COMP 498 Project Proposal: Automated Agents and Network Connectivity in Reddit Threads

Philippe Hébert
Philippe Miriello
Matthew Mongrain
Frances Zsurka

February 8, 2017

We propose to study the relationship between the connectivity of a social network and its susceptibility to incursion by automated agents. To accomplish this, we will gather data from the social network Reddit and use them to construct language models that generate imitations of comment text, and examine to what extent our automated agents, using our language models, can "pose" as real members of a particular community. We will then evaluate how successful our agents have been using Reddit's system of votes, called *karma*, and correlate the amount of karma gained to the strength of the network the agent operates in.

This project comprises several components:

1. **Data collection.** First, we will select a set of individual message boards on Reddit, called *subreddits*, and gather data from them. Reddit threads take the form of nested comment trees which descend from a single parent node, the *post*. The data we collect, via Reddit's JSON API[1], will consist of the content of individual comments, the comment's position within the comment tree, its author, and the amount of karma it received. To reduce the quantity of data we need to process as well as

[1] `https://github.com/reddit/reddit/wiki/JSON`

to obtain higher-quality (and more representative) data, we will focus our attention on posts that score above the average karma per post of the subreddit the post originates from.

2. **Network modeling.** Using the threadedness of the comment data, we will attempt to construct a graph of the social network underlying each thread. We will say that two users are connected if there is a reply chain of length at least 3 involving both participants. In other words, if a user replies to a comment, and the person to whom they replied replies to their comment, we will say the two users are connected in the context of the thread.

3. **Text generation.** Liberally inspired by Andrej Karpathy's `char-rnn`[2] language model, we will train individual language models on the data obtained from each subreddit to allow us to post comments whose content is representative of each subredddit. One language model will be constructed for each subreddit we study. The model we will use is a type of recurrent neural network called a *Long Short Term Memory network* (or LSTM). This model, which is trained at the character- rather than the word-level of a text, is capable of generating high-quality samples of text once trained. To improve the quality of our model, we will weight comments that received more karma more highly. We will then use Reddit's user API[3] to post comments generated by the model in an automated fashion and observe how much karma they receive from (presumably human) Reddit users.

4. **Analysis.** Once we have gathered enough data from each subreddit, we will attempt to correlate certain factors—including the average amount of karma obtained by our automated agent per thread, the length and total karma of resultant comment chains, the sentiment of the thread, and the variety of commenters per thread—to the connectedness of that thread.

[2] `https://github.com/karpathy/char-rnn`
[3] `https://www.reddit.com/dev/api/`

References

[1] Rafal Jozefowicz, Wojciech Zaremba, Ilya Sutskever. *An Empirical Exploration of Recurrent Network Architectures*. Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015.

[2] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. *LSTM: A Search Space Odyssey*. The Swiss AI Lab IDSIA, 2015.

[3] Felix A. Gers and Jürgen Schmidhuber. *Recurrent Nets that Time and Count*. The Swiss AI Lab IDSIA, Lugano, Switzerland. Douglas Guilbeault. *Growing Bot Security: An Ecological View of Bot Agency*. University of Pennsylvania, USA, 2016.

[4] Tim Hwang, Ian Pearce, and Max Nanis. *Socialbots: Voices from the Fronts*. Social Mediator, pp. 3845, 2012bitemnltk

[5] Steven Bird, Edward Loper and Ewan Klein, *Natural Language Processing with Python*, O'Reilly Media, Inc., 2009.

[6] Andrej Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks", `http://karpathy.github.io/2015/05/21/rnn-effectiveness`, 2015.

[7] Yoav Goldberg, "The Unreasonable Effectiveness of Character-level Language Models", `http://nbviewer.jupyter.org/gist/joavg/d76121dfde2618422139`, 2015.

[8] Peter F. Brown, "An Estimate of an Upper Bound for the Entropy of English", Computational Linguistics 18.1, 1992.

[9] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant et al., "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, p. 176182, 2011.