Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Christoph Purschke
University of Luxembourg
christoph.purschke@uni.lu

Abstract

Dialects are one of the main drivers of language variation, a major challenge for natural language processing tools. In most languages, dialects exist along a continuum, and are commonly discretized by combining the extent of several preselected linguistic variables. However, the selection of these variables is theory-driven and itself insensitive to change. We use Doc2Vec on a corpus of 16.8M anonymous online posts in the German-speaking area to learn continuous document representations of cities. These representations capture continuous regional linguistic distinctions, and can serve as input to downstream NLP tasks sensitive to regional variation. By incorporating geographic information via retrofitting and agglomerative clustering with structure, we recover dialect areas at various levels of granularity. Evaluating these clusters against an existing dialect map, we achieve a match of up to 0.77 V-score (harmonic mean of cluster completeness and homogeneity). Our results show that representation learning with retrofitting offers a robust general method to automatically expose dialectal differences and regional variation at a finer granularity than was previously possible.

1 Introduction

People actively use dialects to mark their regional origin (Shoemark et al., 2017a,b), making them one of the main drivers of language variation. Accounting for this variation is a challenge for NLP systems (see for example the failed attempts of people with accents trying to use dialogue systems. Accounting for variation can significantly improve performance in machine translation (Mirkin and Meunier, 2015; Östling and Tiedemann, 2017), geolocation (Rahimi et al., 2017a,b) and help personalize applications and search.

However, regional variation involves a complex set of grammatical, lexical, and phonological features, all of them continuously changing. Consequently, dialects are not static discrete entities, but exist along a continuum in most languages. Variational linguistics and dialectology typically discretize this continuum by using a set of preselected features (Trudgill, 2000), often including outdated vocabulary. The resulting dialect areas are highly accurate, but extremely time-consuming to construct and inflexible (the largest and to date most comprehensive evaluation of German dialects, the Wenker-Atlas (Rabanus et al., 2010) is almost 150 years old and took decades to complete). Work in dialectometry has shown that computational methods, such as clustering (Nerbonne and Heeringa, 1997; Prokić and Nerbonne, 2008; Szmrecsanyi, 2008, inter alia) and dimensionality reduction (Nerbonne et al., 1999; Shackleton Jr, 2005) can instead be used to identify dimensions of variation in manually constructed discrete feature vectors. However, the success of such approaches depends on precise prior knowledge of variation features (Lameli, 2013).

Distributed representations, as unsupervised methods, can complement these methods by capturing similarities between words and documents (here: cities) along various latent dimensions, including syntactic, semantic, and pragmatic aspects. These representations are therefore more compact, less susceptible to data sparsity than latent variable models, and allow us to represent a large number of possible clusters than feature-based representations (cf. Luong et al. (2013)). These properties also allow us to measure similarities on a continuous scale, which makes represen-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4383–4394
Brussels, Belgium, October 31 - November 4, 2018.

tation learning especially useful for the study of regional language variation along several linguistic dimensions.

We use a corpus of 16.8 million anonymous German online posts, cast cities as document labels, and induce document embeddings for these cities via Doc2Vec (Le and Mikolov, 2014). We first show that the resulting city embeddings capture regional linguistic variation at a more fine-grained, continuous regional distinction than previous approaches (Bamman et al., 2014; Östling and Tiedemann, 2017), which operated at a state or language level.[1] We also show that the embeddings can serve as input to a geolocation task, outperforming a bag-of-words model, and producing competitive results.

However, such representations are susceptible to linguistic data bias, ignore geographic factors, and are hard to evaluate with respect to their fit with existing linguistic distinctions. We address these problems by including geographic information via retrofitting (Faruqui et al., 2015; Hovy and Fornaciari, 2018): we use administrative region boundaries to modify the city embeddings, and evaluate the resulting vectors in a clustering approach to discover larger dialect regions.

In contrast to most dialectometric approaches (Nerbonne et al., 1999; Prokić and Nerbonne, 2008), and in line with common NLP practice (Doyle, 2014; Grieve, 2016; Huang et al., 2016; Rahimi et al., 2017a), we also evaluate the clustered dialect areas quantitatively. Rather than testing the geographic extent of individual words against known dialect areas (Doyle, 2014), we compare the match of entire geographic regions to a recent German dialect map (Lameli, 2013). We use cluster evaluation metrics to measure how well our clusters match the known dialect regions.

The results show that our method automatically captures existing (manually determined) dialect distinctions well, and even goes beyond them in that it also allows for a more fine-grained qualitative analysis. Our research shows that representation learning is well suited to the study of language variation, and demonstrates the potential of incorporating non-linguistic information via retrofitting. For an application of our methodology to a larger Twitter data set over multiple languages, see (Hovy et al., In Preparation).

[1] Han et al. (2014) has used city-level representations, but have not applied them to the identification of dialect areas.

