DialectGram: Automatic Detection of Dialectal Changes with Multi-geographic Resolution Analysis

Hang Jiang*, **Haoshen Hong***, **Yuxing Chen***, **Vivek Kulkarni**
Stanford University
{hjian42, haoshen, yxchen28, viveksck}@stanford.edu

Abstract

Several computational models have been developed to detect and analyze dialect variation in recent years. Most of these models typically assume a predefined set of geographical regions. However, dialect variation occurs at multiple levels of geographic resolution ranging from cities within a state, states within a country, and between countries across continents. In this work, we propose a novel approach, DialectGram, to obtain dialect-sensitive sense embeddings without apriori assumptions on the set of regions and apply them to detect dialect variations across multiple geographic resolutions. First, DialectGram automatically infers dialect-sensitive senses with non-parametric Bayesian extension of Skip-gram. For multi-geographic resolution analysis, our model is more superior than the baseline models because DialectGram learn sense embeddings from region-agnostic data with only one-time training, which is used to compose region-specific word embeddings depending on the geographic resolution. On the contrary, the baseline models rely on region tags for training region-specific embeddings, and need to be trained multiple times for different geographic resolutions. Second, in contrast to previous models, DialectGram can predict the proportion of senses in any given region. At last, we address the issue of lacking a standard validation approach for this task. To solve this, we create a new validation set `DialectSim` for evaluating region-specific word embeddings between the United States and the United Kingdom, and establish a quantitative benchmark with our approach. In order to train region-specific word embeddings, we construct a new corpus `Geo-Tweets2019` by collecting English tweets from the US and the UK. Through both qualitative and quantitative analysis, we show that DialectGram can encode rich linguistic variations.

*Equal contribution.

1 Introduction

Studying regional variations of languages is central to the field of sociolinguistics, language change, and dialectology. Traditional approaches (Labov, 1980; Milroy, 1992; Tagliamonte, 2006; Wolfram and Schilling, 2015) focus on rigorous manual analysis of linguistic data collected through time-consuming and expensive surveys and questionnaires. Recently, researchers (Bamman et al., 2014b; Lucy and Mendelsohn, 2018) have been using big corpora from the Internet (Twitter, Reddit, etc.) to study linguistic phenomena. However, it is difficult to scale classical methods to these social media corpora. Therefore, researchers have applied computational approaches with word embeddings to analyze huge corpora and automatically detect linguistic changes.

Using word embeddings to study languages is motivated by the fact that the simple word frequency and syntactic models only utilize limited information from the text. Although the original Skip-gram model (Mikolov et al., 2013a) is efficient to learn high-dimensional word vectors that capture rich semantic relationship between words, it only holds one representation for each word and different usages are mixed. To adopt word embeddings for linguistic variation detection, Kulkarni et al. (2015b) presented `GEODIST` model to learn r region-specific word embeddings for each word, where r is the number of pre-defined regions, based on the work of Bamman et al. (2014a).

There are three main problems to use the previous computational models for capturing the dialectal variations. First, previous models (Bamman et al., 2014a; Kulkarni et al., 2015b) such as Frequency Model, Syntactic Model, and `GEODIST` all rely on pre-defined regional classes to train classifiers and to detect linguistic changes. The use of pre-defined regional classes limits the

flexibility of these baseline models because dialect changes can be observed at various geographic resolution level. Second, previous models only tell limited information such as whether a significant change occur among regions but provide no details about how dialects differ from each other. Third, there is no standard validation method for region-specific word embeddings. In fact, dialectal changes have different categories (i.e. purely lexical variations, syntactic variations, and semantic variations), so it is hard to quantify dialectal changes.

