On the Universality of Deep Contextual Language Models

Shaily Bhatt[1,2] [*]; **Poonam Goyal**[1], **Sandipan Dandapat**[3], **Monojit Choudhary**[2], **Sunayana Sitaram**[2]
[1] Birla Institute of Technology and Science, Pilani, India
[2] Microsoft Research, Bengaluru, India
[3] Microsoft R&D, Hyderabad, India
{f20170040, poonam}@pilani.bits-pilani.ac.in,
{sadandap, monojitc, sunayana.sitaram}@microsoft.com

Abstract

Deep Contextual Language Models (LMs) like ELMO, BERT, and their successors dominate the landscape of Natural Language Processing due to their ability to scale across multiple tasks rapidly by pre-training a single model, followed by task-specific fine-tuning. Furthermore, multilingual versions of such models like XLM-R and mBERT have given promising results in zero-shot cross-lingual transfer, potentially enabling NLP applications in many under-served and under-resourced languages. Due to this initial success, pre-trained models are being used as 'Universal Language Models' as the starting point across diverse tasks, domains, and languages. This work explores the notion of 'Universality' by identifying seven dimensions across which a universal model should be able to scale, that is, perform equally well or reasonably well, to be useful across diverse settings. We outline the current theoretical and empirical results that support model performance across these dimensions, along with extensions that may help address some of their current limitations. Through this survey, we lay the foundation for understanding the capabilities and limitations of massive contextual language models and help discern research gaps and directions for future work to make these LMs inclusive and fair to diverse applications, users, and linguistic phenomena.

1 Introduction

Language Models (LMs) have evolved considerably in the past decade, starting from the introduction of Word2Vec (Mikolov et al., 2013) to the more recent transformer-based deep models like BERT (Devlin et al., 2019) and its successors. When fine-tuned with task-specific data, pre-trained LMs can be adapted to several different settings, i.e., tasks, domains, and even languages, as these LMs have been extended to multiple languages in the multilingual versions like m-BERT and derivatives. These models can be thought of as 'Universal' because of their potential to be utilized 'universally' in several different application scenarios.[1]

The merits of transfer learning or pre-training word representations have been known for a long time. Moreover, the recent advancements in large-scale deep learning have pushed the boundaries of intensive computation and tremendous amounts of data that can be used to pre-train LMs. However, pre-training is resource-intensive and is not carried out for specific scenarios. Instead, massive LMs are deployed into downstream applications with potentially billions of users around the world. This makes 'Universality' a vital characteristic as the models must be inclusive towards a variety of language usage.

The key contributions of this paper are:

- We formally define 'Universality' by selecting seven dimensions- language, multilingualism, task, domain, medium of expression, geography and demography, and time period - that capture a variety of language usage.

- We curate the current empirical and theoretical results that provide evidence of scaling LMs across these dimensions and identify the capabilities and gaps in these models.

- We outline extensions to these models that can help in overcoming current limitations to become truly universal, thus serving a larger number of end-users and scenarios.

[*] Work done while at Microsoft Research, India

[1] Throughout the rest of the paper – "these models", "LMs", "general domain LMs", "contextual LMs", "universal LMs" and all such terms refers to models including but not limited to ELMo, BERT, RoBERTa, GPT their variants, successors and multilingual versions

Proceedings of the 18th International Conference on Natural Language Processing, pages 106–119
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

2 Universality and its Dimensions

Universality can mean many things, and the associated philosophical debate is beyond the scope of this study. Our work aims not to provide a complete or exhaustive list of capabilities expected out of the model but a list of aspects that can be considered as a starting goal to achieve universality.

Our definition of universality spans seven dimensions. These are: language, multilingualism, task, domain, medium of expression, demography and geography, and time period. We selected these dimensions to cover a broad spectrum of language usage and a diverse set of NLP applications. Ideally, a truly universal model should perform strongly, or at least reasonably, across them. We present a detailed analysis of the capabilities and limitations of LMs in these dimensions in the subsequent sections. This is followed by extensions, which are techniques that can be leveraged to overcome the limitations in the particular dimension.

2.1 Reasoning for Selection of Dimensions

Firstly, it is important to re-iterate that this list of seven dimensions does not intend to be an exhaustive one. Rather, these dimensions have been selected so as to cover a broad spectrum of language usage. The reasons why each of these is important for a general-purpose LM are:

Language Massive multilingual models that can support close to 100 languages at a time are quickly becoming standard for building language technologies that cater to a wide and linguistically diverse population. However, while high-resources languages are well served, many low-resource languages are left behind (Joshi et al., 2021). Thus it is important to understand where large scale multilingual LMs stands in terms of availability, evaluation, and performance along the dimension of Language.

