrm{E_D}$	1	10.0M	332.8M
	$\mathrm{M_{WIKI}}$	14	5.1M	199.9M
Bio	$\mathrm{M_D}$	8	3.2M	86.6M
	$\mathrm{E_D}$	1	10.0M	370.6M
	$\mathrm{M_{WIKI}}$	8	6.8M	214.2M

Table 1: A summary of pretraining data used. We use two specific domains—financial (top part) and biomedical (bottom part) as an example in this paper. M stands for Multilingual; E for English; D for Domain-specific; and, Wiki refers to general data, sampled from Wikipedia. The number of tokens are calculated using mBERT cased tokenizer. Note that because languages considered in financial and biomedical domains are not the same , we sample two different $\mathrm{M_{WIKI}}$ covering different languages.

Biomedical domain As `specific` data for the biomedical domain, we use biomedical publications from the PubMed database, in the following languages: *fr*, *en*, *de*, *it*, *es*, *ro*, *ru*, *pt*. For languages other than English, we use the language-specific PubMed abstracts published as training data by WMT, and additionally retrieve all language specific paper titles from the database.[6] For English, we only sample abstracts. We make sure that no translations of documents are included in the pretraining data. The final statistics on biomedical pretraining data can be found in Table 8 in the Appendix, as well as more details about preprocessing the documents. The descriptive statistics of these pretraining data can be found in Table 1.

3.2 Combination of data sources

Recall that `multi-specific` data is usually difficult to obtain, and we explore different strategies to account for this lack. The different compositions of pretraining data are illustrated in Figure 2. We control the size of the resulting corpora by setting a budget of 10 million sentences. This allows a fair comparison across data settings.

With plenty of English `specific` text available, $\mathrm{E_D}$ and $\mathrm{M_D}$+$\mathrm{E_D}$ are composed by simply populating the corpus until reaching the allowance.

[4] Available by request at `https://trec.nist.gov/data/reuters/reuters.html`

[5] `http://people.ischool.berkeley.edu/~khanna/fin10-K`

[6] We use data from a bulk download of `ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline`, version 12/14/20

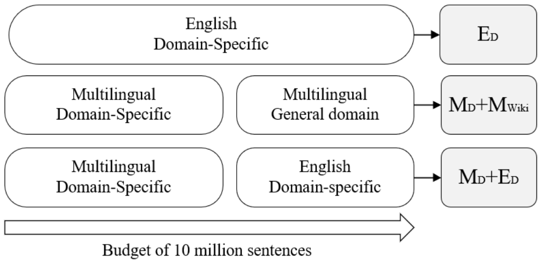
English Domain-Specific
E_{D}
Multilingual Domain-Specific
Multilingual General domain
M_{D}+M_{Wiki}
Multilingual Domain-Specific
English Domain-specific
M_{D}+E_{D}
Budget of 10 million sentences

Figure 2: Composition of pretraining data.

As a resource for `multi-general` data, we use Wikipedia page content, where we ensure the same page is not sampled twice across languages. Up-sampling M_{D}+M_{WIKI} using general domain multilingual data requires a sampling strategy that accounts for individual sizes. Sampling low-resource languages too often may lead to overfitting the repeated contents, whereas sampling high-resource language too much can lead to a model underfit. We balance the language samples using exponentially smoothed weighting (Xue et al., 2020; Conneau and Lample, 2019; Devlin et al., 2019). Following Xue et al., we use a α of 0.3 to smooth the probability of sampling a language, $P(L)$, by $P(L)^{\alpha}$. After exponentiating each probability by α, we normalize and populate the pretraining corpus with Wikipedia sentences according to smoothed values until reaching our budget. Except for English, we up-sample using Wikipedia data. The statistics of the extracted sentences is presented in tables 8 and 9 in the Appendix.

3.3 Pretraining methods

Continue pretraining the whole model We initialize our models with pretrained *base* model weights[7] and then continue pretraining the whole *base* model via the masked language modeling objective. We follow Devlin et al. (2019) in randomly masking out 80% of subtokens and randomly replacing 10% of subtokens. For all models, we use an effective batch size of 2048 via gradient accumulation, a sequence length of 128, and a learning rate of 5e-5. We train all models for 25,000 steps, which takes 10 GPU days.

