9-2019

Demographic Factors as Domains for Adaptation in Linguistic Preprocessing

Sara Morini
The Graduate Center, City University of New York

Demographic Factors as Domains for Adaptation in Linguistic Preprocessing

by

Sara Morini

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2019

Demographic Factors as Domains for Adaptation in Linguistic Preprocessing

by

Sara Morini

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Gita Martohardjono Executive Officer

The City University of New York

ABSTRACT

Demographic Factors as Domains for Adaptation in Linguistic Preprocessing

by

Sara Morini

Advisor: Kyle Gorman

Classic natural language processing resources such as the Penn Treebank (Marcus et al. 1993) have long been used both as evaluation data for many linguistic tasks and as training data for a variety of off-the-shelf language processing tools. Recent work has highlighted a gender imbalance in the authors of this text data (Garimella et al. 2019) and hypothesized that tools created with such resources will privilege users from particular demographic groups (Hovy and Søgaard 2015). Domain adaptation is typically employed as a strategy in machine learning to adjust models trained and evaluated with data from different genres. However, the present work seeks to evaluate whether domain adaptation to demographic groups such as age or gender may be an effective strategy to ameliorate the effects of biased or outdated training corpora in linguistic preprocessing tasks. We find adaptation to demographic groups to be an effective strategy for improving preprocessing performance across all demographic groups.

ACKNOWLEDGMENTS

To my parents, for their trust and confidence; to Evan, for his unshakeable support; to Kyle Gorman, for being an excellent resource and a thoughtful advisor; and to the City University of New York, for providing so many opportunities to me and so many others: thank you.

Contents

List of Tables

List of Figures

1 Introduction and Motivation

As natural language processing and computational linguistics move to embrace and rely on machine learning techniques to process increasingly large and diverse datasets, there is an increased risk of propagating biases that may unfairly advantage one group over another. These biases may be inherent in the training data or result from models' inductive biases and can result in a resource that works better for members of a particular group, which may effectively deny others access. Hovy and Spruit (2016) cite particular risks of exclusion due to demographic biases inherent in text data, arguing that considerations about demographic biases should be present in every stage of research.

As tools created with natural language processing techniques are intended to be universally usable and accessible, the potential for bias against any particular group is cause for concern. Recent literature concerning the identification and mitigation of bias in machine learning and natural language processing models (Bolukbasi et al. 2016, Garg et al. 2018, Shen et al. 2018, Zhao et al. 2018) might be distilled (or overgeneralized) to one sentence:

Models replicate and exacerbate biases present in their training data.

This has been shown in abstract representations of language like word embeddings and in downstream tasks such as sentiment analysis. In contrast, this idea has also appeared in analyses of lower-level linguistic tasks: Hovy and Søgaard (2015) report that part-of-speech taggers trained with older corpora privilege older users, and Garimella et al. (2019) find that not only is the classic Penn Treebank corpus severely gender imbalanced, but the performance part-of-speech taggers and dependency parsers trained on gendered subsets of the corpus varies when evaluated on gender-specific data.

The following thesis addresses questions raised by this literature with the specific goal of not just identifying performance differences based on demographics, but suggesting a path

for improvement. Supervised domain adaptation, formalized by Ben-David et al. (2007), Daumé III (2007), and others, is a machine learning strategy appropriate when there is a large, labeled corpus of source domain data, a small, labeled corpus of target domain data, and a large, unlabeled corpus of target domain data to be labeled. This thesis proposes that domain adaptation could be a viable strategy for mitigating bias: if language data from distinct demographic groups comes from different distributions, we should be able to adapt to these discrete domains and improve performance on the target data.

Specifically, we focus on two linguistic preprocessing tasks. Part-of-speech tagging, a word-category disambiguation task, and sentence boundary detection, the division of blocks of text into discrete sentences, are both well-known tasks for which many off-the-shelf, pre-trained tools exist, as well as models that may be quickly trained by the user. We select these tasks both because they are well-known and because they are tasks that involve the identification and labeling of linguistic units that are widely recognized. Rather than something more nebulous such as sentiment or emotion, native speakers, especially those with some degree of metalinguistic awareness, are reliably able to categorize words into parts of speech and indicate sentence boundaries. Additionally, these tasks present the opportunity to show that bias may be mitigated early in the processing pipeline, avoiding the propagation of errors that may disadvantage groups underrepresented in the training data.

Lynn et al. (2017) also investigates demographic adaptation for linguistic tasks, albeit at an individual level with continuously-valued demographic variables. In the interest of proposing a method for mitigating bias that is at once maximally effective and maximally practical, we test adaptation strategies by adapting to discrete demographic groups.

We begin by describing the data, models, and adaptation strategies used in this thesis.

2 Materials

2.1 Data

Text data used in this experiment include four established, widely used corpora and one novel corpus manually collected and annotated for this thesis. Data was serialized and stored at the document level using Protocol Buffers.[1] Additional details about the corpora are reported in the following sections, and summary statistics about all corpora are reported in Table 1.

2.1.1 OntoNotes

The OntoNotes corpus (Weischedel et al. 2013) comprises text from a variety of genres in English, Chinese, and Arabic. The Wall Street Journal portion consists of English newswire text, a subset of the documents from the Penn Treebank (Marcus et al. 1993), with the exception of documents that were deemed to have too much financial jargon. 71.29% of the total sentences in the Penn Treebank are included in OntoNotes. Documents in OntoNotes have been reannotated to improve accuracy and adhere more closely to LDC guidelines, achieving over 90% interannotator agreement for all levels of annotation (Weischedel et al. 2011). We use the Wall Street Journal portion due to its widespread and canonical use in part-of-speech tagging (Gorman and Bedrick 2019) and sentence boundary detection (Gillick 2009), among other linguistic tasks. Additionally, we make use of a resource made available by Garimella et al. (2019) identifying the author gender of many of the Wall Street Journal articles contained in OntoNotes using historical newspaper databases and author metadata to find author names, then assigning author gender based on historical name popularity.

[1] https://developers.google.com/protocol-buffers/

2.1.2 English Web Treebank

The English Web Treebank is a corpus of English web text annotated for part-of-speech and syntactic information from a variety of sources, including emails, blogs, reviews, answer forums, and newsgroups (Bies et al. 2012). It contains web text gathered between 2002 and 2011, and contains no information about the demographics of the authors.

2.1.3 Switchboard Corpus

The Switchboard Corpus is a collection of spoken telephone conversations occurring between 1990 and 1991 between "previously unacquainted speakers" about a pre-determined set of topics. It is primarily used in speech research. We use the NXT Switchboard Annotations (Calhoun et al. 2009), which contain part-of-speech tags as well as additional annotations to identify the date of birth and gender of each speaker. The process by which speakers' date of birth and gender were collected is not specified.

2.1.4 Brown Corpus

The Brown Corpus is a classic English corpus composed of texts from a variety of genres, all published in 1961. We use the version available in the Natural Language Toolkit (Bird et al. 2009). Though author information for texts included in the Brown Corpus is available,[2] the prevalence of pseudonyms and lack of available information about the authors makes labeling documents with author age or gender a difficult task; it is not attempted here.

2.1.5 Reddit: r/relationships

Reddit.com is a popular social forum and news site composed of user-submitted content organized into 'subreddits', subforums with a unifying topic. For this work, we focus on

[2] http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM

the subreddit r/relationships. This is a space where users solicit advice about interpersonal relationships, romantic or otherwise. At time of writing, the r/relationships subreddit has over 2.6 million subscribers.[3] This subreddit is especially valuable for this work because, unlike other online forums, users typically provide some context for their question by identifying the age and gender of the participants in the conflict, including that of the author. A prototypical post title in this subreddit might read: "`Is my [21M] boyfriend trying to make me [19F] jealous?`" In this case, the author has self-identified as a 19-year-old female and identified their boyfriend as a 21-year-old male. This particular feature of this subreddit makes it a valuable data source in that we are able to to utilize authors' self-identified gender as a variable. This contrasts sharply with gender-identification methodology cited in Hovy (2015), Hovy et al. (2015), and Garimella et al. (2019), where author gender is predicted based on the name associated with a document. This process is inexact and problematic, as it assumes the existence of a gender binary and that each participant identifies within this binary. Following suggestions for a careful treatment of gender put forward by Larson (2017), we aim to avoid the presupposition of a gender binary by utilizing Reddit authors' self-identified gender, which may include identifiers such as 'NB' (nonbinary), 'MTF' (male-to-female) , 'FTM' (female-to-male) or 'trans', and may or may not fall in the categories of 'male' and 'female'.

Though the r/relationships subreddit contains information that will aid in our intended purpose, it is unstructured and unannotated data. The following sections detail the process of creating an annotated corpus of r/relationships posts.

