Georgia Tech ML Qualification Exam (2022) 1-10 Presented 11/30; Submitted 12/13

Unsupervised Transfer for Low-Resource Dialects and Language Varieties

William Held WHELD3@GATECH.EDU
School of Interactive Computing
Georgia Institute of Technology

Abstract

This work explores the background and techniques which form the foundations of my research to build inclusive language technology for low-resource languages and dialects through unsupervised transfer learning. My focus is on *written* language varieties that heavily overlap with Standard American English (SAE), but contain lexical and morphosyntactic variation stemming from sociolinguistic barriers and contact with foreign languages. First, I will describe the work which offers motivation and explores the ethical impacts which guide my research. Then, I will cover the foundational techniques from Multilingual Natural Language Processing which provide insights into paths forward: Explicit Embedding Alignment and Massively Multilingual Pretraining. Third, I will explore existing work on Dialect Natural Language Processing. In conclusion, I will analyze how I believe these works together present a possible path forward for dialectal NLP and how it differs from the one present in today's approaches.

Keywords: Low-Resource NLP, Unsupervised Learning, Inclusive Technology

1. Introduction

The techniques which drive modern Natural Language Processing (NLP) systems have been developed alongside large swathes of annotated data across a broad range of tasks. This reliance on high-cost supervision has concentrated NLP development to 7 out of 7,000 global languages (Joshi et al., 2020). Even within English dialects, current models often explicitly filter everything but Standard American English (SAE) (Gururangan et al., 2022) which leads to downstream failures for speakers of English dialects both abroad (Jurgens et al., 2017) and domestically (Ziems et al., 2022). These discrepancies are a fundamental issue for the development of fair and inclusive NLP systems, as linguistic discrepancies often correlate with protected demographics (Hovy and Spruit, 2016).

Since the distribution of language speakers follows a power law, meaning that most languages have very few speakers, developing universal language technology is prohibitively expensive with the data requirements of existing approaches. Therefore, to enable systems which are both broadly capable and inclusive of the vast diversity of language; future NLP systems will rely increasingly on models which enable knowledge transfer from high-resource environments to low-resource environments. My research focuses on methods in pursuit of this goal by studying how semi-supervised learning, supervised learning with limited data, and unsupervised learning can be applied to learn patterns which are transferable to low-resource domains. My work is driven by impacts (Joshi et al., 2020; Bird, 2022; Hovy and Spruit, 2016), grounded in prior multilingual work on embedding alignment (Grave et al., 2019; Adams et al., 2017; Artetxe et al., 2018) and massively multilingual semisupervised

pretraining (Conneau et al., 2020; Conneau and Lample, 2019; Zoph et al., 2016; Xue et al., 2021), and aims to inform new machine learning techniques using the linguistics-rich knowledge of prior dialectal work (Jørgensen et al., 2016; Blevins et al., 2016; Ziems et al., 2022).

2. Motivations & Impacts

My core research motivations are the ethical, social, and practical impacts of reducing the data requirements of NLP specifically for language varieties and dialects. Hovy and Spruit (2016) explores the harms caused by the status quo of Machine Learning driven NLP and its focus on the ways that the primarily western, educated, industrialized, rich, and democratic users of language technology today speak. Joshi et al. (2020) informs this argument with concrete details, especially surrounding the availability of data resources. Finally, Bird (2022) presents an argument for why language varieties and dialects may be of greater importance for inclusive NLP than other more dominant low-resource scenarios. Together, make the case that inclusive and broadly applicable NLP requires models which incorporate regional and social language variability.

2.1 The Social Impact of Natural Language Processing

Beyond academic interest, Hovy and Spruit (2016) presents a clear case that designing models to account for the diversity of language is essential to include diverse users in language technology applications.

Unlike images, language does not capture a physical signal directly and is instead produced by individuals. This means that each utterance is grounded in the identity of the speaker, which the authors refer to as *situatedness* (Bamman et al., 2014). Prior work has shown that not only can NLP systems predict aspects of identity (Volkova et al., 2015) from lexical features alone, but their performance varies across identity boundaries (Mandel et al., 2012). The situatedness of language, the sensitivity of machine learning models to domain shift, and the data snooping on available benchmarks leads to **exclusion** and **underexposure** problems.

Machine learning assumes that each problem is, at its core, represented as a universal function $f : X \to Y$. While Y is usually clear as part of the task, X is often unclear and selected by the practitioner. The situatedness of language indicates that, for almost all tasks, the identity of a speaker and especially their dialect are a key input to our true function. Probably Approximately Correct learning, which underlies all modern machine learning theory (Haussler, 1990), requires that our training data is sampled from X according to the same distribution D that it will be used on. However, in practice, training data is almost never sampled with minority dialects or language variants in mind. This **exclusion** leads to poor results and demographic bias against those speakers.

Beyond the biases of data, the authors argue researchers themselves are subject to cause bias over time due to **underexposure**. While modern NLP often omits explicitly encoded linguistic knowledge this does not mean that it is truly language agnostic, as even simple n-gram models perform more poorly under different linguistic systems (Bender, 2011). These biases embedded in the models themselves have been carried on to neural counterparts (Ravfogel et al., 2018; Ahmad et al., 2019). Dialect usage is especially impacted by the challenge of underexposure since both data and social biases lead many NLP researchers

to disregard such variation as "bad" English, which does not need to be accounted for in model design.