Adapter-based training In contrast to fine-tuning all weights of the *base* model, adapter-based training introduces a small network between each layer in the *base* model, while keeping the *base* model fixed. The resulting adapter weights, which can be optimized using self-supervised pretraining or later downstream supervised objectives, are usually much lighter than the *base* model, enabling parameter efficient transfer learning (Houlsby et al., 2019). We train each adapter for 1.5M steps, taking only 2 GPU days. We refer readers to Pfeiffer et al. (2020b) for more details of adapter-based training and also describe them in the Appendix D for self-containedness.

		NCBI	PHAR	QUAERO	CLIN	BIORO
		en	*es*	*fr*	*pt*	*ro*
# sents.	train	5,424	8,137	1,540	1,192	1,886
	dev	923	3,801	1,481	336	631
	test	940	3,982	1,413	973	629
# mentions	train	5,134	3,810	4,516	7,600	5,180
	dev	787	1,926	4,123	2,047	1,864
	test	960	1,876	4,086	6,315	1,768
# classes		1	4	10	13	4

Table 2: The descriptive statistics of the biomedical NER datasets.

4 Domain-Specific Downstream Tasks

To demonstrate the effectiveness of our multilingual domain-specific models, we conduct experiments on two downstream tasks—Named Entity Recognition (NER) and sentence classification—using datasets from biomedical and financial domains, respectively.

4.1 NER in the biomedical domain

Datasets We evaluate on 5 biomedical NER datasets in different languages. The French QUAERO (Névéol et al., 2014) dataset, the Romanian BIORO dataset (Mitrofan, 2017), and the English NCBI DISEASE dataset (Doğan et al., 2014) comprise biomedical publications. The Spanish PHARMACONER (Agirre et al., 2019) dataset comprises publicly available clinical case studies, and the Portuguese CLINPT dataset is the publicly available subset of the data collected by Lopes et al. (2019), comprising texts about neurology from a clinical journal. The descriptive statistics of the NER datasets are listed in Table 2, and more details about the datasets can be found in Appendix B. We convert all NER annotations to BIO annotation format, and use official train/dev/test splits if available. For NCBI DISEASE, we use the data preprocessed by Lee et al. (2019). Further preprocessing details can be found in Appendix B.

[7] MBERT: https://huggingface.co/bert-base-multilingual-cased

	OMP *de*	FINNEWS *da*	PHR.BANK *en*
# sentences	10,276	5,134	4,845
# classes	2/9	3	3

Table 3: The descriptive statistics of the financial classification datasets. We frame the German dataset as a binary and a multi-class (9) classification tasks.

NER Model Following Devlin et al. (2019), we build a linear classifier on top of the BERT encoder outputs, i.e. the contextualized representations of the first sub-token within each token are taken as input to a token-level classifier to predict the token's tag. For full model fine-tuning, we train all models for a maximum of 100 epochs, stopping training early if no improvement on the development set is observed within 25 epochs. We optimize using AdamW, a batch size of 32, maximum sequence length of 128, and a learning rate of 2e-5. For adapter-based training, we train for 30 epochs using a learning rate of 1e-4.

4.2 Sentence classification in the financial domain

Datasets We use three financial classification datasets, including the publicly available English FINANCIAL PHRASEBANK (Malo et al., 2014), German ONE MILLION POSTS (Schabus et al., 2017), and a new Danish FINNEWS. The FINANCIAL PHRASEBANK is an English sentiment analysis dataset where sentences extracted from financial news and company press releases are annotated with three labels (Positive, Negative, and Neutral). Following its annotation guideline, we create FINNEWS—a dataset of Danish financial news headlines annotated with a sentiment. 2 annotators were screened to ensure sufficient domain and language background. The resulting dataset has a high inter-rater reliability (a measure of 82.1% percent agreement for raters and a Krippendorff's alpha of .725, measured on 800 randomly sampled examples). ONE MILLION POSTS is sourced from an Austrian newspaper. We use TITLE and TOPIC for two classification settings on this dataset: a binary classification, determining whether a TITLE concerns a financial TOPIC or not; and a multi-class classification that classify a TITLE into one of 9 TOPICS. We list the descriptive statistics in Table 3, and further details can be found in Appendix C.