Collection & Preprocessing We gather posts from the r/relationships subreddit using the API provided at pushshift.io.[4] All posts used here were created between September and November 2018. We used a regular expression to isolate posts whose titles contained the

[3] http://reddit.com/r/relationships
[4] Reddit statistics: http://pushshift.io

type of information we are interested in, which we refer to as "demographic tokens". To ensure that posts were somewhat structured and of a usable length, we selected posts that were determined by the NLTK `TreebankTokenizer` to be between 25 and 75 sentences and 250 and 750 words in length. However, this tokenization was not preserved – all annotation of this corpus is manual and is described in **Manual Annotation**. Next, we faced the task of determining which, if any, of the demographic tokens refer to the post's author. To accomplish this automatically, we isolated the demographic tokens in a given post with a regular expression[5] and examine their local contexts. We extract features with a binary classification task in mind: predicting whether a given demographic token refers to the author of a post or to another specified participant in the conflict. The features used include:

- Whether the demographic token immediately follows "I", "me", or "my"
- Whether the demographic token immediately follows a possessive part-of-speech tag, as determined by the native NLTK part-of-speech tagger (`nltk.pos_tag`)
- Whether the demographic token is adjacent to a noun that refers to another person, such as "wife", "friend", "coworker", etc.

We achieve accuracy of .9286 on this task using logistic regression when evaluating on 1000 manually-labeled post titles. For documents for which we have identified the likely author's demographic information, we select a balanced demographic distribution as follows: for each birth year attested in the Reddit data (calculated using author age and date of post), we select up to 5 posts with authors identifying as male, up to 5 posts with authors identifying as female, and any posts with authors who self-identify outside this gender binary. The resulting set of posts is evenly split by gender and has a uniform distribution for date of birth. The details of these demographic splits are discussed in the **Demographic Groups** section of 2.1.6.

[5] `r'[\[\(\{][a-zA-Z]* */?,*-?  *[0-9][0-9]?\'?[sS]? */?,?[a-zA-Z]?/?[a-zA-Z]*[\]\)\}]'`

Manual Annotation All Reddit data used in these experiments is manually annotated. For sentence boundary detection, all selected documents were hand-annotated by a single annotator. For part-of-speech tagging, a total of 900 sentences were sampled from all Reddit documents. 500 sentences were used for adaptation: 100 for a generic adaptation to the Reddit corpus, and 100 each from the `Male`, `Female`, `Old`, and `Young` demographic partitions. 400 additional sentences were selected for evaluation, with 100 from each potential combination of demographic groups: 100 sentences from `Male/Old`, 100 from `Male/Young`, 100 from `Female/Old`, and 100 from `Female/Young`. At evaluation time, this data was combined into binary groups (i.e., all `Male` data evaluated at once, regardless of age, all `Young` data evaluated at once, regardless of gender). All 900 sentences were tokenized at the word level and annotated for part-of-speech by a single annotator using the Penn Treebank tagging guidelines (Santorini 1990), with the exception of the abbreviation 'TL;DR', which is tagged here as `GW`. This tag is used in EWT and Switchboard for non-final tokens, like an incorrectly hyphenated word or a compound that has been incorrectly split into two tokens. It is used here in an instance where one token stands for multiple words: the abbreviation means "too long; didn't read" and signals a brief summary of the post. To ensure accurate application of the tagging guidelines, we calculate inter-annotator agreement between the Reddit annotator for this work and the annotators of OntoNotes on 100 randomly sampled sentences from OntoNotes, resulting in a Cohen's κ of .9788, a statistic denoting "near perfect agreement" (Landis and Koch 1977).

Anonymization Though Reddit usernames do not necessarily contain identifying information about the user, and posts on the r/relationships subreddit are frequently made with "throwaway" usernames that will not be traced back to a user's main account, Reddit users, especially those discussing interpersonal issues, have not actively consented to the use of their language data in this project. Though the method for collecting, selecting, and annotating

relevant posts has been described here, the tagged data will remain proprietary.

Corpus	Partition	Demographic	Tokens	Sentences
OntoNotes	Train	N/A	662,223	27,179
	Test	N/A	98,277	4,059
	Test	`Male`	45,744	1,864
	Test	`Female`	7,952	341
EWT	Train	N/A	195,906	11,975
	Test	N/A	21,469	1,759
Brown	Train	N/A	937,003	41,686
	Test	N/A	233,714	10,419
Switchboard	Train	N/A	817,397	86,692
	Test	N/A	212,508	22,705
	Test	`Male`	101,084	9.986
	Test	`Female`	111,424	12,719
	Test	`Old`	109,076	11,108
	Test	`Young`	103,423	11,597
Reddit	Test (POS)	N/A	7,529	400
	Test (SBD)	N/A	N/A	8,947
	Test (POS)	`Male`	3,871	200
	Test (SBD)	`Male`	N/A	3,411
	Test (POS)	`Female`	3,658	200
	Test (SBD)	`Female`	N/A	3,494
	Test (POS)	`Old`	3,787	200
	Test (SBD)	`Old`	N/A	3,649
	Test (POS)	`Young`	3,742	200
	Test (SBD)	`Young`	N/A	3,303

Table 1: Number of tokens and sentences per corpus and partition

2.1.6 Corpus Partitions

Train and test For the corpora for which standard training and test splits exist, we use them: for OntoNotes, sections 00–18 are used to train and sections 22–24 are used to test; for EWT, the train and test sets are those established in the CoNNL-U files established in

Silveira et al. (2014). For Brown and Switchboard, we randomly sample documents from each corpus and use an 80%-20% train/test split. The Reddit corpus is not sufficiently large to use as training data, so it is used purely for adaptation and evaluation.

Demographic groups For Switchboard, OntoNotes, and Reddit, data used for testing is partitioned based on author demographics. For OntoNotes, we use the resource released in Garimella et al. (2019) which identifies male- and female-authored articles in the Wall Street Journal section of the Penn Treebank. Using only the documents in the test set which are identified in this resource, we create the `Male` and `Female` subsets of the OntoNotes test set.

For Reddit and the Switchboard test set, we divide the documents by their identified demographic groups: Switchboard from the NXT annotations, and Reddit from automatic author classification. The median date of birth of Switchboard subjects is 1956, and the median date of birth of the Reddit subjects is 1990 (birth years of Reddit users in this corpus range from 1973-2003). For Switchboard, documents written by authors born in or after 1957 are assigned to the `Young` partition; documents whose authors were born before 1957 are assigned to the `Old` partition. For Reddit, the `Old` partition contains documents written by authors born in or before 1990, and the `Young` partition is made up of documents from authors born after 1990. In the same way, we assign male Switchboard and Reddit authors to their respective corpus's `Male` partition and female authors to the `Female` partition. Of the 266 Reddit documents selected for annotation, 133 belong to the `Female`, `Old`, and `Young` partitions, respectively. The `Male` partition contains 131 documents. Two documents were written by authors who identify outside the gender binary, and though these documents are included in their respective age partitions, they are not included in the gendered partitions. Though it is important that self-identified gender information was available for all authors in the Reddit corpus, it is unfortunate that there is insufficient data from users outside the binary to create a separate category, thereby recreating the gender binary we hoped to avoid.

Adaptation For adaptation experiments, we require set amounts of data from each target domain. For part-of-speech tagging experiments where we adapt to corpora which have a training set, namely EWT and Switchboard, we randomly sample the required number of sentences from the training set. For sentence boundary detection while adapting to the same corpora, we randomly sample *documents* until reaching the total required number of sentences. For part-of-speech tagging adaptation using Reddit, we select 100 sentences of tagged data from each demographic group to be used as adaptation data and use the rest for evaluation. For sentence boundary detection, we again sample at the document level. In all cases, data used for adaptation is kept separate and is not used for evaluation.

2.2 Models

For both part-of-speech tagging and sentence boundary detection, we explore a selection of models varying in complexity. Though they are briefly described here, it is outside the scope of this work to discuss the architecture of each model in detail. Rather than an analysis of the performance of the models themselves, this work is an analysis of the performance of these models in conjunction with the proposed adaptation strategies.

2.2.1 Sentence Boundary Detection

The first model selected for the sentence boundary detection task is **Punkt**: an unsupervised, language-independent approach to the task described in Kiss and Strunk (2006). We use the implementation included in NLTK (Bird et al. 2009). It relies on identifying boundary candidates and distinguishing between abbreviations and sentence boundaries based on a number of criteria. The second model, **Perceptronix**,[6] is an averaged perceptron model using simple features including the left and right context of a boundary candidate. **De-**

[6] https://github.com/kylebgorman/perceptronix

tectorMorse,[7] the third model, is another averaged perceptron model which extracts more features, including the identity of the punctuation mark creating the boundary candidate.