Multilingualism LMs are increasingly being deployed into user-facing applications and thus need to deal with real-world language usage. In multilingual (or bilingual) communities, usage of multiple languages at once gives rise to many language variations such as code-mixing, that the model will need to process. For ascertaining how well models can deal with these linguistic phenomena, understanding the capabilities and limitations of the models along this dimension becomes important.

Task LMs are increasingly becoming the standard component of most NLP pipelines. As a result, it is important to study how well they adapt to various different tasks for which they are used.

Domain Typically, LMs are trained on general purpose language, such as that obtained from Wikipedia or the Web. As such, their training data has limited signals for complex vocabulary that is common for a specialized domain such as medical, financial, legal, etc. However, real-world applications of LMs may require it to deal with information from different domains. Thus it is important to understand the limitations of these models when employed in different domain settings.

Medium of Expression Whether the language being processed is from a formal email or from an informal utterance on social media, can make significant difference in its syntactic and semantic properties. LMs being deployed in applications that span across different media are thus bound to come across linguistic variations induced due to the medium of expression. This makes it important to understand how LMs perform across different mediums of expression.

Geography and Demography Most languages in the world, including English, have multiple dialects that are influenced by geographic and demographic factors. The applications that are developed using LMs are intended to be used across the world, spanning users belonging to varied demography. It is hence important that the LMs are inclusive towards different forms of language usage and not just cater to a 'standardized' dialect of the language. Hence, it is important to understand how LMs perform across regional and social language dialects.

Time Period Given the high financial and environmental costs of training language models, a single model can be anticipated to be used for long periods of time. Language, however, changes extremely rapidly. Events happening around the world cause constant changes in the vocabulary and semantics of a language. Thus, it is important for LMs to be robust towards new word senses, sentence structures, etc. It is thus necessary to evaluate models on the dimension of Time Period as they are bound to come across language belonging to different points in the history.

The following sections go over each of these

Table 1: Languages and Tasks covered by different datasets and benchmarks

Datasets	Langs.	Tasks
WikiANN	176[2]	NER
UD v2	90	POS, dependencies
XNLI	15	NLI
XQuAD	11	QA
MLDoc	8	Document classification
MLQA	7	QA
PAWS-X	6	Paraphrase identification
Benchmarks[3]		
XTREME	40	NLI, POS, NER, QA, Paraphrase identification, Sentence retrieval
XGLUE	19	NER, POS, QA, NLI, News classification, QA matching, Paraphrase identification, Query-ad matching, Web page ranking, Question generation, News title generation

dimensions to describe the limitations and capabilities of models along each of them.

3 Language

There are over 7000 languages in the world. There is an increased demand for multilingual systems as information technologies penetrate more lives globally. The largest available LMs include mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), and mT5 (Xue et al., 2020) serve 104, 100, and 101 languages respectively. It is clear that they are far from universal in terms of language coverage compared to the number of languages in the world. Further, there is an expectation that massive multilingual LMs will perform equally, or at least reasonably, well on all the languages they serve.

The limited availability of evaluation benchmarks is a major bottleneck in knowing how LMs perform across the languages they are pre-trained on. Table 1 shows that the largest benchmark, XTREME (Hu et al., 2020), covers less than half of the total number of languages that these LMs are trained on. Moreover, other than datasets for syntactic tasks like NER (Rahimi et al., 2019), and POS (Nivre et al., 2016), the largest available semantic task dataset, XNLI (Conneau et al., 2018), covers only 15 languages. Although LMs are tested on individual tasks or languages that may not be covered in these benchmarks, overall, there is a considerable reliance on standard benchmarks to make modelling choices. Thus, how well LMs perform in the untested languages remains unanswered.

There is a great disparity in performance across the languages that are tested through these benchmarks. A general observation is that the performance of low resource languages continues to be lower than high resource languages. The extent to which cross-lingual transfer helps improve performance varies across languages. Studies that empirically support these claims are:

Cross-script transfer is not equally good across languages in mBERT. Ahmad et al. 2019a find that cross-script cross-lingual transfer is effective in the case of Hindi and Urdu, whereas this is not observed between English and Japanese.

Word order differences across language leads to worse cross-lingual transfer (Ahmad et al., 2019a). The correlation between word ordering distance and cross-lingual transfer is found to be high in the experiments by Ahmad et al. 2019b. K et al. 2020 also find that word order has a significant bearing on transfer.