Contributions In this paper, we make the following contributions to linguistic insights, performance improvements, and algorithmic contributions. We show:

1. how Doc2Vec can be used to learn distributed representations of cities that capture continuous regional linguistic variation. The approach is general and can be applied to other languages and data sets;

2. that the city representations capture enough distinction to produce competitive results in geolocation, even this was not the main focus;

3. that retrofitting can be used to incorporate geographic information into the embeddings, extending the original algorithm's applications;

4. that the clusterings match with a sociolinguistic dialect map (Lameli, 2013), measuring their homogeneity, completeness, and their harmonic mean (V-measure), and reach a V-measure of 0.77, beating an informed baseline;

We publicly release the data, code, and map files for future research at https://github.com/Bocconi-NLPLab.

2 Data

2.1 Source

We use data from the social media app Jodel,[2] a mobile chat application that lets people anonymously talk to other users within a 10km-radius around them. The app was first published in 2014, and has seen substantial growth since its beginning. It has several million users in the German-speaking area (GSA), and is expanding to France, Italy, Scandinavia, Spain, and lately the United States. Users can post and answer to posts within the radius around their own current location. All users are anonymous. Answers to an initial post are organized in threads. The vast majority of posts in Jodel are written in standard German, but since it is conceptually spoken langauge (Koch and Oesterreicher, 1985; Eisenstein, 2013), regional and dialectal forms are common, especially in Switzerland, Austria, and rural areas in Southern Germany. The data therefore reflects current

[2] https://jodel.com/

developments in language dynamics to mark regionality (Purschke, 2018).

We used a publicly available API to collect data between April and June 2017 from 123 initial locations: 79 German cities with a population over 100k people, all 17 major cities in Austria ("Mittel- und Oberzentren"), and 27 cities in Switzerland (the 26 cantonal capitals plus Lugano in the very south of the Italian-speaking area). Due to the 10km radius, posts from other nearby cities get collected as well. We include these additional cities if they have more than 200 threads, thereby growing the total number of locations.[3] Ultimately, this results in 408 cities (333 in Germany, 27 in Austria, 48 in Switzerland). The resulting locations are spread relatively evenly across the entire GSA, albeit with some gaps in parts of Germany with low population density. In total, we collect 2.3 million threads, or 16.8 million posts.

We treat each thread as a document in our representation learning setup, labeled with the name of the city in which the thread took place.

2.2 Preprocessing

We preprocess the data to minimize vocabulary size, while maintaining regional discriminative power. We lowercase the input and restrict ourselves to content words, based on the part-of-speech (nouns, verbs, adjectives, adverbs, and proper names), using the `spacy`[4] tagger.

Prior studies showed that many regionally-distributed content words are topically driven (Eisenstein et al., 2010; Salehi et al., 2017). People talk more about their own region than about others, so the most indicative words include place names (the own city, or specific places within that city), and other local culture terms, such as sports teams. We try to minimize the effect of such regional topics, by excluding all named entities, as well as the names of all cities in our list, to instead focus on dialectal lexical variation.

We use `NLTK`[5] to remove German stop words, and to lemmatize the words. While this step removes the inflectional patterns found in German, which could have regional differences, we focus here on lexical differences, and lemmatization greatly reduces vocabulary size, leading to better representations. While both POS-tagging and NER can introduce noise, they are more flexible and exhaustive than pre-defined word lists.[6] Finally, we concatenate collocations based on the PMI of the adjacent words in the cleaned corpus. The average instance length is about 40 words after cleaning.

2.3 Data Statement

The corpus was selected to represent informal, everyday online speech across the German-speaking area in Europe, and to capture regional distinctions. The data was acquired via the publicly available API. The language is mainly standard German, but with a substantial amount of dialectal entries, mainly from southern German varieties, as well as some French and Italian, which could not be removed without losing dialect. The platform is anonymous, but mainly used by young people, as indicated by a prevalence of college-related topics. It contains spontaneous, written, asynchronous interactions in a chat platform organized by threads. Anonymous reference to prior interlocutors is possible. The app is mainly used to discuss everyday topics, entertainment, flirting, venting, and informal surveys.

3 Methodology

3.1 Representation Learning

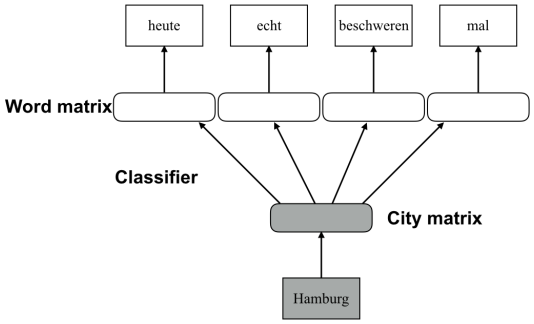

Figure 1: Doc2vec model example for window size 4.