2.2.2 Part-of-Speech Tagging

We select three models for part-of-speech tagging. The first, **TnT** (Brants 2000), is a second-order Markov model which is lightweight and trains and predicts quickly. The second is a part-of-speech tagging implementation of **Perceptronix**, which uses an averaged perceptron model with context and word shape features, including length and suffixes. The third model, **Flair** (Akbik et al. 2018), is the current state-of-the-art model for this task. It is a bidirectional LSTM with conditional random fields and contextual string embeddings, and requires significant computational resources to train and run.

2.3 Adaptation Strategies

For the adaptation experiments, we employ one of two adaptation strategies. The adaptation strategies were selected based on ease of implementation, appropriateness for the task, and potential efficacy.

Naive The first strategy is "naive", adapting to the target domain by simply adding a set amount of target-domain data to the source-domain data used for training. This technique was selected primarily because of its ease of implementation: it requires no access to features or other inner workings of the model. It is a suitable adaptation strategy for all the selected models.

FEDA The second proposed adaptation strategy is described in Daumé III (2007), referred to as Frustratingly Easy Domain Adaptation, here abbreviated to FEDA. FEDA involves

[7] https://pypi.org/project/DetectorMorse/

augmenting feature vectors during training and prediction such that each feature vector contains both a general and either source- or target-specific version of the feature, depending on whether the training example comes from the source or target domain.

For example, vanilla feature extraction for a given sentence will result in a feature vector of length n:

$$\phi = \{x_1, x_2, \ldots, x_n\}$$

Feature extraction of a source-domain sentence using FEDA will result in a feature vector of length $2n$, where one 'copy' of each feature from the original feature vector is appended with a string representing the 'general' domain, and the other 'copy' is appended with a string representing the source domain:

$$\phi_s = \{x_1^\wedge general, x_1^\wedge source, x_2^\wedge general, x_2^\wedge source, \ldots, x_n^\wedge general, x_n^\wedge source\}$$

A similar feature vector is obtained for target-domain sentences:

$$\phi_t = \{x_1^\wedge general, x_1^\wedge target, x_2^\wedge general, x_2^\wedge target, \ldots, x_n^\wedge general, x_n^\wedge target\}$$

At prediction time, data is treated as coming from the target domain and predictions depend more on the target features. This encourages certain trends in the target domain, such as particular usage patterns or token-tag collocations that appear in the target domain but not the source domain, to affect prediction.

Though named because of its relatively simple implementation, FEDA still requires the existence of feature vectors that can be appended to by the user during training and prediction. This isn't the case for Punkt, which is an unsupervised model, for TnT, where the software is already implemented or only available in compiled form, and Flair, where a bidirectional LSTM produces continuously valued rather than discrete features. In light

of these considerations, FEDA is a suitable adaptation strategy only for Perceptronix (for POS-tagging and SBD) and DetectorMorse.

3 Experiments

We conduct four experiments to determine the practicality and effectiveness of domain adaptation with demographic factors for these preprocessing tasks. Experiment 1 establishes baselines for sentence boundary detection (SBD) and part-of-speech (POS) tagging for each model by comparing combinations of training and testing corpora from different domains, both across corpora and across demographic groups. Experiment 2 compares the performance of SBD and POS-tagger models in combination with two domain adaptation strategies while varying the amount of adaptation data. Experiment 3 attempts to resolve a potential source of performance degradation when training and testing across corpora by comparing accuracy of basic and adapted models when evaluated either on the full tagset of the corpus or on a reduced tagset. Experiment 4 explores the efficacy of different methods of adapting to either demographic group, corpus, or both for POS tagging and SBD.

Information about corpus size and partitions used in these experiments can be found in Section 2.1 and Table 1.

3.1 Experiment 1: Cross-domain evaluation

Before addressing the effects of adaptation, we first establish a baseline for all models with selected combinations of training and testing data. Experiment 1a compares model performance when training and testing across corpora, a popular application for domain adaptation. Experiment 1b compares performance when evaluating on text written by members of different demographic groups, our proposed application for domain adaptation.

3.1.1 Experiment 1a: Cross-corpus evaluation

In order to establish a baseline for these models' performance without domain adaptation, we evaluate performance of models trained and tested on separate corpora.

For SBD, we train models using each of the three architectures—Punkt, DetectorMorse (DM), and Perceptronix (PPX)—on each of the four training corpora: OntoNotes, Brown, Switchboard (SWBD) and English Web Treebank (EWT). We then test each of these models on one of three test corpora: OntoNotes, Brown, and the SBD partition of the Reddit data. We compute the F1 score for each model using the number of gold-annotated sentences predicted by the model (recall) and the number of predicted sentences that were gold-annotated sentences (precision). F1 scores for models trained and tested using the same corpus are presented in Table 2, and results of models trained and tested on different corpora are presented in Table 3.

Model	Corpus	
	OntoNotes	Brown
Punkt	.8781	.9638
DM	.9566	.9899
PPX	.9542	.9899

Table 2: Within-corpus training and testing results for each SBD model

All three models perform best when trained and tested on the same corpus. Though there is no training corpus for Reddit, we can see that the best performance on this test set is achieved by models trained on EWT, likely because Reddit and EWT are both composed of text from the internet and include similar vocabulary, abbreviations, or conventions for sentence boundaries. The highest F1 scores overall are achieved by models trained and tested on the Brown Corpus. This is sensible, as the Brown Corpus contains texts that were all

Test Corpus	Model	Training Corpus			
		OntoNotes	Brown	SWBD	EWT
OntoNotes	Punkt		.8912	.6882	.8589
	DM		.9394	.6776	.9110
	PPX		.9069	.6746	.8718
Brown	Punkt	.9534		.9169	.9439
	DM	.9201		.9052	.9707
	PPX	.9794		.9023	.9431
Reddit	Punkt	.9443	.9448	.9284	.9477
	DM	.9298	.9346	.9390	.9550
	PPX	.9374	.9272	.9397	.9499

Table 3: F1 scores for cross-corpus comparison of sentence boundary detection models.

edited for publication and because it contains fewer unfamiliar abbreviations and instances of financial jargon such as those that may be seen in OntoNotes. DetectorMorse is frequently the best-performing model of the three tested here.

A note about Switchboard and EWT in sentence boundary detection Though these models were evaluated on the test sets for Switchboard and EWT, the results are not presented here. Both corpora contain annotations for part-of-speech and sentence boundaries, but certain structural elements of each corpus prevent comparable evaluation in sentence boundary detection.

As EWT contains data from various internet genres, it contains text unlike OntoNotes and Brown in that it has not been copyedited for publication. However, part of what makes annotated sentence boundaries in EWT so hard to recover is the nature of the language contained in the corpus. Of particular note in this regard is the Email genre, composed of emails sent and received by human employees of Enron (Bies et al. 2012).[8] These email documents

[8]These emails were made public by the Federal Energy Regulatory Commission in 2003 following an investigation into the causes of the company's bankruptcy (Grieve 2003).

frequently include greetings and closings, which are annotated as separate sentences as they are separated by newline characters. As a result, there appear to be many "sentences" in the EWT data that end with a comma, or with no punctuation at all. As all three sentence boundary detection models identify candidate boundaries based on punctuation, sentences that end with a comma or no punctuation are virtually impossible to detect. Resulting F1 scores from all three sentence boundary detection models are consistently below .6.

A similar issue is present in the Switchboard corpus, where "sentences" that would appear to belong together are annotated as separate utterances. Situations such as these typically occur at turn boundaries between speakers or instances when both participants are speaking simultaneously. This issue is symptomatic of Switchboard's intended use as a speech corpus, where annotators and transcribers look for "sentence-like units," which may be shorter and not necessarily delimited by punctuation, as opposed to more conventionally-defined sentences (Liu et al. 2005).

It may be of interest to note that while they are not suitable for testing purposes, models *trained* using Switchboard or EWT typically perform comparably to models trained on the other corpora (these results are included in Table 3). This is likely due to the fact that there are sufficient well-formed sentences in Switchboard and EWT to build a model that will fairly consistently identify other well-formed sentences.

To evaluate the performance of POS-tagger models with no adaptation, we train models using each of the three POS-tagger architectures—TnT, Perceptronix, and Flair—using each of the four training corpora listed above. We then test these models on each of five test corpora: OntoNotes, Brown, Switchboard, EWT, and the POS partition of Reddit. Token accuracies for within-corpus tests are reported in Table 4, and token accuracies for between-corpus tests are reported in Table 5.

Model	Corpus			
	OntoNotes	Brown	SWBD	EWT
TnT	.9623	.9564	.9557	.9299
PPX	.9657	.9610	.9676	.9313
Flair	.9790	.9729	.9786	.9673

Table 4: Token accuracies for within-corpus comparison of POS-tagger models.