Contrary to common intuition, K et al. 2020 find that shared vocabulary does not affect Universality or generalization across languages considerably. Artetxe et al. 2020a also find that 'effective vocabulary size per language' affects cross-lingual performance rather than joint or disjoint vocabulary of multiple languages.

In massively multilingual LMs, where typically a joint vocabulary across languages is used, languages tend to compete for the allocations in the shared vocabulary. Siddhant et al. 2020 show that increasing number of languages may worsen performance compared to models with fewer languages. This is similar to the findings of Wu and Dredze 2019. Thus, limiting pre-training to only the required languages needed for the downstream tasks may be more beneficial. Conneau et al. 2020a coined the term *"curse of multilinguality"* for this phenomenon and pointed to the trade-off between model performance and language coverage. This result is also shown in MuRIL, a BERT model trained on 17 Indian languages, which outperforms mBERT on the XTREME benchmark significantly across all languages (MuRIL, 2021). Similarly, clustering languages and using different multilingual model for each group, rather than one massive model, gives better performance in Neural Machine Translation (Tan et al., 2019).

Wu and Dredze 2019 observe that mBERT does

[2] Balanced version

[3] Each task may cover only a subset of languages

not transfer well between distant languages. Further, they conclude that while mBERT may perform very well in cross-lingual transfer compared to other models, it still falls short of models that have been trained with cross-lingual signals like bitext, bilingual dictionaries, or limited target language supervision.

Answering whether all languages in mBERT are equally well represented, Wu and Dredze 2020 find that mBERT does not learn high-quality representations for all languages, especially for low resource languages. The bottom 30% languages in terms of data size perform even worse than a non-BERT model for NER. For low resource languages, the combined effect of less data for pre-training and annotated data for fine-tuning compounds together leading to worsening of their performance. On the other end of the spectrum, the top 10% languages are hurt by joint training as mBERT performs worse than monolingual baselines of NER.

To summarize, universality in the language dimension has three levels. At the highest level, the largest models available today span only around 100 of the 7000+ languages globally and thus are far from universal in terms of language coverage. Secondly, out of the languages that these models are trained on, not all of them are evaluated, implying that we do not have enough information to make generalized claims of universality in performance for the languages that the LMs support. Finally, at the lowest level, the performance is not uniform across the tested languages. The performance of lower resourced languages is lower than that of higher-resourced languages. Increasing the number of languages hurts performance at both ends of the spectrum, and cross-lingual transfer is non-uniform and dependent on many factors.

Extensions: Monolingual models learn generalizable representations and can be adapted to new languages without joint training. Conneau et al. 2020b show that the representations learned by monolingual models without any shared vocabulary align with each other and can be adapted to a new language. Similarly, Artetxe et al. 2020b study the transfer of monolingual representations to new languages without using shared vocabulary or joint training. They propose a zero-shot cross-lingual transfer technique where the resultant model is a monolingual LM adapted to a new language. Tela et al. 2020 study adaptation to the extremely low resourced language, Tigrinya. They find that English XLNet generalizes better than BERT and mBERT, which is surprising given that mBERT is trained in multiple languages. Thus, the adaptation of monolingual models may help in extending LMs to new low-resource languages.

Wang et al. 2020 enlarge mBERT's vocabulary and continue pre-training on 27 target languages, out of which 11 are new. They observe performance improvement in zero-shot cross-lingual NER for all 27 languages. The extension benefits both the existing and newly added languages. The drawback is that the base model (mBERT) is biased towards the target languages, downgrading performance on non-target languages.

The data and compute cost of training LMs from scratch poses a major limitation, especially for low-resourced languages. Tran 2020 propose a data and compute efficient technique to circumvent the need of training language-specific models from scratch. They learn target language word-embeddings from an English LM while keeping the pre-trained encoder layer fixed. The English and target language LMs are then both fine-tuned to obtain a bilingual LM. This technique performs better than mBERT and XLM-R on XNLI in 5 of 6 languages with different amounts of resources.

Chi et al. 2020 combine the cross-lingual transfer of a multilingual LM with a task-specific monolingual LM to improve zero-shot cross-lingual classification. The source-language monolingual 'teacher' model provides supervision for the downstream task, and the multilingual model acts as a 'student'. The method outperforms direct multilingual fine-tuning for zero-shot cross-lingual sentiment analysis and XNLI in most of the languages.

