Project Proposal - Reddit Analyzer

CS 4701: Practicum in Artificial Intelligence

Chirag Bharadwaj (cb625) Shantanu Gore (sg937)

26 April 2017

Progress Report[1]

Scope

Given our original proposal and in light of recent feedback, we have decided that the proposed scope for this project was appropriate. In particular, given our current progress, we have found that collecting several hundreds of gigabytes (GBs) worth of Reddit comments dating as far back as December 2005 is not only feasible (albeit on external memory sources), but in fact provides a wealth of information on which the model can be trained to achieve near-realistic results. Our current progress suggests that using just bigrams is insufficient to properly generate comments that coherently convey ideas beyond simplistic ones, so the original plan of using trigrams or 4-grams (beyond which there is diminishing returns due to overfitting) as the model and a CFG as a forcing function seems within reach.

Evaluation

As a result of our reconsideration, we have decided to revise our proposal in terms of the evaluation methods employed. We now realize that our initial plan of evaluating each generated comment via its score may have been too short-sighted in terms of usefulness. Furthermore, classifying the user interactions as positive or negative experiences would be a cumbersome process. As a result, as was suggested during a round of feedback, we would like to also introduce a third metric for evaluation: human feedback. That is, we would like to provide our bot's profile page to a number of third parties and ask for their opinions. We plan on doing this as follows:

- We will give our bot a neutral name so that its purpose cannot be ascertained from its moniker.
- We will run two different experiments: one with a control group and one with an experimental group. The experimental group will be informed of the bot's purpose and be asked to evaluate its efficacy while the control group will simply be asked what they think of the particular Reddit account belonging to the bot without any further information. We hope that information will not disseminate between the groups as they evaluate.
- We could also introduce a blind or double-blind factor by randomly selecting comments and providing it to both sets of groups so that we can minimize the bias.

We plan on running this evaluation via Google Forms for at least a week so that we can gain optimal amounts of feedback. Of course, as is standard practice, we will inform the control group of our motives (if they have not ascertained it by other means) at the end of the evaluation process. We would like to use this human evaluation component in conjunction with (rather than as a replacement for) the upvote ratio and our personal evaluations of online user interaction/feedback.

[1] The original proposal is included on pages 3 and 4.

Revised Timeline

As we began to work on the project, we realized that our original proposed timeline may not have been as true to the distribution of work as we would have expected. Our current progress is as follows:

- Data has been collected.
- A basic model using bigrams has been created and generates (somewhat grammatically correct but mostly incoherent) comments. Here are some example outputs:

```
Subreddit: reddit.com

Text: That's the popular demand," but to CNN "liberal pundits."
```

```
Subreddit: reddit.com

Text: Paul's massive scale. C task and also invented to worry that I can't spin to the factory
workers by the tribe of history, it up with. You know, a ham sandwich, is your opponent is it
has confirmed that you extra $0.50 a secret executive branch predictor of the common solutions
are the legislators: Penn Teller have followed by it. For several things work, where you're
concerned with US companies without reason? Or in the question will leave you have any interest
then surely be. And you are digital, in the work, the programming or lack jobs and point is
x, y such as though still stands. The article on their best at least not alowed to be careful
to an option, but at a Mac Classic Maddox.
```

```
Subreddit: reddit.com

Text: I'm "defensive" and relatively liberal media every language support this.
```

As mentioned earlier, the generated comments **almost** make sense, but they are still missing something.

As a consequence, we have put forth the following revised timeline:

- Implement full N-gram system - 4/27 - 5/1
- Start uploading comments to Reddit and monitor feedback programatically - 5/1 - 5/5
- Implement improvements (such as CFGs, etc.) - 5/5 - 5/8
- Evaluation period - 5/8 - 5/15

We are unsure whether we will be able to do some of our proposed secondary analyses, but our primary objectives are quite feasible in the remaining time, we believe.

Original Proposal

Introduction

We intend to make a Natural Language Processing (NLP) system to interact with the Reddit online community. We hope to incorporate several connected features into this system, such as writing comments, as well as analyzing comments both temporally and topically (by subreddit).

Reddit is a massive online forum with over 16 million users. It features thousands of ”subreddits”, each focused on a different topic. These subreddits span a wide range of subjects, from `r/news`, a subreddit dedicated to world news and `r/science`, a subreddit dedicated to advancements in science, to `r/jokes` (focused on humor), and `r/aww`, a subreddit dedicated to cute content, often of cats. Each of these communties is independently moderated, which gives each one a different atmosphere (the usefulness of which will become apparent later in this document). Furthermore, as a forum that has existed for over ten years, there are millions of ”posts” (top level threads) and ”comments” (replies to posts) that have been accumulated. The combination of these factors makes Reddit, and the data that can be collected from it, a very useful dataset for data science and analysis.