Classifier Following Devlin et al. (2019), we built a classification layer on top of the [CLS] token. We perform simple hyperparameter tuning with the baseline monolingual model on each dataset separately. The parameter setting is selected on a coarse grid of batch-sizes $[16, 32]$ and epochs $[2, 4, 6]$. The best-performing hyperparameters on each dataset are then used in experiments using other pretrained models. All experiments follow an 80/20 split for train and testing with an equivalent split for model selection.

5 Results

To measure the effectiveness of multilingual domain adaptive pretraining, we compare the effectiveness of our models trained with MDAPT on downstream NER and classification, to the respective monolingual baselines (`mono-general`), and to the base multilingual model without MDAPT (Table 4). Where available, we also compare to the respective monolingual domain-specific models (`mono-specific`).

Baseline models As `mono-general` baselines, we use English BERT (Devlin et al., 2019), Portuguese BERT (Souza et al., 2020), Romanian BERT (Dumitrescu et al., 2020), BETO (Cañete et al., 2020) for Spanish, FlauBert (Le et al., 2020) for French, German BERT (Chan et al., 2020), and Danish BERT.[8] `Mono-specific` baselines exist only for a few languages and domains, we use EN-BIO-BERT (Lee et al., 2019) as English biomedical baseline, and EN-FIN-BERT (Araci, 2019) as English financial baseline. To the best of our knowledge, PT-BIO-BERT (Schneider et al., 2020) is the only biomedical model for non-English language, we use it as Portuguese biomedical baseline, see Appendix A for more details.

5.1 Main results

The main results for the biomedical NER and financial sentence classification tasks are presented in Table 4. We report the evaluation results for the `mono`-BERT baselines in the respective languages and the performance difference of the multilingual models compared to these monolingual baselines. We also consider two domain adaptive pretraining approaches: full model training, reported in the upper half of the table, and adapter-based training in the lower half.

[8] `https://github.com/botxo/nordic_bert`

	Biomedical NER					Financial Sentence Classification			
	QUAERO	BIORO	PHAR	NCBI	CLINPT	OMP-2	OMP-9	FINNEWS	PHR.BANK
	fr	*ro*	*es*	*en*	*pt*	*de*	*de*	*da*	*en*
	Full model pretraining								
mono-specific-BERT	-	-	-	88.1	72.9	-	-	-	87.3
mono-BERT	**61.9**	**75.5**	88.2	85.1	72.6	**91.4**	71.5	**65.2**	**85.0**
mBERT	-3.7	-1.6	+0.2	+1.0	-0.2	-0.6	-0.4	-2.4	-2.6
+ E_{D}	-3.6	-1.6	**+0.6**	+1.5	-0.6	-0.3	0	-2.5	-1.2
+ M_{D}+E_{D}	-2.7	-0.9	+0.5	**+2.1**	**+0.1**	-0.2	**+0.1**	-1.6	-1.1
+ M_{D}+$\mathrm{M}_{\mathrm{WIKI}}$	-2.1	-1.4	+0.3	+1.8	0.0	-0.1	**+0.1**	-1.6	-1.4
	Adapter-based pretraining								
mono-BERT	**58.6**	**73.2**	86.6	82.6	63.5	90.5	69.1	**66.0**	**85.3**
mBERT	-4.5	-4.5	-0.3	+0.1	-3.7	0.0	+0.8	-3.1	-3.1
+ E_{D}	-2.9	-2.0	+1.5	+1.4	+1.8	+0.7	+1.5	-4.9	-3.5
+ M_{D}+E_{D}	-1.3	-1.9	**+1.9**	+1.4	**+2.7**	**+0.9**	**+3.8**	-1.7	-2.6
+ M_{D}+$\mathrm{M}_{\mathrm{WIKI}}$	-1.4	-2.6	+1.0	**+1.8**	+1.6	+0.6	+2.6	-1.9	-3.2

Table 4: Evaluation results on biomedical NER and financial sentence classification tasks. We report the results—span-level micro F_1 for NER and sentence-level micro F_1 for classification—on the monolingual BERTs. Performance differences compared to the monolingual baselines are reported for multilingual BERTs, with and without mDAPT. All experiments are repeated five times using different random seeds, and mean values are reported.

