Learning to Recognize Dialect Features

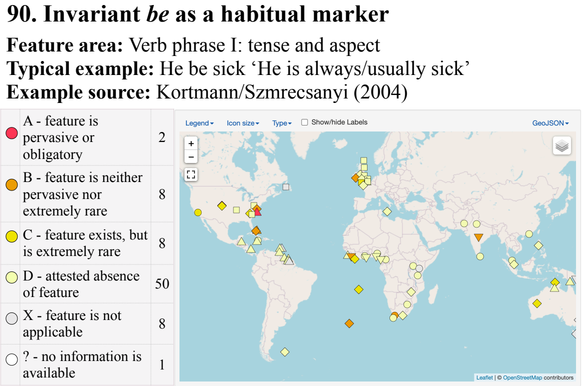
Dorottya Demszky[1]* Devyani Sharma[2] Jonathan H. Clark[3]
Vinodkumar Prabhakaran[3] Jacob Eisenstein[3]
[1]Stanford Linguistics [2]Queen Mary University of London [3]Google Research
ddemszky@stanford.edu
d.sharma@qmul.ac.uk
{jhclark,vinodkpg,jeisenstein}@google.com

Abstract

Linguists characterize dialects by the presence, absence, and frequency of dozens of interpretable features. Detecting these features in text has applications to social science and dialectology, and can be used to assess the robustness of natural language processing systems to dialect differences. For most dialects, large-scale annotated corpora for these features are unavailable, making it difficult to train recognizers. Linguists typically define dialect features by providing a small number of *minimal pairs*, which are paired examples distinguished only by whether the feature is present, while holding everything else constant. In this paper, we present two multitask learning architectures for recognizing dialect features, both based on pretrained transformers. We evaluate these models on two test sets of Indian English, annotated for a total of 22 dialect features. We find these models learn to recognize many features with high accuracy; crucially, a few minimal pairs can be nearly as effective for training as thousands of labeled examples. We also demonstrate the downstream applicability of our dialect feature detection model as a dialect density measure and as a dialect classifier.

1 Introduction

Dialect variation is a pervasive aspect of language, and understanding and accounting for it is essential if we are to build robust natural language processing (NLP) systems that serve everyone. Linguists characterize dialects by the presence, absence, and frequency of dozens of interpretable *dialect features*, of the type shown in Figure 1. This feature-based perspective has several advantages: (1) developing precise characterizations of dialects and differences and similarities among them, allowing accurate predictions for (and interpretations of) variable language use; (2) disentangling the grammatical constructions that make up the dialect from the content that is frequently discussed in the dialect; (3) supporting the construction of interpretable and socially meaningful dialect density measures, which are powerful tools for understanding the role of language in society (e.g., Craig and Washington, 2002; Voigt et al., 2017); (4) making it possible to test the performance of NLP systems across dialect features, thus helping to ensure that NLP systems perform adequately even on examples that differ systematically from "high-resource" varieties such as U.S. English.


90. Invariant *be* as a habitual marker

Feature area: Verb phrase I: tense and aspect
Typical example: He be sick 'He is always/usually sick'
Example source: Kortmann/Szmrecsanyi (2004)

●	A - feature is pervasive or obligatory	2
●	B - feature is neither pervasive nor extremely rare	8
●	C - feature exists, but is extremely rare	8
○	D - attested absence of feature	50
○	X - feature is not applicable	8
○	? - no information is available	1

Figure 1: An example dialect feature from the Electronic World Atlas of Varieties of English (eWAVE), https://ewave-atlas.org/parameters/90#2/7.4/24.2

The main challenge for recognizing dialect features is the lack of labeled data and the difficulty of obtaining such data. Annotating dialect features requires linguistic expertise and it is highly time-consuming given the large number of features and their sparsity. In dialectology, large scale studies are limited to only those features that can be detected using regular expressions of surface forms

[*] Work done while at Google Research.

and parts-of-speech, e.g., `PRON be ADJ` for the HABITUAL *be* feature in Figure 1; several syntactic and semantic features cannot be detected with such patterns (e.g. object fronting, extraneous article). Furthermore, part-of-speech tagging is unreliable in language varieties that differ from the training data, such as regional and minority dialects (Jørgensen et al., 2015).