This project aims to detect linguistic variations in English by presenting a novel and efficient computational model DialectGram, and to create a standard validation set `DialectSim` for evaluating the region-specific word embeddings. The contributions of our paper are as follows:

- **Multi-resolution Model:** We introduce DialectGram, a method to study the geographic variation in language across multiple levels of resolution without assuming knowledge of the geographical resolution apriori.
- **Sense Distribution:** DialectGram predicts how likely each sense of a word is used in a context. This feature enables us to know the sense proportion at different regions. On the contrary, the previous models only tell whether a shift of meaning occurs or not.
- **Corpus and Validation Set:** We build a new English Twitter corpus `Geo-Tweets2019` for training dialect-sensitive word embeddings. Furthermore, we construct a new validation set `DialectSim` for evaluating the quality of English region-specific word embeddings between UK and USA.

2 Related Work

Linguistic variations. In the past, sociologists and linguists have been studying linguistic changes by designing experiments to manually collect data (Labov, 1980; Milroy, 1992) and conducting variation analysis (Tagliamonte, 2006). Recently, studies (Gonçalves and Sánchez, 2014; Bamman et al., 2014a,b; Kulkarni et al., 2015a) in linguistic variation on social media have shown that social media is a reliable source for researchers to build corpora. Many researchers (Eisenstein et al., 2010; Gulordava and Baroni, 2011; Kim et al., 2014; Kulkarni et al., 2015b; Kenter et al., 2015; Lucy and Mendelsohn, 2018) have used different computational models to study dialect variations with respective to geography, gender, and time.

Eisenstein et al. (2010) is one of the first to tackle the linguistic variation problem with computational models. They design a multi-level generative model that uses latent topic and geographic variables to analyze the lexical dialectal variations in English. This latent variable model is able to generate an author's geographic location based on the author's text. To quantitatively evaluate the models, they compute the physical distance between the prediction and the true location. Different from the previous work (Eisenstein et al., 2010), Gonçalves and Sánchez (2016) apply K-means method to cluster the geographic lexical superdialtects. Donoso and Sanchez (2017) move a step forward from Gonçalves and Sánchez (2016) by proposing two metrics to calculate the linguistic distance between geographic regions. That is, instead of using the physical distance between the predicted and the true location, they compute cosine similarities or Jensen-Shannon Divergence (JSD) to evaluate the model quantitatively. However, none of these works use word embeddings models, which motivates us to move forward to the following works.

Geo-Specific Word Embeddings. As many studies show, word embeddings share some non-linear patterns, which linear mapping usually fail to capture. Bamman et al. (2014a) introduce a new approach to learn the vector representations of words that are sensitive to the speakers' geographical locations – for every word, the model learns a common, main representation and distributed representations associated with 51 US states. Based on Bamman et al. (2014a); Kulkarni et al. (2015a), Kulkarni et al. (2015b) design a new model `GEODIST` that learns global and differential embeddings jointly from region-tagged texts, which together compose a deeper joint embedding for each region and produce better results in similarity measure. Nevertheless, a pre-defined set of regions is required for the model to update region-specific embeddings. This approach has three disadvantages. First, it does not help researchers easily detect how language varies on a continuous map. For instance, Kulkarni et al. (2015b) assume that English are different in US an UK, and train the network to learn two set of word embeddings

for the two regions. A model trained on US and UK data does not reveal information about how language varies within US states. In fact, dialects should not be segregated by "pre-defined" regions and dialects exist at different geographical levels. Second, this approach requires multiple trainings to conduct multi-resolution analysis. To learn how English changes within the states, Kulkarni et al. (2015b) has to tag each US tweet with a state name and train the model again. Third, the model cannot predict how a word is used in a given context. It simply encodes different senses of a word into one mixed embedding in each region and does not reveal what senses a word contains not how which sense a word is used in a context.