As in sentence boundary detection, the best performance for all models is consistently achieved when models are trained and tested using data from the same corpus. Flair always produces the best result, consistent with its status as the state-of-the-art model for part-of-speech tagging. Indeed, we are able to replicate published state-of-the-art performance for Flair and TnT using the Penn Treebank, and we replicate performance on OntoNotes reported by Gorman and Bedrick (2019). When testing on the Reddit corpus, TnT achieves the best performance when trained on EWT, but the other two models perform best when trained on OntoNotes. This result illustrates both Reddit's similarity to EWT (which provided the best performance when used as a training set for SBD) and OntoNotes's applicability as a good universal training set, since it generalizes well to Reddit. However, this may also be attributed to the fact that the Reddit corpus is tagged in the style of OntoNotes with a very similar tagset (with the exception of the `GW` tag; further discussion can be found in Section 2 or Experiment 3). Questions of variability in tagsets and tagging conventions and their effect on performance are revisited in Experiment 3.

The Penn Treebank remains one of the most widely used corpora in various NLP tasks, including standard training and testing data for part-of-speech tagging. Therefore, because OntoNotes is a more well-annotated portion of this corpus, we use OntoNotes as the training data for all models in the remaining experiments for this paper, unless otherwise indicated.

Test Corpus	Model	Training Corpus			
		OntoNotes	Brown	SWBD	EWT
OntoNotes	TnT		.9295	.9069	.8782
	PPX		.9329	.8841	.8772
	Flair		.9498	.9386	.9204
Brown	TnT	.9206		.9083	.8547
	PPX	.9257		.8953	.8555
	Flair	.9422		.9294	.8841
SWBD	TnT	.8653	.8979		.8339
	PPX	.8699	.8987		.8348
	Flair	.9306	.9317		.8908
EWT	TnT	.9295	.9073	.9137	
	PPX	.9187	.9016	.9053	
	Flair	.9606	.9333	.9447	
Reddit	TnT	.9167	.9093	.8943	.9200
	PPX	.9177	.9119	.8977	.9162
	Flair	.9430	.9288	.9231	.9418

Table 5: Token accuracies for cross-corpus comparison of POS-tagger models.

3.1.2 Experiment 1b: Cross-demographic evaluation

After examining the effect of varying training and testing data by corpus, we consider whether testing on data from different demographic groups produces similar impacts on model performance. For sentence boundary detection, each of the three model architectures were trained on the OntoNotes training set and evaluated on demographic partitions of the OntoNotes test set and the Reddit corpus. Details about these partitions can be found in Section 2.1 and Table 1. We restrict this analysis to the OntoNotes and Reddit corpora, as the Brown Corpus does not contain information about gender and EWT and Switchboard are unsuitable for the sentence boundary detection task (as mentioned above). Results for this experiment are reported in Table 6.

	Female	Male
OntoNotes	.8977	.8899
	.9630	.9663
	.9659	.9581
Reddit	.9457	.9457
	.9242	.9330
	.9380	.9367

(a) Cross-gender comparison

	Old	Young
Reddit	.9279	.9356
	.9181	.9276
	.9349	.9482

(b) Cross-age comparison

Table 6: F1 scores for cross-demographic comparison of sentence boundary detection models.

No clear pattern emerges from the cross-gender comparison for sentence boundary detection: for all models, performance is not consistently better when testing on `Female` data compared to `Male` data. For the cross-age comparison, all models perform better when testing on `Young` authors as compared to `Old` authors. We hypothesize that higher performance on the data from `Young` users may result from younger users' stricter adherence to grammatical convention: in an effort to conform and be accepted as a member of the r/relationships forum and have their questions answered, these users may be less likely than older users to omit sentence-final punctuation or deviate from convention in other ways, resulting in sentences that are more easily automatically detected.

For part-of-speech tagging, the performance of each of the three model architectures was compared when trained on OntoNotes and evaluated on demographic-based corpus partitions. Token accuracies for these comparisons are reported in Table 7. Comparisons where 95% Wilson score confidence intervals for each partition do not overlap are indicated in bold.

As in SBD, there are no significant differences between performance on `Male` and `Female` data. In the cross-age comparison, however, five of the six comparisons show that performance on data from `Old` authors is significantly better than that of `Young` authors at the 95% confidence level, the opposite finding of the SBD experiment. One possible cause for this

	Female	Male
OntoNotes	.9604	.9602
	.9663	.9624
	.9810	.9777
SWBD	.9056	.9083
	.8835	.8848
	.9392	.9380
Reddit	.9196	.9140
	.9172	.9181
	.9464	.9398

(a) Cross-gender comparison

	Old	Young
SWBD	**.9099**	**.9036**
	.8898	**.8782**
	.9414	**.9357**
Reddit	**.9292**	**.9041**
	.9287	**.9065**
	.9464	.9396

(b) Cross-age comparison

Table 7: Token accuracy for cross-demographic comparison of POS-tagger models.

discrepancy is the relative ages of the authors compared to the authors of the data used in OntoNotes. As the journalists whose work was included in the corpus were writing in 1989, they are closer in age to the authors in the `Old` partitions of Switchboard and Reddit than the `Young` partition, though even authors in the `Old` partition of Reddit are younger than the `Young` Switchboard authors. This difference in ages may result in differences in vocabulary and usage that have an effect on POS-tagging results, but not necessarily conventions about punctuation and abbreviation that would lead to different styles of sentences.

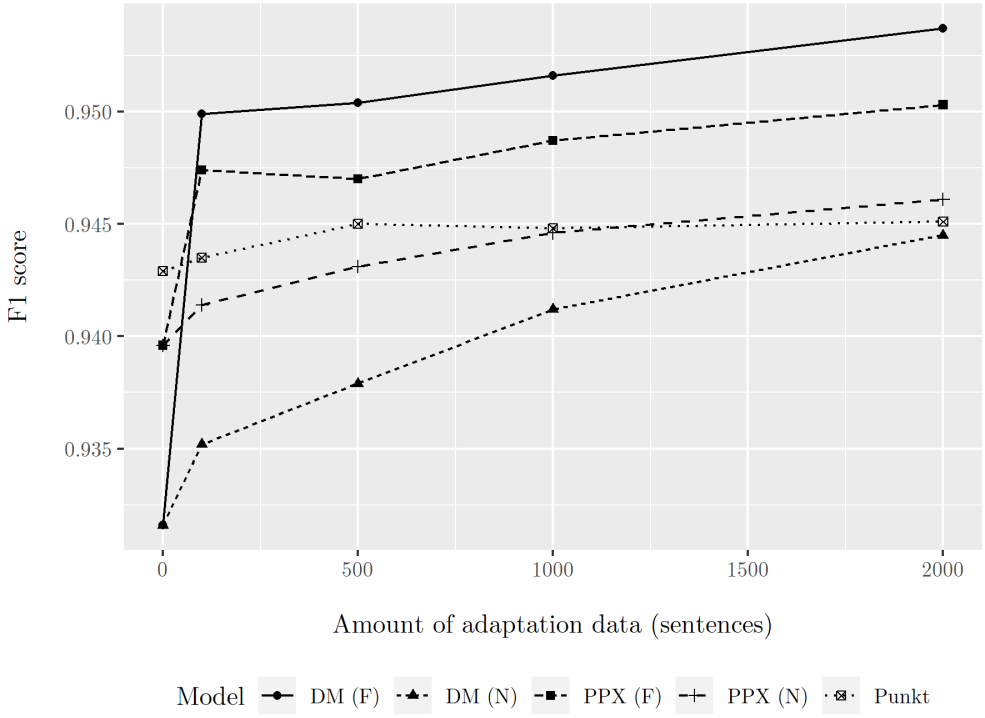
Though cross-demographic performance differences are smaller than those observed in cross-corpus performance, we proceed with further analysis of the applicability and effectiveness of demographic factors as domains for adaptation in these two tasks.

3.2 Experiment 2: Amount of adaptation data

In order to assess the practicality and effectiveness of the proposed adaptation strategies, we evaluate the performance of each model and adaptation strategy combination with increasing amounts of adaptation data. For sentence boundary detection, we adapt models trained on

OntoNotes with data from Reddit in increments of 100, 500, 1000, or 2000 sentences of manually annotated data. The F1 scores of these models are shown in Figure 1, and the absolute increase in F1 score of each model, relative to no adaptation, is reported in Table 8. Bold results in the table indicate comparisons between the indicated model and an analogous model with less adaptation data (or no adaptation data, when the indicated model has 100 sentences of adaptation data) which are **not** significant ($p < .001$) according to the two-sided, mid-p variant (Fagerland et al. 2013) of the McNemar test (Gillick and Cox 1989). We compute the McNemar statistic using the notion that a hypothesized sentence is counted as a 'win' if it is also a gold-annotated sentence.

Figure 1: F1 scores of sentence boundary detection models while varying amount of adaptation data.