In this paper, we propose and evaluate learning-based approaches to recognize dialect features. We focus on Indian English, given the availability of domain expertise and labeled corpora for evaluation. First, we consider a standard multitask classification approach, in which a pretrained transformer (Vaswani et al., 2017) is fine-tuned to recognize a set of dialect features. The architecture can be trained from two possible sources of supervision: (1) thousands of labeled corpus examples, (2) a small set of *minimal pairs*, which are hand-crafted examples designed to highlight the key aspects of each dialect feature (as in the "typical example" field of Figure 1). Because most dialects have little or no labeled data,[1] the latter scenario is more realistic across most dialects. We also consider a novel multitask architecture, which learns across multiple features by encoding the feature names and descriptions, similar to recent work on few-shot or zero-shot multitask learning (Logeswaran et al., 2019; Brown et al., 2020).

Empirical evaluations of these models are described in Sections 4 and 5. To summarize our main findings:

- It is possible to learn to detect individual dialect features: several features can be recognized with high accuracy given a large corpus of about 10k labeled examples, and the best model achieves a macro-AUC of 0.857 across a set of ten grammatical features for which a large test set is available.
- Comparable performance (macro-AUC=0.827) can be achieved with little annotation, using roughly five minimal pairs per feature.
- Our dialect feature recognizers can be combined into a dialect density measure (DDM), allowing for the ranking of transcripts by dialect density.
- The dialect feature recognizers can also be used as document classifiers, determining whether conversational transcripts are in Indian or U.S. English.

2 Data and Features of Indian English

We develop methods for detecting 22 dialect features associated with Indian English. Even though India has over 125 million English speakers — making it the world's second largest English-speaking population — there is relatively little NLP research focused on Indian English. However, our methods are not designed around any specific properties of Indian English, and many of the features that are associated with Indian English are also present in other dialects of English.

Our training data comes from two main sources: an annotated corpus (§ 2.1) and minimal pairs (§ 2.2). For evaluation, we use corpus annotations exclusively. The features are described in Table 1, and our data is summarized in Table 2.

2.1 Corpus Annotations

The International Corpus of English (ICE; Greenbaum and Nelson, 1996) is a collection of corpora of world varieties of English, organized primarily by the national origin of the speakers/writers. We focus on annotations of spoken dialogs (S1A-001 – S1A-090) from the Indian English subcorpus (ICE-India). The ICE-India subcorpus was chosen in part because it is one of the only corpora to receive large-scale annotations of dialect features. In addition, to contrast Indian English with U.S. English (§ 4), we use the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) that constitutes the ICE-USA subcorpus of spoken dialogs.

We work with two main sources of dialect feature annotations in the ICE-India corpus:

Lange Features. The first set of annotations come from Lange (2012), who annotated 10 features in 100 transcripts. They chose a feature set that emphasizes the distinctive syntactic properties of Indian English related to topicalization and fronting. In our experiments, we use half of this data for training (50 transcripts, 9392 unique examples), and half for testing (50 transcripts, 9667 unique examples).

[1] While data is frequently labeled at the *language* level and sometimes at the level of nation-level *locales*, finding dialect-level annotations is far less common.

Feature	Example	Count of instantiations	
		Lange (2012)	Our data
article omission	*(the) chair is black*		59
direct object pro-drop	*she doesn't like (it)*		14
focus *itself*	*he is doing engineering in Delhi itself*	24	5
focus *only*	*I was there yesterday only*	95	8
habitual progressive	*always we are giving receipt*		2
stative progressive	*he is having a television*		3
lack of inversion in wh-questions	*what you are doing?*		4
lack of agreement	*he do a lot of things*		23
left dislocation	*my father, he works for a solar company*	300	19
mass nouns as count nouns	*all the musics are very good*		13
non-initial existential	*every year inflation is there*	302	8
object fronting	*minimum one month you have to wait*	186	14
PP fronting with reduction	*(on the) right side we can see a plate*		11
preposition omission	*I went (to) another school*		17
inversion in embedded clause	*I don't know what are they doing*		4
invariant tag *(isn't it, no, na)*	*the children are outside, isn't it?*	786	17
extraneous article	*she has a business experience*		25
general extender *and all*	*then she did her schooling and all*		7
copula omission	*my parents (are) from Gujarat*	71	
resumptive object pronoun	*my old life I want to spend it in India*	24	
resumptive subject pronoun	*my brother, he lives in California*	287	
topicalized non-argument constituent	*in those years I did not travel*	272	

Table 1: Features of Indian English used in our evaluations and their counts in the two datasets we study.