Word Sense Disambiguation. To overcome the drawbacks of the GEODIST approach, we reconstruct the automatic dialect detection problem as a variation of word sense induction and disambiguation task by encoding dialect usage into senses. Word ambiguity such as polysemy and homonymy is an important feature of natural languages. For instance, the word "pants" usually refer to "underwear" in the US versus "trousers" in the UK. Reisinger and Mooney (2010) is the first paper that modifies the single "prototype" vector space model to obtain multi-sense word embeddings with average cluster vectors as prototypes. Many works (Huang et al., 2012; Neelakantan et al., 2014; Tian et al., 2014; Chen et al., 2014) are later dedicated to combine Skip-gram, clustering algorithm, and linguistic knowledge to learn word senses and embeddings jointly. Bartunov et al. (2016) follow a more principled nonparametric Bayesian approach and propose the Adaptive Skip-gram (AdaGram) model, which is able to induce word senses without assuming any fixed number of prototypes. In the task of automatic dialect detection, we also do not want to assume the number of senses, and AdaGram seems to be a ideal candidate to induce dialect senses and detect dialect changes.

3 Data

3.1 Geo-Tweets2019 Corpus

First of all, we created the new corpus called `Geo-Tweets2019`, which consists of tweets during April and May in 2019 from the United States and the United Kingdom. For this project, since we are interested in the dialectal change between the English speakers in the US and the UK, we only consider English tweets. Each tweet includes the user ID, the published time, the geographic location, and tweet text. We have around 2M tweets from the US and 1M from the UK.

We first crawled English tweets from the US and the UK with geographic bounding boxes through Tweepy[1]. After that, we preprocessed the tweets with the tweet tokenizer from Eisenstein et al., 2010 and regular expressions. At this step, we filtered out URL's, emojis, and other irregular uses of English to shrink the size of vocabulary and to facilitate the training of word vectors. Statistics can be seen in Table 2.

Number	US	UK	Total
tweet	2,075,394	1,088,232	3,163,626
token	41,637,107	22,012,953	63,650,060
term	865,784	469,570	1,167,790

Table 2: Statistics of Geo-Tweets2019

3.2 DialectSim Validation Set

To evaluate the models, we come up with the new validation set `DialectSim`, which comprises of words with same or shifted meanings in the US and the UK. To build this validation set, we first crawled a list of words that show different meanings from the Wikipedia page[2] and pick 341 words that appear more than 20 times in our corpus in the UK and the US. Table 1 presents three examples in the dataset. In order to generate balanced positive and negative samples, we sample another 341 negative examples randomly from our `Geo-Tweets2019` dataset. A minimum frequency of 20 is also used for negative sampling. We manually evaluated the negative examples and agreed that 90% of the negative examples stay the same meaning across UK and US. At last, we split them into training set with 511 samples (75%) and testing set with 171 samples (25%).

`DialectSim` can also be applied to different corpora by picking overlapping positive examples and sampling negative examples from their corpora as described above. By doing so, they can easily build benchmarks with their models. Overall, this validation set makes it easy for later researchers to compare their works.

[1] http://docs.tweepy.org/en/v3.5.0/api.html

[2] https://en.wikipedia.org/wiki/Lists_of_words_having_different_meanings_in_American_and_British_English

Word	US Meaning	UK Meaning
flat	smooth and even; without marked lumps or indentations	apartment
flyover	flypast, ceremonial aircraft flight	elevated road section
pants	trousers	underwear
lift	elevator	raise
football	soccer	American football

Table 1: Examples of words that have different meanings in American and British English

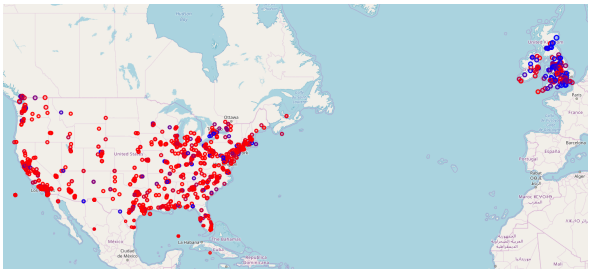

Figure 1: Word heat map of *gas*. Tweets that contain *gas* with predicted sense "gaseous substance" are illustrated as blue circles; tweets that contain *gas* with predicted sense "gasonline" are plotted as red circles.