Pfeiffer et al. 2020 propose an adaptor based modular framework that mitigates the curse of multilinguality and adapts a multilingual model to arbitrary tasks and languages using language and task-specific adaptors. Their method gives state-of-the-art results for cross-lingual transfer among typologically diverse languages across tasks including NER, causal commonsense reasoning, and QA.

4 Multilingualism

In multilingual communities, several linguistic phenomena lead to variation in language usage. While some of these are well-known and studied, others do not get enough attention. Universal LMs should be able to deal with these phenomena as we can expect them to encounter such forms of language when deployed in user-facing applications.

Code-mixing, or using two more languages in a single utterance, is common in multilingual communities. LMs may not perform optimally in the presence of such code-mixing. Multilingual models like mBERT are not pre-trained with mixed language data, which leaves the model under-prepared for code-switched settings resulting in sub-optimal performance (Khanuja et al., 2020). This can be overcome to a certain extent by training using other data, such as social media, but it is unlikely to cover all the forms of code-switching produced by multilinguals.

Romanization of languages to the Latin script has increased with the advent of digitization of communication. While some models like XLM-R (Conneau et al., 2020a) and MuRIL (MuRIL, 2021) use romanized versions of some languages in training, the effectiveness of these LMs on Romanized (or more generally transliterated) text is still unclear.

Diglossia is a kind of multilingualism where a single community uses a substantially different language dialect in different communication settings (Ferguson, 1959). Since data from the internet is used in training, it is likely that LMs cannot handle diglossia. However, there are no studies that concretely prove (or disprove) this.

To summarize, apart from code-mixing, there has been very little work in recognizing, understanding, or improving LMs for different phenomena arising due to multilingualism, making the dimension under-represented in the study of LMs.

Extensions: Efforts have been made to improve LMs, particularly mBERT to deal with code mixed data. Khanuja et al. 2020 present a modified version of mBERT which performs better than standard mBERT in English-Hindi and English-Spanish code mixed data using synthetically generated code-mixed data for continued pre-training.

5 Task

NLP applications range from syntactic tasks, like POS, NER, etc. to semantic tasks like NLI, QA, etc. LMs learn task-agnostic representations and can be fine-tuned with task-specific data or used in task-specific architectures as features. Thus, Universal LMs should adapt well to a wide variety of tasks.

Like languages, the extent of evaluation on different NLP tasks is constrained by the availability of benchmarks that span various tasks. As shown in Table 1, only a small fraction of the large number of tasks studied in NLP are evaluated by benchmarks. While there are many other task-specific datasets, the success of LMs is associated with performance on these benchmarks rather than a wider variety of tasks.

Universal Language Model Fine-tuning for Text classification (ULMFiT) uses discriminative fine-tuning, gradual unfreezing of layers, and slanted triangular learning rates for target-specific fine-tuning, giving better performance on multiple tasks (Howard and Ruder, 2018). This work also explicitly defines the term 'universal', in their context as referring to – applicable to all tasks in text classification, using a single training process and architecture, usable without feature-engineering, and not requiring additional in-domain data.

Masked language modeling (MLM) is the most generalizable pre-training objective for the extent of transfer among twelve pre-training objectives for nine target tasks (Liu et al., 2019).

Data size is important in effective pre-training of LMs (Liu et al., 2019) but transfer gains between source and target tasks are also possible with smaller source datasets (Vu et al., 2020b).

Similarity between source and target tasks is important for transfer gains. Liu et al. 2019 find that closeness in pre-training objective and target task is important for transfer. Peters et al. 2019 find that while feature extraction and fine-tuning of LMs give similar performance, exceptions occur when the source and target tasks are either very similar or very dissimilar. Vu et al. 2020b also find that similarity is important, especially in low-resource scenarios, but, exceptions of transfer gains between dissimilar tasks are possible.

To summarize, LMs are universal in the task dimension owing to task-specific architectures or fine-tuning. However, the success of LMs is often associated with few benchmarks which cover limited tasks. The similarity between tasks, data size, and pre-training objectives are keys to transfer gains, which are important for Universality.

Extensions: Pattern Exploiting Training (PET) (Schick and Schütze, 2020a) reformulates tasks as cloze questions,[4] making them the same as the MLM objective, requiring less data with no additional fine-tuning or a task-specific architecture to achieve remarkable zero-shot and few-shot perfor-

[4] Cloze questions are statements with exactly one masked token.

mance. Using PET with ALBERT few-shot performance competitive to GPT-3, which is 780 times larger in terms of the number of parameters compared to ALBERT, is obtained (Schick and Schütze, 2020b). The recently introduced T5 model (Raffel et al., 2020) leverages a text-to-text framework to enabling a single architecture to perform multiple tasks and achieving state-of-art results. These are concrete steps to enable universality towards tasks, as a single model can be built to generalize across solving multiple tasks.

