Modeling Language Games

Chris Proctor, Veronica Lin

cproctor@stanford.edu, vronlin@stanford.edu
Stanford University, cs229

1 Introduction

The last five years have seen great advances in natural language processing (NLP). Word2Vec (Mikholov et al., 2013) uses an embedding which maps words to high-dimensional vectors such that nearby words have similar meaning. GloVe (Pennington et al., 2014) takes a different approach, considering global word covariance statistics. Both models are effective for a wide range of NLP tasks (Baroni et al., 2014).

Training these models relies on the fact that word meanings are based on their use. However, the results frame word meaning as fixed and global, whereas learning theories view word meaning as dynamically shaped through use in local "language games" (Wittgenstein, 1953). Talk is an important practice through which identities are performed; learning is described as increasingly central participation in the practices of a community (Lave & Wenger, 1991) and as "the process of identity formation in figured worlds" (Boaler & Greeno, 2000).

> Communities of practice and other "figured worlds" provide examples of the types of people that it is possible for participants in those communities and worlds to become– street vendor, rap artist, lover of books, chess master, public speaker, basketball player, technology whiz, team leader. And they provide opportunities for participants to begin to enact new identities, to take on and to adapt sanctioned ways of behaving, interacting, valuing, and believing. (Hull & Greeno, 2006)

NLP methods could be profoundly useful in analyzing discourse in education and related fields if they could model contextualized linguistic meaning, and how it affects participation. This project proposes steps toward that goal with two research questions:

1. Can we use the speech of participants in a discourse community to predict their trajectories of participation?
2. Can we characterize how word meanings change between everyday language and within the discourse community?

2 Related Work

We extend the work of Danescu-Niculescu-Mizil et al. (2013), who explore the relationship between change in online linguistic communities and relative change in individual users' language. They found evidence of Labov's language stability assumption: after a period of linguistic adolescence, users' linguistic practices tend to stabilize. As community norms change, users drift toward the periphery of the community before abandoning it. Using this phenomenon, the authors were able to predict a user's community lifespan based on the pace of her linguistic maturation within the community. Using monthly snapshots of community practices in the form of bigram probabilities, Danescu-Niculescu-Mizil et al. then compute the cross-entropy of individual posts to determine the alignment of the post and community norms. Although this approach provides a distance

metric between a user's language use and community norms, it does not offer further insight into the relationship. We replicate their approach as a baseline, and extend it using word vector language models that capture more semantic information.

There have been numerous approaches to identifying bias and stereotyping in speech. Using hand-selected features of dialogue from police body cameras, Voight (2017) showed that officers consistently speak less respectfully to black people than white. Wu (2017) used distributions of words and topics around males and females to show sexism in an online Economics forum. These examples demonstrate the potential for fine-grained linguistic analysis.

Because characterizing changes in high-dimensional spaces is difficult, we analyze changes within subspaces. Arora et al. (2015) argue that the linear structures observed within word embeddings, such as the well-known example "king - man + woman = queen," should be understood as semantic relations, where projecting changes in word positions onto a subspace captures the way a word's meaning changes over time with respect to the specified relation. Kusner et al. (2015) compare documents using 'word mover distance,' a measure of document similarity using pointwise distance between their words. Building on this, Bolukbasi et al. (2016) used a subspace based on gender to identify bias in word vectors. We applied similar approaches to characterize differences in the embeddings.

3 Methods

3.1 Data Collection

Hacker News is a discussion forum affiliated with the Bay Area startup incubator Y Combinator. The site is organized as a list of posts, each of which has an attached comments thread. Table 1 provides summary statistics from the dataset, which was obtained using Hacker News's public API and includes all comments from the site's inception in 2007 through August 2017. The data was preprocessed using Python's NLTK.

Table 1. Hacker News comments dataset statistics

Posts	12,166,758
Users	315,634
Users with more than 50 posts	31,274
Median sentences per post	2.0 (std: 3.0)
Median words per post	48.0 (std: 85.4)

3.2 Predicting User Trajectories with Linguistic Features

Partially replicating Danescu-Niculescu-Mizil et al. (2013), we predicted whether or not a user will soon depart the community (fewer than m subsequent posts), given a user's first w posts with at least n total posts. Following one case from the prior paper, we use $w = 20$, $m = 30$, and $n = 200$. The distribution of users joining, leaving, staying, and bouncing (joining and leaving in the same year) follows similar trends as the datasets analyzed in the earlier paper.

We created two monthly snapshot language models to generate features. The first used kenlm (Heathfield, 2011) to construct a bigram language model with modified Kneser-Ney smoothing (Heathfield et al., 2013). The second used the word2vec negative sampling skipgram algorithm (Mikholov et al., 2013) with vectors in $\Re^{300}$. For each word, the model predicts the likelihood

of seeing each word in a surrounding window as well as the likelihood of seeing several negative samples drawn from the vocabulary. After the model was initialized with a pretrained embedding on 100bn words from Google News (representative of everyday language), the word vectors were iteratively trained over each month's comments using gensim's word2vec implementation (Rehurek & Sojka, 2010) with a constant learning rate of 0.1 and 10 epochs per month.