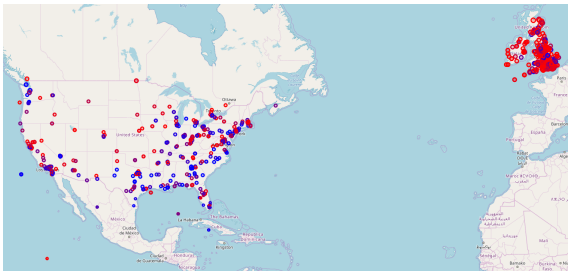

Figure 2: Word heat map of *flat*. Tweets that contain *flat* with predicted sense "apartment" are illustrated as red circles; tweets that contain *flat* with predicted sense "smooth and even" are plotted as blue circles.

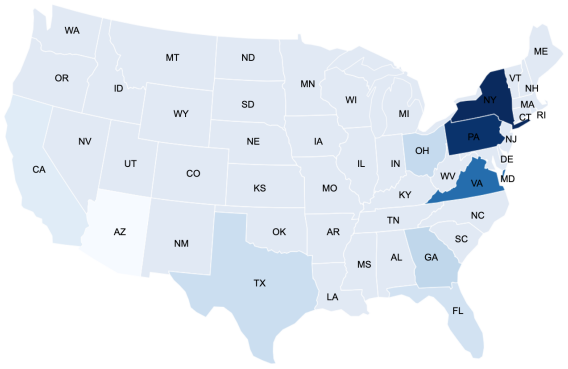

Figure 3: Word heat map of *buffalo* in state-level. Blue means *Buffalo city* and white means *buffalo sauce*. Grey means that this state contains no tweet using the word *buffalo* in our corpus.

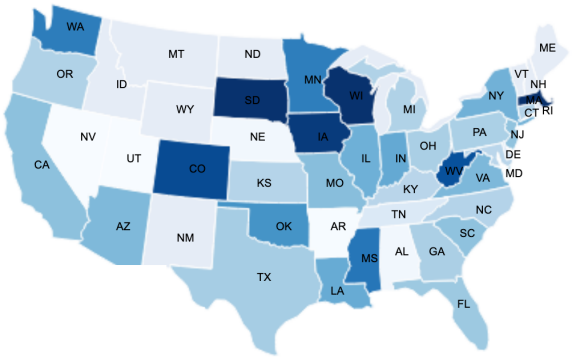

Figure 4: Word heat map of *cracking* in state-level. Blue means *exciting* and white means *under such a lot of emotional strain that they become mentally ill*. Grey means that this state contains no tweet using the word *cracking* in our corpus.

4 Models

Our hypothesis is that the word-embedding methods are more powerful in capturing the dialectal changes on social media than the traditional word-frequency and the Syntactic Models. In comparison, we run experiments on two baseline models (i.e. Frequency Model and Syntactic Model) and two word-embedding models (i.e. GEODIST model and DialectGram model).[3]

4.1 Baseline Models

Frequency Model. One baseline method to detect whether there are significant changes between usage in two regions is to count the occurrence of a word in the US and the UK tweets. We have implemented this Frequency Model as described in Kulkarni et al. (2015b).

Syntactic Model. Extended from the previous Frequency Model, the Syntactic Model takes Part-of-Speech (POS) tag into consideration as well. More specifically, if a word is used equally frequently in both countries, but the their POS usages are different, then we consider the meaning of two words as different between two countries. We use NLTK[4] as well as CMU ARK Twitter Part-of-Speech Tagger[5] for POS tagging. The latter one

[3] Our datasets, models and code are available at: https://github.com/yuxingch/DialectGram

[4] https://www.nltk.org/

[5] http://www.cs.cmu.edu/~ark/TweetNLP/

provides more reasonable POS tags and we choose it for our Syntactic Model.