Our work is motivated by the finding that domain adaptive pretraining enables models to better solve domain-specific tasks in monolingual scenarios. The first row in Table 4 shows our re-evaluation of the performance of the three available domain adaptive pretrained mono-specific-BERT models matching the domains investigated in our study. We confirm the findings of the original works, that the domain-specific models outperform their general domain mono-BERT counterparts. This underlines the importance of domain adaptation in order to best solve domain-specific task. The improvements of PT-BIO-BERT over PT-BERT are small, which coincides with the findings of Schneider et al. (2020), and might be due to the fact that the CLINPT dataset comprises clinical entities rather than more general biomedical entities.

Full model training Recall that the aim of mDAPT is to train a single multi-specific model that performs comparable to the respective mono-general model. Using full model pretraining, we observe that the domain adaptive pretrained multilingual models can even outperform the monolingual baselines for *es* and *en* biomedical NER, and *de* for financial sentence classification. On the other hand, we observe losses of the multilingual models over the monolingual baselines for *fr* and *ro* NER, and *da* and *en* sentence classification. In all cases, mDAPT narrows the gap to monolingual performance compared to mBERT, i.e. multilingual domain adaptive pretraining helps to make the multilingual model better suited for the specific domain.

	mBERT	mDAPT	¬ mDAPT
QUAERO	58.2	59.8	58.0
BIORO	73.9	74.5	73.4
NCBI	86.0	87.2	85.9
CLIN	72.4	72.7	71.8
PHAR	88.5	88.9	87.8
PHR.BANK	82.4	83.9	82.5
FINNEWS	62.8	63.6	62.2
OMP-2	90.8	91.3	91.0
OMP-9	71.1	71.6	71.0/71.7

Table 5: Cross-domain control experiments. We report two control results for OMP-9 since two mDAPT-setting achieved the same averaged accuracy.

Adapter-based training Adapter-based training exhibits a similar pattern: mDAPT improves mBERT across the board, except for the *da* and *en* sentence classification tasks, where mDAPT is conducted using only *en*-specific data. For most tasks, except *da* and *en* sentence classification, the performance of adapter-based training is below the one of full model training. On *pt* NER dataset, the best score (66.2) achieved by adapter-based training is much lower than the one (72.7) by the full model training.

Comparison of combination strategies After we observe a single multi model can achieve competitive performance as several mono models, the next question is how do different combination strategies affect the effectiveness of mDAPT? As a general trend, the pretraining corpus composed of multilingual data—$\mathrm{M_D}$+$\mathrm{E_D}$ and $\mathrm{M_D}$+$\mathrm{M_{WIKI}}$—achieves better results than $\mathrm{E_D}$ composed by only *en* data. This is evident across both full - and adapter-based training. $\mathrm{M_D}$+$\mathrm{E_D}$ performs best in most cases, especially for the adapter-based training. This result indicates the importance of multilingual data in the pretraining corpus. It is worth noting that even pretraining only on $\mathrm{E_D}$ data can improve the performance on non-English datasets, and for *en* tasks, we see an expected advantage of having more *en*-specific data in the corpus.

5.2 Cross-domain evaluations

To make sure that the improvements of mDAPT models over mBERT stem from observing multilingual domain-specific data, and not from exposure to more data in general, we run cross-domain experiments (Gururangan et al., 2020), where we evaluate the models adapted to the biomedical domain on the financial downstream tasks, and vice versa. The results are shown in Table 5, where we report results for the best mDAPT model and its counterpart in the other domain (¬ mDAPT). In almost all cases, mDAPT outperforms ¬ mDAPT, indicating that adaptation to the domain, and not the exposure to additional multilingual data is responsible for mDAPT's improvement over mBERT. For the OMP datasets, ¬ mDAPT performs surprisingly well, and we speculate this might be because it requires less domain-specific language understanding to classify the newspaper titles.