Hovy and Spruit (2016) primarily argues for the existence these challenges from a philosophical and observational standpoint. While the existence of these disparities and causes is apparent to those in the field, the work itself does not provide strong quantitative evidence and measurements of the data and design disparities it discusses, which is tackled in the next work.

2.2 The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Joshi et al. (2020) provides a taxonomy and quantitative analysis in support of the problems introduced by Hovy and Spruit (2016). While they omit discussions of language varieties and dialects, they provide a useful lens to understand the ways research advances within NLP. First, they provide a taxonomy of language resource distributions, separating languages into 5 classes ranging from languages which have minimal labeled and unlabeled data to those which have a wealth of both, such as Standard American English. They then study the typographic features of languages which are in the lowest resource classes of 0 and 1. Finally, they use 3 quantitative analysis techniques to understand the research trends surrounding linguistic diversity in NLP conferences.

The first section of the paper quantifies and categorizes **exclusion** from existing data resources using a taxonomy. The resource taxonomy is split into two dimensions: unlabeled data and labeled task-specific data. Unlabeled data is frequently treated as a given in NLP research, however for many languages without strong internet presence or written traditions even unlabeled data resources are lacking. In a decision which limits the applicability of this analysis to dialects, the authors use the number of Wikipedia pages in the language as a metric of unlabeled data. Notably, while Wikipedia is commonly used for language modeling, it is a formal and edited medium which may under-represent the occurrence of languages and dialects which see frequent casual use.

For labeled data, they use the Linguistic Data Consortium[1]and the European Language Resources Association[2] catalogues. While incomplete, these language resource catalogues are more dialect inclusive than the unsupervised data metric, with non-English dialect benchmarks being hosted on each(Kilany, Hanaa et al., 1997; Iskra et al., 2004). The classes are as follows: class 0 has almost no data of either category, class 1 has only small amounts of unlabeled data, class 2 has small amounts of both labeled and unlabeled data, class 3 has lots of unlabeled data but little labeled data, classes 4 and 5 both data types with class 5 differentiated by a greater degree of labeled data.

For our purposes, dialects are most strongly associated with class 3, termed "Rising Stars". While they are poorly represented on Wikipedia, unlabeled data is increasingly available through online social media (Blodgett et al., 2016) but have a dearth of labeled data for downstream tasks. The term "Rising Star" highlights the reason my research is timely, as recent trends of semi-supervised and unsupervised transfer methods are most applicable to this category. Dialects of high-resource languages have a second advantage over existing class 3 languages - namely a high degree of typological similarity with the

1. https://catalog.ldc.upenn.edu/
2. http://catalog.elra.info/en-us/

dominant dialect. The work highlights that typological similarity yields additional support for the effectiveness of unsupervised zero-shot transfer.

Finally, the paper shows that the easily measured data imbalances do lead to researcher **underexposure** in the form of measurably lower interest and diverging modeling approaches. The authors first sort each language l in year y of conference c in decreasing order using the number of papers which mention it $M_{(c,l)} = \sum_{y=1950}^{2020} M_{(c,l,y)}$. The Mean Reciprocal Rank of a class of languages is used to quantify interest in that class. Barring the International Workshop on Semantic Evaluation (SEMEVAL), the most underexposed languages are the most excluded from data resources.

However, the authors show that interest in linguistic diversity is increasing across the NLP community. Using the counts from above, the authors use the language entropy of a conference $e_{(c,y)} = \sum_{l \in L} p_{(c,y)}(l) \log p_{(c,y)}(l)$ as a metric of diversity. The probability is estimated using a frequentist approach, with $p_{(c,y)}(l) = \frac{M_{(c,l,y)}}{\sum_{l \in L} M_{(c,l,y)}}$. This metric shows increasing interest in studying linguistic diversity in recent years, especially for newly introduced conferences with less established norms. However, this interest is only measured across discrete languages, not for variation within them. As the next work cites, this focus on new languages can be academically interesting, while focusing on challenging problems where user benefits are not necessarily maximized.

2.3 Local Languages, Third Spaces, and other High-Resource Scenarios

Bird (2022) argues that current targets of research focus on low-resource languages often overlooks areas where the same populations could be better impacted. Amongst other points, the work uses language ecology (Haugen, 2001) to highlight that dialects and language variations are perhaps the most impactful research areas to address empirical harms of **exclusion** and **underexposure** due to the realities of language use by many low-resource language speakers.

While traditional low-resource language research focuses on a *language-centric* world view, the practical impacts shown in Hovy and Spruit (2016) are *community-centric*. Bird (2022) encourages researchers to look at the communities instead, which are more often than not already multilingual (Grosjean, 2021). Within these communities, language use is often diglossic - a low-register is used for intra-community communication while a standardized high-register language is used for external communication.

