Investigating Continued Pretraining for Zero-Shot Cross-Lingual Spoken Language Understanding

Samuel Louvan[1,2], **Silvia Casola**[1], **Bernardo Magnini**[1]
1. Fondazione Bruno Kessler, Italy
2. University of Trento, Italy
slouvan@fbk.eu, scasola@fbk.eu, magnini@fbk.eu

Abstract

Spoken Language Understanding (SLU) in task-oriented dialogue systems involves both intent classification (IC) and slot filling (SF) tasks. The *de facto* method for zero-shot cross-lingual SLU consists of fine-tuning a pretrained multilingual model on English labeled data before evaluating the model on unseen languages. However, recent studies show that adding a second pretraining stage (*continued pretraining*) can improve performance in certain settings. This paper investigates the effectiveness of continued pretraining on unlabeled spoken language data for zero-shot cross-lingual SLU. We demonstrate that this relatively simple approach benefits either SF and IC task across 8 target languages, especially the ones written in Latin script. We also find that discrepancy between languages used during pretraining and fine-tuning may introduce training instability, which can be alleviated through code-switching.

1 Introduction

In task-oriented dialogue systems, a Spoken Language Understanding (SLU) component typically involves intent classification (IC) and slot filling (SF) (Tur and De Mori, 2011) tasks. For example, in "*Show me the fares for Delta flights from Dallas to San Francisco*", the intent is ASKING AN AIRFARE and its corresponding slots are *Delta* (AIRLINE-NAME), *Dallas* (CITY-ORIGIN), and *San Francisco* (CITY-DESTINATION). Scaling SLU models to other languages is still challenging, especially when there is limited or no labeled data available in the target language (Louvan and Magnini, 2020).

To approach this problem, previous work studies IC and SF tasks in a zero-shot cross-lingual setting (Schuster et al., 2019; Upadhyay et al., 2018; Xu et al., 2020), where it is assumed that a labeled dataset is only available for a high resource language (e.g., English). With the rise of pretrained multilingual language models (LMs) (Devlin et al., 2019; Lample and Conneau, 2019) the most common approach is by fine-tuning the pretrained multilingual model on the English labeled data, and then evaluate the model directly on the target language data that are not seen during fine-tuning.

While direct fine-tuning serves as a strong baseline, pretrained LMs are not necessarily *universal* and they may need domain-specific adaptation. Recent works have shown that adding a second pretraining stage (or *continued pretraining*) before fine-tuning can give positive impact on the model performance (Beltagy et al., 2019; Lee et al., 2020; Gururangan et al., 2020). During continued pretraining, we continue training the pretrained language model using a *domain-specific* or *task-specific* unlabeled dataset, with the same masked language model objective. This stage is useful to alleviate the *domain mismatch* between the original pretraining and the target task data. By continued pretraining on domain specific unlabeled data, the model acquires prior knowledge which is expected to be helpful in the fine-tuning stage. This approach has shown promising results on text classification, typically on English. However, it remains unclear whether it is applicable in the context of *zero-shot cross-lingual* SLU.

In contrast to previous work which has mostly focused on English classification tasks, we investigate the effectiveness of continued pretraining for zero-shot cross-lingual SLU tasks on eight target languages. Our study reveals that the existing continued pretraining method (Gururangan et al.,

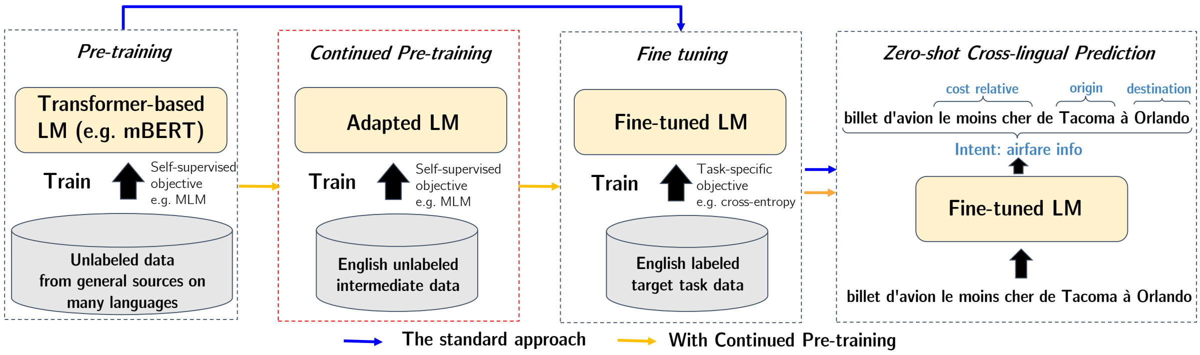


Figure 1: The overall stages of zero-shot cross lingual SLU using a pretrained multilingual model. The standard approach follows the stages marked with blue arrows (*direct fine-tuning*). We investigate the effectiveness of adding a continued pretraining stage (red dashed box) in the overall pipeline.

