Community-Conditioned Language Models for Community-Level Linguistic Variation

Bill Noble & Jean-Philippe Bernardy

December 7 , 2022

University of Gothenburg, Centre for Linguistic Theory and Studies in Probability (CLASP)
Presented at the NLP+CSS workshop

What is variation?

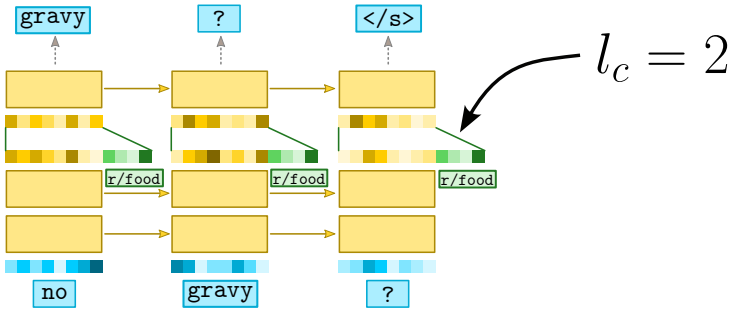
- Variation occurs different levels of the classical linguistic hierarchy.
  - phonetic variants $\rightarrow$ phonemes
  - syntactic or morphological constructions $\rightarrow$ semantic interpretation
  - words $\rightarrow$ lexical semantic interpretation
- It can happen across:
  - time, within a community (change)
  - individuals (linguistic style)
  - social or pragmatic contexts (register, polysemy, etc.)
  - macro-social categories (gender, age, etc.)
  - speech community (sociolinguistic variation)

Research questions

- Do conditional language models capture linguistic variation?
- How does variation measured in this way relate to social network connections between communities?
- Does the architecture of the language model matter?

Community-conditioned language models (CCLMs)

- 3-layer left-to-right autoregressive language models
- Transformer and LSTM
- A community embedding conditions the model by community (concatenated to the model's hidden state)
- Different layer depths for the community embedding $l_c \in \{0, 1, 2, 3\}$



Perplexity

For a language model LM, let $H_{\mathrm{LM}}(M)$ be the model's cross-entropy loss, averaged over tokens in m. We define the perplexity on a set of messages, M, to be the exponential of the model's average cross-entropy loss:

$$\mathrm{Ppl}_M = e^{\mathrm{average}_{m \in M}\, H(m)}$$

Information gain

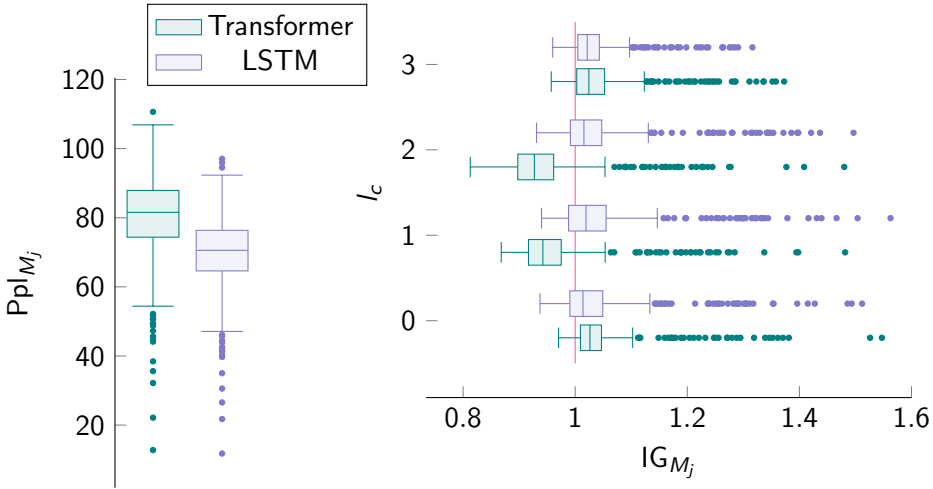
We can measure the information gain afforded by the community embedding by comparing the perplextiy of the model with a baseline model with no community embedding (LM):

$$H_{\mathrm{LM}}(m) - H_{\mathrm{CCLM}}(m)$$

For a set of messages, M, we consider the average information gain in exponential space (as a ratio of perplexities):

$$\begin{aligned} \mathsf{IG}_M &= \frac{e^{\mathsf{average}_{m \in M}(H_{\mathrm{LM}}(m))}}{e^{\mathsf{average}_{m \in M}(H_{\mathrm{CCLM}}(m))}} \\ &= e^{\mathsf{average}_{m \in M}(H_{\mathrm{LM}}(m) - H_{\mathrm{CCLM}}(m))} \end{aligned}$$

Model performance by community



Left: Perplexity of the baseline model **Right:** Information gain of the CCLM

Social embedding

- We compare the CCLM-learned community embeddings with a social network-bosed community embedding (Kumar et al., 2018)
- Trained by negative sampling optimization on the author-community co-occurance matrix
- S (social embedding) – no linguistic information, purely based on shared memebers
- L (CCLM embedding) – no community membership information, only text

Social embedding

- We compare the CCLM-learned community embeddings with a social network-bosed community embedding (Kumar et al., 2018)
- Trained by negative sampling optimization on the author-community co-occurance matrix
- S (social embedding) – no linguistic information, purely based on shared memebers
- L (CCLM embedding) – no community membership information, only text

We want to see how correlated the different CCLM embeddings are with the social embedding.

$$||L - S||_F = \sum_i (L_i - S_i)$$

Problem: Even if several dimensions of L and S are correlated, they will not coincide in the *representation* of embeddings.