Model	Adaptation Data (sentences)			
	100	500	1000	2000
Punkt (Naive)	**.0005**	.0021	**.0018**	**.0021**
DM (Naive)	.0037	.0063	.0096	.0129
DM (FEDA)	.0183	**.0188**	**.0200**	.0221
PPX (Naive)	**.0018**	**.0035**	**.0050**	**.0064**
PPX (FEDA)	.0078	.0074	.0090	**.0107**

Table 8: Absolute F1 score increase for sentence boundary detection models trained with varying amounts of adaptation data

All models show increases in F1 scores after the addition of just 100 sentences of adaptation data, regardless of adaptation strategy. DetectorMorse adapted with FEDA benefits the most from adaptation and achieves the best performance overall, followed by Perceptronix adapted with FEDA. Performance does not increase monotonically for Punkt and Perceptronix adapted with FEDA, but all models show a net increase in performance as more adaptation data is added.

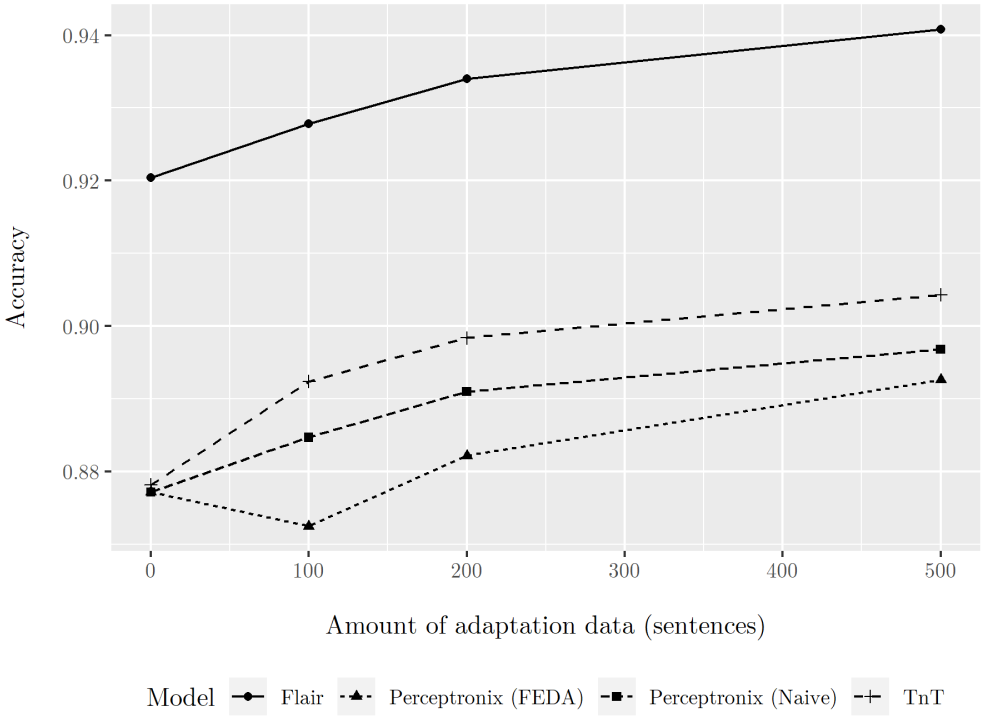
For POS-tagging, we test adaptation performance in two directions: adapting models trained on OntoNotes to EWT, and adapting models trained on EWT to OntoNotes. All models are tested using the corpus to which they have been adapted. Token accuracy for models trained on OntoNotes and adapted to EWT with 100, 200, or 500 sentences of adaptation data are shown in Figure 2. Absolute and relative error reduction achieved by the adapted models, with respect to no adaptation, is reported in Table 9. We calculate absolute and relative error reduction as follows, where x is the accuracy with the 'old' system and $\hat{x}$ is the accuracy with the 'new' system:

$$AbsoluteErrorReduction = \hat{x} - x \tag{1}$$

$$RelativeErrorReduction = 1 - \frac{1 - \hat{x}}{1 - x} \tag{2}$$

The results shown in bold indicate models which are not significantly better than models with less adaptation, according to the mid-p McNemar test (all other comparisons were significant at $p < .001$ according to the same test). Analogous results for models trained on EWT and adapted to OntoNotes are reported in Figure 3 and Table 10.

Figure 2: Token accuracy of POS-tagger models trained on OntoNotes, adapted to EWT



As with SBD, all models achieve a net increase in accuracy with the addition of more adaptation data, though the performance of the Perceptronix model adapted to EWT with FEDA suffers after the initial adaptation of 100 sentences. In a departure from the SBD results, POS-tagger models using the naive adaptation strategy outperform FEDA-adapted models for nearly all amounts of adaptation data. However, because Perceptronix is the only model for which FEDA can be implemented, this finding may not hold when more data points are considered.

Model	Adaptation Data (sentences)					
	100		200		500	
	AER	RER	AER	RER	AER	RER
TnT (Naive)	.0142	11.63	.0202	16.56	.0261	21.42
PPX (Naive)	.0075	6.08	.0137	11.19	.0196	15.97
PPX (FEDA)	**−.0047**	**N/A**	.0050	4.10	.0154	12.52
Flair (Naive)	.0075	9.25	.0136	17.10	.0204	25.59

Table 9: Absolute and relative error reduction for POS-tagger models trained on OntoNotes, adapted to EWT.

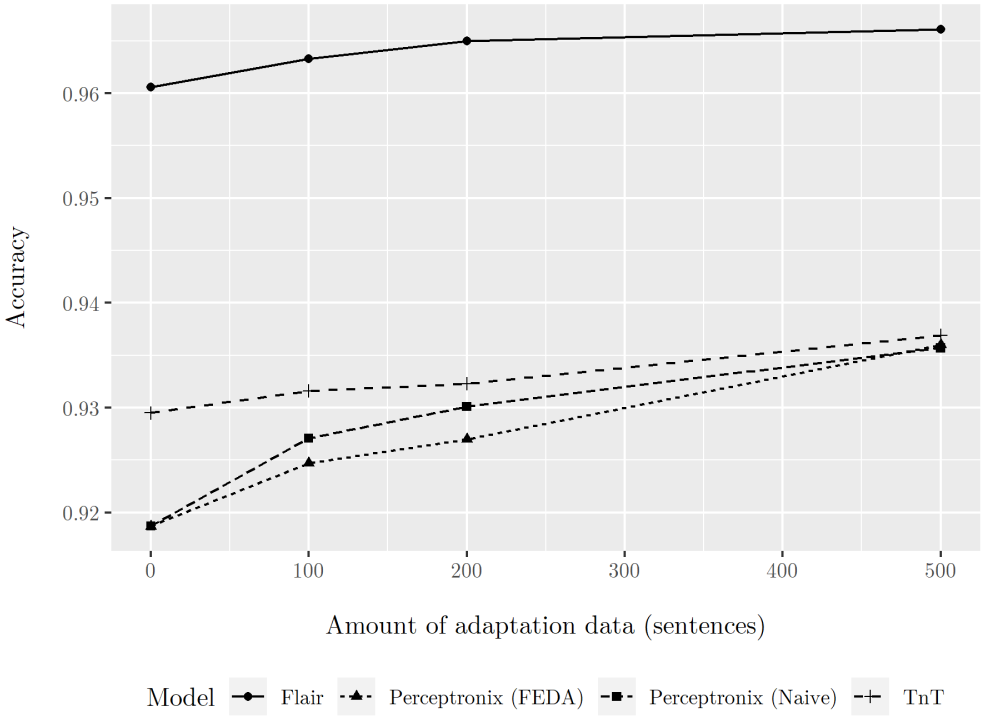
Model	Adaptation Data (sentences)					
	100		200		500	
	AER	RER	AER	RER	AER	RER
TnT (N)	.0021	2.93	.0028	4.03	.0074	10.48
PPX (N)	.0084	10.31	**.0114**	**14.01**	.0171	20.99
PPX (F)	.0060	7.42	**.0083**	**10.23**	.0173	21.29
Flair (N)	.0027	6.84	.0044	11.25	.0055	13.96

Table 10: Absolute and relative error reduction for POS-tagger models trained on EWT, adapted to OntoNotes.

This experiment finds that utilization of either selected adaptation strategy results in increased performance for both sentence boundary detection and part-of-speech tagging. However, especially for POS tagger models adapted with small amounts of data (100 sentences), the naive adaptation strategy is more reliable – FEDA requires more adaptation data to see comparable gains in performance.