6 Domain

NLP models are applied to many real-world applications in different domains like medical, scientific, legal, financial etc. A universal model in this dimension should adapt to different domains or scenarios without loss of performance.

Processing domain-specific language often requires the processing of specialized vocabulary and language usage. Even though LMs learn some implicit clusters of domains (Aharoni and Goldberg, 2020), this may not be enough and specialized domain-specific LMs are needed (Lee et al., 2020; Chalkidis et al., 2020; Beltagy et al., 2019; Huang et al., 2019; Araci, 2019; Gu et al., 2020).

Despite the success of general domain LMs, it is found that pre-training LM on in-domain data improves performance across high and low resource settings (Gururangan et al., 2020).

Lee et al. 2020 introduced BioBERT, learned using continued pre-training of BERT on medical text, which performed better than BERT on biomedical tasks. In contrast, Gu et al. 2020 introduce PubMedBERT and challenge the benefits of out-of-domain data in pre-training by showing that pre-training LM from scratch on in-domain data (if available) is better than mixed or continual pre-training. Huang et al. 2019 propose the clinicalBERT model that is pre-trained on clinical notes text corpora, which learns better relationships among medical concepts and outperforms general domain LMs in clinical tasks.

Chalkidis et al. 2020 introduce Legal-BERT and observe that continual pre-training of BERT or training it from scratch with legal data both perform similarly and significantly better than using BERT off the shelf. Beltagy et al. 2019 release SciBERT, trained from scratch on scientific publications giving better performance on scientific tasks. Araci 2019 use domain adaptation and transfer learning to develop FinBERT, achieving state-of-the-art performance in the financial domain.

Performance of domain-specific LM can degrade on general domain tasks (Gu et al., 2020; Xu et al., 2020; Thompson et al., 2019; Rongali et al., 2020). This phenomenon is also known as catastrophic forgetting and prevents the LM from being truly universal.

To summarize, LMs are not universal in the domain dimension, and different domain-specific LMs have been introduced to cater to this requirement. Domain-specific LMs are either trained from scratch or by mixed or continual pre-training of existing LMs. While none of these techniques are clear winners, performance degradation in general domain tasks is observed in many cases.

Extensions: Vu et al. 2020a study adversarial masking strategies to learn specific target domain vocabulary along with continual pre-training by carefully selecting tokens to be masked, leading to better domain adaptation performance across multiple source and target domains.

MuTSPad (Multi-Task Supervised Pretraining and Adaptation) (Meftah et al., 2020) leverages hierarchical learning of a multi-task model on high-resource domain followed by fine-tuning on multiple tasks on the low-resource target domain.

Ben-David et al. 2020 extend the pivot-based transfer learning to transformer-based LMs by developing PERL (Pivot-based Encoder Representation of Language), that uses continual pre-training with MLM to learn representations that reduce the gap between source and target domains followed by fine-tuning for the downstream classification task. The pivot is selected such that the source and target domain labels have greater mutual information to facilitate a good transfer.

Jiang and Zhai 2007 propose several heuristics like removing misleading examples from the source domain, assigning more weight to target domain instances, and augmenting target training instances with predicted labels for better domain adaptation from a distributional perspective.

Various methods that are computationally efficient (Poerner et al., 2020), use more effective adversarial training (Ma et al., 2019), and reduce the requirement of annotated data in low-resource settings (Hazen et al., 2019) have been proposed for computation and data efficient domain adaptation.

Rongali et al. 2020 overcome catastrophic forgetting, out-performing domain-specific LMs while maintaining performance on general domain tasks.

7 Medium of Expression

Language varies with the medium of expression. There are syntactic and semantic alterations like the use of ill-formed sentence or grammatical structure, inflections, slang, compressions, and abbreviations due to limited space, familiarity with the audience, and communicative intent. Universal LMs should be robust to such variations. Often these specialized settings are simply treated as a different domain (Qudar and Mago, 2020; Nguyen et al., 2020). However, in this work, we treat 'domain' as specialized fields of expertise. The current discussion pertains to the medium in which language is used.

Language in social media and texting may not follow conventions of written language. Sentences are often shorter or grammatically ill-formed and may not be coherent enough for LMs that rely on contextual information (Eisenstein, 2013; Han et al., 2013; Choudhury et al., 2007).