Dialect features		**Unique annotated examples**	
Feature set	Count	Corpus examples	Min. pair examples
Lange (2012)	10	19059	113
Extended	18	367	208

Table 2: Summary of our labeled data. All corpus examples for the Lange features are from ICE-India; for the Extended feature set, examples are drawn from ICE-India and the Sharma data.

Extended Features. To test a more diverse set of features, we additionally annotated 18 features on a set of 300 turns[2] randomly selected from the conversational subcorpus of ICE-India, as well as 50 examples randomly selected from a secondary dataset of sociolinguistic interviews (Sharma, 2009) to ensure diverse feature instantiation. We selected our 18 features based on multiple criteria: 1) relevance to Indian English based on the dialectology literature, 2) coverage in the data (we started out with a larger set of features and removed those with fewer than three occurrences), 3) diversity of linguistic phenomena. The extended features partially overlap with those annotated by Lange, yielding a total set of 22 features. Annotations were produced by consensus from two of the authors. To measure interrater agreement, a third author independently re-annotated 10% of the examples, with Cohen's $\kappa = 0.79$ (Cohen, 1960).

2.2 Minimal Pairs

For each of the 22 features, we created a small set of minimal pairs. The pairs were created by first designing a short example that demonstrated the feature, and then manipulating the example so that the feature is absent. This "negative" example captures the *envelope of variation* for the feature, demonstrating a site at which the feature could be applied (Labov, 1972). In this way, negative examples in minimal pairs carry more information than in the typical annotation scenario, where absence of evidence does not usually imply evidence of absence. In our minimal pairs, the negative examples were chosen to be acceptable in standard U.S. and U.K. English; minimal pairs can thus be viewed as situating dialects against a standard variety. Here are some example minimal pairs:

article omission: *chair is black* → *the chair is*

[2] We manually split turns that were longer than two clauses, resulting in 317 examples.

black

focus *only*: *I was there yesterday* only → *I was there just yesterday.*

non-initial existential: *every year inflation* is there → *every year there is inflation.*

For most features, each minimal pair contains exactly one positive and one negative example. However, in some cases where more than two variants are available for an example (e.g., for the feature INVARIANT TAG *(isn't it, no, na)*), we provide multiple positive examples to illustrate different variants. For Lange's set of 10 features, we provide a total of 113 unique examples; for the 18 extended features, we provide a set of 208 unique examples, roughly split equally between positives and negatives. The supplement provides more details, including a complete list of minimal pairs.

3 Models and training

We train models to recognize dialect features by fine-tuning the BERT-base transformer architecture (Devlin et al., 2019). We consider two strategies for constructing training data, and two architectures for learning across multiple features.

3.1 Sources of supervision

We consider two possible sources of supervision:

Minimal pairs. We apply a simple procedure to convert minimal pairs into training data for classification. The positive part of each pair is treated as a positive instance for the associated feature, and the negative part is treated as a negative instance. Then, to generate more data, we also include elements of other minimal pairs as examples for each feature: for example, a positive example of the RESUMPTIVE OBJECT PRONOUN feature would be a negative example for FOCUS *only*, unless the example happened to contain both features (this was checked manually). In this way, we convert the minimal pairs into roughly 113 examples per feature for Lange's features and roughly 208 examples per feature for the extended features. The total number of unique surface forms is still 113 and 208 respectively. Given the lack of labeled data for most dialects of the world, having existing minimal pairs or collecting a small number of minimal pairs is the most realistic data scenario.

Corpus annotations. When sufficiently dense annotations are available, we can train a classifier based on these labeled instances. We use 50 of the ICE-India transcripts annotated by Lange, which consists of 9392 labeled examples (utterances) per feature. While we are lucky to have such a large resource for the Indian English dialect, this high-resource data scenario is by no means typical across dialects.