Though based on the trends in the charts it appears as though models with more adaptation data would continue to perform better, the experiments reported here are limited to 500 sentences of adaptation data for POS-tagging and 2000 sentences of adaptation data for sentence boundary detection, in an attempt to limit exploration to amounts of data

Figure 3: Token accuracy POS-tagger models trained on EWT, adapted to OntoNotes



which could be feasibly annotated by hand. For the following experiments, we maintain this limitation in the name of practicality: the amount of data used for adaptation is 100 sentences for POS-tagging and 1000 sentences for SBD. In preparing the Reddit data for these experiments, manual annotation of this amount of data typically took about 90 minutes for either task. This is similar to results reported by Garrette and Baldridge (2013), in which around 100 sentences of the Penn Treebank are hand-annotated for part-of-speech in 2 hours. Though the actual amount of time will vary with the annotator's familiarity with the annotation guidelines, this benchmark of 100 sentences of part-of-speech data and 1000 sentences of sentence boundary detection data intends to be a reasonable compromise between practicality and potential effectiveness.

3.3 Experiment 3: Full vs. Universal tagset

The Universal Tagset, described by Petrov et al. (2012), consists of 14 universal part-of-speech categories that cover the most frequent parts of speech in most languages. Among other goals, it aims to generalize well across languages and enable the evaluation of multiple corpora or treebanks on a single tagset. Compared to a typical English tagset, it is much simpler. For example, distinctions between types of verbs, such as VBG (gerund), VBZ (3SG), VBD (past tense), VBN (participle) and VB (bare verb), are all collapsed into one VERB tag in the Universal tagset.

Our motivation for using the Universal tagset here is as follows: if much or most of the errors made by POS taggers when training and testing on different corpora is due to different tags used in each corpus, we should see some of these differences disappear when we evaluate using the Universal tagset instead. In addition, we expect accuracy when evaluating on the full tagsets to increase once the models have been adapted to the target domain, as the models will learn the new tags and be exposed to new applications of familiar tags during training. We do not expect a similar increase in accuracy after adaptation when evaluating using the Universal tagset, as the tags that differed in the full tagset should have mapped down to the same Universal tag.

To evaluate on the Universal tagset, we create tag mappings to the Universal tagset from the OntoNotes, Switchboard, EWT, and Reddit tagsets. A mapping from the Penn Treebank to the Universal tagset is available,[9] but not so for the other corpora. Tag dictionaries for the remaining corpora were built by modifying the PTB mapping and consulting the respective tagging guidelines for each corpus. All tagsets map to the Universal tagset used in UD version 2.[10] All tag mappings, as well as their corresponding parts of speech, can be found in Appendix A.

[9] https://github.com/slavpetrov/universal-pos-tags
[10] https://universaldependencies.org/u/pos/

On differences in tagging guidelines Though the tagging guidelines for OntoNotes, Switchboard, and EWT are all based on the Penn Treebank guidelines described by Santorini (1990), there are several notable differences. The Penn Treebank tag mapping contains 68 tags, including rare tags like "NN|NNS", indicating that the selected token is ambiguous between a singular and plural noun. The EWT tagset contains 50 tags, none of which indicate ambiguity. In addition, EWT contains tokens such as email addresses and web URLs that are not present in OntoNotes training data (though these are given their own tag: ADD). The Switchboard tagset, obtained by compiling all gold part-of-speech tags used in the Switchboard corpus, contains 94 tags, some of which also indicate ambiguity. While most of the tags are shared with OntoNotes (as is the case with EWT), there are several tags that are only present in the Switchboard tagset. Many of these include the caret symbol (ˆ), an indication that there is a typo in the transcription of the corresponding token (Calhoun et al. 2009). This presents an additional difficulty in evaluating on the Switchboard corpus: typos in the transcription make the correct prediction by any of the POS tagger models extremely difficult. Below are three examples from the Switchboard tagging guidelines:

> right/ˆVB a book about it
>
> He one/ˆVBD the race
>
> know/ˆDT matter where you build it

All three examples present difficulties to taggers. Though it is imaginable that the correct part of speech may be predicted from the surrounding words and their parts of speech, predicting unseen (and unattested) token/tag pairs like those shown in the second and third example is highly unlikely.

There are two logically possible strategies for evaluation on the Universal tagset: mapping the full tagset to the Universal tagset before training (in which case all predicted tags will

use the Universal tagset), or training on the full tagset and mapping predicted tags to the Universal tagset. Results for an experiment which compares evaluation on the full tagset to these two strategies are reported in Table 11. It reports results for training and evaluating on the full tagset (F/F), training and evaluating on the Universal tagset (U/U), and training on the full tagset and evaluating on the Universal tagset (F/U). All results reported in Table 11 were trained with the OntoNotes training set.

Test Corpus	Model	F/F	U/U	F/U
OntoNotes	TnT	.9623	.9634	.9713
	PPX	.9675	.9723	.9732
	Flair	.9790	.9845	.9845
EWT	TnT	.8782	.8915	.8999
	PPX	.8772	.8968	.8971
	Flair	.9204	.9369	.9389
SWBD	TnT	.9069	.9226	.9379
	PPX	.8841	.9115	.9100
	Flair	.9386	.9604	.9617

Table 11: Token accuracies for POS-tagger models when training and evaluating on the Full (F) or Universal (U) tagset.

Though evaluating on the Universal tagset always results in increased token accuracy, the best results are typically obtained in the case where the models are trained using the full tagset and predicted tags are mapped down to the Universal tagset. This finding is also reported in Petrov et al. 2012, which reasons that "the transition model based on the universal POS tagset is less informative" (2019). For the remaining experiments in this section which evaluate on the Universal tagset, we adopt the strategy of training on the full tagset and evaluating on the Universal tagset.

Table 12 reports, for each POS tagger model, the token accuracy with no adaptation when evaluating on the full and Universal tagsets. It also reports the absolute and relative

error reduction when comparing no adaptation to naive adaptation with 100 sentences of data from the test domain. We evaluate on EWT and Switchboard, whose tagsets diverge from OntoNotes, and Reddit, tagged in the same style as OntoNotes (with the addition of the GW tag for the abbreviation "TL;DR").

Test Corpus	Model	Full Tagset			Universal Tagset		
		Acc	AER	RER	Acc	AER	RER
SWBD	TnT	.9069	.0103	11.07	.9379	.0078	12.58
	PPX	.8841	.0263	22.68	.9100	.0258	28.65
	Flair	.9386	.0064	10.47	.9617	.0024	6.26
EWT	TnT	.8782	.0074	9.25	.8999	.0145	14.51
	PPX	.8772	.0075	6.07	.8971	.0070	6.84
	Flair	.9204	.0142	11.63	.9389	.0037	6.10
Reddit	TnT	.9167	.0016	1.94	.9398	.0017	2.87
	PPX	.9177	.0011	1.39	.9370	.0016	2.53
	Flair	.9430	.0025	4.46	.9590	.0019	4.53

Table 12: Token accuracy for no adaptation and absolute and relative error reduction for models with 100 sentences of adaptation when evaluated on full and Universal tagsets.

While not a universal finding (except for evaluation on Reddit), there is typically a greater relative error reduction when evaluating on the Universal tagset than on the full tagset, in line with the stated predictions: discrepancies in the full tagset that map down to the same Universal tag are eliminated. Reddit reliably shows the most modest increases in relative error reduction when comparing the results of evaluation on the full and universal tagsets. As the Reddit data was annotated using the OntoNotes tagset, this result is expected: rather than resulting from learning the correct use of new tags, the error reduction is only gained through resolving incorrect tags that map to the same universal tag. These results show that while evaluating on the Universal tagset does result in higher accuracy as fine distinctions are collapsed together, models still benefit greatly from adaptation. However, adapting with

more data may be beneficial, as 100 sentences of adaptation data does not always appear to be enough to encourage learning new tags or token-tag pairs well.

Table 13 reports the five most frequent gold/predicted missed tags across all models when evaluated using the full or Universal tagset, with and without adaptation.

Test Corpus	Most Frequent Incorrect Tags			
	Full Tagset		Universal Tagset	
	None	100	None	100
SWBD	BES/VBZ	BES/VBZ	INTJ/PROPN	INTJ/ADV
	UH/NNP	UH/RB	INTJ/ADV	ADV/ADP
	UH/RB	RB/IN	ADV/ADP	DET/ADP
	RB/IN	UH/NNP	INTJ/NOUN	INTJ/NOUN
	UH/NN	IN/RB	DET/ADP	ADV/ADJ
EWT	NN/NNP	NN/NNP	NOUN/PROPN	NOUN/PROPN
	NNP/NN	NNP/NN	PROPN/NOUN	PROPN/NOUN
	JJ/NNP	JJ/NNP	ADJ/PROPN	ADJ/PROPN
	,/:	,/:	ADV/ADP	ADV/ADP
	RB/IN	RB/IN	PUNCT/NOUN	ADJ/VERB
Reddit	RB/IN	RB/IN	ADV/ADP	ADV/ADP
	VBP/VB	VBP/VB	VERB/NOUN	ADV/ADJ
	PRP/PRP$	PRP/PRP$	ADJ/ADV	VERB/NOUN
	JJ/RB	JJ/RB	ADV/ADJ	ADJ/ADV
	NN/JJ	NN/JJ	PRON/DET	PRON/DET

Table 13: Most frequently mispredicted tags for full and Universal tagsets with and without adaptation

Some of the most frequent incorrect token-tag pairs in the full tagset disappear immediately upon evaluation with the Universal tagset, notably the BES/VBZ distinction in evaluation on Switchboard (OntoNotes does not use the BES tag). Other popular incorrect token-tag pairs persist in evaluation on either tagset, such as the noun/proper noun distinction issue in EWT. Besides the distinctions that collapse in evaluation on the Universal tagset, the ranking of frequently incorrect tags largely remains the same.