LMs perform sub-optimally on non-standard language as compared to specialized LMs. BERTweet (Nguyen et al., 2020), a Roberta-based LM trained on English tweets, outperforms both RoBERTa and XLM-R base models even though RoBERTa and XLM-R use 2 times and 3.75 times more data, respectively. TweetBERT (Qudar and Mago, 2020), another LM trained on tweets outperforms seven general domain LMs.

To summarize, the language used in different media of expression is substantially different. The limited amount of investigation in this direction indicates that LMs are not universal to these variations and perform sub-optimally on the language used in social media and texting.

Extensions: Dai et al. (2020) propose cost-effective training by appropriate selection of additional data for training a LM from a Twitter corpus.

8 Geography and Demography

Standard and non-standard dialects, both social and regional, lead to varied word, and language use (Labov, 1980; Milroy, 1992; Tagliamonte, 2006; Wolfram and Schilling, 2015; Nguyen et al., 2016). Regional dialect refers to the varied usage of the same language across different places. For example, the use of 'wicked' can refer to bad or evil ("he is a wicked man"), or as an intensifier to adverbs ("my son is wicked smart") (Bamman et al., 2014a; Kulkarni et al., 2016). Sociolects, or social dialects, similar to regional dialects, are language dialects dependant on social variables like age, race, gender, socio-economic status, ethnicity, etc. (Nguyen et al., 2016). Geographic and demographic variations stem from grammatical, phonological, syntactic, lexical, semantic features, or any combinations, making it difficult to capture and evaluate.

Since LMs are trained on standard language dialects, non-standard dialects spoken by millions of people are largely ignored. Such a system can result in bias against specific cultural or geographical communities in user-facing applications leading to ethical implications in building fair NLP systems.

A Universal LM should be sensitive to semantic shifts arising from its users' demographic or geographical diversity. Taking these variations into consideration has resulted in improved performance and personalization of applications like conversational agents, sentiment analysis, word prediction, cyber-bullying detection, and machine translation (Östling and Tiedemann, 2017; Rahimi et al., 2017; Mirkin and Meunier, 2015; Hovy, 2015; Hovy and Søgaard, 2015; Stoop and van den Bosch, 2014; Volkova et al., 2013).

Kulkarni et al. 2016 present a novel approach to quantify semantic shift that is statistically significant across geographical regions and propose a new measure of dialect similarity to establish how close the language in two regions is. Demszky et al. 2020 focus on Indian English and show that dialect features can be learned given very limited data with strong performance.

In the 2020 VarDial evaluation task for Romanian Dialect Identification, an SVM ensemble based on word and character n-grams outperformed fine-tuned Romanian BERT model. These results are consistent with the earlier evaluation results of VarDial where shallow models outperformed deep models. In Social Media Variety Geolocation, predicting geolocation (coordinates) from text, the best performance was obtained by a BERT architecture with a double regression classification output. In contrast, the next two best models were both shallow. (Gaman et al., 2020).

Demographic features like age, gender have been predicted from language usage Peersman et al. (2011); Morgan-Lopez et al. (2017), whereas social class or ethnicity receive less attention (Mohammady Ardehaly and Culotta, 2015). Prediction of demographic features from language use quantifies the correlation between social variables and social dialects. While individuals may intentionally or naturally digress from such conventions, these statistical patterns are a cornerstone for studying the interaction between society and language in computational sociolinguistics (Nguyen et al., 2016).

There are some worrisome findings of bias of performance across age, ethnicity, or gender in contextual LMs. (Hovy and Søgaard, 2015) find that a POS tagger trained on the English Penn treebank performed better on texts written by older authors. Tan et al. 2020 show bias against nonstandard English (in this case Singaporean English) in BERT. Bhardwaj et al. 2020 expose gender bias in BERT by showing that the model assigns lower (or higher) scores consistently for sentences that contain words indicating gender in cases where gender information should have no bearing.

To summarize, the influence of geography and demography on language usage is well known, and LMs must be sensitive and inclusive of such variation. However, there has been limited, albeit now growing, attention to these factors. In some cases, shallow models have outperformed deep models in recognizing semantic shifts, and there is evidence of bias against particular social groups.

Extensions: Bamman et al. 2014b learn language representations that take geographical situations or variations into account by enriching Vector space word representations (word2vec) with geographical knowledge from metadata about authors. Hovy and Purschke 2018 employ retrofitting for including geographic information to capture regional variation in continuous regional distribution and at a fine-grained level using online posts in German and the corresponding cities of their authors as labels to create document embeddings. While these techniques do not involve any contextual LMs, such representations and retrofitting can be extended to contextual LMs.