To learn both word and city representations, we use the Doc2Vec implementation of para-

[3]The number of threads differs widely even between cities, ranging from dozens to over 40k in cities like Munich, Vienna, or Berlin.

[4]https://spacy.io/

[5]http://www.nltk.org/

[6]Note that stopwords and place names are more reliably detected in their standard form than in regional variants of abbreviations, meaning the standard forms are more reliably excluded if posts are written in High German, than if posts are written in dialect. This may lead to higher coherence for regions with a higher amount of non-standard tokens (as in Switzerland), thereby actually supporting our goal of detecting regional variation.

graph2vec (Le and Mikolov, 2014) in `gensim`.[7] The model is conceptually similar to `word2vec` (Mikolov et al., 2013), but also learns document label representations (in our case city names), embedded in the same space as the words. We use distributed bag-of-words (DBOW) training. The model parameters are fitted by predicting randomly sampled context words from a city vector. The objective is to maximize the log probability of the prediction,

$$y = \arg\max_{W} \log \sum_{i=1}^{N} log(p(w_i|k))$$

where k is a city, and $W = w_{i...N}$ a sequence of N randomly sampled words from the thread (see Figure 1 for a schematic representation).

During training, semantically similar words end up closer together in vector space, as do words "similar" to a particular city, and cities that are linguistically similar to each other.

Due to the nature of our task, we unfortunately do not have gold data (i.e., verified cluster labels) to tune parameters.We therefore follow the settings described in (Lau and Baldwin, 2016) for the parameters, and set the vector dimensions to 300, window size to 15, minimum frequency to 10, negative samples to 5, downsampling to 0.00001, and run for 10 iterations.

3.2 Visualization

In order to examine whether the city embeddings capture the continuous nature of dialects, we visualize them. If our assumption holds, we expect to see gradual continuous change between cities and regions.

We use non-negative matrix factorization (NMF) on the 300-dimensional city representation matrix to find the first three principal components, normalize them each to values 0.0–1.0 and interpret those as RGB values.[8] I.e., we assume the first principal component signals the amount of red, the second component the amount of green, and the third component the amount of blue. This triple can be translated into a single color value. E.g., 0.5 red, 0.5 green, and 0.5 blue translates into medium gray. This transformation translates city representations into color values that preserve linguistic similarities. Similar hues correspond to similar representations, and therefore, by extension, linguistic similarity.

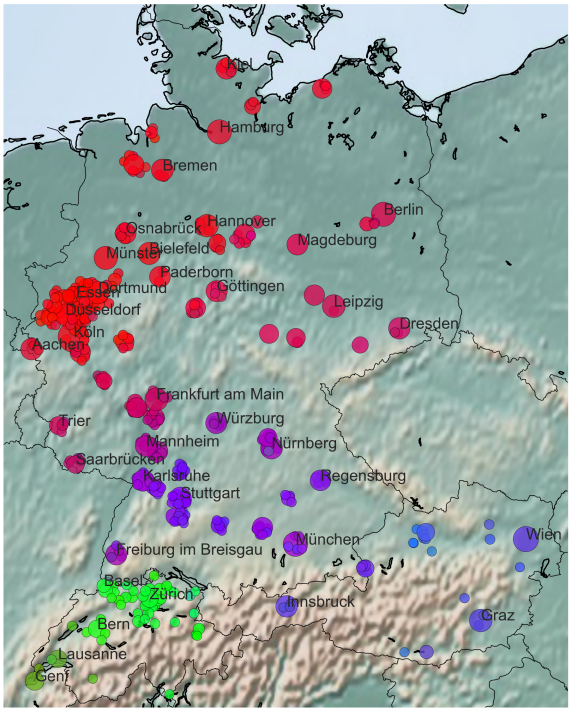

Figure 2: Gradient color map of first three components of city embeddings, interpreted as RGB, for all cities above 200 threads. Color reflects linguistic similarity.

NMF tries to find a decomposition of a given i-by-k matrix W into d components by a i-by-d row-representation V and a d-by-k column representation H. In our case, $d = 3$. Since we are only interested in a reduced representation of the cities, V, we discard H.

The result is indeed a continuous color gradient over the cities over 200 threads, see Figure 2. The circle size for every city indicates the relative number of threads per location.

In order to get reliable statistics, we restrict ourselves to cities with more than 200 observed conversations (about 2.1M conversations: 1.82M in Germany, 173k in Austria, and 146k in Switzerland). Including cities with fewer conversations adds more data points, but induces noise, as many of those representations are based on too little data, resulting in inaccurate vectors.

Even without in-depth linguistic analysis, we can already see differences between Switzerland (green color tones) and the rest of the GSA. Within

[7] https://radimrehurek.com/gensim/models/doc2vec.html

[8] Note that instead learning 3-dimensional embeddings would not amount to the same, as those are likely *not* equivalent of the three first principal components, and thus not as useful. 300 dimensions capture other degrees of variation, increasing the chance to capture meaningful latent dimensions.