6 Analysis

Our experiments suggest that mDAPT results in a pretrained model which is better suited to solve domain-specific downstream tasks than mBERT, and that mDAPT narrows the gap to monolingual model performance. In this section, we present further analysis of these findings, in particular we investigate the quality of domain-specific representations learned by mDAPT models compared to mBERT, and the gap between mono- and multilingual model performance.

Domain-specific multilingual representations Multilingual domain adaptive pretraining should result in improved representations of domain-specific text in multiple languages. We evaluate the models' ability to learn better sentence representations via a cross-lingual sentence retrieval task, where, given a sentence in a source language, the model is tasked to retrieve the corresponding translation in the target language. To obtain a sentence representation, we average over the encoder outputs for all subtokens in the sentence, and retrieve the k nearest neighbors based on cosine similarity. As no fine-tuning is needed to perform this task, it allows to directly evaluate encoder quality. We perform sentence retrieval on the parallel test sets of the WMT Biomedical Translation Shared Task 2020 (Bawden et al., 2020). The results in Table 6 show that mDAPT improves retrieval quality, presumably because the models learned better domain-specific representations across languages. Interestingly, with English as target language (upper half), the model trained on English domain-specific data works best, whereas for English as source language, it is important that the model has seen multilingual domain-specific data during pretraining.

	mBERT	+$\mathrm{E_D}$	+ $\mathrm{M_D}$+$\mathrm{E_D}$	+ $\mathrm{M_D}$+$\mathrm{M_W}$
es → *en*	86.7	**91.9**	89.4	87.2
pt → *en*	**87.3**	77.1	77.5	83.9
de → *en*	79.4	**88.7**	83.9	80.9
it → *en*	85.6	**90.9**	87.4	87.1
ru → *en*	67.5	**84.4**	76.5	74.6
en → *es*	86.7	84.7	**90.5**	87.4
en → *pt*	89.4	78.2	**90.4**	86.8
en → *de*	79.4	79.6	**87.8**	81.2
en → *it*	83.9	82.9	**88.1**	86.1
en → *ru*	70.3	81.6	**90.8**	89.5

Table 6: Precision@1 for biomedical sentence retrieval. Best score in each row is marked in bold. The upper half shows alignment to English, the lower half alignment from English.

Effect of tokenization Ideally, we want to have a mDAPT model that performs close to the corresponding monolingual model. However, for the full fine-tuning setup, the monolingual model outperforms the mDAPT models in most cases. Rust et al. (2020) find that the superiority of monolingual over multilingual models can partly be attributed to better tokenizers of the monolingual models, and we hypothesize that this difference in tokenization is even more pronounced in domain-specific text. Following Rust et al. (2020), we measure tokenizer quality via *continued words*, the fraction of words that the tokenizer splits into several subtokens, and compare the difference between monolingual and multilingual tokenizer quality on `specific` text (the train splits of the downstream tasks), with their difference on `general` text sampled from Wikipedia. Figure 3 shows that the gap between monolingual and multilingual tokenization quality is indeed larger in the `specific` texts (green bars) compared to the `general` texts (brown bars), indicating that in a specific domain, it is even harder for a multilingual model to outperform a monolingual model. This suggests that methods for explicitly adding representations of domain-specific words (Poerner et al., 2020; Schick and Schütze, 2020) could be a promising direction for improving our approach.

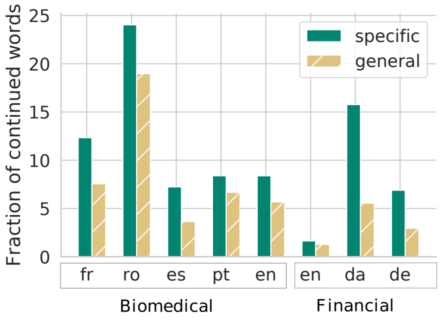

Figure 3: Difference in fraction of continued words between `mono`- and `multi`-lingual tokenizers on general and specific datasets. The bars indicate improvement of the monolingual tokenizer over the multilingual tokenizer.

