word2diavec: Modeling second-person pronouns in vector space

Margaret Anne Rowe
MLC '19
Georgetown University
mar299@georgetown.edu

Abstract

We use FastText and word2vec vector space models to study the second-person plural pronouns of English as used in informal textual communication over the social networking website Twitter. Though model evaluation was quite poor due to a corpus size of 13k tweets, we found notable cosine similarity for all pronouns in comparison to the baseline "you" but little similarity between individual pronouns, suggesting that speakers may use them in different contexts to index different meanings.

1 Objective

Semantic vector space models (VSMs) represent words as numbered weights across hundreds of dimensions of space. VSMs allow for researchers to quantify the relationship between words on an abstract level, drawing from Harris' (1954) theory of distributional semantics to understand meaning based on a word's context.

What can VSMs tell us about the spread of dialectal language forms over social media? Though social media is conducted in a largely standardized written medium, meaning it loses many nonstandard features of dialect speech, it is home to a host of stylistic and textual innovations that can more closely mirror informal speech and dialectal vocabulary. Its online, asynchronous nature means that many linguistic forms once unknown outside regional boundaries have the opportunity to spread and be adapted around the country.

In this paper, we use VSMs to model second-person plural pronouns in dialectal English as used in casual conversations on the social network Twitter. We hypothesize that there will be distinct differences in their semantic meaning encoded within their offsets.

Furthermore, dialectal maps have proven to be viral successes in the 2010s, garnering millions of pageviews and bringing greater linguistic awareness to the broader public (Vaux and Goulder, 2003; Vaux and Jøhndal, n.d.; Katz and Andrews, 2013; Katz, 2016). Studying, visualizing, and mapping pronouns gives us as researchers the opportunity to make abstract relations relatable to non-specialists. Rather than use dialect maps, this study makes an exploratory attempt to use visualizations of VSMs instead.

2 Dataset

This project uses two data sources: a .tsv file of geographically-sorted Twitter data and Mikolov et al.'s (2013) analogy dataset.

Twitter is a microblogging service that allows users to post publicly-viewable *tweets* of up to 240 characters. These data were downloaded using the Roesslin's (2009) Tweepy implementation of the Twitter API, which allows for approximately 450 tweets to be downloaded per 15-minute period. The script was run for approximately 10 hours on nine search terms to download 10,000 unique tweets and associated metadata per term. Due to formatting issues, it is difficult to ascertain how many tweets were initially saved; after filtering, 14,384 tweets remained.

Filtering consisted of matching user-provided location fields against a gazeteer of major U.S. cities, states and abbreviations, and regional nicknames. Tweets were saved if their given locations were found in the gazeteer and they did not contain certain emoji and stopwords that potentially compromised the accuracy of their data (e.g., ➡️ or *to*, implying that the tweeter had moved cities; or *she/her* or ⚧, implying that the tweeter has used the location field to index their identity). Though Twitter does allow users to automatically tag their tweets with their current geographic location, this feature is rarely used; however, Pavalanathan and Eisenstein (2015) found a high degree of correlation between self-reported and auto-generated location data, while Hecht et al. (2011) found that 66% of location fields did contain “real” geographic data.

Mikolov et al.’s (2013) evaluation dataset consists of 14 categories of 19,495 analogies of syntactic, geographic, and gendered terms in relation to one another. It is used to evaluate whether a VSM captures morphological relations and world knowledge (e.g., *sing : singing :: rain : raining*; *Atlanta : Georgia :: Richmond : Virginia*).

3 Background

The field of dialectology encompasses the study of regional varieties of speech shared by groups of speakers, documenting the diverse sociocultural, theoretical, and historical heritage of human behavior (Wolfram and Schilling, 2016). As informal communication has moved online, so have regional dialects; while many writing systems disprefer regional forms that differ from a codified standard, the ubiquity of online communities in day-to-day life has led to notable diversity in its textual styles (Eisenstein, 2018).

Computational dialectology is a small yet growing subfield of interest within dialectology, taking advantage of the vast quantities of natural communication online to quantitatively explore the extent to which geographical variation can be observed in online writing, how it matches spoken geographical variation, and how it changes over time (Eisenstein, 2018).