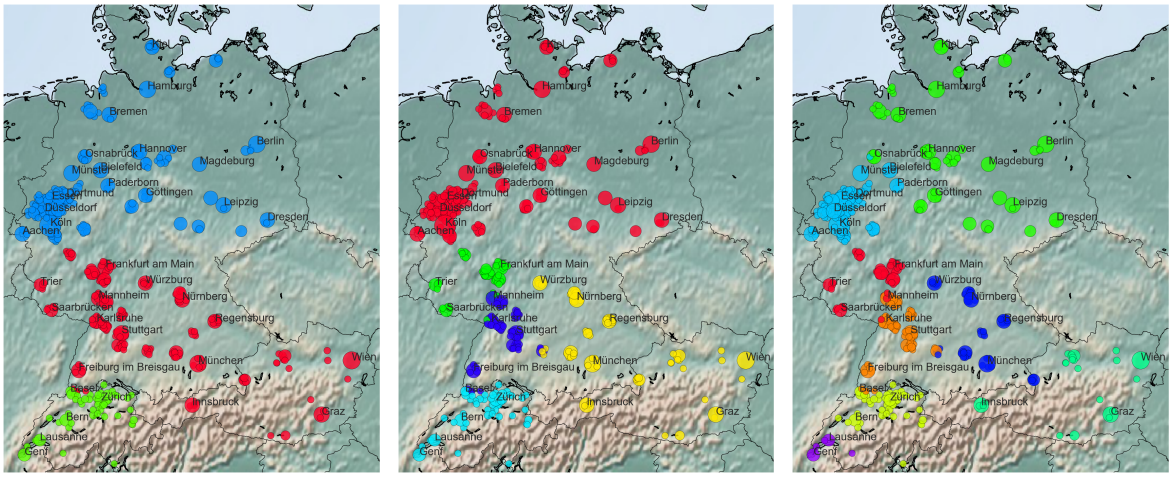

Figure 3: Clustering solutions of retrofit city embeddings for entire GSA with 3, 5, and 8 clusters. Colors denote clusters, assigned randomly.

Switzerland, we see a distinction between the German (lighter green) and the French-speaking area around Lausanne and Geneva (darker tones). On the other hand, we find a continuous transition from red over purple to bluish colors in Germany and Austria. These gradients largely correspond to the dimensions North→South(East): red→blue and West→East: intense tones →pale tones. These dimensions mirror the well-known strong linguistic connection between the southeast of Germany and Austria, and between most cities in the north of Germany.

3.3 Clustering

The visualization in the last section already suggests that we capture the German dialect continuum, and the existence of larger dialect areas. However, in order to evaluate against existing dialect maps, we need to discretize the continuous representation. We use hierarchical agglomerative clustering (Ward Jr, 1963) with Ward linkage, Euclidean affinity, and structure to discover dialect areas. We compare the agglomerative clustering results to a k-means approach.

Agglomerative clustering starts with each city in its own cluster, and recursively merges pairs into larger clusters, until we have reached the required number. Pairs are chosen to minimize the increase in linkage distance (for Ward linkage, this measure is the new cluster's variance). We use cities with 50–199 threads (66 cities) to tune the clustering parameters (linkage function and affinity), and report results obtained on cities with more than 200 threads.

Since the city representations are indirectly based on the words used in the respective cities, the clustering essentially captures regional similarity in vocabulary. If the clusters we find in our data match existing dialect distinctions, this provides a compelling argument for the applicability of our methodology.

3.4 Including geographic knowledge

While we capture regional variation by means of linguistic similarities here, it does include a geographic component as well. The embeddings we learn do not include this component, though. This can produce undesirable clustering results. Large cities, due to their "melting-pot" function, often use similar language, so their representations are close in embedding space. This is an example of Galton's problem (Naroll, 1961): Munich and Berlin are not linguistically similar because they belong to the same dialect, but due to some outside factor (in this case, shared vocabulary through migration).

To address geography, we experiment with two measures: clustering with structure, and retrofitting (Faruqui et al., 2015; Hovy and Fornaciari, 2018).

Structure To introduce geographic structure into clustering, we use a connectivity matrix over the inverse distance between cities (i.e., geographically close cities have a higher number), which is used as weight during the merging. This weight makes close geographic neighbors more likely to be merged before distant cities are.

Note, though, that this geographic component

does *not* predetermine the clustering outcome: geographically close cities that are linguistically different still end up in separate clusters, as we will see. The Spearman ρ correlation between the geographic distance and the cosine-similarity of cities is positive, but does not fully explain the similarities (Austria 0.40, Germany 0.42, Switzerland 0.72). The stronger correlation for Switzerland suggests a localized effect of regional varieties. Geographic structure in clustering does, however, provide speedups, regional stability, and more stable clustering solutions than unstructured clustering. We will see this in comparison to k-means.