Error analysis on financial sentence classification To provide a better insight into the difference between the `mono` and `multi` models, we compare the error predictions on the Danish FINNEWS dataset, since results in Table 4 show that the `mono` outperforms all `multi` models with a large margin on this dataset. We note that the FINNEWS dataset, which is sampled from tweets, contains a heavy use of idioms and jargon, on which the `multi` models usually fail. For example,

- Markedet lukker: **Medvind** til bankaktier på en rød C25-dag [POSITIVE]

  English translation: *Market closes: **Tailwind** for bank shares on a red C25-day*

- Nationalbanken tror ikke særskat får den store betydning: Ekspert kaldet det **"noget pladder"** [NEGATIVE]

  English translation: *The Nationalbank does not think special tax will have the great significance: Expert called it **"some hogwash"***

Pretraining data for the `mono` DA-BERT includes Common Crawl texts and custom scraped data from two large debate forums. We believe this exposes the DA-BERT to the particular use of informal register. By contrast, the pretraining data we use are mainly sampled from publications. This could be an interesting direction of covering the variety of a language in sub-domains for a strong MDAPT model.

7 Related Work

Recent studies on domain-specific BERT (Lee et al., 2019; Alsentzer et al., 2019; Nguyen et al., 2020), which mainly focus on English text, have demonstrated that in-domain pretraining data can improve the effectiveness of pretrained models on downstream tasks. These works continue pretraining the whole *base* model—BERT or ROBERTA—on domain-specific corpora, and the resulting models are supposed to capture both generic and domain-specific knowledge. By contrast, Beltagy et al. (2019); Gu et al. (2020); Shin et al. (2020) train domain-specific models from scratch, tying an in-domain vocabulary. Despite its effectiveness, this approach requires much more compute than domain adaptive pretraining, which our work focuses on. Additionally, we explore an efficient variant of domain adaptive pretraining based on adapters (Houlsby et al., 2019; Pfeiffer et al., 2020b), and observe similar patterns regarding pretraining a multilingual domain-specific model.

Several efforts have trained large scale multilingual language representation models using parallel data (Aharoni et al., 2019; Conneau and Lample, 2019) or without any cross-lingual supervision (Devlin et al., 2019; Conneau et al., 2020; Xue et al., 2020). However, poor performance on low-resource languages is often observed, and efforts are made to mitigate this problem (Rahimi et al., 2019; Ponti et al., 2020; Pfeiffer et al., 2020b). In contrast, we focus on the scenario that the NLP model needs to process domain-specific text supporting a modest number of languages.

Alternative approaches aim at adapting a model to a specific target task within the domain directly, e.g. by an intermediate supervised fine-tuning step (Pruksachatkun et al., 2020; Phang et al., 2020), resulting in a model specialized for a single task.

Domain adaptive pretraining, on the other hand, aims at providing a good base model for different tasks within the specific domain.

8 Conclusion

We extend domain adaptive pretraining to a multilingual scenario that aims to train a single multilingual model better suited for the specific domain. Evaluation results on datasets from biomedical and financial domains show that although multilingual models usually underperform their monolingual counterparts, domain adaptive pretraining can effectively narrow this gap. On seven out of nine datasets for document classification and NER, the model resulting from multilingual domain adaptive pretraining outperforms the baseline `multi-general` model, and on four it even outperforms the `mono-general` model. The encouraging results show the implication of deploying a single model which can process financial or biomedical documents in different languages, rather than building separate models for each individual language.

Acknowledgements

We thank PwC for providing the data and thank the experts for their annotation support and valuable feedback. Mareike Hartmann was funded by the Lundbeck foundation for research in the context of the BrainDrugs project and the German Federal Ministry of Research (BMBF) in the context of the XAINES project. Xiang Dai and Desmond Elliott are supported by Innovation Fund Denmark in the context of AI4Xray project.

References

Aitor Gonzalez Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in*

Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online. Association for Computational Linguistics.

Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. Finbert: pre-trained model on sec filings for financial natural language tasks.