Tan et al. 2020 propose an adversarial approach to make models like BERT more robust to nonstandard forms of English using inflectional morphology perturbations.

Debiasing techniques such as the ones studied in (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019) can remove gender stereotypes from pretrained word embeddings.

9 Time Period

Language evolves continuously, and individual word meanings can change significantly over the years (Cook et al., 2014; Kim et al., 2014; Hamilton et al., 2016). Most of the data used in LM pre-training is from the late 20th century. Thus, whether these models can handle word senses across timescales is a pertinent question. Universal LMs should appropriately deal with language with a nuanced understanding of diachronic semantic change (DSC) because, when deployed in downstream applications, such variations may be encountered and misinterpreted. Some of the studies we mention below are not strictly focused on contextual LMs. Nevertheless, we find it important to note such research as we hope it can be extended to contextual LMs in the future.

The intensity of the change of meaning of different words is different – some are more subtly changed than the others. Kim et al. 2014 trace a period from 1900-2009, obtain year-specific word embeddings on the Google Books N-grams corpus, and pinpoint the extent and time-period of occurrence of semantic shift.

Hamilton et al. 2016 evaluate static word embeddings for known historical changes using corpora spanning four languages and two historical periods. They create diachronic embeddings by learning separate representations across the time periods followed by alignment over different time scales.

DSC can be detected by clustering word sensed. Mitra et al. 2014 organize words in a time-period specific graph where its nearest neighbors are cooccurring words, and word-senses are clustered. The shift in word sense or the emergence of a new word sense can be identified by the change of cluster for a particular word. Giulianelli et al. 2020 perform clustering over usage types in BERT and use the contextual property of LM to quantify semantic change instead of relying on a specific set of word senses.

DSC evaluation lacks standardization. Schlechtweg et al. 2019 perform a large scale evaluation on German, revealing the best set of parameters for optimal performance, compare various state-of-the-art methods, and outline improvements for better performance.

Shallow models can outperform contextual LMs in identifying semantic shift. Schlechtweg et al. 2019 show that a shallow skip-gram model with negative sampling, orthogonal alignment, and cosine distance performs best in identifying DSC in German. Kaiser et al. 2020 reconfirm this by using a similar model to obtain the first position in the DIACR-Ita shared task (Basile et al., 2020) on Italian DSC. Similar findings of only limited success in contextual LMs are reported in the shared task on Unsupervised Lexical Semantic Change Detection in 5 languages hosted at SemEval 2020 (Schlechtweg et al., 2020).

While the success in identifying these shifts may be limited, (Rodina et al., 2020) find that DSC identified by contextual LMs can have a strong correlation with human judgment of change.

Within contextual LMs, BERT and ELMo perform similarly for Russian. Rodina et al. 2020 show that neither BERT nor Elmo outperform each other when fine-tuned using historical text in Russian to detect semantic change. Moreover, results from the shared task on Unsupervised Lexical Semantic Change Detection in 5 languages hosted at SemEval 2020 (Schlechtweg et al., 2020) show that systems performing well over one language may not perform as well for other languages.

To summarize, time period is under-studied and there is little understanding of whether contextual LMs can handle such nuanced language variation. For the closely related task of DSC, shallow models can outperform deep LMs, and performance can vary greatly across languages.

Extensions: Rudolph and Blei 2017 develop dynamic word-embeddings with an attribute of time that captures the semantic shifts in word meanings in sequential historical data on top of Bernoulli embeddings such that representations are shared within specific time periods rather than throughout the corpus. Similarly, Bamler and Mandt 2017 use timestamped data to build static probabilistic representation for tracing semantic change.

To mitigate the problem of using different representations of words over different time periods, Hu et al. 2019 propose a framework for tracking and representing word senses by leveraging pre-trained BERT embeddings and Oxford dictionary data to learn fine-grained senses.

10 Conclusion

Deep Contextual LMs are being applied today to various different applications due to their perceived 'Universality'. In this work, we attempt to holistically define 'Universality' to encompass a wide variety of scenarios and linguistic phenomena.

We define Universality using seven dimensions: Language, Multilingualism, Task, Domain, Medium of Expression, Geography and Demography, and Time Period. These dimensions result in unique variations in language usage that are commonly encountered in real-life scenarios. We aim for this definition to be sound rather than complete. That is, a model should strive to achieve Universality in these dimensions, but they are in no way a complete, exhaustive list of everything the model needs to be capable of.