Current zero-shot approaches often focus on the lower-resourced intra-community languages due to the interesting challenges they present. However, in a community-centric view, research is more effectively aimed at the languages used by a community for external communication which are often dialects, creoles, and codeswitched variations of much higher resource languages. In many low-resource situations, speakers "will not be hampered by the lack of language technology in their local language, but by the lack of support for their variety of the contact language" (Bird, 2022). Despite a higher degree of similarity, this is often true for English dialects "simply because local spoken varieties of English are still not well supported" (Bird, 2022).

This community-centric reasoning ties together my vision with a strong reason to focus on dialects and language variants, rather than low-resource languages broadly. The advances of multilingual NLP hold a wealth of technical insights and my work is to understand how

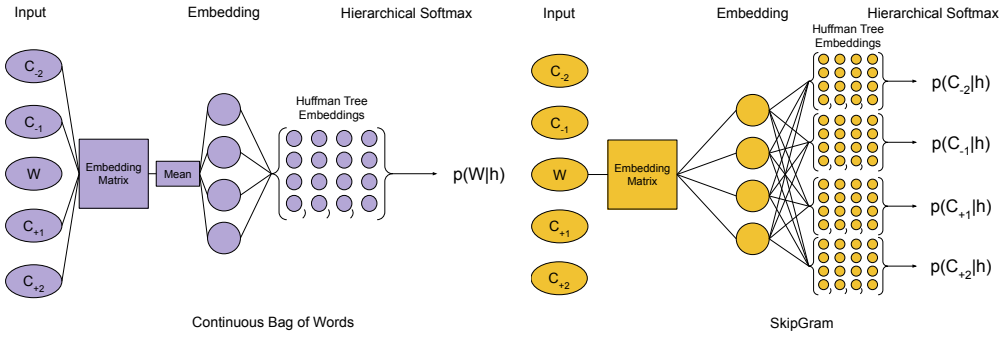

Figure 1: The two single-layer neural network embedding setups from Mikolov et al. (2013).

to extend these insights to language variants where they can benefit the most excluded speakers in the most direct and impactful way.

3. Transfer Learning Techniques from Multilingual NLP

Cross-lingual transfer learning has long aimed to learn common structures which are transferable across languages to support this vision. In some ways, dialectal and regional variation have stronger ties to transfer learning, as many speakers' ability to understand other varieties (Sebba, 1997) indicates a high degree of transferability. While early works often exploited commonality through the use of discrete pivot representations, created either by translation (Mann and Yarowsky, 2001; Tiedemann et al., 2014; Mayhew et al., 2017) or language-agnostic task formulations (Zeman, 2008; McDonald et al., 2011), modern methods often use the embedding space of deep neural networks as a continuous pivot representation.

Two major lines of methodology which have led to significant advances in transfer learning hold potential for low-resource varieties and dialects: embedding space alignment and semi-supervised multilingual pretraining.

3.1 Embedding Alignment

Embedding alignment aims to explicitly produce a pivot representation in embedding space with the goal of making monolingual models applicable to many languages. Adams et al. (2017) first illustrated that alignment methods driven by bilingual lexicons were applicable to low-resource languages and in fact improved performance on downstream language modeling tasks. In parallel, Artetxe et al. (2018) and Grave et al. (2019) introduced methods which allowed for high quality alignment to be done without seed lexicons with novel initialization procedures and self-learning. These alignment methods often perform better on high similarity between languages, making them well suited for dialects and language varieties.

3.2 Embedding Background

Each of the following works depends on an understanding of the SkipGram and Continuous-Bag-Of-Words Word2Vec methods of Mikolov et al. (2013). In Figure 1, we illustrate each variant. In both setups, a center word W is selected with a surrounding context window $C_w = \{C_{-R}, C_{-R+1} \dots C_{+(R-1)} C_{+R}$ of size R. In Continuous Bag-Of-Words, the hidden

word W is predicted using the average embedding of C_w. For SkipGram, each word in C_w is predicted using the embedding for W.

Traditionally, the probabilities would be predicted using the softmax of the inner product with our embedding and the embedding of each candidate: $P(c) = \sigma(h_w^\top v_c)$. However, for dimensionality D and vocabulary size V the $O(DV)$ runtime becomes intractable for large vocabularies. Therefore, in each variant, the probability is instead computed using the hierarchical softmax, which breaks the probability computation down as the walk to reach the word at depth $L(c)$ through a Huffman Coding Tree. This probability is computed through a series of inner products with inner node embeddings $v_{n_{(c,i)}}$: $p(c) = \prod_{i=0}^{L(c)-1} \sigma(h_w^\top v_{n_{(c,i)}})$. This reduces the complexity to $O(H \log(V))$.

With either variant, the resulting embeddings are dense representations which appear to effectively capture the meaning of each word. Words with similar meanings are neighbors and the space has a geometry which allows for semantic math (Mikolov et al., 2013). However, since the points themselves are only anchored to other embeddings from the same training run. This means that embeddings trained on different languages, or even simply different corpora of the same language, are not interchangeable. Each of the following works addresses this challenge.

3.2.1 Cross-Lingual Word Embeddings for Low-Resource Language Modeling

Traditionally, Word embedding methods learn from large corpora of unlabeled data in a semi-supervised fashion. Adams et al. (2017) aims to use bilingual dictionaries to produce cross-lingual word embeddings with the express goal of aiding transfer to languages where such large corpora do not exist.