3.2 Architectures

We consider two classification architectures:

Multihead. In this architecture, which is standard for multitask classification, we estimate a linear *prediction head* for each feature, which is simply a vector of weights. This is a multitask architecture, because the vast majority of model parameters from the input through the deep BERT stack remain shared among dialect features. The prediction head is then multiplied by the contextualized embedding for the [CLS] token to obtain a score for the applicability of a feature to a given instance.

DAMTL. Due to the few-shot nature of our prediction task, we also consider a novel architecture that attempts to exploit the natural language descriptions of each feature. This is done by concatenating the feature description to each element of the minimal pair. The instance is then labeled for whether the feature is present. This construction is shown in Figure 2. Prediction is performed by learning a single linear prediction head on the [CLS] token. We call this model *description-aware multitask learning*, or DAMTL.

Model details. Both architectures are built on top of the BERT-base model, and both are trained by cross-entropy. We use batches of size 32, perform fine-tuning for 100 epochs (due to the small size of the training data), and apply the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-5}. Extensive hyperparameter search is left for future work.

3.3 Regular expressions

In dialectology, regular expression pattern matching is the standard tool for recognizing dialect features (e.g., Nerbonne et al., 2011). For the fea-

y	x
1	[CLS] article omission [SEP] Chair is black. [SEP]
0	[CLS] article omission [SEP] The chair is black. [SEP]
0	[CLS] article omission [SEP] I was there yesterday only. [SEP]
...	...
1	[CLS] focus only [SEP] I was there yesterday only. [SEP]
0	[CLS] focus only [SEP] I was there just yesterday. [SEP]
0	[CLS] focus only [SEP] Chair is black. [SEP]
...	...

Figure 2: Conversion of minimal pairs to labeled examples for multi-task learning, using two minimal pairs.

tures described in Table 1, we were able to design regular expressions for only five.[3] Prior work sometimes relies on regular expressions that include both surface forms and part-of-speech (e.g., Bohmann, 2019), but part-of-speech cannot necessarily be labeled automatically for non-standard dialects (Jørgensen et al., 2015; Blodgett et al., 2016), so we consider only regular expressions over surface forms.

4 Results on Dialect Feature Detection

In this section, we present results on the detection of individual dialect features. Using the features shown in Table 1, we compare supervision sources (corpus examples versus minimal pairs) and classification architectures (multihead versus DAMTL) as described in § 3. To avoid tuning a threshold for detection, we report area under the ROC curve (ROC-AUC), which has a value of 0.5 for random guessing and 1 for perfect prediction.

4.1 Results on Lange Data and Features

We first consider the 10 syntactic features from (Lange, 2012), for which we have large-scale annotated data: the 100 annotated transcripts from the ICE-India corpus are split 50/50 into training and test sets. As shown in Table 3, it is possible to achieve a Macro-AUC above 0.85 overall with multihead predictions on corpus examples. However, with only a few minimal pairs per feature, we can approach this performance, with a Macro-AUC of 0.83. This is promising, because it suggests the possibility of recognizing dialect features for which we lack labeled corpus examples – and such low-data situations are by far the most common data scenario among the dialects of the world. On both sources of supervision, the conventional multihead architecture outperforms the DAMTL architecture. Regarding specific features, it is unsurprising that the lexical features (e.g., FOCUS *itself*) are easiest to recognize. The more syntactical features (e.g., COPULA OMISSION, RESUMPTIVE OBJECT PRONOUN) are more difficult, although some movement-based features (e.g., LEFT DISLOCATION, RESUMPTIVE SUBJECT PRONOUN) can be recognized accurately.

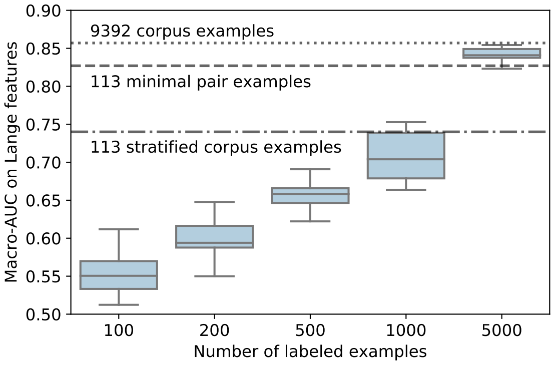


Figure 3: Performance of the multihead model as the number of corpus examples is varied. Box plots are over 10 random data subsets, showing the 25th, 50th, and 75th percentiles; whiskers show ±1.5 times interquartile range.