While the Universal Tagset is an applicable tool for considering differences in tagging guidelines and an excellent resource for cross-lingual parsing and other tasks, it obscures some more fine-grained distinctions that may be made between parts of speech, and we do not recommend it to be used as a substitute for richer tagsets.

3.4 Experiment 4: Adaptation to demographic group

After establishing demographic factors as potential domains for adaptation in POS-tagging and sentence boundary detection, we explore the effectiveness of different styles of adaptation to demographic factors.

In the following experiments, we evaluate models and adaptation strategies in combination with one of three styles of adaptation:

- Adaptation to **corpus**: for a given target domain, we adapt with data from the target domain, maintaining the overall demographic distribution of the target corpus.
- Adaptation to **opposite demographic group**: for a given target domain, we adapt with data from the same domain but from the "opposite" demographic (e.g., evaluate `Male` data with `Female` adaptation).
- Adaptation to **corpus and demographic group**: for a given target domain, we adapt directly to the target domain in corpus *and* demographic distribution.

Tables 14 and 15 show the absolute increase in F1 score (a proxy for absolute error reduction) of applicable model and adaptation strategy combinations relative to no adaptation for test corpora divided by gender and age, respectively. All models were trained using OntoNotes and adapted with 1000 sentences in the given adaptation style. For each model, results for the adaptation style which produces the greatest difference in F1 score are indicated in bold.

Test Corpus	Model	Adaptation to		
		Corpus	Opposite demo	Demo + Corpus
Reddit - `Female`	Punkt (N)	**.0036**	.0002	−.0002
	DM (N)	**.0193**	.0089	.0070
	DM (F)	**.0373**	.0303	.0262
	PPX (N)	**.0145**	.0056	.0099
	PPX (F)	**.0237**	.0117	.0180
Reddit - `Male`	Punkt (N)	−.0028	**.0013**	.0011
	DM (N)	.0050	.0076	**.0095**
	DM (F)	**.0192**	.0141	.0186
	PPX (N)	.0017	**.0073**	.0023
	PPX (F)	**.0066**	.0019	.0059

Table 14: Absolute F1 score increase for sentence boundary detection models with adaptation to gender.

For text written by `Female` Reddit users, adapting to the Reddit corpus overall provides better performance than adapting to either `Male` or `Female` Reddit data. For `Male` Reddit users, there is no clear adaptation style that produces the best results over all models. Evaluation on both age-divided subsets of the Reddit corpus shows that adapting to the overall corpus tends to provide the best performance.

Results for the analogous POS-tagging experiments are reported in Tables 16 and 17. These tables report relative error reduction with respect to to no adaptation for each model and adaptation style. Entries marked Not Applicable indicate instances where the token accuracy of the model without adaptation exceeded that of the model with adaptation, resulting in no relative error reduction. For each model, the adaptation style which produced the highest relative error reduction is indicated in bold.

Test Corpus	Model	Adaptation to		
		Corpus	Opposite demo	Demo + Corpus
Reddit - `Old`	Punkt (N)	.0081	.0014	.0019
	DM (N)	**.0210**	.0074	.0119
	DM (F)	.0228	.0240	**.0275**
	PPX (N)	**.0157**	.0044	.0077
	PPX (F)	**.0211**	.0089	.0126
Reddit - `Young`	Punkt (N)	.0007	**.0021**	.0009
	DM (N)	**.0143**	.0076	.0063
	DM (F)	**.0312**	.0244	.0204
	PPX (N)	.0117	**.0067**	.0040
	PPX (F)	**.0205**	.0144	.0136

Table 15: Absolute F1 score increase for sentence boundary detection models with adaptation to age.

For `Female` data in both Switchboard and Reddit, adapting to corpus and demographic group together proves to be the most effective strategy. Especially for Switchboard, whose tagset differs more substantially from that of the training data, adaptation to corpus in general provides substantial relative error reduction for all models.

Though most pronounced in comparisons on Switchboard, a general finding is that adapting to the demographic group that is least represented in the training corpus provides the greatest relative error reduction for all demographic groups. Adapting to `Young` Switchboard speakers results in the greatest reduction of error when testing on `Young` *and* `Old` Switchboard users, and the same pattern is present in Reddit and for gender comparisons – adapting to `Female` data frequently results in the greatest reduction of error on `Male` and `Female` data. These results suggest that sentence boundary detection in general is less sensitive to demographic differences, and that for part-of-speech tagging, the greatest benefits when evaluating on any partition are seen when models are adapted not just to the test corpus, but to the demographic group within that corpus that is least represented in the

training data. As OntoNotes is a predominantly `Old` (especially when compared to Reddit) and predominantly `Male` corpus (Garimella et al. 2019), adaptation to `Female` data improves performance for both gendered subgroups for Reddit and Switchboard, and adaptation to `Young` data produces the same result for both corpora.

Test Corpus	Model	Adaptation to		
		Corpus	Opposite demo	Demo + Corpus
Reddit - Female	TnT (N)	3.06	2.38	**5.10**
	PPX (N)	5.28	3.30	**7.92**
	PPX (F)	N/A	N/A	N/A
	Flair (N)	13.25	7.35	**16.89**
Reddit - Male	TnT (N)	0.90	1.50	**2.40**
	PPX (N)	N/A	**4.10**	1.26
	PPX (F)	N/A	N/A	N/A
	Flair (N)	**6.01**	3.00	4.72
SWBD - Female	TnT (N)	12.10	11.75	**14.18**
	PPX (N)	23.09	24.30	**25.83**
	PPX (F)	32.25	**35.68**	35.07
	Flair (N)	10.52	9.97	**14.52**
SWBD - Male	TnT (N)	**9.97**	9.88	9.89
	PPX (N)	22.18	**23.01**	22.30
	PPX (F)	29.13	31.37	**31.38**
	Flair (N)	10.39	**11.14**	8.55

Table 16: Relative error reduction for POS-tagger models with adaptation to gender

Test Corpus	Model	Adaptation to		
		Corpus	Opposite demo	Demo + Corpus
Reddit - `Old`	TnT (N)	1.87	**2.61**	2.24
	PPX (N)	0.0	**1.48**	0.37
	PPX (F)	N/A	N/A	N/A
	Flair (N)	**3.46**	0.97	0.02
Reddit - `Young`	TnT (N)	1.95	2.51	**3.62**
	PPX (N)	2.57	2.57	**6.86**
	PPX (F)	N/A	**0.86**	N/A
	Flair (N)	**4.43**	3.54	3.54
SWBD - `Old`	TnT (N)	10.38	**11.50**	9.45
	PPX (N)	21.74	**23.05**	20.93
	PPX (F)	29.55	**32.92**	30.44
	Flair (N)	11.03	**14.68**	12.18
SWBD - `Young`	TnT (N)	11.80	10.78	**13.69**
	PPX (N)	23.53	22.40	**24.28**
	PPX (F)	31.93	32.79	**35.91**
	Flair (N)	9.89	12.17	**15.15**

Table 17: Relative error reduction for POS-tagger models with adaptation to age.

4 Discussion

The four experiments presented in the preceding sections have attempted to investigate the appropriateness, efficacy, and practicality of adaptation to demographic groups for part-of-speech tagging and sentence boundary detection. Here we evaluate the results as a whole and make recommendations for future work.

Much of this work has been premised on recent literature which suggests that demographic characteristics of the authors of text data have a role to play in the evaluation of this data using language processing tools and models, especially when these tools are created using resources that are outdated or potentially biased in other ways. We attempt to address this first in Experiment 1, which finds that without adaptation, there are not significant performance differences across demographic groups for either POS-tagging or sentence boundary detection. However, we do see statistically significant performance differences when comparing POS-tagging performance on data from `Old` authors with `Young` authors, consistent with conclusions presented by Hovy (2015). The goal of Experiment 2 was to assess both practicality and effectiveness of the adaptation strategies in general, finding that both the naive strategy and FEDA were effective at improving model performance with as few as 100 sentences of target domain data for either task.