This work builds on top of Duong et al. (2016), which performs dynamic data augmentation of CBOW using a bilingual dictionary. During training, if the center word W is an entry within the bilingual dictionary an additional data point is added where W is replaced with its synonym from the bilingual dictionary. To account for polysemy across languages, the authors only replace W with the synonym with the highest cosine similarity at that point in training. This simple data augmentation technique ties the two embedding spaces together such that even words which do not occur in the bilingual dictionary are drawn near to similar words in the paired language.

In Adams et al. (2017), they evaluate this technique in a setting where a high-resource source language is paired with a low-resource target language with varying sizes of unlabeled data. The authors show that, when paired with 5 million examples from a high-resource language, cross-lingual word embeddings are able to yield similar correlations with human judgement of word similarity with 10 times less data.

However, this method contains several constraints which make it difficult to apply in the context of dialect. While bilingual dictionaries are common for across the standard dialects of different languages, they are uncommon within dialects of English. Secondly, the technique requires an intervention during training time. As NLP has moved increasingly towards pretrained models, this makes the intervention costly in the modern day. To address both of these challenges, we study two approaches for fully unsupervised embedding alignment which follow the self-learning algorithm in Algorithm 1.

Algorithm 1 A Generic Algorithm for Self-Learning of Embedding Alignment

$X \leftarrow$ compute_and_norm_embeddings($lang_x$)
$Y \leftarrow$ compute_and_norm_embeddings($lang_y$)
$D \leftarrow$ initialize_dictionary(X, Y)
for $t \in T$ **do**
  $Q \leftarrow$ compute_mapping(X, Y, D)
  $Q \leftarrow$ orthogonalize_mapping(Q)
  $D \leftarrow$ compute_new_dictionary(X, Y, Q)
end for
return Q

3.2.2 A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings

Artetxe et al. (2018) augments prior minimally supervised work (Artetxe et al., 2017) with a novel unsupervised initialization and a set of regularization procedures to overcome noisy local minima in embedding alignment.

First, they provide a dictionary initialization technique which makes a strong assumption of isometry. To understand this, we will expand on how isometry naturally leads to their initialization. Let X be the embedding matrix for a source language and let Y be an embedding matrix for a target language we are mapping onto. If our mapping Q is isometric, then for any two vectors X_i and X_j, the following is true for any metric d: $d(QX_i, QX_j = d(X_i, X_j)$. If we then look at the self similarity matrix $X' = XX^\top$, we see that for each entry $X'_{(i,j)} = d(X_i, X_j)$ and the same is true for $(QX)' = d(QX_i, QX_j)$. Therefore, $X' = (QX)'$ as long as Q is isometric.

If we then assume that there is some isometric mapping Q such that $DY = QX$, then we know that $(DY)' = X'$ without solving for Q. Therefore, each row Y_i is simply a shuffled version of a row in X. To resolve this, each row is sorted independently and then rows sorted by magnitude. If Q is isometric, these sorted matrices will be identical. However, since the mapping between languages is not exactly isometric, Artetxe et al. (2018) solves for D using nearest neighbors in the L_2 metric. This process replaces the initialize_dictionary method in Algorithm 1.

Following that, the procedure of Artetxe et al. (2017) is followed. compute_mapping and orthogonalize_mapping are performed jointly using the solution to the orthogonal procrustes problem which is $Q = UV^\top$[3] for the singular value decomposition $USV^\top = XDY^\top$. compute_new_dictionary is performed using nearest neighbors retrieval between X and QY resulting in a new D. The process is repeated until convergence. Additionally, Artetxe et al. (2018) adds the following regularization methods to avoid local optima:

1. **Annealed Dropout:** The authors apply annealed dropout (Rennie et al., 2014) to the matrix $X(QY)^T$ during dictionary induction in a method they term *stochastic dictionary induction*.

3. I use the notation Q to standardize between papers. Artetxe et al. (2018) splits Q into W_x and W_y and applies them separately while my definition of $Q = W_x W_y^\top$ and is applied only to Y.

2. **Training Data Truncation:** The authors truncate the vocabulary to the most frequent words in both languages. The intuition is that far less frequent words are less likely to have exact synonyms and only introduce noise.

3. **Regularized Similarity:** To account for local minima caused by performing nearest neighbor retrieval in high dimensional space, the authors regularize similarity by accounting for the mean similarity of each candidate: $\text{similarity}(x, y) = 2\cos(x, y) - \frac{1}{k}\sum_{y_k \in KNN(x)} \cos(x, y_k) - \frac{1}{k}\sum_{x_k \in KNN(y)} \cos(x_k, y)$ (Lample et al., 2018). Interestingly, this adjusted lookup method is equivalent to hard negative sampling using KNN.

4. **Bidirectional Dictionary Induction:** To account for polysemy, the nearest neighbor dictionary induction is performed in both directions allowing many-to-one mappings for both languages.

This work is the culmination of many empirical iterations of methodology by Artetxe and his co-authors and results in an approach which is unsupervised, quickly converges, and provides strong empirical results even for distant language pairs. However, the method itself lacks grounding in the wealth of literature in optimal transport theory (Peyré et al., 2019). In parallel, the next work arrived at similar empirical methods, while grounding them in sound mathematics.