GEODIST (Skip-gram) Model. The main idea of GEODIST model (Bamman et al., 2014a; Kulkarni et al., 2015b) is to learn region-specific word embeddings using the Skip-gram framework proposed by Mikolov et al. (2013a). Instead of learning a single vector to represent a word, this model aims to jointly learn a global embedding $\delta_{\text{MAIN}}(w)$ as well as (multiple) differential embeddings $\delta_{r_i}(w)$ for each word w in the vocabulary with $R = (r_1, r_2, \dots)$ geographical regions. The region-specific embedding is therefore defined as the sum of the global embedding and the differential embedding for that region: $\phi_{r_i}(w) = \delta_{\text{MAIN}(w)} + \delta_{r_i}(w)$. The objective function is to minimize the negative log-likelihood of the context word given the center word conditioned on the region. Negative-sampling mentioned in the work of Mikolov et al. (2013b) has been applied during the training process as an alternative to the hierarchical softmax. We use stochastic gradient descent method (Bottou, 1991) to update the model parameters. We implement our own GEODIST model in PyTorch.

4.2 DialectGram Model

We construct a new model for detecting dialectal changes which we called DialectGram for Dialectal Adaptive Skip-gram. The model first adopts Adaptive Skip-gram (AdaGram) to learn multi-sense word embeddings with (Bartunov et al., 2016) through training on the region-agnostic corpus. According to the regions we want to compare, the model compose region-specific word embeddings by taking weighted average of sense embeddings. At last, the model calculates the distance between region-specific word embeddings of the same word to determine whether a significant change occurs.

At the first step of our DialectGram, we apply Adaptive Skipgram (AdaGram) to learn multi-sense word embeddings. Specifically, we train AdaGram on our Twitter corpus that contains both tweets from the United Kingdom and the United States to induce and update multiple senses of each word. It introduces a latent variable z that encodes the index of active meaning: $p(v|z = k, w, \theta) = \prod_{n \in \text{path}(v)} \sigma(\text{ch}(n)\text{in}_{wk}^T \text{out}_n)$. Instead of assuming any fixed number of prototypes or relying on high-quality linguistic resources like WordNet, it uses Dirichlet process (DP) to automatically determine the required number of prototypes for the word. Following Beta distribution, the prior probability of word w's k-th meaning is expressed as:

$$p(z = k|w, \beta) = \beta_{wk} \prod_{r=1}^{k-1} (1 - \beta_{wr}) \quad (1)$$

$$p(\beta_{wk}|\alpha) = \text{Beta}(\beta_{wk}|1, \alpha), k = 1, 2, \dots \quad (2)$$

Usually, a larger α leads to more meanings in general. Therefore, the DialectGram is formatted as the following where $Z = \{z_i\}_{i=1}^N$ is the set of senses for words:

$$p(Y, Z, \boldsymbol{\beta}|X, \alpha, \theta) = \prod_{w=1}^{V} \prod_{k=1}^{\infty} p(\beta_{wk}|\alpha) \prod_{i=1}^{N} \Big[p(z_i|x_i, \boldsymbol{\beta}) \prod_{j=1}^{C} p(y_{ij}|z_i, x_i, \theta) \Big] \quad (3)$$

At inference time, the posterior predictive over context words y given input word x is:

$$p(y|x, D, \theta, \alpha) = \int \sum_{z=1}^{T} p(y|x, z, \theta) p(z|\theta, x) q(\theta) d\theta \quad (4)$$

We expect that DialectGram can encode some information of dialectal variations into multi-sense word embeddings.

Algorithm 1 Use DialectGram to Compose Region-specific Embeddings

Input: w word
Output: e_r weighted region embedding for w

```
Load the trained DialectGram model
Build Index_r on Corpus from region r
for s, p ∈ GetSensePriors(w) do
    S_c[s] ← 0, S_p[s] ← p ▷ Note: S_c is sense counts, S_p is sense priors
end for
for all c ∈ GetContexts(w) do
    s ← Disambiguate(w, c)
    S_c[s] ← S_c[s] + 1
end for
e_r ← GetWeightedVector(S_c, S_p)
```