English is notable for its lack of a second person plural pronoun (SPPP), used to refer to a group of addressees. This is due to the historical quirk of *you*, originally a plural form, gradually being used in the same manner as French *vous*, which can be either plural or formal. *You* had replaced singular *thou* outright by 1575 and left English with it as the only second-person pronoun for both singular *and* plural addressees (Harper, 2019). English speakers worldwide have filled the semantic gap by innovating dialectal SPPPs, none of which have become standard and most of which are stigmatized.

In the United States, Wolfram and Schilling (2016) note multiple SPPPs, including:

- **you guys** – common throughout the U.S., but especially in the Midwest, Northeast, and West Coast. Debatably gender-neutral
- **yall / y’all** – Southern dialects and African American English, but used throughout the country (Eisenstein, 2018)
    - *y’all* is a contracted form of **you all**, which is used throughout the US, but notably in Kentucky
- **yinz** – Pittsburgh, Pennsylvania dialect
- **youse** – Northern, especially Northeastern dialects

Despite their regional origins, these pronouns have often spread throughout the United States, as visible in dialect surveys (Vaux, 2002; Katz, 2016). This is likely in part due to the internet making once region-locked forms more visible nationally. Whether this is at the expense or advantage of smaller regional forms is up for debate.

4 Methodology

Once downloaded and filtered, tweets were extracted from the .tsv file, preprocessed to split emojis and other special characters into individual tokens while normalizing collocations and contractions of interest (*you guys* becomes *youguys*, *y’all* becomes *yall*), and tokenized using NLTK (Bird et al., 2009).

Once tokenized, the corpus was used to train 100-dimension word2vec and FastText VSMs for 50 epochs using Gensim (Řehůřek and Sojka, 2010), creating a vocabulary of the 20,969 tokens that appeared at least twice in the corpus.

word2vec is a word-based model, while FastText is a subword-based model that uses character ngrams to enrich the vector space, making it better for sparse data. Though word2vec seems to encode semantic similarity better than FastText (Jain, 2016), Twitter’s range of colloquial and innovative textual styles may lead to lower accuracy when trained at the word level. As such, it was decided to use both models and compare results, using averaged cosine similarity between models’ shared vocabulary words as a baseline.

Evaluation of the models’ performance was done using the offset methods for solving word analogies outlined in Linzen (2016). Estimating the offset using one pair of related words should also allow us to discover other word pairs that are related in the same way. So, the offset method traditionally solves the analogy

$$a : a^* :: b : \underline{\quad\quad} \tag{1}$$

by finding the word closest in cosine similarity to the landing point; formally,

$$x^* = \operatorname{argmax} \cos(x',\ a^* - a + b) \tag{2}$$

or, informally,

$$yinz - Pittsburgh + Atlanta = yall \tag{3}$$

since

$$Pittsburgh : yinz :: Atlanta : yall \tag{4}$$

as they are both regional forms of plural *you*.

However, this “vanilla” model is imprecise, as it may still give the correct answer even if the offsets are inconsistent; the closest word to *x'* may just be *b*’s nearest neighbor. As such, Linzen proposes a battery of other functions. We evaluate our model here using

- The standard **Vanilla** model,
- **Only-B**, which returns the nearest neighbor of *b*,
- and **Ignore-A**, returning the word most similar to both *a** and *b*.

These metrics were applied to the corpus using Mikolov et al.’s (2013) dataset and Pedregosa et al.’s (2011) cosine similarity function in scikit-learn, scoring one point for each accurate *x'* out of the total number of examples. The same metrics were also run using a small corpus of pronoun analogies meant to capture number and location.

5 Results

Overall performance of the model was quite poor. The model’s average Linzen scores on the Mikolov et al. data for both the entire dataset and the syntactic categories can be seen in Table 1, as well as the average scores for the syntax. This is almost certainly due in part to the miniscule corpus size, as well as stylistic and tonal differences in the gathered Twitter data that may disprefer certain categories included in the more formal analogy set (e.g., currencies, world capitals). Somewhat surprisingly given Jain’s (2016) reported semantic advantage of word2vec, FastText outperformed word2vec in every area, with scores for both rising slightly when tested on semantic and morphological data rather than geographic data, which is more likely to be out-of-vocabulary. Furthermore, we see scores rising for each respective Linzen metric, with Ignore-A producing the highest accuracy rates.