Orthogonal procrustes

The Orthogonal Procrustes problem: find the *minimum* distance between L_i and S_i, for any orthogonal matrix Ω applied to L:

$$d(L, S) = \text{argmin}_{\Omega} ||\Omega L - S||_F$$

Ω gives a map from linguistic embeddings to social embeddings

Orthogonal procrustes

The Orthogonal Procrustes problem: find the *minimum* distance between L_i and S_i, for any orthogonal matrix Ω applied to L:

$$d(L, S) = \text{argmin}_{\Omega} ||\Omega L - S||_F$$

Ω gives a map from linguistic embeddings to social embeddings

Solution:

$$d(L, S) = n - \text{Tr}(\Sigma)$$

where the matrix Σ is obtained by the singular value decomposition (SVD)
$U^T \Sigma V = LS^T$.

- $d(L, S)$ ranges from 0 (perfect correlation) to n (510 in this case)

Testing for significance

To test if the $d(L, S)$ indicates a significant correlation:

- Sample random embeddings L' (Monte Carlo)
- Measure $d(L', S)$
- Observe the mean ($\mu_d = 413.39$) and standard deviation ($s_d = 2.9$)

Testing for significance

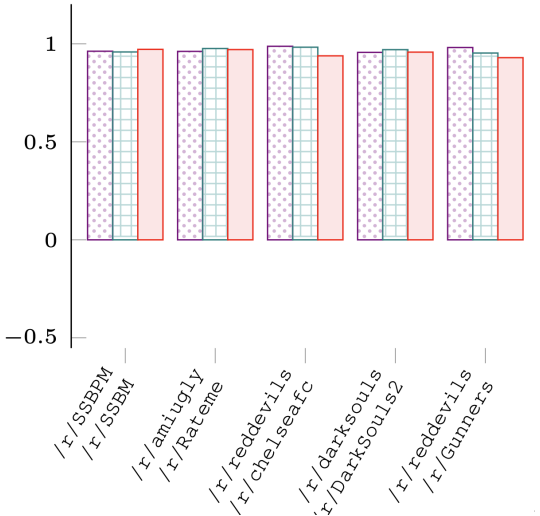
To test if the $d(L, S)$ indicates a significant correlation:

- Sample random embeddings L' (Monte Carlo)
- Measure $d(L', S)$
- Observe the mean ($\mu_d = 413.39$) and standard deviation ($s_d = 2.9$)

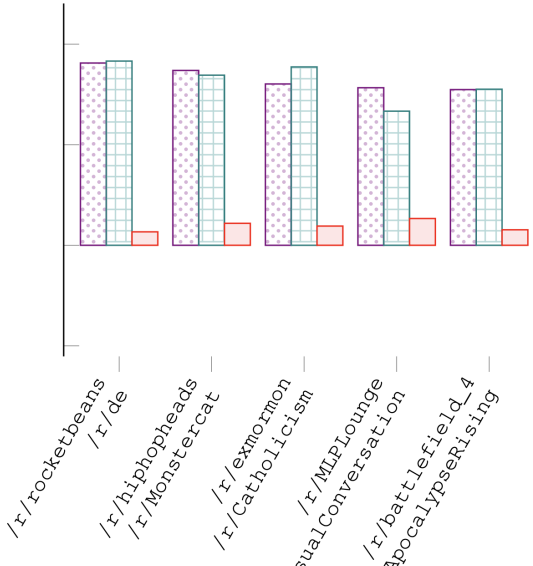
		LSTM	Transformer
l_c	0	254.06 (61.21)	239.41 (66.79)
	1	245.14 (64.29)	**232.18** (68.54)
	2	249.17 (62.90)	233.47 (68.32)
	3	**241.13** (65.67)	237.74 (66.84)

Table 1: Distance between different CCLM embeddings and the social embedding, as measured by $d(L, S)$. In parentheses is the number of standard deviations from the mean distance of our random embedding samples.

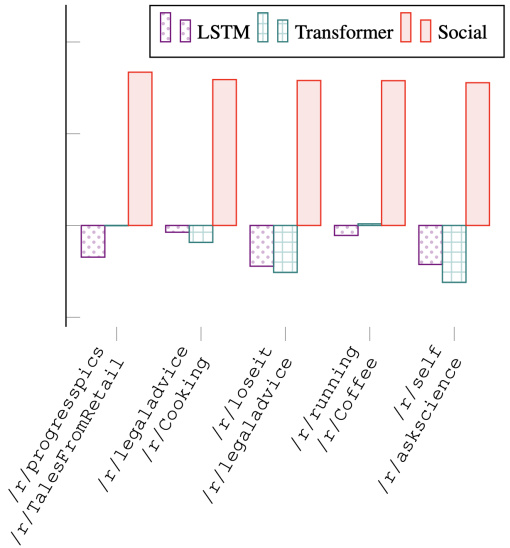
Comparing communities across embeddings



Comparing communities across embeddings



Comparing communities across embeddings



Conclusions & Future work

- Linguistic representations of communites are correlated with social representations, but also capture different information
- In general, this method can be used to study sociolinguistic variation across communities, but we need to be careful to test for significance and how we interpretet the results!

Conclusions & Future work

- Linguistic representations of communites are correlated with social representations, but also capture different information
- In general, this method can be used to study sociolinguistic variation across communities, but we need to be careful to test for significance and how we interpretet the results!
- Which members of a community rely on community context the most?
- Do communities converge/drift apart in their language use over time?
- What sorts of linguistic variation to CCLM embeddings capture?

Conclusions & Future work

- Linguistic representations of communites are correlated with social representations, but also capture different information
- In general, this method can be used to study sociolinguistic variation across communities, but we need to be careful to test for significance and how we interpretet the results!
- Which members of a community rely on community context the most?
- Do communities converge/drift apart in their language use over time?
- What sorts of linguistic variation to CCLM embeddings capture?

Thank you!

bill.noble@gu.se