The convergence of language in online communities

Devin Gaffney & Zach Wehrwein

23 March 2016

Methods

The data employed in this paper consists of a complete archive of 1,659,467,343 comments posted to Reddit between October 2007 and May 2015. Each comment on Reddit is in the format of a JSON object, that, along with the actual text of the comment, contains various metadata about the comment, including the Unix time the comment was posted, the subreddit that the comment was posted in, the name of the user that posted the comment, and importantly, the id of the comment or post to which this comment was replying. Inherent to Reddit's design is a tree-like structure of interaction - all comments are replies to other comments or replies to a post, which serves functionally as the root of each tree (Cite whatever foundational reddit paper). Taken together in the context of a single subreddit, the series of tree networks of comments on the posts constitute a network of interaction, where users respond to one another across posts over time in the subreddit.

To measure the cultural specificities of any given subreddit, there must be a baseline estimate of what constitutes Reddit's general, background culture, and what is distinct between that and the culture observed in any particular subreddit. In order to achieve this, a random sample of 500,000 comments was selected from the entire sample of Reddit content. These comments were then broken up into their 1-gram components, and each word w_i of every comment was counted by frequency, then divided by the sum of words S_r in the sample of the general Reddit comment database. Then, the vector X_r represents the proportion of usage for all words on the sample of comments. This represents an approximate estimation of the general proportional language use on Reddit with a sensitivity of observing a word proportional to the sample size of the comments used to generate the vector.

Then, the same process is conducted, but restricted only to the full corpus of comments from a given subreddit - the value S_s denotes the total number of words posted in comments in a subreddit, and the vector X_s represents the proportion of word use (or $\frac{w_i}{S_s}$ for each distinct

w_i word). Then, for any given w_i, $M_{s_i} = \frac{X_{s_i}}{X_{r_i}}$ provides a valuable ratio - the proportion of observed use of a word within the context of a subreddit over the proportion of observed use of a word on Reddit generally. This new vector M_s, containing all ratios for observed words in the subreddit, constitutes a metric approximating the uniqueness of any particular terms use in the subreddit, as we would expect very common words to appear roughly equivalently both on Reddit generally and on the subreddit specifically, and would expect that specific terms referring to the topics of interest within the subreddit specifically to be used more frequently in the subreddit than Reddit generally. The lowest ratios in this list, then, would be words more commonly used on Reddit than compared to this particular subreddit.

This metric M_s, or the subreddit-wide ratios of uniqueness of word use as compared to general Reddit comments, can then be used in a temporal analysis of the subreddit. Starting with the first comment on the subreddit, all comments posted in a given day on a given subreddit are collected, and a new M_d is generated in the same way that M_s was generated, but the sample of comments sourced for that generation is restricted to the day. Then, sorted by days, these new M_d vectors can provide the raw material with which to assess the degree to which a subreddit appears to be most "subreddity" over time. Initially, the data points collected were a series of Spearman's ρ for each M_d when comparing to M_s, but the sensitivity to misordered but otherwise highly used words created a considerable amount of noise in the data, where even a few more frequent or less frequent posting of a particular set of words could make the metric behave erratically. Ultimately, for each day of data, we measure the proportion of words that appear highly in M_s that also appear highly in M_d. This, of course, introduces another problem – if there are more words posted on a given day, the likelihood of words stochastically appearing in that set increases. The proportionality of the approach, however, ensures that the words can't just occur – they have to occur very often.