Community-Conditioned Language Models for Community-Level Linguistic Variation

Bill Noble and
Jean-Philippe Bernardy

Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science

UNIVERSITY OF GOTHENBURG

linguistic variation

Variation among speakers of the "same" language can be observed at all levels of linguistic analysis: pronounciation, vocabulary, syntactic structure, norms of interaction, etc.

Linguistic communication requires common ground, which can be based on joint membership in a particular *speech community*.

The linguistic practices of speakers vary across macro-social features like gender and geography. Variation is also observed across speech communities.

community-conditioned language models (CCLMs)

Given a message $m = (w_1, ..., w_n)$, a language model estimates the probability of the message:

$$\mathrm{LM}(m) = \prod_{i \leq n} p(w_i \mid w_0, ..., w_{i-1})$$

Given a message and a community, c, a CCLM estimates the same probability, given that it came from that community:

$$\mathrm{CCLM}(m, c) = \prod_{i \leq n} p(w_i \mid w_0, ..., w_{i-1}, c)$$

information gain

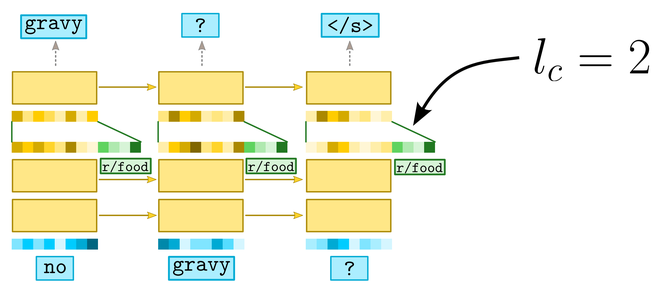
Given a LM and a CCLM, we can estimate the *information gain* resulting from knowing the community that a message came from as the difference in the entropy of the two models:

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

models

We experiment with 3-layer LSTM and Transformer auto-regressive language models.

Community information is incorporated as a learned embedding, injected by concatenating the embedding with the hidden state at layer l_c



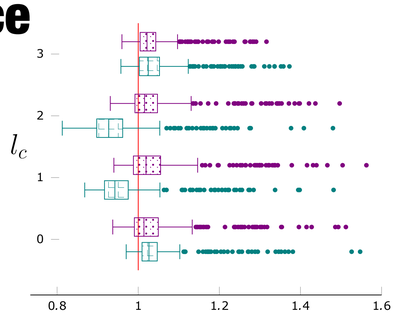
data

messages: Reddit comments
communities: 510 dfifferent sub-reddits
42 000 messages total (all from 2015)

model performance

Mean perplexity of the un-conditioned language models, stratifed by community of origin.

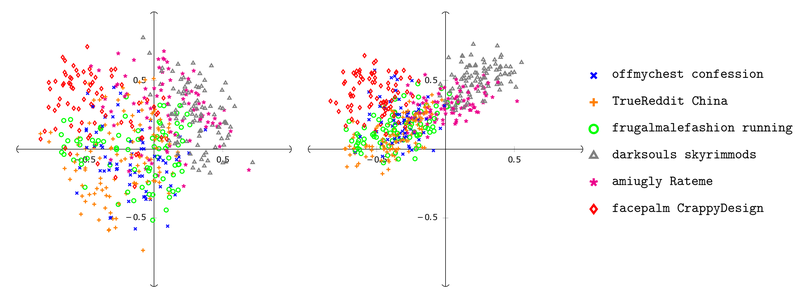
Note: Perplexity is defined here as the exponential of the entropy, averaged over a set of messages



Mean information gain (as a ratio of perplexities) for models with different community embedding depths, stratified by community.

Values above 1 indicate that the CCLM outperforms the LM, for that community. As expected, this is usually the case.

community embeddings



PCA projection of the community embedding (LSTM with $l_c = 2$ LEFT) and an embedding learned from the user-community co-occurance matrix (RIGHT). Colors are determined by k-means clustering on the communities on the RIGHT embedding.

After aligning the embeddings with orthogonal Procrustes, we see a high degree of correlation between the embeddings, suggesting that in this dataset, language use at the community level is correlated with community composition.

Please see our recorded talk for further observations on this comparison, significance testing of embedding correlations, and potential applications to social science and sociolinguistic questions.

CLASP centre for linguistic theory and studies in probability