2020), that is successful in English text classification tasks, does not always generalize to the context of zero-shot cross-lingual SLU. We focus on the following research questions:

(Q1) *Is continued pretraining effective for zero-shot cross-lingual SLU tasks?*
↪ Our experiments on the MultiATIS++ dataset (Xu et al., 2020) reveal that incorporating continued pretraining on intermediate English data can improve performance over direct fine-tuning for all languages, on zero-shot SLU. The performance gain is especially evident for languages with Latin script writing system. The benefit of continued pretraining diminishes as we inject cross-lingual supervision in the fine-tuning stage, even with simple data augmentation through code-switching.

(Q2) *What are the factors that influence the effectiveness of the continued pretraining stage?*
↪ Using the target language for continued pretraining before fine-tuning on English can be detrimental to the overall performance. However, this can be largely alleviated by code-switching the fine-tuning data. We also observe that performance improvement are not obtained by merely adding more continued pretraining data; higher domain similarity between the continued pretraining data and the fine-tuning data is indeed more important.

2 Continued Pretraining in Zero-Shot SLU

Figure 1 shows a comparison between the standard direct fine-tuning approach with the continued pretraining approach. The main difference is the additional intermediate pretraining stage (second block in Figure 1), in which we continue training the model on an intermediate unlabeled data using the same masked language modeling objective. As the original pretraining data is relatively far from the task-oriented dialogues used in SLU, we hypothesize that continued pretraining can alleviate the *domain* mismatch and ingest a better prior knowledge that will be useful during fine-tuning.

Intermediate Data for Continued Pretraining. We define several criteria for the intermediate pretraining data for the continued pretraining stage. First, their domain should be relatively close to the target dataset. We interpret the term domain as a multidimensional *variety space* (Ramponi and Plank, 2020; Plank, 2016): a domain comprises multiple aspects (style, topic, and genre (van der Wees et al., 2015)) that contribute to the text variation. Using this perspective and considering the target domain of a task-oriented dialogue system, we require that the intermediate data comprises text that presents a *spoken language dialog* style and covers a *broad range of topics*. Second, the dataset should be several magnitudes larger in size than the target task dataset. Finally, it must be available in many languages to support our study of continued pretraining with the target language.

3 Experimental Setup

In this section, we describe the experimental settings related to models, evaluation metrics, and datasets.

3.1 Models

For all of our experiments, we use a transformer-based model (Vaswani et al., 2017), namely multilingual BERT (mBERT) (Devlin et al., 2019), as the pretrained model. This model was pretrained on Wikipedia articles covering 104 languages, and we use the *bert-base-multilingual-cased* version.

Continued Pretraining. For the continued pretraining stage, we further train mBERT with unlabeled intermediate data using only the Masked Language Modeling (MLM) objective for 12.5K steps, and mostly adopt the hyperparameters in Gururangan et al. (2020). We compare the following configurations: (i) $\mathrm{DAPT}_{\mathrm{Tgt}}$ a continued domain adaptive pretraining (DAPT) of mBERT on intermediate unlabeled data on the target language. (ii) $\mathrm{DAPT}_{\mathrm{En}}$ a continued DAPT of mBERT on intermediate unlabeled data on English.

Fine-Tuning. As the baseline model, without any adaptation (No DAPT), we use the joint IC and SF model architecture (Chen et al., 2019). This model is the state-of-the-art for IC and SF (Louvan and Magnini, 2020), and it is often used as one of the baselines in recent zero-shot cross-lingual SLU studies (Xu et al., 2020; Li et al., 2021). The model is trained on the English dataset; as the setup is zero-shot cross-lingual and we use the model's last epoch for zero-shot evaluation following Xu et al. (2020). We evaluate the effectiveness of each of the DAPT configurations when applied to the following fine-tuning scenarios:

- Fine-tuning on English (FINETUNE-EN). This is the standard fine-tuning scenario, where we take mBERT either with DAPT or no DAPT, fine-tune it on the English IC and SF data, and then perform zero-shot prediction to all target language data.
- Fine-tuning on the English *code-switched* data (FINETUNE-CS). In this scenario, we perform data augmentation on the English fine-tuning dataset via code-switching. We follow the approach from Qin et al. (2020), where we replace the English words with their translation in the target language using the Panlex bilingual dictionary (Kamholz et al., 2014). Given a training batch, we randomly sample sentences and tokens to replace. We use the same hyperparameter used by Qin et al. (2020), that defines both sentence and word ratio to control the word replacement. We include FINETUNE-CS because we want to study the benefits of DAPT when adding stronger cross-lingual supervision in the fine-tuning stage.

We did not experiment with more complex models models as our main goal is to investigate the effect of the the continued pretraining stage, rather than achieving the state of the art performance per se.

Implementation & Model Evaluation metric. For the intent and SF models, we adapt the implementation from Qin et al. (2020) in which they make it publicly available (https://github.com/kodenii/CoSDA-ML). The sentence and token ratio replacement for code-switching is set to 1.0 and 0.9 respectively. For training, the learning rate is set to 10^{-5}, batch size is set to 32, number of epoch is set to 20. We did not do extensive hyperparameter tuning, as this is a zero-shot cross lingual case where the target dataset is not available, we use the same hyperparameters as Xu et al. (2020). For the continued pretraining we use the language modeling script from Huggingface (Wolf et al., 2019). We use the `bert-base-multilingual-cased`, hidden state size is 768, we apply dropout probability of 0.1. The number of training steps is 12,500 following Gururangan et al. (2020), the batch size is set to 16.

3.2 Dataset

SF and IC Dataset. We use the MultiATIS++ (Xu et al., 2020) dataset, which contains nine languages (Table 1). The dataset is derived from the original ATIS English dataset (Hemphill et al., 1990), widely used as a benchmark for IC and SF for task-oriented dialogue systems. Utterances are related to conversations of a user asking for flight information to a system.

Language	#train / #dev /#test	#slot	#intent
English (EN)	4.4K/ 490 / 893	83	24
German (DE)	4.4K / 490 / 892	83	24
Spanish (ES)	4.4K/ 490 / 893	83	24
French (FR)	4.4K / 490 / 893	83	24
Portuguese (PT)	4.4K / 489 / 892	83	24
Hindi (HI)	1.4K / 160 / 888	74	22
Japanese (JA)	4.4K / 490/ 886	83	24
Chinese (ZH)	4.4K / 490 / 893	83	24
Turkish (TR)	0.6K / 60/ 715	70	21

Table 1: Multi-ATIS++ (Xu et al., 2020) statistics.

Continued Pretraining Dataset. We use the OpenSubtitle (OpenSub) (Lison and Tiedemann, 2016) (Table 2) dataset for the continued pretraining stage for several reasons. First, the dataset is constructed from movies and TV series containing *spoken language* in dialogue settings covering a broad range of topics. Second, OpenSubtitle covers all the *languages* that we use on the downstream tasks, which enables us to evaluate not only

DAPT_{En} but also DAPT_{Tgt}. Third, the dataset is large in size, thus ideal for continued pretraining. Typically, the dataset used for continued pretraining is larger than that used for fine-tuning. For our experiments we randomly sampled 100K sentences for each language in the OpenSub dataset, resulting in a dataset around 20 times larger than the downstream task dataset.

Language	Total Tokens
English (EN)	734,302
German (DE)	691,039
Spanish (ES)	711,264
French (FR)	739,551
Portuguese (PT)	676,789
Hindi (HI)	688,675
Japanese (JA)	747,780
Chinese (ZH)	611,700
Turkish (TR)	554,709

Table 2: OpenSub (Lison and Tiedemann, 2016) dataset statistics. Each language has 100 K utterances.