Project Overview

Generating comments

We intend to focus on a few different things during our project, the biggest of which is an automated comment generator. By looking at comments that have been posted historically, we hope to create an AI that can write comments relevant to the discussion (although understanding existing comments and replying to them is outside the scope of this project), perhaps even ones that cannot be distinguished from comments written by humans. In addition, we plan to be able to post in a variety of different subreddits, using the ”lingo” of that particular subreddit.

Technical Approach

We have a few different ideas on how to best create this system, and we plan to try each of them out and build a system that uses bits and pieces from each one, depending on what works best. The first thing we intend to do is create a system that generates comments based on N-grams. N-grams, or strings of words, are a very commonly used technique in NLP to model human language. An 1-gram, an N-gram with N = 1, at its core is an utterance, such as ”Artificial” or ”Intelligence”. As N increases, it represents groups of utterances, such as ”Artificial Intelligence” or ”Ice Cream”. This technique is so popular mainly because words in a sentence are not independent of one another. Specifically, a word is very indicative of the word that comes after it, as well as before it - for example, `cream` would be much more likely to follow `ice` than, say, `backpack`.

This concept of N-grams lends itself very well to generating random sentences - we can first generate a random 1-gram (a word), based on a probability distribution. This probability distribution can be different for different comments; our current idea is to choose this distribution based on the subreddit we are commenting in, as well as the current post we’re commenting on. For example, if we were in the `r/sports` subreddit we would be much more likely to use the word `baseball` than `iPhone`, and if we were commenting on a post about the Superbowl we would be much more likely to use the word `touchdown` than `basketball`. After we choose this first word, we can use higher order N-grams to pick the next word. For example, if our first word is `cell` we can create a marginal distribution over all the 2-grams that start in `cell`, and pick a word from there, such as `phone`.

One downside of this pure N-gram approach is that it makes no regard to proper grammar. While perfect grammar is not required (or even expected, in certain subreddits), not considering grammar at all would lead to completely incomphrensible sentences. As such, we aim to find some way to incorporate grammar into our generated sentences, perhaps using methods such as Context Free Grammars (CFGs).

Comment Analysis

In addition to our main goal of generating comments, we think that it would be very interesting to perform some analysis on temporal and topical aspects of Reddit comments. Time permitting, we hope to be able to analyze how language usage on online forums changes over time, such as common phrases used. We anticipate that the overall structure of comments will remain largely the same, because language simply does not change fast enough to really have an impact over the course of 10-15 years, but we think that the specific phrases that people use are much more shortlived. We hope to prove or disprove this hypothesis by analyzing a variety of subreddits over the past years, and computing various statistics (such as N-gram usage) across them.

In addition to temporal analysis, we hope to also perform analyses on different subreddits, such as which phrases are used commonly in one but not in another, etc. One cool extension of this would be to cluster subreddits based on comments written, although we are currently unsure whether that is within the scope of this project or not.

Evaluation

We will evaluate our system by the quality of the comments it generates. We will measure comment quality by two main metrics: upvotes and user interaction. One of our objectives will be to make funny / relevant comments, which will be measured by the number of upvotes that each comment receives. This is a purely quantitive metric, and as such is optimal for analysis and guiding improvement. We will also consider qualitative metrics, such as how relevant our comments are to the discussion (which will be evaluated by us on a per-comment basis), and how other users interact with our comments (such as by replying to them, etc.).

We also aim to make controversial comments - comments that may not recieve many upvotes, but will spur discussion. Our success in doing so would be measured by the number of replies that are posted to that comment, as well as perhaps the number of upvotes those replies get. We think that this is a valid metric because from observation, comments that reply to a controversial comment well get a significant number of upvotes from people that agree with that view.

Timeline

We intend to proceed with the project with the following self imposed deadlines. Due to the nature of our project, we will require time to collect data about comments which we make, and as such aim to keep at least 2 weeks at the end solely for evaluation.

- Collect Data - 3/25 - 4/1
- Implement basic N-Gram system and research improvements - 4/1 - 4/15
- Implement improvements (such as CFGs, etc.) - 4/15 - 4/30
- Evaluation period - 4/30 - 5/15

Throughout this period, we intend to work on our secondary goal of comment analysis.