Low-resource evaluations. The minimal pair annotations consist of 113 examples; in contrast, there are 9392 labeled corpus examples, requiring far more effort to create. We now consider the situation when the amount of labeled data is reduced, focusing on the Lange features (for which labeled training data is available). As shown in Figure 3, 5000 labeled corpus examples are required to match the performance of roughly 5 minimal pairs per feature.

One reason that subsampled datasets yield weak results is that they lack examples for many features. To enable a more direct comparison of corpus examples and minimal pairs, we created a set of "stratified" datasets of corpus examples, such that the number of positive and negative examples for each feature exactly matches the mini-

[3] Features: FOCUS *itself*, FOCUS *only*, NON-INITIAL EXISTENTIAL, INVARIANT TAG *(isn't it, no, na)*, and GENERAL EXTENDER *and all*. Table 6 lists all regular expressions.

Supervision:	Corpus examples		Minimal pairs	
Architecture:	DAMTL	Multihead	DAMTL	Multihead
focus *itself**	0.905	0.953	0.836	0.911
focus *only**	0.984	0.936	0.992	0.929
invariant tag *(isn't it, no, na)*	0.995	0.986	0.969	0.899
copula omission	0.620	0.681	0.622	0.699
left dislocation	0.884	0.896	0.740	0.890
non-initial existential*	0.998	0.993	0.906	0.842
object fronting	0.786	0.827	0.618	0.764
resumptive object pronoun	0.627	0.606	0.654	0.829
resumptive subject pronoun	0.876	0.900	0.694	0.867
topicalized non-argument constituent	0.731	0.789	0.464	0.638
Macro-AUC	0.840	0.857	0.750	0.827

Table 3: ROC-AUC results on the Lange feature set, averaged across five random seeds. Asterisk (*) marks features that can be detected with relatively high accuracy (> 0.85 ROC-AUC) using regular expressions.

mal pair data. Averaged over ten such random stratified samples, the multihead model achieves a Macro-AUC of 0.740 ($\sigma = 0.031$), and the description-aware model achieves a Macro-AUC of 0.671 ($\sigma = .026$). These results are considerably worse than training on an equivalent number of minimal pairs, where the multihead model achieves a Macro-AUC of 0.827 and the DAMTL model achieves a Macro-AUC of 0.750.. Minimal pairs are thus especially useful as training data for recognizing dialect features, in comparison with typical labeled examples.

4.2 Results on Extended Feature Set

Next, we consider the extended features, for which we have sufficient annotations for testing but not training (Table 1). Here we compare the DAMTL and multihead models, using minimal pair data in both cases. As shown in Table 4, performance on these features is somewhat lower than on the Lange features, and for several features, at least one of the recognizers does worse than chance: DIRECT OBJECT PRO-DROP, EXTRANEOUS ARTICLE, MASS NOUNS AS COUNT NOUNS. These features seem to require deeper syntactic and semantic analysis, which appears to be difficult to learn from a small number of minimal pairs. On the other extreme, features with a strong lexical signature are recognized with high accuracy: GENERAL EXTENDER *and all*, FOCUS *itself*, FOCUS *only*. These three features can also be recognized accurately by regular expressions, as can non-initial existential.[4] However, for a number of other features, it is possible to learn a fairly accurate recognizer from just five minimal pairs: ARTICLE OMISSION, INVERSION IN EMBEDDED CLAUSE, LEFT DISLOCATION, LACK OF INVERSION IN WH-QUESTIONS. Unlike the Lange features, DAMTL performs slightly better than the multihead recognizer on the extended features, suggesting that feature descriptions may help in this case.

4.3 Summary of Dialect Feature Detection

Many dialect features can be automatically recognized with reasonably high discriminative power, as measured by area under the ROC curve. However, there are also features that are difficult to recognize in this way: this particularly includes features of omission (such as DIRECT OBJECT PRO-DROP and PREPOSITION OMISSION), as well as more semantic features such as MASS NOUNS AS COUNT NOUNS. While some of the features that can be recognized by a classifier can also be identified through regular expressions (e.g., FOCUS *only*), there are many features that can be learned but cannot be recognized by regular expressions. We now move from individual features to aggregate measures of dialect density.