We survey research across all the dimensions and find that: First, while dimensions like language, task, and domain are more widely studied, other dimensions, especially multilingualism, geography and demography, and time period receive less attention. Second, limited evaluation benchmarks constrain the complete understanding of capabilities even in the more studied dimensions. Third, language variation arising in specific scenarios of demography, geography, time period, multilingualism, and medium of expression is often studied in an isolated manner.

The dimensions we survey are a starting point that LMs can aim to be inclusive towards in order to serve a diverse set of users and scenarios. Large contextual LMs may not be the optimal choice for all scenarios, with shallow, task-specific models sometimes leading to better outcomes. Overall, 'Universality' is yet to be fully understood, studied, and achieved. We hope that this work will lay the foundation to understanding the capabilities and limitations of LMs and spur further research into making models more inclusive and fair.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019a. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019b. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020a. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389. PMLR.

David Bamman, Chris Dyer, and Noah A. Smith. 2014a. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

David Bamman, Chris Dyer, and Noah A. Smith. 2014b. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. Diacr-ita@ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org*.

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.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2020. Investigating gender bias in bert.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

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.

Zewen Chi, Li Dong, Furu Wei, Xianling Mao, and Heyan Huang. 2020. Can monolingual pretrained models help cross-lingual classification? 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 12–17, Suzhou, China. Association for Computational Linguistics.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3-4):157–174.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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.

Dorottya Demszky, Devyani Sharma, Jonathan H Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2020. Learning to recognize dialect features. *arXiv preprint arXiv:2010.12707*.

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.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, 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.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):1–27.

Timothy J Hazen, Shehzaad Dhuliawala, and Daniel Boies. 2019. Towards domain adaptation from limited data for question answering using deep neural networks. *arXiv preprint arXiv:1911.02655*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. 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 483–488, Beijing, China. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. *International Conference on Machine Learning*, pages 4411–4421.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. Op-ims@ diacr-ita: Back to the roots: Sgns+ op+ cd still rocks semantic change detection. *arXiv preprint arXiv:2011.03258*.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vivek Kulkarni, Bryan Perozzi, Steven Skiena, et al. 2016. Freshman or fresher? quantifying the geographic variation of language in online social media. In *ICWSM*, pages 615–618.

William Labov. 1980. *Locating language in time and space*. Academic Press New York.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Sara Meftah, Nasredine Semmar, Mohamed-Ayoub Tahiri, Youssef Tamaazousti, Hassane Essafi, and Fatiha Sadat. 2020. Multi-task supervised pretraining for neural domain adaptation. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 61–71, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

James Milroy. 1992. *Linguistic variation and change: On the historical sociolinguistics of English*. B. Blackwell.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

Ehsan Mohammady Ardehaly and Aron Culotta. 2015. Inferring latent attributes of Twitter users with label regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195, Denver, Colorado. Association for Computational Linguistics.

Antonio A Morgan-Lopez, Annice E Kim, Robert F Chew, and Paul Ruddle. 2017. Predicting age groups of twitter users based on language and metadata features. *PloS one*, 12(8):e0183537.

MuRIL. 2021. Multilingual Representations for Indian Languages. https://tfhub.dev/google/MuRIL/1. Accessed: 2021-01-29.

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.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for 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. 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.

Mohiuddin Md Abdul Qudar and Vijay Mago. 2020. Tweetbert: A pretrained language representation model for twitter text analysis. *arXiv preprint arXiv:2010.11091*.

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*, 21:1–67.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. 2020. Elmo and bert in semantic change detection for russian. *arXiv preprint arXiv:2010.03481*.

Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Improved pretraining for domain-specific contextual embedding models. *arXiv preprint arXiv:2004.02288*.

Maja Rudolph and David Blei. 2017. Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *AAAI*, pages 8854–8861.

Wessel Stoop and Antal van den Bosch. 2014. Using idiolects and sociolects to improve word prediction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 318–327, Gothenburg, Sweden. Association for Computational Linguistics.

Sali A Tagliamonte. 2006. *Analysing sociolinguistic variation*. Cambridge University Press.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: The case of tigrinya. *arXiv preprint arXiv:2006.07698*.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Thuy-Trang Vu, Dinh Phung, and Gholamreza Haffari. 2020a. Effective unsupervised domain adaptation with adversarially trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6163–6173, Online. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020b. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

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

Walt Wolfram and Natalie Schilling. 2015. *American English: dialects and variation*. John Wiley & Sons.

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

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

Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. 2020. Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.