Retrofitting Faruqui et al. (2015) introduced retrofitting of vectors based on external knowledge. We take the idea proposed for word vectors and semantic resources and extend it following Hovy and Fornaciari (2018) to apply it to city representations and membership in geographic regions. We construct a set Ω with tuples of cities (c_i, c_j) such that there exists a region R where $c_i \in R$ and $c_j \in R$. We use the NUTS2 regions (Nomenclature of Territorial Units for Statistics, a EuroStats geocoding standard) to determine R. In Germany, NUTS2 has 39 regions, corresponding to government regions.

To include the geographic knowledge, we retrofit the existing city embeddings C. The goal is to make the representations of cities that are in the same region more similar to each other than to cities in other regions, resulting in a retrofit embeddings matrix $\hat{C}$. For a retrofit city vector $\hat{c}_i$, the update equation is

$$\hat{c}_i = \alpha c_i + \beta \frac{\sum_{j:(i,j)\in\Omega} \hat{c}_j}{N}$$

where $\hat{c}_i$ is the original city vector, and α and β are tradeoff parameters to control the influence of the geographic vs. the linguistic information. See Faruqui et al. (2015) and Hovy and Fornaciari (2018) for more details.

4 Evaluation

In order to evaluate our methodology, we measure both its ability to match German dialect distinctions, and the performance of the learned embeddings in a downstream geolocation task.

Figure 3 provides examples of different clustering solutions after retrofitting. Note that colors are assigned randomly and do *not* correspond to the linguistic similarity from Figure 2. Switzerland immediately forms a separate cluster (the 2-cluster solution separates Switzerland vs. everything else), and further clusters first separate out more southern German varieties before distinguishing the northern varieties. This is in line with sociolinguistic findings (Plewnia and Rothe, 2012) about ubiquity of dialect use (more common in the south, therefore more varied regions, reflected in our clustering). Due to space constraints, we have to omit further clustering stages, but find linguistically plausible solutions beyond the ones shown here. For an in-depth qualitative analysis of the different clustering solutions and the sociodemographic and linguistic factors, see Purschke and Hovy (In Preparation).

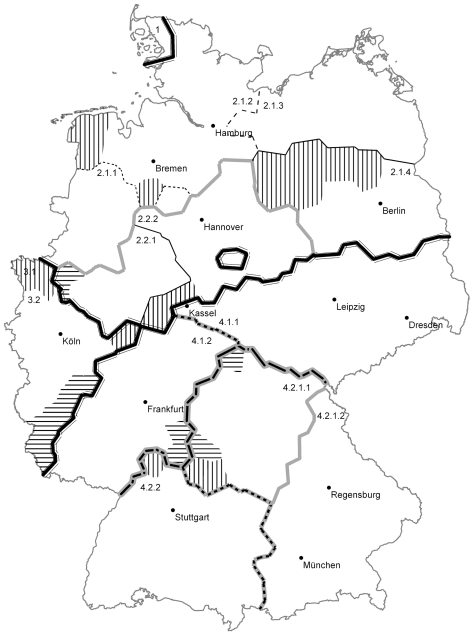

Figure 4: German dialect Regions after Lameli (2013). Shaded areas denote dialect overlap.

Dialect match We use the map of German dialects and their regions by Lameli (2013) (see Figure 4) and its 14 large-scale areas[9] as gold standard to measure how well the various clustering-solutions correspond to the dialect boundaries. This map is based on empirical quantitative analysis of German dialects, albeit based on data from the 19th century, and therefore naturally on different domains and media than our study.

Note that we can only assess the cities within modern-day Germany (clusters formed in Austria or Switzerland are not covered). We therefore rerun the clusterings on the subset of German cities, so results differ slightly from the clusters induced

[9] Some areas partially overlap with each other.