4 Results

The main goal of our experiment is to answer research question (**Q1**). Table 3 compares the zero-shot performance for SF and IC across languages. In terms of language (by column in Table 3), we observe that all languages improve over No-DAPT in at least one DAPT setting, suggesting that DAPT is effective across languages. Observing the results per task, SF benefits from either DAPT_{En} or DAPT_{Tgt} for German, Spanish, French, Portuguese, and Turkish, which all are languages with Latin scripts writing system. For these languages, the margin obtained from DAPT when fine-tuning on English (FINETUNE-EN) is higher than when we apply DAPT on code-switched data (FINETUNE-CS). The margin of DAPT when applied on FINETUNE-CS diminishes because FINETUNE-CS uses a stronger supervision signal in the fine-tuning stage, thus providing a higher baseline. For languages with non-Latin script writing system, continued pretraining is less useful; we only observe marginal improvement on Japanese when applying DAPT_{En} and FINETUNE-EN. Similar to Lauscher et al. (2020), we believe that performance is also affected by typological language proximity such as the subject, verb, and object ordering, phonology features or other aspect related to the original size of the pre-training data of mBERT. We leave this for future work.

DAPT is less effective for IC than for SF. The only language that consistently benefits from continued pretraining in both fine-tuning scenarios is Turkish. We found that it is harder to improve the model performance of languages with Latin script through DAPT because the baseline is relatively high; a stronger supervision signal would thus be needed. The performance gain is small even for those languages that do benefit from DAPT. We also observe that using a different language between continued pretraining and fine-tuning stages, DAPT_{Tgt} and FINETUNE-EN, may hamper performance.

5 Analysis and Discussion

To answer the research question (**Q2**), we analyze our results focusing on the performance variation when using different languages in DAPT and fine-tuning (§5.1) and the effect of domain distribution in different sources for DAPT_{En} (§5.2).

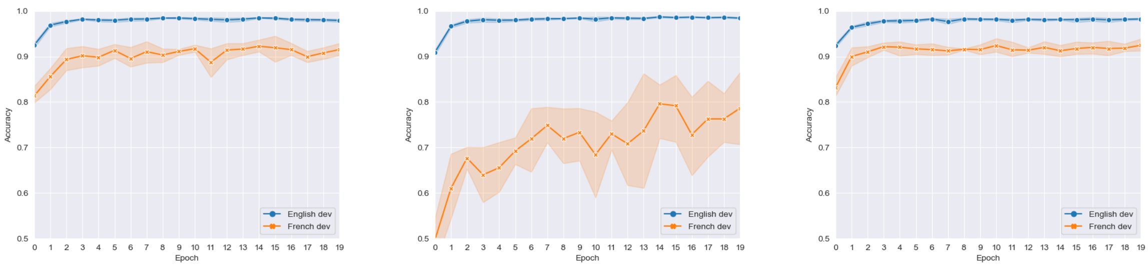
5.1 Performance Variation when Applying DAPT

As we have noticed in Section §4, there are cases where performance drop when we use DAPT_{Tgt} and FINETUNE-EN, especially for IC. This behaviour holds even for languages relatively close to English, such as German and French. One possible reason for the drop in accuracy is that the language difference introduces instability in fine-tuning. Our post-hoc analysis shows that the target language performance during training on the dev set has a large deviation and continues fluctuating even after the English dev performance has stabilized. This observation resonates with a previous study from Keung et al. (2020), which shows that, for zero-shot text classification, English dev performance often does not correlate with those of the target language. Using DAPT_{Tgt} and FINETUNE-EN pronounces the disagreement of performance between the English and the target dev set. Figure 2 shows the comparison of the IC performance during training across continued pretraining strategies when fine-tuning on English for French. However, for the SF task, we do not observe a large performance variation even with a language mismatch: this might indicate that text classification is more susceptible to instability than sequence tagging. The variability caused by DAPT_{Tgt} is largely alleviated when we use DAPT_{En}. For the FINETUNE-CS scenario, the system is relatively stable even when combined

SF F1								
	DE	ES	FR	PT	HI	JA	ZH	TR
FINETUNE-EN								
No-DAPT	65.3	71.3	64.0	61.9	47.5	62.2	66.3	27.4
$\Delta\mathrm{DAPT}_{\mathrm{Tgt}}$	+4.0	−2.4	−7.7	−0.6	−12.9	−9.7	−0.6	+18.5
$\Delta\mathrm{DAPT}_{\mathrm{En}}$	+2.1	+0.9	+5.9	+1.4	−4.5	+0.8	−0.2	−5.8
FINETUNE-CS								
No-DAPT	75.5	80.8	71.9	72.0	58.1	67.1	81.6	72.0
$\Delta\mathrm{DAPT}_{\mathrm{Tgt}}$	−0.2	−0.4	+0.5	+1.1	−3.9	−6.3	−1.2	−10.9
$\Delta\mathrm{DAPT}_{\mathrm{En}}$	+0.4	+0.1	+4.6	+1.2	−13.9	−8.4	−0.7	−15.8