3.2.3 Unsupervised Alignment of Embeddings with Wasserstein Procrustes

(Grave et al., 2019) also follow the Algorithm 1. However, where Artetxe et al. (2018) rely on intuitions driven from strong assumptions, they instead utilize prior findings from Optimal Transport theory.

For initialize_dictionary, they also utilize the self-similarity matrices X' and Y', formalizing the problem as finding D which minimizes $||X'D - DY'||_2^2$. This is equivalent to the initialization Artetxe et al. (2018) aims to solve approximately using nearest neighbor retrieval over sorted self-similarity matrices. Instead, this works solves a convex relaxation of the problem, relaxing the dictionary D from a permuation matrix to a doubly stochastic matrix. This makes the problem convex, allowing them to solve it with the convex combination algorithm - although the set of doubly stochastic matrices is a cone so the problem can be solved more efficiently by conic solvers such as SCS (O'Donoghue et al., 2016).

In their self learning procedure, they again choose to trade efficiency for theoretically grounded solutions to the optimal transport problems within embedding alignment. compute_new_dictionary is replaced by a fast approximate solution to the optimal transport problem using the Sinkhorn Algorithm (Cuturi, 2013). However, the trade-off of this approach is that the Sinkhorn algorithm is more costly than the nearest neighbor computation of Artetxe et al. (2018). This cost is the primary motivator to utilize a batch-based optimization of Q using stochastic gradient descent for compute_mapping and projecting back onto the set of orthogonal matrices using $Q^{i+1} = UV^\top$ where $USV^\top = Q$. However, this stochastic optimization approach also can be interpreted as an alternate form of annealing (Bottou, 1991). This work also finally performs retrieval using the same regularized similarity metric from Lample et al. (2018).

3.3 Multilingual Pretraining

As with many areas of NLP, Large Pretrained Language models have become the dominant approach for transfer in low-resource cross-lingual transfer. The ability of large models to implicitly learn pivot representations which could benefit low-resource languages was first shown in Neural Machine Translation by Zoph et al. (2016). (Conneau and Lample, 2019) incorporated this insight into the pretraining of masked language models, leading to models which could perform zero-shot transfer for many tasks without explicit translation. (Conneau et al., 2020) and (Xue et al., 2021) scaled up this paradigm to produce large and effective multilingual models for hundreds of languages. In monolingual models, on the other hand, non-standard and dialectal language is often explicitly filtered out as low-quality (Gururangan et al., 2022) leaving a clear path for thoughtfully re-incorporating dialectal data in pretraining.

3.3.1 Transfer Learning for Low-Resource Neural Machine Translation

While Zoph et al. (2016) does not fall under the current paradigm of pretraining on language modeling, it formed an early foundation to show the potential of high-resource languages providing strong benefits to neural models in low-resource settings. Importantly, it highlights the ability of large models to learn alignment implicitly when provided the right inductive biases.

This work begins with an Encoder-Decoder model which has 6 components: Source Language Embeddings, an Encoder RNN, a Decoder RNN, a Decoder Attention Module, Target Language Input Embeddngs, and Target Language Output Embeddings. All 6 of these components are trained on a high-resource language pair, such as French-English. When transferred to a new low-resource language pair, such as Hausa-English, the source language embeddings are reinitialized and the target language output embeddings are frozen. The model is then trained again on this low-resource pair, but this process leads to a 40% improvement in translation performance on the low-resource pair. Using a pretrained model as a starting point is intuitively beneficial, as it provides the model structural inductive biases which are not innately encoded in the architecture (Papadimitriou and Jurafsky, 2020). Similarly, model freezing has strong grounding, as it acts as strong regularization and vastly reduces the VC dimension of a model learning on a reduced dataset.

More surprising is the fact that random initialization of the input embeddings converges to the same results as initialization based on a dictionary mined from parallel alignment. While the experiment is a somewhat minor detail in the original work, this paper is the first to illustrate that pretraining leads to inductive biases that allow the model to implicitly learn embeddings that are as empirically effective as explicitly aligned multilingual embeddings without any explicit alignment loss. This finding was later cemented explicitly by Wada and Iwata (2018) who showed that training monolingual embeddings with a shared encoder model resulted in effectively aligned input embeddings.

3.3.2 Cross-lingual Language Model Pretraining

In Conneau and Lample (2019), crosslingual models were fully cemented with a large pretrained model as both massively effective zero-shot learners, but also strong initialization points for low-resource languages with minimal supervision. Unlike prior methods, the work shows that multilingual embedding models are capable of downstream transfer across many languages and many tasks, without any specialized finetuning approaches. This approach