When applying the Linzen metrics to a custom dataset of 200 pronoun analogies (e.g., *you : yall :: I : we*), all results plummeted (save for Vanilla word2vec), as shown in Table 2. Such a decrease, on an analogy set designed to measure the offsets of the pronouns that the data were specifically cherrypicked for, was a surprise.

Metric	FastText	word2vec
Vanilla (all)	0.0174	0.0012
Vanilla (syn)	0.0315	0.0017
Only-B (all)	0.0491	0.0055
Only-B (syn)	0.0852	0.0058
Ignore-A (all)	0.1055	0.0138
Ignore-A (syn)	0.1891	0.0215

Table 1: FastText and word2vec accuracy scores on Linzen’s (2016) metrics when tested on Mikolov et al.’s (2013) analogies. all represents the averaged score across the full dataset, while syn represents syntactic and morphological categories only.

Metric	FastText	word2vec
Vanilla	0.0100	0.0300
Only-B	0.0250	0.0000
Ignore-A	0.0100	0.0000

Table 2: FastText and word2vec accuracy scores on Linzen’s (2016) metrics when tested on 200 pronoun analogies.

Pronoun	cos sim.	Pronoun	cos sim.
you	-0.0604	*yall*	0.0236
youse	-0.0246	*youguys*	-0.0315
yinz	0.010	*youall*	0.0731

Table 3: Cosine similarity between models for the six pronouns of interest.

Model similarity between pronoun tokens was also quite low, with models ranging from -0.06 to 0.07 cosine similarity. This likely has much to do with the way they were trained; however, their proximity to 0, indicating almost no relation between the two terms at all, was unexpected.

Where the models actually succeeded was their ability to encode relationships between the pronouns of interest to this study, as shown in Tables 4a-b. Though FastText suggested stronger relations between pronouns than word2vec, the difference was fairly small. Furthermore, we see that all pronouns are similar to *you* across models, ranging from 0.56 cosine similarity in word2vec with *youse* to a high 0.79 with *y’all*. Most notable among the dialectal pronouns is *youguys* and *yall*, the two “standard” dialectal pronouns, which have a score 0.73 similarity. Also fascinating is the difference between *yall* and *youall* – despite the former being a contraction of the latter, they rank among the most dissimilar of all pronouns.

However, visually representing this model in a Pyplot graph (Hunter, 2007) somewhat distorts these relationships, as seen in Figure 1. In reducing a VSM’s 100 dimensions down to two using scikit-

pronoun	*youse*	*yinz*	*yall*	*youguys*	*youall*
you	0.6966	0.6781	0.7798	0.7917	0.7266
youse		0.6954	0.5937	0.5641	0.5214
yinz			0.6704	0.6168	0.5422
yall				0.7290	0.5993
youguys					0.7523

Table 4a: Pronoun cosine similarity in FastText.

pronoun	*youse*	*yinz*	*yall*	*youguys*	*youall*
you	0.5651	0.6444	0.7628	0.7167	0.6401
youse		0.7007	0.4959	0.4523	0.3748
yinz			0.6420	0.5716	0.5222
yall				0.5969	0.5192
youguys					0.5969

Table 4b: Pronoun cosine similarity in word2vec.

learn's implementation of the PCA algorithm (Pedregosa et al., 2010), it necessarily loses some of the information encapsulated in the original data. As such, the pronouns are fairly spaced out; while still unusually spaced out, we see intriguing proximity between *youall* and *youguys*, and *yall, yinz,* and *youse*.

6 Discussion

What went wrong with these models? To start, a lack of data. Though Twitter offers researchers impossibly vast quantities of data, the limits it places on its free API greatly limit our ability to easily access that data. When further filtered for "Americans using pronouns," the amount of available data drops further.

Despite this, with the data we do have, we can see that dialectal pronouns do appear to be used differently on Twitter. Though they all share some degree of similarity to the base form *you*, their cosine scores drop when compared to each other, suggesting that each is used in different contexts with a different meaning. What's more, we can see that FastText is indeed a better model for encoding sparse datasets, despite word2vec's semantic superiority.

Furthermore, this study suggests that the subword-level FastText does model social media data better than word2vec. While it was expected that FastText would outperform word2vec due to its character-level encoding, it was surprising that word2vec's advantages with semantic similarity did not help it perform better.