Compared to the GEODIST model which needs the geo-label to update the region-specific embeddings, DialectGram learns multi-sense word embeddings on our dataset without knowing the region label of each tweet. For instance, we expect

DialectGram to automatically induce and learn the two dialectal meanings of the word *flat* such as *apartment* in the UK and *even* in the US from our dataset which simply consists of tweets from the countries but not the country labels. As the algorithm shown above, we first train the model on our Geo-Tweets2019 corpus to learn word sense embeddings with Julia implementation[6] and then convert the trained model with python implementation[7] for building the pipeline. To obtain a word's region-specific embedding in a place, we first use DialectGram to predict the dominant sense for the word in each tweet from a region and use weighted average of the sense embeddings as the region-specific word embedding e_r.

After fine-tuning DialectGram with different sets of hyperparameters, we found the following hyperparameter setting works the best to obtain good word sense embeddings: `min_freq` $=$ 20, `window_size` $=$ 10, `dimension` $=$ 100, `maximum_prototype` $=$ 30, $\alpha = 0.1$, `epoch` $= 1$, `sense_threshold` $= 1e-17$. It is worth noting that a large α (the underlying Dirichlet process) may lead to too many senses for some words and a small α, on the contrary, results in too few senses.

To measure the significance of the dialectal change, Kulkarni et al. (2015b) propose an unsupervised method to detect words with statistically significant meaning changes, which is sensitive to the choice of words. However, with the new `DialectSim` dataset, we can evaluate the models on the list of annotated words in a simple and standardized way. We evaluate both Skip-gram models (i.e. GEODIST and DialectGram) by calculating the Manhattan distance[8] between a word's region-specific embeddings.

5 Results

5.1 Qualitative Result

We investigate the words that `GEODIST` model predicts to have a significant dialectal change between the two regions. For example, the word *mate* is one of the top 20 words in our vocabulary if we sort the vocabulary by the Manhattan distance between the US and the UK embeddings from high to low. However, words like *draft* are predicted to have different regional meanings but not labelled as "significant" in `DialectSim`. We further discuss this issue in section 6.4.

We select some words with significantly different meanings between the UK and the US. In our DialectGram model, we select the most frequent 2 senses, which usually account for more than 99% usage variation of a word, and plot a heat map on world map[9].

The word maps in Figure [1, 2] suggest that the usage of *gas* and *flat* are different in the UK and in the US. *Gas* is used commonly as petrol and related to gas station in the US, but in the UK, *gas* usually refers to air and natural gas. *Flat* could refer to *apartment* but in the US this meaning is not as common as in the UK.

5.2 Quantitative Results

For each model, we have defined a `score` function that takes in one word and return a real number denoting its difference in meanings between the UK and the US. We search for a threshold that maximizes the accuracy on training set. Then we test the model performance on testing set. The results are recorded in Table 4. We further discuss the quantitative results in section 6.

6 Analysis

6.1 Dataset

Our training corpus `Geo-Tweets2019` has over three million tweets from US and UK. However, we still find that data points for some words are sparsely located in the world map, which make it difficult for us to conduct linguistic variation analysis on a micro level. Therefore, we only present the country-level and state-level analysis in the paper. If we have a larger training corpus, we will have more data points in city or town level, which allows us to even exam the linguistic changes in more geographic resolution levels.

The novel validation set `DialectSim` has 341 high-quality positive examples (words that have shifted meanings in UK and US). To validate a model, we need to sample negative samples from the training corpus. This can introduce some noise to the validation set. However, we manually verified the negative examples and agreed that over

[6] https://github.com/sbos/AdaGram.jl

[7] https://github.com/lopuhin/python-adagram

[8] We tried euclidean and cosine distance as well, but Manhattan distance produces the best result out of the three.