	ORIGINAL						RETROFIT					
	K-MEANS			AGGLOMERATIVE			K-MEANS			AGGLOMERATIVE		
#	V-score	H	C	V-score	H	C	V-score	H	C	V-score	H	C
2	0.41	0.27	0.89	0.41	0.27	0.83	0.43	0.28	0.94	0.44	0.28	0.95
3	0.53	0.39	0.84	0.46	0.33	0.73	0.57	0.42	0.87	0.54	0.40	0.85
4	0.61	0.49	0.80	0.59	0.48	0.76	0.66	0.53	0.86	0.68	0.56	0.88
5	0.61	0.50	0.79	0.63	0.54	0.74	0.69	0.59	0.83	0.71	0.62	0.84
6	0.65	0.56	0.76	0.64	0.58	0.72	0.72	0.64	0.82	0.72	0.64	0.82
7	0.64	0.57	0.74	0.66	0.61	0.72	0.72	0.65	0.80	0.69	0.64	0.76
8	0.62	0.56	0.70	0.66	0.61	0.71	0.70	0.67	0.74	0.73	0.70	0.76
9	0.70	0.65	0.76	0.70	0.68	0.72	0.70	0.67	0.73	0.73	0.70	0.75
10	0.68	0.66	0.70	0.70	0.68	0.72	0.71	0.70	0.72	0.74	0.72	0.75
11	0.69	0.67	0.71	**0.72**	0.71	0.72	0.74	0.75	0.74	0.74	0.74	0.74
12	0.66	0.65	0.67	0.70	0.72	0.68	0.71	0.72	0.69	0.75	0.78	0.72
13	0.67	0.68	0.66	0.70	0.73	0.67	0.73	0.75	0.71	0.74	0.78	0.70
14	0.67	0.68	0.66	0.69	0.73	0.65	0.71	0.76	0.66	0.74	0.79	0.70
15	0.66	0.68	0.64	0.71	0.77	0.66	0.74	0.80	0.70	0.76	0.82	0.70
16	0.67	0.71	0.64	0.71	0.78	0.66	0.73	0.80	0.67	**0.77**	0.85	0.71
17	0.67	0.70	0.63	0.70	0.78	0.64	0.73	0.81	0.67	0.76	0.85	0.68
18	0.65	0.68	0.62	0.70	0.78	0.64	0.74	0.83	0.66	0.75	0.85	0.66
19	0.64	0.68	0.59	0.70	0.79	0.63	0.72	0.82	0.64	0.75	0.87	0.67
20	0.65	0.71	0.59	0.69	0.80	0.61	0.74	0.85	0.66	0.75	0.87	0.66

Table 1: Evaluation of the fit of various cluster solutions against the reference dialect map Lameli (2013). k-means results averaged over 5 runs. Agglomerative clustering with structure. Retrofitting on NUTS2 regions. Baseline: 0.74 V-score, 0.93 homogeneity, 0.62 completeness.

on the entire GSA.

We report *homogeneity* (whether a cluster contains only data points from a single region) and *completeness* (how many data points of a NUTS region are in the same cluster), as well as their harmonic mean, the *V-score*. This corresponds to precision/recall/F1 scores used in classification. Note that we will not be able to faithfully reconstruct Lameli's distinctions, since Lameli's map contains overlapping regions, whose data points therefore already violate perfect homogeneity.

The outline of dialect regions in Lameli's map is based on the NUTS2 regions, so we compare all clustering solutions to an informed baseline that assigns each city the NUTS2 region it is located in. Except for regions in dialect overlaps, each NUTS region is completely contained in one dialect region, so the baseline can achieve almost perfect homogeneity.

Downstream task geolocation For the geolocation task, we randomly select 100 cities with at least 200 threads from each country (7 in Austria, 82 in Germany, 11 in Switzerland). We then collect threads with at least 100 words from these cities for each country (11,240 threads from Austria, 137,081 from Germany, and 18,590 from Switzerland). Each thread is a training instance, i.e., we have 166,911 instances. We use the Doc2Vec model from before to induce a document representation for each instance and use the vector as input to a logistic regression model that predicts the city name.

For testing, we sample 5,000 threads from the same cities (maintaining the same proportional distribution and word count constraint), but from a *separate* data set, collected two months after the original sample. We again use the Doc2Vec model to induce representations, and evaluate the classifier on this data.

We measure accuracy, accuracy at 161km (100 miles), and the median distance between prediction and target. We compare the model with Doc2Vec representations to a bag-of-words (BOW) model with the same parameters. Since the representation here is based on words, we can not apply retrofitting. As baseline, we report the

most-frequent city prediction.

5 Results

Dialect match Table 1 shows the results of clustering solutions up to 20 clusters for both retrofit and original embeddings. Irrespective of the clustering approach, retrofit representations perform markedly better.

Homogeneity increases substantially the more clusters we induce (in the limit, each data point becomes a single cluster, resulting in perfect homogeneity), whereas completeness decreases slightly with more clusters (they increase the likelihood that a region is split up into several clusters). We achieve the best V-score, 0.77, with 16 clusters.

Averaged k-means (over 5 runs) is much less consistent, due to random initialization, but presumably also because it cannot incorporate the geographic information. For few clusters, its performance is better than agglomerative clustering, but as the number of clusters increases (and the geographic distribution of the cities becomes more intricate), k-means stops improving.

The baseline achieves almost perfect homogeneity, as expected (the only outliers are NUTS regions in overlap areas). Completeness is lower than almost all clustering solutions, though. The V-score, 0.74, is therefore lower than the best clustering solution.

Both the cluster evaluation metrics and the visual correspondence suggest that our method captures regional variation at a lexical level well.

MODEL	↑ACC	↑ACC@161	↓MED. DIST.
baseline	0.03	0.31	269.33
BOW	0.21	0.50	156.17
D2V	**0.26**	**0.52**	**145.16**

Table 2: Geolocation performance for city embeddings and bag-of-word vectors on held-out data set. Baseline predicts most frequent city from training data.