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.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

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.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

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.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

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.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in Portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Maria Mitrofan. 2017. Bootstrapping a romanian corpus for medical named entity recognition. In *RANLP*, pages 501–509.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vuli?, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *EMNLP*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 7654–7673, Online. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Edoardo M Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2020. Parameter space factorization for zero-shot learning across tasks and languages. *arXiv preprint arXiv:2001.11453*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Baseline models

Table 7 is a comparison between baseline `mono` models and the `multi` model. For the NER tasks, we use the `cased` versions for all experiments. For sentence classification, we use `uncased` versions for DA-BERT and EN-BERT.

enclose

B Biomedical data

Preprocessing pretraining data For the English abstracts, we sentence tokenize using NLTK and filter out sentences that do not contain letters. For the WMT abstracts, we filter out lines that start with #, as these indicate paper ID and author list. We determine the language of a document using its metadata provided by PubMed. We transliterate Russian PubMed titles (in Latin) back to Cyrillic using the `transliterate` python package (`https://pypi.org/project/transliterate/`).

Downstream NER data The French QUAERO (Névéol et al., 2014) dataset comprises titles of research articles indexed in the biomedical MEDLINE database, and information on marketed drugs from the European Medicines Agency. The Romanian BIORO (Mitrofan, 2017) dataset consists of biomedical publications across various medical disciplines. The Spanish PHARMACONER (Agirre et al., 2019) dataset comprises publicly available clinical case studies, which show properties of the biomedical literature as well as clinical records, and has annotations for pharmacological substances, compounds and proteins. The English NCBI DISEASE (Doğan et al., 2014) dataset consists of PubMed abstracts annotated for disease names. The Portuguese CLINPT dataset is the publicly available subset of the data collected by Lopes et al. (2019), and comprises texts about neurology from a clinical journal.

Prepocessing NER data We convert all annotations to BIO format. The gaps in discontinuous entities are labeled. We sentence tokenize at line breaks, and if unavailable at fullstops. We word tokenize all data at white spaces and split off numbers and special characters. If available, we use official train/dev/test splits. For BIORO, we produce a random 60/20/20 split. For CLINPT, we use the data from volume 2 for training and development data and test on volume 1.

C Financial data

Preprocessing pretraining data Sentences are tokenized using NLTK. For languages not cover by the sentence tokenizer, we split by full stops. Additionally, a split check of particular large sentences, filtering out sentences with no letters, and HTML and tags have been removed.

Downstream classification data

FINMULTICORPUS The corpus consists of PwC publications in multiple languages made publicly available on PwC websites. The publications cover a diverse range of topics that relates to the financial domain. The corpus is created by extracting text passages from publications. Table 2 describes the number of sentences and the languages that the CPT corpus cover.

FINNEWS The financial sentiment dataset is curated from financial newspapers headline tweets. The motivation was to create a Danish equivalent to FINANCIAL PHRASEBANK. The news headlines are annotated with a sentiment by 2 annotators. The annotators were screened to ensure sufficient domain and educational background. A description of *positive*, *neutral*, and *negative* was formalized before the annotation process. The dataset has an 82.125% rater agreement and a Krippendorff's alpha of .725 measured on 800 randomly sampled instances.

ONE MILLION POSTS (Schabus et al., 2017) The annotated dataset includes user comments posted to an Austrian newspaper. We use the TITLE (newspaper headline) and TOPICS, i.e., 'KULTUR', 'SPORT', 'WIRTSCHAFT', 'INTERNATIONAL', 'INLAND', 'WISSENSCHAFT', 'PANORAMA', 'ETAT', 'WEB'. With the dataset, we derive two downstream tasks. The binary classification task OMP $_{binary}$ that deals with whether a TITLE concerns a financial TOPICS or not. Here we merge all non-financial TOPICS into one category. The multi-class classification OMP $_{multi}$ seeks to classify a TITLE into one of the 9 TOPICS.