Further work is needed to understand how semantic shift and geographic spread affects SPPPs, synchronically and diachronically. The models of these data did not adequately capture the rich context needed to understand pronouns in context; future work would be wise to supplement such quantitative methods with a more qualitative corpus study, as well as to use transfer learning to fine tune a VSM trained over gigabytes of text with the gathered data in order to retrieve more robust representations. Additionally, this work regrettably leaves out *you'uns*, an Appalachian variant of plural *you*, as well as *u*, an abbreviation of *you* commonly used in casual online conversation. Non-canonical forms such as *yallll* or *youz*, are also not considered here. To more fully understand what SPPPs mean in the United States, and how they contribute to the rich stylistic and dialectal diversity both in and out of regional boundaries, we must more thoroughly consider these variants (of what are, already, variants).

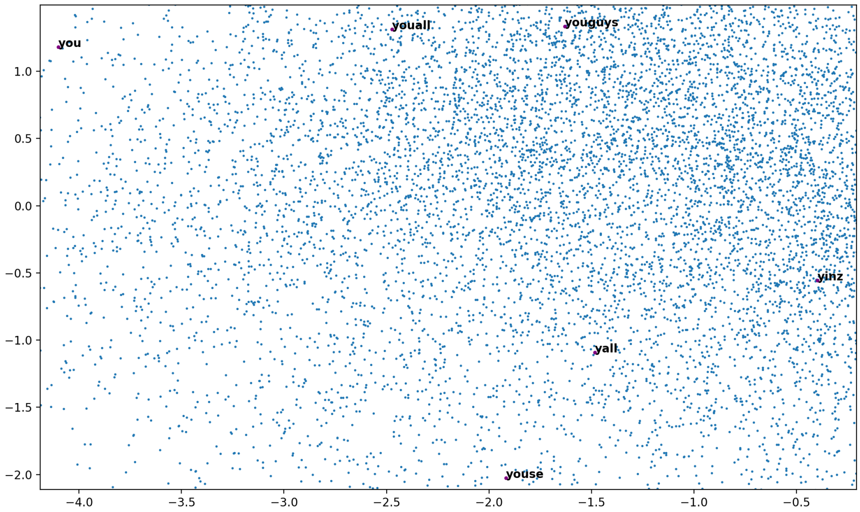

Figure 1: The FastText model in two dimensions, reduced using scikit-learn's PCA module.

Acknowledgments

Special thanks goes to Emily Beaman (BA'16, MS'18) for patiently answering all of the vector space questions I couldn't find on StackExchange.

References

S. Bird, E. Loper, and E. Klein 2009. *Natural language processing with Python*. O'Reilly Media Inc., Cambridge, MA.

D. Harper. (2019). you (pron.). *Online etymology dictionary.* https://www.etymonline.com/word/you

Z. S. Harris. 1954. Distributional structure. *Word*, *10*(2-3), 146–162.

B. Hecht, L. Hong, B. Suh, and E. H. Chi. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of Human Factors in Computing Systems (CHI)*, pages 237–246. http://misrc.umn.edu/Papers/Research%20Papers/bhecht_chi2011_location.pdf.

J. Jain. 2016. FastText and Gensim word embeddings. *Rare Technologies blog.* https://rare-technologies.com/fasttext-and-gensim-word-embeddings/

J. Katz and W. Andrews. 2013. How y'all, youse and you guys talk. *The New York Times.* https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html

J. Katz. 2016. *Speaking American: How y'all, youse, and you guys talk: A visual guide*. Houghton Mifflin Harcourt.

T. Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP.* doi:10.18653/v1/W16-2503

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*. https://arxiv.org/pdf/1301.3781.pdf.

U. Pavalanathan and J. Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 2138–2148. https://aclweb.org/anthology/D15-1256.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderpas, A. Passos, D. Corpeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valetta, Malta.

J. Roesslin. 2009. *Tweepy: An easy-to-use Python library for accessing the Twitter API.* http://www.tweepy.org/.

B. Vaux and S. A. Goulder. 2003. *The Harvard dialect survey.* http://dialect.redlog.net/

B. Vaux and M. Jøhndal. n.d. *The Cambridge online survey of World Englishes.* http://www.tekstlab.uio.no/cambridge_survey

W. Wolfram and N. Schilling. 2016. *American English: Dialects and variation* (3rd ed.). John Wiley & Sons, Chichester, UK.