5 Measuring Dialect Density

A dialect density measure (DDM) is an aggregate over multiple dialect features that tracks the vernacularity of a passage of speech or text. Such measures are frequently used in dialectological and education research (e.g., Craig and Washington, 2002; Van Hofwegen and Wolfram, 2010), and are also useful as predictors and dependent

[4] `\band all\b`, `\bitself\b`, `\bonly\b`, `\bis there\b|\bare there\b`

Dialect feature	DAMTL	Multihead
article omission	0.664	0.719
direct object pro-drop	0.478	0.547
extraneous article	0.494	0.536
focus *itself**	1.000	0.693
focus *only**	0.997	0.514
habitual progressive	0.553	0.588
invariant tag *(isn't it, no, na)*	0.958	0.696
inversion in embedded clause	0.643	0.729
lack of agreement	0.544	0.526
lack of inversion in wh-questions	0.656	0.554
left dislocation	0.672	0.797
mass nouns as count nouns	0.407	0.548
non-initial existential*	0.737	0.718
object fronting	0.576	0.678
preposition omission	0.553	0.631
PP fronting with reduction	0.548	0.663
stative progressive	0.595	0.585
general extender *and all**	0.963	0.897
Macro-AUC	0.669	0.645

Table 4: ROC-AUC results on the extended feature set, averaged across five random seeds. Because labeled corpus examples are not available for some features, we train only on minimal pairs. Asterisk (*) marks features that can be detected with relatively high accuracy ($>$ 0.85 ROC-AUC) using regular expressions.

variables in other social science research (e.g., Voigt et al., 2017). Recently, a DDM was used to evaluate the performance of speech recognition systems by the density of AAVE features (Koenecke et al., 2020). The use of DDMs reflects the reality that speakers construct individual styles drawing on linguistic repertoires such as dialects to varying degrees (Benor, 2010). This necessitates a more nuanced description for speakers and texts than a discrete dialect category.

Following prior work (e.g., Van Hofwegen and Wolfram, 2010) we construct dialect density measures from feature detectors by counting the predicted number of features in each utterance, and dividing by the number of tokens. For the learning-based feature detectors (minimal pairs and corpus examples), we include partial counts from the detection probability; for the regular expression detectors, we simply count the number of matches and dividing by the number of tokens. In addition, we construct a DDM based on a document classifier: we train a classifier to distinguish Indian English from U.S. English, and then use its predictive probability as the DDM. These DDMs are then compared on two tasks: distinguishing Indian and U.S. English, and correlation with the density of expert-annotated features. The classifier is trained by fine-tuning BERT, using a prediction head on the [CLS] token.

5.1 Ranking documents by dialect density

One application of dialect feature recognizers is to rank documents based on their dialect density, e.g. in order to identify difficult cases for model evaluation. We correlate the dialect density against the density of expert-annotated features from Lange (2012), both measured at the transcript-level. We report the Spearman rank-correlation r.

As shown in Table 5, the document classifier performs poorly: learning to distinguish Indian and U.S. English offers no information on the density of Indian dialect features, suggesting that the model is attending to other information, such as topics or entities. The feature-based model trained on labeled examples performs best, which is unsurprising because it is trained on the same type of features that it is now asked to predict. Performance is weaker when the model is trained from minimal pairs. Minimal pair training is particularly helpful on rare features, but offers far fewer examples on the high-frequency features, which in turn dominate the DDM scores on test data. Regular expressions perform well on this task, because we have regular expressions for the high-frequency features, and because the precision issues are less problematic in aggregate when the DDM is not applied to non-dialectal transcripts.

5.2 Dialect Classification

Another application of dialect feature recognizers is to classify documents or passages by dialect (Dunn, 2018). This can help to test the performance of downstream models across dialects, assessing dialect transfer loss (e.g., Blodgett et al., 2016), as well as identifying cases of interest for manual dialectological analysis. We formulate a classification problem using the ICE-India and the Santa Barbara Corpus (ICE-USA), which contain transcripts of spoken dialogs in Indian English and U.S. English respectively. Each corpus is divided into equal-size training and test sets.