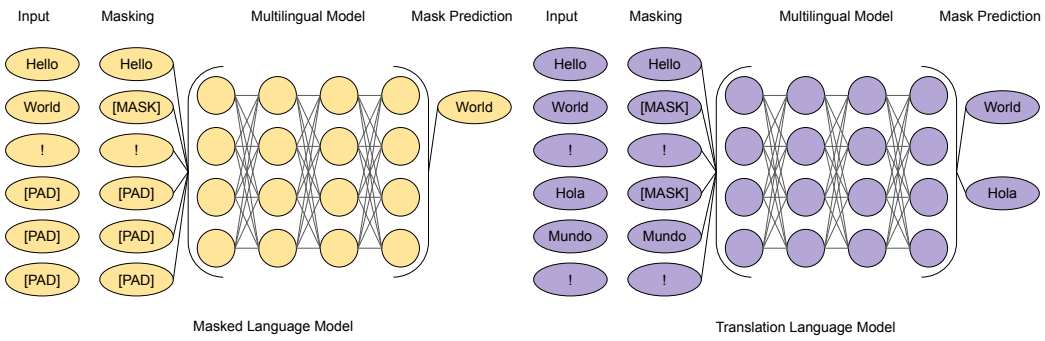

Figure 2: Language Modeling Objectives used in Large Pretrained Models

of providing linguistic robustness solely through pretraining is especially promising, given that dialects and linguistic variants primarily suffer from a lack of labeled data as opposed to unlabeled data as mentioned in Section 2.2.

Conneau and Lample (2019) develops principles to extend the masked language modeling objective (Devlin et al., 2019)[4] to mixed streams of Multilingual data. Given a corpus, the masked language modeling objective constructs a semi-supervised denoising problem by corrupting 15% of all tokens to become either a [MASK] or a randomly sampled incorrect token. The model is then trained to be able to reconstruct the corrupted token as shown in Figure 2 using cross-entropy loss. Since the model uses sub-word tokenization, the full softmax can be computed. In this work, the 15% of tokens is selected before training and the corruption is produced as a pre-processing step.

To produce a cross-lingual model, N languages are chosen each with n_l sentences. Each is assigned a probability by adjusting the frequentist probability of each language in the training data $p_l = \frac{n_l}{\sum_{i=0}^{N} n_i}$ with temperature τ[5] to increase the representation of low-resource languages: $q_l = \frac{p_l^{1/\tau}}{\sum_{i=0}^{N} n_i^{1/\tau}}$. Batches are monolingual - with each batch being entirely made up of language l with probability q_l. By increasing temperature ρ, the entropy in the sampling distribution increases leading to higher representation of low-resource languages.

Additionally, the authors propose a stronger form of cross-lingual supervision when parallel data is available called Translation Language Modeling, shown in Figure 2. Translation Language Modeling is identical, but samples from pairs of cross-lingual parallel sentences. Since the model can attend to tokens across both languages, this training procedure provides stronger supervision for alignment.

This simple procedure resulted in a massive leap forward in multilingual NLP, achieving state-of-the-art results across downstream sequence classification, both supervised and unsupervised machine translation, and word-level translation. Impressively, it outperforms even prior state-of-the-art alignment approaches for word-level translation despite having much weaker inductive biases for word-level alignment. Translation language modeling fur-

4. While the paper also discusses causal language modeling, it consistently shows that causal language modeling is consistently outperformed by its masked counterpart.
5. The paper uses $\alpha = \frac{1}{\tau}$. Here I have opted for τ to use the traditional temperature analogy.

ther improves these results, which has been explored in further works incorporating explicit alignment in pretraining such as (Hu et al., 2021)

3.3.3 Unsupervised Cross-lingual Representation Learning at Scale

The limits of unsupervised model scaling were the focus of Conneau et al. (2020), rather than novel training paradigms. This model termed *XLM-R* scales the training data and number of languages from previous related counterpart.

First, the model dynamically masks tokens throughout training allowing a richer space of semi-supervised examples from the same amount of training data (Liu et al., 2019). Second, they scale the available training data by scraping the entire common crawl, rather than just Wikipedia. They do this without significant quality loss by filtering documents during preprocessing which leads to the largest perplexity spike for a simple language model (Wenzek et al., 2020). Notably, this filtering method removes text in a normative fashion, filtering out non-standard dialects which are less represented within monolingual data. Finally, they scales the number of languages within a single model from 15 to 100.

The scaling approach further improves on the results of the closely related XLM model of the prior work, improving the state-of-the-art on 3 downstream multilingual tasks. However, the scale leads to a number of interesting empirical phenomena such as interference, where model performance degrades as more languages are added. The focus on scaling unsupervised data for low-resource languages, rather than scaling labeled data, illustrates how cross-lingual transfer fundamentally shifts the priorities of developing language resources, especially for dialects.

3.3.4 mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Xue et al. (2021) extends the applicability of multilingual training, not only to multiple languages, but to nearly any task through the use of text-to-text pretraining. A limitation of XLM-R and other encoder only models is that they are focused only on the representation of text. For many user facing tasks, such as summarization and question-answering, models must also understand how to generate text in response to user inputs. While encoder-only models are able to effectively align embeddings, they require the use of language-specific data for novel generation tasks.

To address this, the mT5 model uses a text-to-text pretraining process (Raffel et al., 2020) in which an encoder-decoder model is fed masked text and tasked with predicting masked tokens causally, rather than using a head on top of the mask hidden state. Since the causal generation process does not require fixed length masks, this allows the pretraining process to provide more difficult supervision via masked span prediction, rather than simply masked token prediction.

