Conditional language models for linguistic variation and change.

Bill Noble

October 1, 2020

A language model estimates the probability of a sequence by predicting the next word, given the sequence so far. A *conditional* language model takes additional context, c, into account.

$$P(w_1, ...w_n \mid c) = \prod_{i=1}^{n} P(w_i \mid w_1, ...w_{i-1}; \tag{1}$$

A common neural language modeling technique based on (1) concatenates a vector representation of the sequence $w_1, ..., w_{i-1}$ (e.g., the hidden layer of a recurrent network) with a vector representation of c. This approach has been used in various text generation domains such as image captioning (Vinyals et al., 2015) and machine translation (Kalchbrenner and Blunsom, 2013).

We propose to use conditional language models for linguistic variation and change by conditioning on speech community or time period. In preliminary work, we trained LSTM and transformer models on comments from 46 different English-language forums of the social media website Reddit, conditioning on the community of origin. We found that the two architectures had opposite preferences for how deep in the network the community representation should be injected to maximize information gain over an unconditioned model, suggesting that they may pick up on different aspects of variation.

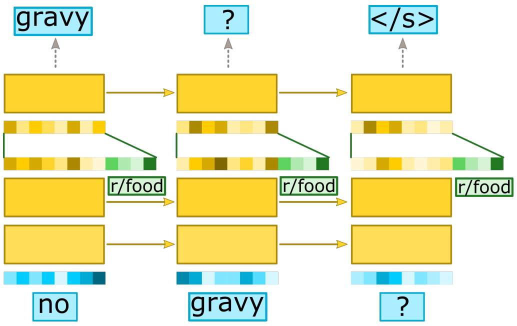

Figure 1: Three-layer RNN language model with a community embedding concatenated between the second and third layers.

In ongoing work, we use this approach to study linguistic change by conditioning on time period instead of (or in addition to) community of origin. We are interested in how the output of the conditioned language model, as well as the vector representations of community or time period and can be used to study variation and change. At this early stage, we are still trying to determine where this approach fits in with previous work in semantic change detection, and what historical linguistic questions it may help answer.

References

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, page 10, Seattle, Washington.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. *arXiv:1411.4555 [cs]*.