We construct such a dialect classifier by building on the components from Section 5.1. For the test set, we measure the D' ("D-prime") statistic,

$$D' = \frac{\mu_{\text{IN}} - \mu_{\text{US}}}{\sqrt{\frac{1}{2}(\sigma^2_{\text{IN}} + \sigma^2_{\text{US}})}}. \quad (1)$$

	ranking	classification	
Dialect density measure	r	D'	acc.
Document classifier	-0.17	14.52	1
Multihead, corpus examples	0.83	2.31	1
Multihead, minimal pairs	0.45	2.24	0.83
Regular expressions	0.71	1.61	0.80

Table 5: Performance of dialect density measures. The D' measures the ability to distinguish Indian and U.S. English; the Spearman r measures the correlation with the ground truth dialect density, as measured by the annotated Lange features. Accuracy is computed by choosing a threshold to balance the number of false positives and false negatives.

This statistic, which is closely related to a Z-score, quantifies the extent to which each metric distinguishes between the two populations. We also report classification accuracy; lacking a clear way to set a threshold, we choose a value that balances the number of false positives and false negatives.

As shown in Table 5, both the document classifier and the multihead feature detection model (trained on labeled examples) achieve perfect accuracy at discriminating U.S. and Indian English. The D' discriminability score is higher for the document classifier, which is trained on a cross-entropy objective that encourages making confident predictions. Both feature-based models achieve similar discriminability, but the model trained on minimal pairs is less accurate. Regular expressions perform relatively poorly: they suffer from low precision because they respond to surface cues that may be present in U.S. English, even when the dialect feature is not present (e.g., the word *only*, the phrase *is there*).

6 Background

Dialect classification. Prior work on dialect in natural language processing has focused on distinguishing between dialects (and closely-related languages). For example, the VarDial 2014 shared task required systems to distinguish between nation-level language varieties, such as British versus U.S. English, as well as closely-related language pairs such as Indonesian versus Malay (Zampieri et al., 2014), and later evaluation campaigns expanded this set to other varieties (Zampieri et al., 2017). In general, participants in these shared tasks have approached the problem as a form of text classification; neural architectures have appeared in the more recent editions of these shared tasks, but with a few exceptions (e.g., Bernier-Colborne et al., 2019), they have not outperformed classical techniques such as support vector machines. Our work differs by focusing on a specific set of known dialect features, rather than document-level classification between dialects.

Discovering and detecting dialect features. Another line of research uses machine learning feature selection techniques to discover dialect features from corpora. For example, Dunn (2018, 2019) induces a set of *constructions* (short sequences of words, parts-of-speech, or constituents) from a "neutral" corpus, and then identifies constructions with distinctive distributions over the geographical subcorpora of the International Corpus of English (ICE). Social media is also used for this purpose: for example, features of African American Vernacular English (AAVE) can be identified by correlating linguistic frequencies with the aggregate demographic statistics of the geographical areas from which geotagged social media was posted (Eisenstein et al., 2011; Stewart, 2014; Blodgett et al., 2016).

Closer to our contribution is work that attempts to detect predefined dialect features features. Jørgensen et al. (2015) and Jones (2015) designed lexical patterns to identify non-standard spellings that match known phonological variables from AAVE (e.g., *sholl* 'sure'), demonstrating the presence of these variables in social media posts from regions with high proportions of African Americans. Blodgett et al. (2016) use the same geography-based approach to test for phonological spellings and constructions corresponding to syntactic variables such as habitual *be*; Hovy et al. (2015) show that a syntactic feature of Jutland Danish can be linked to the geographical origin of product reviews. In general, these approaches have focused on features that could be recognized directly from surface forms, or in some cases, from a combination of surface forms and part-of-speech. In contrast, we focus on learning to recognize features from examples, enabling the recognition of features for which it is difficult or impossible to craft patterns.