Replicating the analysis in Danescu-Niculescu-Mizil et al. (2013) on linguistic flexibility, we used a sample of 10000 users with at least 50 posts of 30+ tokens. Posts from each decile of the user's lifespan in the community were used to compute linguistic flexibility, as represented by the Jaccard coefficient of a post with respect to the previous ten posts. A post's distance from the community was calculated using each snapshot language model for the month in which the post was written, with cross-entropy for the bigram language model and log-likelihood for the word vector model.

Using the two language models, the following features were generated. Each feature is computed for each of a user's first $w = 20$ posts, and then the posts are grouped into bins of size 5 and the feature values are averaged.

1. Frequency, the average time between comments
2. Month, the month in which the comment was posted
3. BigramCE, the cross-entropy of the post (monthly bigram model)
4. WordVectorLL, the log-likelihood of the post (monthly word vector model)
5. DiffLL, the difference between the log-likelihood of the post with respect to the monthly word vector model and the initial "everyday language" model.

The first two features are known to be highly predictive of user retention (Yang et al., 2010). Using these features, we trained a logistic regression model on 60% of the users, reserving 20% for a development set and 20% for a test set. We report precision, recall, and F_1 for each combination of features.

3.3 Modeling Word Meaning Change Over Time

The trained monthly word vector models allow us to measure linguistic changes in word meanings. The line between two anchor words form a relational axis; points corresponding to other words can then be projected onto this axis to show the extent to which they are associated with the anchor words. If E is the embedding matrix, E_{word} is the word vector for a given word, and the anchor word vectors are E_0 and E_1, then the magnitude of a word's projection onto a normalized relational axis is given by:

$$|Projection| = \frac{(E_{word} - E_0) \cdot (E_1 - E_0)}{(E_1 - E_0) \cdot (E_1 - E_0)}$$

Using the same anchor words for each monthly language model, we compute how words' meanings change over time with respect to a relational axis, even though the words are moving through $\Re^{300}$ space from month to month. With this approach, we produce a list of the words whose meanings change the most over time, and as the community's word meanings diverge from everyday meanings.

4 Results

4.1 Predicting User Trajectories with Linguistic Features

Unlike Danescu-Niculescu-Mizil et al. (2013), our language models show a steady distancing from the community over lifespan (Figures 1 and 2), and also that new users in Hacker News do

not yield a clear trend in language flexibility (excluded due to space). These two findings tell a coherent story: because there is no period of adaptation to the community's linguistic norms, users start the process of marginalization from the moment they join.

We hypothesize several reasons for this: relative to the RateBeer community, the Hacker News community is an order of magnitude larger. It may also be that users offer less "onboarding" to new members. Additionally, Hacker News may have a better-defined cultural niche for hackers and people interested in startups, such that they arrive at the community already part of the discourse community. Hacker News started as an online extension of a face-to-face community, so some of the language adaptation hypothesized by Danescu-Niculescu-Mizil et al. (2013) may have happened in the offline space, or perhaps there are lurkers who start the adaptation process prior to posting.

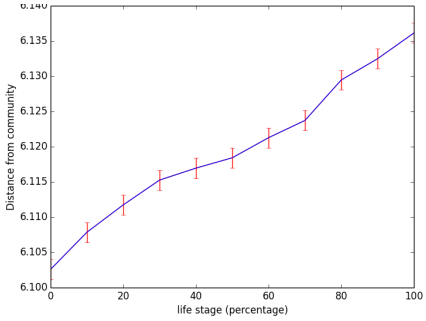

Fig. 1. Distance from community (bigram cross-entropy)

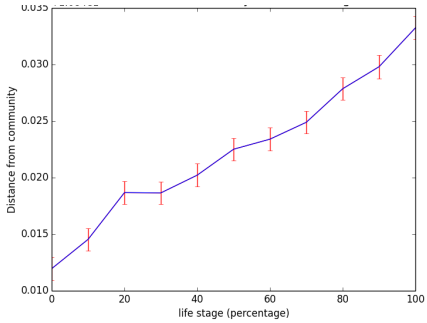

Fig. 2. Distance from community (word vector log-likelihood

Finally, Table 2 shows the results of the prediction task using various feature sets (the first two rows are from Danescu-Niculescu-Mizil et al. (2013)). Implementing the same algorithms, the Hacker News dataset achieved a much higher recall than RateBeer from the prior paper. This is at least partly due to the higher percentage of 'living' users sampled in the prior paper. More important for our research question is that none of the linguistic features derived from the language models contributed significantly to the F_1 score. This is consistent with the earlier finding that Hacker News users do not appear to have an early period of linguistic adolescence. The fact that activity is much more predictive in the Hacker News dataset may also be a factor, as there may be less signal remaining for the linguistic features to pick up.