IC ACCURACY								
	DE	ES	FR	PT	HI	JA	ZH	TR
FINETUNE-EN								
No-DAPT	90.0	91.9	92.1	92.8	81.1	83.0	87.1	61.2
$\Delta\mathrm{DAPT}_{\mathrm{Tgt}}$	−10.8	+0.5	−13.3	−1.6	−13.3	−1.9	−2.9	+8.1
$\Delta\mathrm{DAPT}_{\mathrm{En}}$	−0.8	−0.1	+0.1	−0.6	−2.5	−0.5	−2.4	+8.3
FINETUNE-CS								
No DAPT	95.1	96.4	96.6	94.2	85.6	85.1	88.0	66.2
$\Delta\mathrm{DAPT}_{\mathrm{Tgt}}$	−1.1	−0.2	−0.5	+1.3	+0.6	−2.4	+0.3	+3.9
$\Delta\mathrm{DAPT}_{\mathrm{En}}$	−1.6	−0.2	−0.2	+0.4	−0.8	−2.6	−7.3	+12.3

Table 3: Performance comparison on the test set for SF and IC. Scores for No DAPT are the average slot F1 and intent accuracy over five runs. The $\Delta\mathrm{DAPT}_{\mathrm{Tgt}}$ and $\Delta\mathrm{DAPT}_{\mathrm{En}}$ indicate the delta between DAPT and No DAPT.



(a) No DAPT + FINETUNE-EN (b) $\mathrm{DAPT}_{\mathrm{Tgt}}$+ FINETUNE-EN (c) $\mathrm{DAPT}_{\mathrm{En}}$+ FINETUNE-EN

Figure 2: Post-hoc analysis: *development set* performance variation on IC between English and French, using FINETUNE-EN and applying different DAPT strategies.

with $\mathrm{DAPT}_{\mathrm{Tgt}}$ or $\mathrm{DAPT}_{\mathrm{En}}$.

5.2 Domain Relevance for $\mathrm{DAPT}_{\mathrm{En}}$

We aim at investigating whether the improvement from the continued pretraining comes from the domain relevance of the intermediate data. For this purpose, we selected a few *written text* datasets instead of spoken language, which are focused on a *specific topic*. Specifically, we use the European Medicines Agency (EMEA) and European Central Bank corpus (ECB) from Tiedemann (2012). EMEA contains articles about human, veterinary, or herbal medicines extracted from the EMEA website. ECB contains financial documents that are extracted from the website and documentation of the European Central Bank. In order to check that EMEA and ECB are more distant in terms of domain from MultiATIS than OpenSub, we compute the Jensen Shannon Divergence (JSD) measure of the term distribution (Dai et al., 2020; Ruder and Plank, 2017). We compute the JSD between the MultiATIS English dataset that is used for fine-tuning and each English intermediate dataset. Based on the JSD measure, EMEA and ECB are more distant to MultiATIS than OpenSub (Table 4).

For each intermediate dataset, we randomly sample 100K sentences and use them for continued pretraining. We compare the SF performance of $\mathrm{DAPT}_{\mathrm{En}}$ with FINETUNE-EN on Open-

	OpenSub	EMEA	ECB
JSD	0.419	0.391	0.397

Table 4: Domain similarity between MultiATIS and each of the intermediate data.

Lang.	No DAPT	$\Delta\text{DAPT}_{\text{En}}$		
		OpenSub	EMEA	ECB
DE	65.3	+2.1	−2.5	−9.5
ES	71.3	+0.9	+0.9	+1.3
FR	64.0	+5.9	+2.0	+0.7
PT	61.9	+1.4	−0.3	−9.1
Avg		+2.5	+0.005	−4.1

Table 5: Comparison of SF performance with different intermediate data.

Sub, EMEA, and ECB in Table 5. We focus on languages that belongs to Indo-European family which mostly obtain benefit from DAPT on SF (Table 3) Overall, we see that DAPT using OpenSub obtains improvements over No-DAPTin all cases. The DAPT performance using EMEA and ECB are lower than OpenSub in most cases. Even for DE and PT languages, DAPT with ECB obtains substantially lower performance than No-DAPT. However, there are cases when EMEA or ECB match or even perform better than OpenSub i.e., for Spanish. These cases indicate that performing *data selection* before continued pretraining could be beneficial to construct more optimal DAPT dataset. It would be interesting also to observe how continued pre-training would work using smaller unlabeled pre-training data but more task relevant. We leave this possibility for future work.