This initializes a language agnostic encoder and a multilingual decoder that can generate output in all training languages. This general utility has made mT5 a powerful baseline model, as the same finetuning procedure leads to powerful zero-shot transfer for classification, structured prediction, and generation tasks, while encoder-only models required custom architectures for each of these categories. Follow-up works have exploited this task-agnostic architecture by training the model to perform multi-task learning which exhibits both task and language data efficiency.

3.4 Relevance To Dialectal NLP

Despite massive multilingual pretraining leading to strong benefits for multilingual NLP, dialectal NLP in English has primarily focused on interventions at the level of supervised data. By constructing new resources for a dialect of interest, practitioners are able to see an immediate improvement over Standard American English pretrained models for their task of interest. The effectiveness of cross-lingual transfer, even over distant pairs of languages, highlights an alternate and underexplored path using unsupervised transfer learning. Interestingly, current state-of-the-art multilingual models do not even require explicit labels of which language is being modeled - meaning that unsupervised transfer using such models does not require high accuracy dialect classifiers. However, current approaches to training large models often explicitly remove non-standard text (Gururangan et al., 2022; Wenzek et al., 2020). Tackling this challenge requires developing new preprocessing methodologies for text which are dialect inclusive.

4. NLP For English Dialects

Existing work on NLP for English Dialects has focused primarily on data gathering and augmentation. (Jørgensen et al., 2016) used online lexicons to provide weak supervision through partial labeling and constrained bootstrapped learning. (Blevins et al., 2016) manually annotates a small dataset and uses simple domain adaptation methods to enable transfer. Going beyond lexicon, (Ziems et al., 2022) manually constructs a set of morphosyntactic transformations for use in data augmentation, in addition to a dictionary induction method designed for codeswitching.

4.1 Learning a POS tagger for AAVE-like language

In order to demonstrate a discrepancy on the part-of-speech tagging task, (Jørgensen et al., 2016) labels a corpus of subtitles from African American characters on the show *The Wire*, rap lyrics from African American rappers, and tweets regionally associated with African American Vernacular. They first show that, even when trained on colloquial Twitter data, a Perceptron Part-Of-Speech tagger augmented with beamsearch struggles with Out-Of-Vocabulary tokens which harm performance.

In their best performing system, the paper then mines lexicons from 4 online community dictionaries which are focused on "slang", but often capture dialectal terminology. They use these lexicons to produce ambiguous tag dictionaries which provide a set of possible labels for each out-of-vocabulary token. Using unlabeled dialectal data, they produce "ambiguously labeled" data where each token has a set of possible tags given by the tag dictionary. A penalty is imposed *only* if the predicted label does not lie within the set of possible labels. This utilizes the knowledge of the community dictionaries to provide weak supervision allowing these OOV tokens to be used similar to Standard American English words with similar tag sets.

This approach is limited, however, to tasks for which even ambiguous labels can be extracted from online resources. This property is somewhat unique to purely linguistic tasks such as part-of-speech tagging, limiting the applicability to the most high impact areas of NLP which involve higher level processing.

4.2 Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression.

As part of a larger project studying grief in gang-involved youth in Chicago, Blevins et al. (2016) tackles several challenges in processing the English variant prevalent in their target community. Beyond the technical contributions, this work highlights the need for dialectal NLP in order to make greater contributions to Computational Social Science. The work explores two tasks: POS tagging and Emotion Classification.

For POS tagging, they also use a perceptron with beamsearch over tag predictions. However, rather than rely only on weak supervision they manually label 616 utterances in their target English variety. In order to combine their training data with existing Twitter training data, they perform domain adaptation using a dummy feature indicating the domain of input (Daumé III, 2007). This form of regional adaptation has been shown to be valuable to large pretrained NLP systems, even with their much larger expressive capacity (Hofmann et al., 2022). They show that an SAE model suffers a 15% performance drop when applied to the target English variety and their intervention is able to recover 4% of that overall performance.

For emotion classification, they manually translate 400 inputs their target dialect into Standard American English and use Viterbi alignment to produce a word translation table between their target and Standard American English (**?**). They then use this translation table to apply an emotion glossary to their target dialect. This phase would massively benefit from the larger scaled word translation mining of the methods described above, as the glossary they produce is severely limited by the number of translations they were able to manually produce.

This work shines a spotlight for the applicability of modern unsupervised methods to important aspects of dialectal NLP. Blevins et al. (2016) had reasonable quantities of unlabeled data for their high level social task of understanding the sentiment of each tweet - however their alignment procedure required manually intensive translation which limited them to a small scale. With highly effective techniques for mixed pretraining and alignment across embedding spaces, works of this kind can be rapidly accelerated.

4.3 VALUE: Understanding Dialect Disparity in NLU

The VALUE project (Ziems et al., 2022) constructs a set of morphosyntactic transformations from SAE to AAVE, using linguist attested patterns and a dependency parsing model designed for SAE. Since these rules operate using controllable rule-based morphosyntactic changes, they are guaranteed to be label preserving - avoiding introducing noise into the label space. The authors use this to demonstrate consistent harms across all but 1 task of the GLUE benchmark, but also address these harms by using their approach as data augmentation.