Downstream Evaluation: Geolocation Table 2 shows the results of the geolocation downstream task. Despite the fact that the representation learning setup was *not* designed for this task and excluded all the most informative words for it (Salehi et al., 2017), the induced embeddings capture enough pertinent regional differences to achieve reasonable performance (albeit slightly below state of the art, which typically has a median distance around 100km, and an accuracy@161 of 0.54, see cf. Rahimi et al. (2017b)) and decidedly outperform the BOW model and most-frequent-city baseline on all measures.

6 Analysis

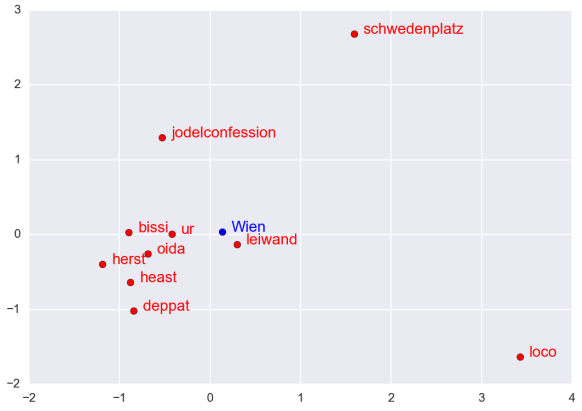

Figure 5: Visualization of city representation for Wien (Vienna) and its 10 nearest word neighbors in two dimensions. The closest seven words are all Austrian dialect words

Because both words and cities are represented in the same embeddings space (at least before retrofitting), we can compare the vectors of cities to each other (asking: which cities are linguistically most similar to each other, which is what we have done above) and words to cities (asking: which words are most similar to/indicative of a city). The latter allows us to get a qualitative sense of how descriptive the words are for each city.

Figure 5 shows an example of word and city similarity for the city representation of Vienna.

We can also use the cluster centroid of several city vectors to represent entire regions. The new vector no longer represents a real location, but is akin to the theoretic linguistic center of a dialect region. We can then find the most similar words to this centroid. For the solution with 3 clusters (cf. Figure 3), we get the solutions in Table 3. As expected, the regional prototypes do not overlap, but feature dialectal expressions in the south, and general standard German expressions in the north.

Again, for an in-depth qualitative analysis and discussion of the socio-linguistic correlations, see Purschke and Hovy (In Preparation).

7 Related Work

Dialectometric studies, exploring quantitative statistical models for regional variation, range from

CLUSTER	PROTOTYPES	TRANSLATION
Switzerland	esch, ond, vell, gaht, wüki, nöd, besch, emmer, nor, au nöd	*is, and, many, goes, really, not, (you) are, always, just, also not*
Northern Germany	ja gut, erstmal, sieht, drauf, vielleicht, mehr, gut, sehen, schonmal, Ahnung	*well yes, first, sees, onto, maybe, more, good, see, already, idea*
Southern Germany & Austria	afoch, voi, nd, i a, oda, möppes, nimma, is a, mei, gscheid	*easy, full, and, me too, or, girl (*SLANG*), no more, is also, well, smart*

Table 3: Prototypical words (10 nearest neighbors) for each of 3 clusters.

work on dialect data in Dutch (Nerbonne and Heeringa, 1997; Prokić and Nerbonne, 2008; Wieling et al., 2011, inter alia) and British English (Szmrecsanyi, 2008), to Twitter-based approaches for American dialect distinctions (Grieve et al., 2011; Huang et al., 2016) and the regional differentiation of African American Vernacular English (Jones, 2015). While these papers rely on existing dialect maps for comparison, they rarely quantitatively evaluate against them, as we do.

Recently, NLP has seen increased interest in *computational sociolinguistics* (Nguyen et al., 2016). These works examine the correlation of socio-economic attributes with linguistic features, including regional distribution of lexical and phonological differences (Eisenstein et al., 2010; Doyle, 2014; Bamman et al., 2014), syntactic variation (Johannsen et al., 2015), diachronic variation (Danescu-Niculescu-Mizil et al., 2013; Kulkarni et al., 2015; Hamilton et al., 2016), and correlation with socio-demographic attributes (Eisenstein et al., 2011; Eisenstein, 2015). Other have further explored regional variation on social media, and showed the prevalence of regional lexical variants (Hovy et al., 2015; Hovy and Johannsen, 2016; Donoso and Sánchez, 2017). Several works include quantitative comparisons to measure the empirical fit of their findings (Peirsman et al., 2010; Han et al., 2014; Huang et al., 2016; Grieve, 2016; Kulkarni et al., 2016), albeit not on entire existing dialect maps.