6 Related Work

Zero-Shot Cross-Lingual SLU. Before the advent of the pre-trained multilingual transformer models, most approaches relied on pre-trained cross-lingual embeddings to perform zero-shot SLU. Upadhyay et al. (2018) uses cross-lingual embedding (Bojanowski et al., 2017) to perform zero-shot SLU while Schuster et al. (2019) uses multilingual embedding (Cove) from pre-trained multilingual bi-LSTM encoder used in Neural Machine Translation (NMT). Liu et al. (2019) leverages transferable latent variables to improve the sentence representation across languages. More recently, as pre-trained multilingual transformer models show potential in zero-shot settings, most approaches focus on improving their multilingual representation through augmentation and alignment methods. Qin et al. (2020) proposes multilingual code-switching using a bi-lingual dictionary to improve mBERT's multilingual representation. Xu et al. (2020) introduces soft alignment of slots between English and the target language produced by a machine translation system that eliminates the need for an annotation projection pipeline. Kulshreshtha et al. (2020) study the effect of various cross-lingual alignment methods to improve mBERT representation.

Continued Pre-training Domain adaptation is a long-studied problem in the NLP community (Daumé III, 2007; Blitzer et al., 2007), in which we assume data in the target domain might be hard to obtain while being abundant in source domains. Continued pre-training – where the model is trained on relevant data using the same pre-training objective – is used for mitigating the distribution mismatch between the pre-training and the fine-tuning data in terms of *domain* (Logeswaran et al., 2019; Han and Eisenstein, 2019; Gururangan et al., 2020; Beltagy et al., 2019), *task* (Gururangan et al., 2020), and *language* (Pfeiffer et al., 2020). A complementary approach performs a first fine-tuning on related auxiliary tasks (for which training data are easy to obtain) before the final fine-tuning on the downstream task (Arase and Tsujii, 2019; Garg et al., 2020; Khashabi et al., 2020). Our work is in line with Gururangan et al. (2020) where we investigate further the effectiveness of continued pre-training in the context of zero-shot cross-lingual SLU.

7 Conclusion

We systematically study the effectiveness of continued pre-training of a multilingual model on intermediate English unlabeled spoken language data for zero-shot cross-lingual tasks, namely intent classification and slot filling, on 8 languages. Our results show that the domain knowledge learned in English is transferable to other languages. The gain from continued pre-training diminishes as we inject cross-lingual supervision in the fine-tuning stage. There are several factors that influence the effectiveness of the continued pre-training: (i) Using different language between pre-training and fine-tuning can hamper performance and introduce instability in the model training, which can be alleviated with code switching. (ii) Domain similarity is important. The more similar – in terms of data distribution – the intermediate data to the target dataset yields better performance.

References

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. 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 5393–5404, Hong Kong, China, November. 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, November. Association for Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 440–447, Prague, Czech Republic, June. Association for Computational Linguistics.

P. Bojanowski, E. Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. ArXiv, abs/1902.10909.

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, November. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transf ormer models for answer sentence selection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):7780–7788, Apr.

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, July. 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, November. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, USA, June 24-27, 1990. Morgan Kaufmann.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3145–3150, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Phillip Keung, Y. Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't Use English Dev: On the Zero-Shot Cross-Lingual Evaluation of Contextual Embeddings. In EMNLP.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1896–1907, Online, November. Association for Computational Linguistics.

Saurabh Kulshreshtha, José Luis Redondo García, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In EMNLP.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499, Online, November. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36:1234 – 1240.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2950–2962, Online, April. Association for Computational Linguistics.

Pierre Lison and J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In LREC.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303, Hong Kong, China, November. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3449–3460, Florence, Italy, July. Association for Computational Linguistics.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 480–496, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. 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, November. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. arXiv preprint arXiv:1608.07836.

L. Qin, Minheng Ni, Y. Zhang, and W. Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In IJCAI.

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), December. International Committee on Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 372–382, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 3795–3805. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may. European Language Resources Association (ELRA).

Gokhan Tur and Renato De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6034–6038.

Marlies van der Wees, Arianna Bisazza, Wouter Weerkamp, and Christof Monz. 2015. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 560–566, Beijing, China, July. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman

Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5052–5063, Online, November. Association for Computational Linguistics.