D Adapter-based training

Recall that the main component of a transformer model is a stack of transformer layers, each of which consists of a multi-head self-attention network and a feed-forward network, followed by layer normalization. The idea of adapter-based

	Training data	Vocab size	# parameters
DE-BERT (Chan et al., 2020)	OSCAR (Common Crawl), OPUS (Translated web texts), Wikipedia, Court decisions [163.4G]	30.0K	109.1M
DA-BERT	Common Crawl, Wikipedia, Debate forums, OpenSubtitles [9.5G, 1.6B]	31.7K	110.6M
EN-BERT (Devlin et al., 2019)	English Wikipedia, Books [3.3B]	29.0K	108.3M
EN-BIO-BERT (Lee et al., 2019)	Initialized with EN-BERT; continue on PubMed, PMC [18B]	29.0K	108.3M
EN-FIN-BERT (Araci, 2019)	Initialized with EN-BERT; continue on News articles [29M]	30.5K	109.5M
ES-BERT (Cañete et al., 2020)	OPUS, Wikipedia [3B]	31.0K	109.9M
FR-BERT (Le et al., 2020)	24 corpora, including Common-Crawl, Wikipedia, OPUS, Books, News, and data from machine translation shared tasks, Wikimedia projects [71G, 12.7B]	68.7K	138.2M
PT-BERT (Souza et al., 2020)	brWaC (web text for Brazilian Portuguese) [2.6B]	29.8K	108.9M
PT-BIO-BERT (Schneider et al., 2020)	Initialized with MBERT; continue on PubMed and Scielo (scholarly articles) [16.4M]	119.5K	177.9M
RO-BERT (Dumitrescu et al., 2020)	OSCAR, OPUS, Wikipedia [15.2G, 2.4B]	50.0K	124.4M
MBERT (Devlin et al., 2019)	Wikipedia [72G]	119.5K	177.9M

Table 7: A comparison between baseline mono models and the multi model: MBERT. We use total file size (Gigabyte) and the total number of tokens to represent the training data size.

training (Houlsby et al., 2019; Stickland and Murray, 2019; Pfeiffer et al., 2020a) is to add a small size network (called *adapter*) into each transformer layer. Then during the training stage, only the weights of new adapters are updated while keeping the base transformer model fixed. Different options regarding where adapters are placed, and its network architecture exist. In this work, we use the bottleneck architecture proposed by Houlsby et al. (2019) and put the adapters after the feed-forward network, following (Pfeiffer et al., 2020a):

$$\text{Adapter}_l\,(h_l, r_l) = U_l\,(\text{ReLU}\,(D_l\,(h_l))) + r_l$$

where r_l is the output of the transformer's feed-forward layer and h_l is the output of the subsequent layer normalisation.

Lang	PM abstracts	PM titles	M_D	M_{WIKI}
fr	54,047	681,774	735,821	872,678
es	73,704	312,169	385,873	939,452
de	31,849	814,158	846,007	831,257
it	14,031	265,272	279,303	923,548
pt	38,716	79,766	118,482	811,522
ru	43,050	576,684	619,734	908,011
ro	0	27,006	27,006	569,792
en	227,808	0	227,808	903,706
Total	483,205	2,756,829	3,240,034	6,759,966

Table 8: Number of sentences of multilingual domain-specific pre-training data for biomedical domain. Upsampling for *en* was done from PM abstracts instead of Wikipedia.

Lang	RCV2	PwC	M_D	M_{WIKI}
zh	222,308	1,466	223,774	470,111
da	72,349	192,352	264,701	465,044
nl	15,131	34,344	49,475	391,750
fr	863,911	51,500	915,411	143,427
de	1,104,603	71,382	1,175,985	0
it	138,814	22,499	161,313	467,680
ja	88,333	20,936	109,269	450,352
no	92,828	19,208	112,036	451,799
pt	57,321	35,323	92,644	439,942
ru	192,869	48,388	241,257	468,466
es	936,402	51,100	987,502	95,691
sv	132,456	25,336	157,792	467,050
en	0	346,856	346,856	444,532
tr	0	34,990	34,990	362,685
Total	3,917,325	955,680	4,873,005	5,118,529

Table 9: Number of sentences of multilingual domain-specific pretraining data for financial domain. Upsampling for *en* used the TRC2 corpus instead of Wikipedia.