Though we demonstrate that adaptation to demographic groups with manually annotated data may be effective and lead to increased performance, it is important to consider that such a strategy is not always possible or practical. The Reddit corpus presented here is unique in that its authors have self-identified their gender and age, which allows us to proceed with a careful and informed representation of each variable. However, a situation like this one is rare, as age and gender are seldom recorded or available in collections of text corpora or in web text that may be collected for processing. There is also the separate fact that, even if demographic information for the proposed target corpus is available, it may need to be

hand-annotated. While we show in Experiments 2 and 4 that 100 sentences with part-of-speech annotations is enough adaptation data for these benefits to be visible, this process requires some linguistic expertise or significant time investment that may not be possible, especially when pretrained models for these tasks are so readily available.

Though adaptation to demographic groups may not always be a feasible suggestion, especially for preprocessing tasks that are easier to accomplish with pretrained models, we hope that the illustration here, that even 100 sentences of target data can facilitate a statistically significant performance improvement, is useful. Rather than outline a protocol for identifying how to adapt preprocessing tools to demographic distributions, we hope that the strategies demonstrated here illustrate that not only is adaptation like this possible, but that it is relatively accessible, even simple under certain circumstances.

5 Conclusion

This thesis has explored whether adapting linguistic preprocessing models to specific demographic distributions can improve performance when evaluating with the same or other demographic distributions. We report that distributions based on age or gender are both appropriate and effective target domains for part-of-speech tagging, while adapting to the target corpus as a whole, rather than to a specific demographic distribution, is the most effective strategy for sentence boundary detection. Additionally, we report that the most effective strategy to improve part-of-speech tagging performance over all demographic groups is to adapt using data from the target corpus that is written by members of the least-represented demographic group in the source corpus. This is a powerful finding, arguing that considering which demographic groups are represented in training data for language processing tasks can increase performance for all groups, not just a minority that may be particularly disadvantaged.

One goal of this work is to provide further evidence that biases that occur when using imbalanced or outdated corpora are real, whether we choose to use adaptation as a bias mitigation strategy or not. Whether these strategies chosen to address these biases are those explored here or those proposed in future work, we hope this work can be a motivating factor in considering training corpus demographics, even (or especially) in preprocessing tasks. As attention in the field has in part shifted away from the creation of high-quality training and evaluation resources in favor of improving the state-of-the-art on well-established tasks and developing new model architectures, it is of value to consider the potential effects of these aging resources and evaluate strategies to combat them.

A Appendix A

Table 18: Corpus-specific tags and Universal tag mappings

OntoNotes	EWT	SWBD	Universal	Part of Speech
$	$	$	SYM	Dollar Sign
”	”	”	PUNCT	Right Quotes
,	,	,	PUNCT	Comma
-LRB-	-LRB-	-LRB-	PUNCT	Left Round Bracket
-RRB-	-RRB-	-RRB-	PUNCT	Right Round Bracket
.	.	.	PUNCT	Period
:	:	:	PUNCT	Colon
	ADD		X	Web Address
AFX	AFX	AFX	X	Affix
		BES	VERB	3SG Be
CC	CC	CC	CCONJ	Coordinating Conjunction
CD	CD	CD	NUM	Number
DT	DT	DT	DET	Determiner
EX	EX	EX	PRON	Expletive Pronoun
FW	FW	FW	X	Foreign Word
	GW	GW	X	Non-final Token
		HVS	VERB	3SG Have
HYPH	HYPH	HYPH	PUNCT	Hyphen
IN	IN	IN	ADP	Preposition
JJ	JJ	JJ	ADJ	Adjective
JJR	JJR	JJR	ADJ	Comparative Adjective
JJS	JJS	JJS	ADJ	Superlative Adjective
		JJ|RB	ADJ	
LS	LS	LS	X	List Item
MD	MD	MD	VERB	Modal Verb
NFP	NFP	NFP	SYM	Non-Final Punctuation
NN	NN	NN	NOUN	Noun
NNP	NNP	NNP	PROPN	Proper Noun
NNPS	NNPS	NNPS	PROPN	Plural Proper Noun
NNS	NNS	NNS	NOUN	Plural Noun
		NNSˆPOS	NOUN	
PDT	PDT	PDT	DET	Predeterminer
POS	POS	POS	PART	Possessive
PRP	PRP	PRP	PRON	Pronoun
PRP$	PRP$	PRP$	DET	Possessive Pronoun

Continued on next page

Table 18 – *Continued from previous page*

OntoNotes	**EWT**	**SWBD**	**Universal**	**Part of Speech**
RB	RB	RB	ADV	Adverb
RBR	RBR	RBR	ADV	Comparative Adverb
RBS	RBS	RBS	ADV	Superlative Adverb
RP	RP	RP	ADP	Particle
SYM	SYM	SYM	SYM	Symbol
TO	TO	TO	PART	"to"
		TO|IN	ADP	
UH	UH	UH	INTJ	Interjection
		UH|IN	INTJ	
VB	VB	VB	VERB	Base Form Verb
VBD	VBD	VBD	VERB	Past Tense Verb
VBG	VBG	VBG	VERB	Gerund
VBN	VBN	VBN	VERB	Past Participle
		VBN|VBD	VERB	
VBP	VBP	VBP	VERB	Present Tense Verb
VBZ	VBZ	VBZ	VERB	3SG Present tense
WDT	WDT	WDT	DET	Wh- Determiner
WP	WP	WP	PRON	Wh- Pronoun
WP$	WP$	WP$	DET	Possessive Wh- Pronoun
WRB	WRB	WRB	ADV	Wh- Adverb
	XX	XX	X	Partial Word
``	``	``	PUNCT	Left Quotes
		^BES	VERB	
		^CC	CCONJ	
		^CD	NUM	
		^CD^NN	NOUN	
		^CD^NNS	NOUN	
		^DT	DET	
		^DT^NN	NOUN	
		^EX	PRON	
		^GW	X	
		^IN	ADP	
		^IN^IN	ADP	
		^IN^NN	NOUN	
		^JJ	ADJ	
		^JJS	ADJ	
		^NN	NOUN	
		^NNP	NOUN	

Continued on next page

Table 18 – *Continued from previous page*

OntoNotes	EWT	SWBD	Universal	Part of Speech
		^NNPS	NOUN	
		^NNP^NNP	PROPN	
		^NNP^POS	PROPN	
		^NNS	NOUN	
		^NNS^POS	NOUN	
		^NN^NNS	NOUN	
		^PDT	DET	
		^PRP	PRON	
		^PRP$	DET	
		^PRP^VBP	VERB	
		^RB	ADV	
		^RP	ADP	
		^TO	PART	
		^UH	INTJ	
		^VB	VERB	
		^VBD	VERB	
		^VBG	VERB	
		^VBN	VERB	
		^VBP	VERB	
		^VBP^RB	VERB	
		^VBP^RP	VERB	
		^VBZ	VERB	
		^VB^RP	VERB	
		^WP	PRON	
		^WRB	ADV	

References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English Web Treebank LDC2012T13.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Brants, T. (2000). TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231, Seattle, Washington. ACL.

Calhoun, S., Carletta, J., Jurafsky, D., Nissim, M., Ostendorf, M., and Zaenen, A. (2009). NXT Switchboard Annotations LDC2009T26.

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

Fagerland, M. W., Lydersen, S., and Laake, P. (2013). The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(1):91.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Gillick, D. (2009). Sentence boundary detection and the problem with the US. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244.

Gillick, L. and Cox, S. J. (1989). Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 532–535. IEEE.

Gorman, K. and Bedrick, S. (2019). We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Grieve, T. (2003). *The decline and fall of the Enron empire*. https://www.salon.com/2003/10/14/enron_22/. Accessed September 1, 2019.

Hovy, D. (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.

Hovy, D., Johannsen, A., and Søgaard, A. (2015). User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee.

Hovy, D. and Søgaard, A. (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.

Hovy, D. and Spruit, S. L. (2016). 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.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Liu, Y., Stolcke, A., Shriberg, E., and Harper, M. (2005). Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 451–458. Association for Computational Linguistics.

Lynn, V., Son, Y., Kulkarni, V., Balasubramanian, N., and Schwartz, H. A. (2017). Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey. European Languages Resources Association (ELRA).

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Shen, J. H., Fratamico, L., Rahwan, I., and Rush, A. M. (2018). Darling or babygirl? Investigating stylistic bias in sentiment analysis. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Silveira, N., Dozat, T., De Marneffe, M.-C., Bowman, S. R., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *LREC*, pages 2897–2904.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, chapter OntoNotes: A Large Training Corpus for Enhanced Processing. Springer.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., and Houston, A. (2013). OntoNotes Release 5.0 LDC2013T19.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.