Minimal pairs in NLP. A distinguishing aspect of our approach is the use of minimal pairs rather than conventional labeled data. Minimal pairs are well known in natural language processing from the Winograd Schema, in which the referent for a pronoun is manipulated by changing a single word in the text (Levesque et al., 2012). The Winograd Schema pairs are used for evaluation, but Kocijan et al. (2019) show that fine-tuning on a related dataset of minimal pairs can improve performance on the Winograd Schema itself. A similar idea arises in counterfactually-augmented data (Kaushik et al., 2019) and contrast sets (Gardner et al., 2020), in which annotators are asked to identify the minimal change to an example that is sufficient to alter its label. However, those approaches use counterfactual examples to *augment* an existing training set, while we propose minimal pairs as a replacement for large-scale labeled data. Minimal pairs have also been used to design controlled experiments and probe neural models' ability to capture various linguistic phenomena (Gulordava et al., 2018; Ettinger et al., 2018; Futrell et al., 2019; Gardner et al., 2020; Schuster et al., 2020). Finally, Liang et al. (2020) use contrastive explanations as part of an active learning framework to improve data efficiency. Our work shares the objective of Liang et al. (2020) to improve data efficiency, but is methodologically more similar to probing work that uses minimal pairs to represent specific linguistic features.

7 Conclusion

We demonstrate that it is possible to construct dialect feature recognizers using only a small number of minimal pairs: in most cases, just five positive and negative examples per feature. This opens the door to better understanding the many dialects for which labeled data does not exist. Future work will extend this approach to multiple dialects, focusing on cases in which features are shared across two or more dialects. This lays the groundwork for the creation of dialect-based "checklists" (Ribeiro et al., 2020) to assess the performance of NLP systems across the diverse range of linguistic phenomena that may occur in any given language.

Acknowledgments. Thanks to Claudia Lange for sharing her annotations, and for discussion of this research. Thanks to Axel Bohmann for sharing information about his work on recognizing dialect features with regular expressions. Valuable feedback on this research was provided by Jason Riesa, Kristina Toutanova, Dan Jurafsky and especially Vera Axelrod for valuable feedback. Devyani Sharma is supported in part by a Google Faculty Research Award.

References

Sarah Bunin Benor. 2010. Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics*, 14(2):159–183.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving cuneiform language identification with BERT. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 17–25, Ann Arbor, Michigan. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. 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.

Axel Bohmann. 2019. *Variation in English worldwide: Registers and global varieties*. Cambridge University Press, Cambridge.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Holly K Craig and Julie A Washington. 2002. Oral Language Expectations for African American Preschoolers and Kindergartners. *American Journal of Speech-Language Pathology*, 11(1):59–70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara Corpus of Spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Jonathan Dunn. 2018. Finding variants for construction-based dialectometry: A corpus-based approach to regional cxgs. *Cognitive Linguistics*, 29(2):275–311.

Jonathan Dunn. 2019. Modeling global syntactic variation in English using dialect classification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, Oregon, USA. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets.

Sidney Greenbaum and Gerald Nelson. 1996. The international corpus of English (ICE) project. *World Englishes*, 15(1):3–15.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 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.

Taylor Jones. 2015. Toward a description of african american vernacular english dialect regions using "black twitter". *American Speech*, 90(4):403–440.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania Press.

Claudia Lange. 2012. *The syntax of spoken Indian English*. John Benjamins Publishing Company, Amsterdam.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Weixin Liang, James Zou, and Zhou Yu. 2020. Alice: Active learning with contrastive natural language explanations. *arXiv preprint arXiv:2009.10259*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap-a web application for dialectology. *Dialectologia: revista electrònica*, pages 65–89.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403.

Devyani Sharma. 2009. Typological diversity in New Englishes. *English World-Wide*, 30(2):170–195.

Ian Stewart. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the*

Association for Computational Linguistics, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.

Janneke Van Hofwegen and Walt Wolfram. 2010. Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, 14(4):427–455.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *National Academy of Sciences*, 114(25):6521–6526.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

A Regular Expressions

Table 6 shows the regular expressions that we used for the five features, where such patterns were available.

Feature	**Regular expression**
focus *itself*	`\bitself\b`
focus *only*	`\bonly\b`
non-initial existential	`\bis there\b|\bare there\b`
invariant tag *(isn’t it, no, na)*	`\bisn’t it\b|\bis it\b|\bno\b|\bna\b`
general extender *and all*	`\band all\b`

Table 6: Regular expressions we used, for the features that such patterns were available.