The transformations can then be applied automatically to test data to evaluate potential discrepancies and to training data as a form of data augmentation. Unlike the prior works, this methodology can be applied to any task rather than a specific task and can even be used as a data augmentation technique during pretraining, which has been shown to reduce other representational harms of language models (Qian et al., 2022). This work provides a clear step to move beyond the incremental data annotation approach used previously

and towards one that exploits both effective methods from multilingual NLP and expert linguistic resources about dialects and other language variants as discussed by Bird (2022).

Presently, I have joined the VALUE project to extend and evaluate the methodology for multiple dialects at once, aiming to drive multi-dialectal NLP in the direction of mulitlingual models described above. Due to the heavy overlap of acceptable patterns across many languages, this multi-dialectal form of VALUE enables feature sharing across many dialects, rather than simply relying on SAE transfer to all others. Since each above work highlights the importance of typographic similarity across transfer languages, we believe this approach will be key.

5. Conclusion

Throughout this outline I have covered:

1. **Motivations:** Language technology which is linguistically inclusive has strong ethical and practical impacts (Hovy and Spruit, 2016). However, current NLP focuses primarily on the standard dialects of high-resource languages (Joshi et al., 2020). Amongst low-resource linguistic phenomena, dialects are impactful and understudied (Bird, 2022).

2. **Possible Methods:** The highly studied area of multilingual NLP has provided a wealth of techniques with applicability to dialect. Explicit word alignment (Adams et al., 2017; Artetxe et al., 2018; Grave et al., 2019) has the potential to efficiently tackle lexical variation. More extensively, joint pretraining(Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021) is underexplored for English dialects, despite its broad adoption elsewhere, with English dialects often explicitly excluded by filtering procedures.

3. **Prior English Dialect Approaches:** Prior work has focused on labeled data, using expert resources to provide weak supervision, perform translation, or to create data augmentation techniques. In my work, I aim to adapt and improve unsupervised methods from multilingual NLP for the dialectal setting to instead exploit unlabeled data. Amongst prior approaches, (Ziems et al., 2022) is most applicable to this direction as it produces a task agnostic resource.

These foundational works cover at a high level the reasons I am interested in dialect, the technique space I aim to explore, and the methods I aim to directly utilize and compare against.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-1088`.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1253. URL https://aclanthology.org/N19-1253.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL https://aclanthology.org/P17-1042.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://aclanthology.org/P18-1073.

David Bamman, Chris Dyer, and Noah A. Smith. 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, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2134. URL https://aclanthology.org/P14-2134.

Emily M. Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

Steven Bird. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.539. URL https://aclanthology.org/2022.acl-long.539.

Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1207.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120.

Léon Bottou. Stochastic gradient learning in neural networks. 1991.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1033.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1136. URL https://aclanthology.org/D16-1136.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

François Grosjean. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press, 2021.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection, 2022. URL https://arxiv.org/abs/2201.10474.

Einar Haugen. The ecology of language. *The ecolinguistics reader: Language, ecology and environment*, pages 57–66, 2001.

David Haussler. *Probably approximately correct learning*. University of California, Santa Cruz, Computer Research Laboratory Santa . . . , 1990.

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B Pierrehumbert, and Hinrich Schütze. Geographic adaptation of pretrained language models. *arXiv preprint arXiv:2203.08565*, 2022.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://aclanthology.org/P16-2096.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.284. URL https://aclanthology.org/2021.naacl-main.284.

Dorota J Iskra, Rainer Siemund, Jamal Borno, Asuncion Moreno, Ossama Emam, Khalid Choukri, Oren Gedge, Herbert S Tropf, Albino Nogueiras, Imed Zitouni, et al. Orientel-telephony databases across northern africa and the middle east. In *LREC*, 2004.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1130. URL https://aclanthology.org/N16-1130.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2009. URL https://aclanthology.org/P17-2009.

Kilany, Hanaa, Gadalla, H, Arram, Howaida, Yacoub, A, El-Habashi, Alaa, and McLemore, C. Egyptian colloquial arabic lexicon, 1997. URL https://catalog.ldc.upenn.edu/LDC99L22.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the Second Workshop on Language in Social Media*, pages 27–36, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `https://aclanthology.org/W12-2104`.

Gideon S. Mann and David Yarowsky. Multipath translation lexicon induction via bridge languages. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. URL `https://aclanthology.org/N01-1020`.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1269. URL `https://aclanthology.org/D17-1269`.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 62–72, 2011.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL `http://arxiv.org/abs/1301.3781`.

Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016. URL `http://stanford.edu/~boyd/papers/scs.html`.

Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.554. URL `https://www.aclweb.org/anthology/2020.emnlp-main.554`.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

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

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5412. URL https://aclanthology.org/W18-5412.

Steven J. Rennie, Vaibhava Goel, and Samuel Thomas. Annealed dropout training of deep networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 159–164, 2014. doi: 10.1109/SLT.2014.7078567.

Mark Sebba. *Contact languages: Pidgins and creoles*. Bloomsbury Publishing, 1997.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1614. URL https://aclanthology.org/W14-1614.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. Inferring latent user properties from texts published in social media. In *AAAI Conference on Artificial Intelligence*, 2015.

Takashi Wada and Tomoharu Iwata. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306*, 2018.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.494.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Daniel Zeman. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.258. URL https://aclanthology.org/2022.acl-long.258.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, 2016.