The use of representation learning is new and relatively limited, especially given its prevalence in other areas of NLP. Bamman et al. (2014) have shown how regional meaning differences can be learned from Twitter via distributed word representations between US states, but not for individual cities. More recently, Kulkarni et al. (2016); Rahimi et al. (2017a) and Rahimi et al. (2017b) have shown how neural models can exploit regional lexical variation for geolocation, while also enabling dialectological insights, whereas our goals are exactly reversed. Östling and Tiedemann (2017) have shown how distributed representations of entire national languages capture typological similarities that improve translation quality. Most of these papers focus on downstream performance that accounts for regional variation, rather than on explicitly modeling variation. We include a downstream performance, but also evaluate the cluster composition quantitatively.

8 Conclusion

We use representation learning, structured clustering, and geographic retrofitting on city embeddings to study regional linguistic variation in German. Our approach captures gradual linguistic differences, and matches an existing German dialect map, achieving a V-score of 0.77. The learned city embeddings also capture enough regional distinction serve as input to a downstream geolocation task, outperforming a BOW baseline and producing competitive results. Our findings indicate that city embeddings capture regional linguistic variation, which can be further enriched with geographic information via retrofitting. They also suggest that traditional ideas of regionality persist online. Our methodology is general enough to be applied to other languages that lack dialect maps (e.g., Switzerland), and to other tasks studying regional variation. We publicly release our data and code.

Acknowledgements

We would like to thank the anonymous reviewers of this paper and Barbara Plank, who helped to strengthen and clarify our findings.

References

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed Representations of Geographically Situated Language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 828–834. Proceedings of ACL.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM.

Gonzalo Donoso and David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. *VarDial 2017*, page 16.

Gabriel Doyle. 2014. Mapping Dialectal Variation by Querying Social Media. In *EACL*, pages 98–106.

Jacob Eisenstein. 2013. What to do about bad language on the Internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

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.

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-Volume 1*, pages 1365–1374. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Jack Grieve. 2016. *Regional variation in written American English*. Cambridge University Press.

Jack Grieve, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(2):193–221.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*, pages 1489–1501. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Dirk Hovy and Tommaso Fornaciari. 2018. Increasing In-Class Similarity by Retrofitting Embeddings with Demographic Information. In *Proceedings of the 2018 conference on Empirical Methods in Natural Language Processing*.

Dirk Hovy and Anders Johannsen. 2016. Exploring Language Variation Across Europe–A Web-based Tool for Computational Sociolinguistics. In *Proceedings of LREC*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*.

Dirk Hovy, Afhsin Rahimi, Tim Baldwin, and Julian Brooke. In Preparation. Visualizing Regional Language Variation Across Europe on Twitter. In Stanley D. Brunn and Roland Kehrein, editors, *Handbook of the Changing World Language Map*. Dordrecht: Springer.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe-Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. 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. 2016. Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. In *ICWSM*, pages 615–618.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54. Walter de Gruyter.

Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. page 78.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

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

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

Raoul Naroll. 1961. Two solutions to Galton's problem. *Philosophy of Science*, 28(1):15–39.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*, pages 11–18.

John Nerbonne, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. *Time Warps, String Edits and Macromolecules: The theory and practice of sequence comparison*, 15.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*.

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

Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

Albrecht Plewnia and Astrid Rothe. 2012. Sprache – Einstellungen – Regionalität. In Ludwig M. Eichinger, Albrecht Plewnia, Christiane Schoel, Dagmar Stahlberg, and Gerhard Stickel, editors, *Sprache und Einstellungen. Spracheinstellungen aus sprachwissenschaftlicher und sozialpsychologischer Perspektive. Mit einer Sprachstandserhebung zum Deutschen*.

Jelena Prokić and John Nerbonne. 2008. Recognising groups among dialects. *International journal of humanities and arts computing*, 2(1-2):153–172.

Christoph Purschke. 2018. Language regard and cultural practice: Variation, evaluation, and change in the German regional languages. In Betsy Evans, Erica Benson, and James Stanford, editors, *Language regard: Methods, variation, and change*, pages 245–261. Cambridge University Press, Cambridge.

Christoph Purschke and Dirk Hovy. In Preparation. Lörres, Möppes, and the Swiss. (Re)Discovering Regional Patterns in Anonymous Social Media Data. *Journal of Linguistic Geography*.

Stefan Rabanus, Roland Kehrein, and Alfred Lameli. 2010. Creating digital editions of historical maps. *Language and space*, 2:375–385.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017a. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017b. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL2017)*. Association for Computational Linguistics (ACL2017).

Bahar Salehi, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and hockey teams: Names can reveal your location. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 116–121.

Robert G Shackleton Jr. 2005. English-American speech relationships: A quantitative approach. *Journal of English Linguistics*, 33(2):99–160.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2017a. Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of VarDial Workshop*. Association for Computational Linguistics.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017b. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of EACL*. Association for Computational Linguistics.

Benedikt Szmrecsanyi. 2008. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):279–296.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Martijn Wieling, John Nerbonne, and R Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, 6(9):e23613.