Table 2. User trajectory prediction task

Data set	Features	Precision	Recall	F_1	Departed	Living
RateBeer	Activity	0.737	0.193	0.305	261	465
RateBeer	Activity + BigramCE			0.374	261	465
HackerNews	Activity	0.769	0.803	0.786	1977	1602
HackerNews	Activity + BigramCE	0.770	0.805	0.787	1977	1602
HackerNews	Activity + WordVectorLL	0.768	0.804	0.785	1977	1602
HackerNews	Activity + DiffLL	0.771	0.807	0.788	1977	1602
HackerNews	Activity + WordVectorLL + DiffLL	0.769	0.805	0.787	1977	1602

4.2 Modeling Word Meaning Change Over Time

Extending relational axes over time offered many insights into how word meanings shifted. Due to space limitations, Figures 3 and 4 illustrate words' change in position over time only for two relational axes, man-woman and black-white. The selected words are from the 20 words that change the most over time on the relational axis. This method can be used in future work to analyze other relations.

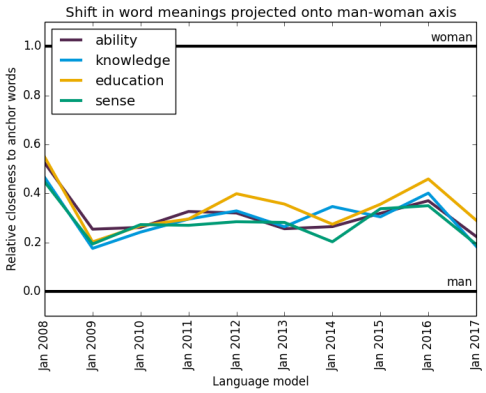

Fig. 3. Change on man-woman relational axis

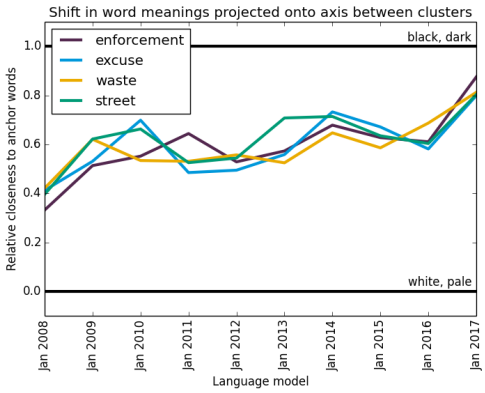

Fig. 4. Change on black-white relational axis

While we cannot draw any firm conclusions about this data here, some of the often-discussed stereotypes about women in technology are clearly visible. Margolis & Fisher (2003) and many others have shown that such stereotypes are pervasive and that they have a strong effect on womens' participation in computer science learning environments. The potential contribution of this method of modeling is to show that sexism is more than a problem of sexists; sexism is woven into the meanings of a discourse community's words. In future work, we hope to more rigorously quantify these findings and show that the orientation of users' language on relational axes of gender and other identity categories is predictive of their future participation trajectories.

5 Conclusions

This paper offers the first practical means of modeling the semantic content of participation trajectories within evolving linguistic communities. While the word2vec-based linguistic features did not improve performance on the prediction task, the model's characteristics are very similar to the bigram model, which Danescu-Niculescu-Mizil et al. (2013) found to significantly improve prediction of users' future participation trajectories. Our interpretation of why Hacker News may function differently from RateBeer is consistent with research on how learners move from peripheral to more central participation (Munter & Ma, 2014). Finally, using relational axes to analyze changes in the space of word meanings is a promising avenue for future research.

In addition to methodological improvements discussed above, future work will involve further analysis of the Hacker News community, supported by a more substantial theoretical framework. Additionally, we plan to use these methods to model in-person discourse communities.

6 References

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2015). Rand-walk: A latent variable model approach to word embeddings. arXiv preprint arXiv:1502.03520.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1) (pp. 238-247).

Boaler, J., & Greeno, J. G. (2000). Identity, agency, and knowing in mathematics worlds. Multiple perspectives on mathematics teaching and learning, 171-200.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (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 (pp. 307-318). ACM.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 187-197). Association for Computational Linguistics.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In ACL (2) (pp. 690-696).

Hull, G. A., & Greeno, J. G. (2006). Identity and agency in nonschool and school worlds. Counterpoints, 249, 77-97.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In International Conference on Machine Learning (pp. 957-966).

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge university press.

Leskovec, J., & Sosi, R. (2016). Snap: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1), 1.

Ma, J. Y., & Munter, C. (2014). The spatial production of learning opportunities in skateboard parks. Mind, Culture, and Activity, 21(3), 238-258.

Margolis, J., & Fisher, A. (2003). Unlocking the clubhouse: Women in computing. MIT press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 201702413.

Wittgenstein, L. (1953). Philosophical investigations (GEM Anscombe, trans.).

Wu, A. (2017). Gender stereotyping in academia: Evidence from Economics Job Market Rumors Forum.

Yang, J., Wei, X., Ackerman, M. S., & Adamic, L. A. (2010). Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. ICWSM, 10, 186-193.