[9] Please use color print since our maps contains multiple colors.

word	sense 1 neighbors	sense 2 neighbors
gas	industrial, masks, electric	car, station, bus
flat	kitchen, shower, window	shoes, problems, temperatures
buffalo	syracuse, hutchinson	chicken, fries, seafood
subway	starbucks, restaurant, mcdonalds	1mph, commercial, 5kmh

Table 3: Neighbors of sense embeddings for selected words. This shows DialectGram is able to learn semantic variations of words.

Model	Acc	Prec	Recall	F1
Frequency	0.5600	0.5600	0.5887	0.5568
Syntactic	0.5263	0.5714	0.4828	0.5233
GEODIST	0.6432	**0.7424**	0.5810	0.6518
DialectGram	**0.6667**	0.6837	**0.6438**	**0.6632**

Table 4: Test performance. Acc, Prec means accuracy and precision. DialectGram has better accuracy, recall, and F1 score than GEODIST.

90% of the negative examples indeed hold the same meanings in UK and USA.

6.2 Frequency Model

We observed that Frequency Model is more sensitive to word difference between two countries: *football* in the UK is same as *soccer* in the US, causing an imbalanced frequency of term *football* between both countries. However, it can not detect some semantic changes of words if the semantic change preserves frequency for both countries: *flat* has similar frequency in both countries, despite the fact that *flat* could mean *apartment* in the UK, whereas this usage is uncommon in the US. This model does not suffer from overfitting problem, because the model is fairly simple and the parameter space is quite small. The performance of this Frequency Model is better than Syntactic Model, from which we can infer that the frequency of a word might be more correlated with the word sense change than Syntactic Model.

6.3 Syntactic Model

Syntactic Model performs the worst among all the models. It still gets slightly higher precision than the Frequency Model on test set because it gets some dialectal syntactic changes correct. There are two reasons for its bad performance. First, it is challenging for the POS tagger to tag every word correctly because tweets are super casual and informal. Compared with the NLTK POS tagger, the CMU ARK Twitter POS Tagger improves the model's performance significantly, which proves the importance of having high-quality POS tags. Since the CMU tagger is not perfect in tagging the tweets, its low performance impairs the performance of the Syntactic Model. Besides, many word sense changes do not alter POS tags in our corpus. For example, *pants* refers to *underwear* in the UK while it refers to *jeans* in the US, and both of them are noun. Syntactic Model is not good at identifying these changes.

6.4 GEODIST Model

As mentioned in Section 5.1, GEODIST model is able to detect dialect changes. The accuracy on the test set beats the previous two baseline models (0.6432 versus 0.5600 and 0.5263), as shown in Table 4. It also outperforms the baseline models in terms of precision and F1 score. In fact, GEODIST model has the highest precision among all models, including the DialectGram model that will be discussed in the next section. We also notice that the recall on the test set is the lowest. The high precision with low recall indicates that for those changes that GEODIST model thinks are significant are indeed significant, but the model is very conservative and picky. It misses more words that actually have significant dialectal changes. For example, the difference between the two region-specific embeddings of the word *pants* is predicted to be not significant, while *pants* does have different meanings in the UK and the US (Table 1).

6.5 DialectGram Model

DialectGram outperforms the GEODIST model in accuracy, recall, and F1 score. However, its precision is lower than that of the GEODIST and Frequency Model. This is already impressive given the fact that DialectGram does not use the country label during the word embedding training.

The sensitivity of α is the main reason for its lower performance in precision compared to GEODIST model. If α is not set appropriately for training, some senses have overlapped meanings. For the word *gas* in Table 3, we sometimes have a sense 3 which is accompanied by words such as

air, house, pipe. This sense seems to be a mix of sense 1, gaseous substance, and sense 2, gasoline. We picked α based on the model's performance on the training set, but this α can be further fine-tuned if more computing resources are available.

With the trained DialectGram model on US and UK aggregate data, we are able to easily configure it to conduct inter-state comparisons. For the word *buffalo*, we pick the most two dominant senses where the *Buffalo City* sense is blue and the *buffalo sauce* sense is white. We can observe that the heat map is more blue around the New York state area. In a different heat map for word *cracking*, we observe that the Midwest area is more blue, indicating people are more likely to use the word for *excellent*, while people in other areas like to say *cracking up* to mean *under such a lot of emotional strain that they become mentally ill*. We normalized the data points by filtering out states whose tweet number is less than 15 since a small number of data points can suffer from high variance. In the future, we can crawl more data from Twitter such that we have enough data points for each area.

Word	Neighbors of the Dominant Sense
color	girls, wearing, yellow, shirts, colour
colour	skin, wear, hair, yellow, color, wearing
advice	request, ask, advise, call
advise	advice, share, request, provide
theater	theatre, cinema, movie, #avengersengame
theatre	theater, cinema, movie, marvel

Table 5: Examples of purely lexical variations

Additionally, we found DialectGram is able to encode lexical, syntactic and semantic variations of words. Table 5 presents the purely lexical variations for pairs *color-colour*, *advice-advise* and *theater-theatre* captured by DialectGram. As listed in the second column of Table 5, *color* and *colour* share some neighboring words like *wear(ing)* and *yellow*. More interestingly, *colour* is one of the nearest neighbors of *color* and vice versa. From our world knowledge, *color* and *colour* are the two purely lexical variations of the same concept and DialectGram results are congruous with this prior knowledge. This happens for all other pairs as well, which demonstrates the model's ability to detect the lexical variations across regions.

Syntactic variations over regions are demonstrated in word heat maps in section 5.1, and semantic variations over regions are quantitatively measured as discussed in section 5.2 and 6.

7 Conclusion

In this work, we proposed a novel method to detect linguistic variations on multiple resolution levels. In our new approach, we use DialectGram to train multiple sense embeddings on region-agnostic data, compose region-specific word embeddings, and determines whether there is a significant dialectal variation across regions for a word. This is superior to the baseline models since DialectGram does not rely on the region-labels for training multi-sense word embeddings. The use of region-agnostic data allows DialectGram to conduct multi-resolution analysis with one-time training, whereas the `GEODIST` model requires multiple trainings for different resolution levels. It's also worth noting that the dataset for training is not restricted to social media corpus – we may use any geo-tagged data to train DialectGram. Besides, since DialectGram does not require pre-defined labels to learn sense embeddings, this method can be directly applied to temporal or gender analysis of language at multi-resolution levels.

Besides, we constructed `Geo-Tweets2019`, a new corpus from online Twitter users in the UK and US for training word embeddings. To validate the work, we also contributed a new validation set `DialectSim` for explicitly measuring the performance of our models in detecting the linguistic variations between the US and the UK. This validation set allows for more precise comparison between our method (DialectGram) and previous methods including Frequency Model, Syntactic Model, and `GEODIST` model. On `DialectSim`, our method achieves better performance than the previous models in accuracy, recall, and F1 score. Through linguistic analysis, we also found that DialectGram model learns rich linguistic changes between British and American English in word semantic, lexical, and syntactic changes. In the future, we want to include tweets from more English-speaking countries like India and Australia.

Acknowledgments

We would like to express our special thanks for Cindy Wang and Christopher Potts, who gave precious advice and comments to our project.

References

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

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, pages 130–138.

Léon Bottou. 1991. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *EMNLP*.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in twitter. *arXiv preprint arXiv:1702.06777*.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074.

Bruno Gonçalves and David Sánchez. 2016. Learning about spanish dialects through twitter. *Revista Internacional de Lingüística Iberoamericana*, pages 65–75.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1191–1200. ACM.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015a. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015b. Freshman or fresher? quantifying the geographic variation of internet language. *arXiv preprint arXiv:1510.06786*.

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

Li Lucy and Julia Mendelsohn. 2018. Using sentiment induction to understand variation in gendered online communities. *arXiv preprint arXiv:1811.07061*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *HLT-NAACL*.

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

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *COLING*.

Walt Wolfram and Natalie Schilling. 2015. *American English: dialects and variation*, volume 25. John Wiley & Sons.