COMP-551: Applied Machine Learning
Project 1: A Multilingual Dialogue Dataset

Samuel Whaite
Student number:260656949
McGill University
Montreal, Canada
Email: samuel.whaite@mail.mcgill.ca

Salman Aman Memon
Student number: 260726294
McGill University
Montreal, Canada
Email: salman.memon@mail.mcgill.ca

Yawar Ismaeel Khalid
Student number:260742619
McGill University
Montreal, Canada
Email: yawar.khalid@mail.mcgill.ca

I. Introduction

This dataset was created using conversations from the popular internal social forum Reddit. The language of choice is French and it contain 5,031 conversations.

The code for the dataset is available on https://github.com/salman306/appliedMLfrench.git

II. Contents and Description

A. Reddit

The source of this dataset was Reddit. Reddit is a highly popular social media forum that allows users to share, discuss and vote on content they share. The site offers a wide variety of communities, or "subreddits" dedicated to discussing particular topics [1].

Although the primary language of correspondence is English, there are a few subreddits where the medium of communication is French. For this particular dataset, the following subreddits were queried:

- France: this popular subreddit has a over 160,000 followers. Most of the posts are in French with the exception of certain international topics. All manner of topics are discussed on this subreddit ranging from politics to technical science discussion.',
- Quebec: This subreddit has over 15,800 very active subscribers discussing a variety of topics related to Quebec. Having a Qubcois source of French conversations also gives us access to this popular dialect.
- ETS: The cole de technologie suprieure is a subreddit which is a part of the University of Quebec network. It specializes in discussions related to science and technology in French.
- PolyMTL: A subreddit of the Polytechnique de Montral, this thread caters to discussions about the university and Montreal.
- BiereQc: a French subreddit dedicated to discussions encompassing the Quebec beer scene.

This wide range of discussion boards gave us a broad spectrum of conversations and topics to sift through. Moreover, it also allowed us to cater to the Quebec specific dialect simply by targeting those specific subreddits.

Mining conversations from Reddit gives the dataset a number of advantages. Since we are using actual conversations between people, our dataset could help train a conversational agent that sounds more natural. Moreover, Reddit is a very rich source with highly active users making it one of the largest sources of conversations especially compared to news sites and other discussion boards. Lastly, it gave us access to French as well the popular Qubec dialect along with all the colloquil terms that people typically use while conversing with each other.

However, human to human conversations also have a few challenges when it comes to using them as training data for conversation agents [4]. The variation in style and the tendency to reference topics outside of the conversation makes it more challenging, although having a large training set can certainly help cater to some of these issues.

Our dataset has a high variation in terms of topics as well as styles of conversation. Utterances vary between long sentences and short replies as is natural with conversations. However, the highlight of the dataset has to be the large number of training examples we were able to extract. As stated in [4], a larger training set can help mitigate some of the limitations caused by high variation in convesations sourced from human to human interactions. Figure 1 shows a summary of the vital stats.

Total conversations	5031
Total utterances	15713
Total words	1674857
Average utterances per conversation	4.7
Average utterance length	21.76
Most common words	'de', 'la', 'le', 'que', 'pas'

Fig. 1. Overview of dataset

III. Data Acquisition

Having chosen the subreddits to mine conversations from, we outlined two methods of doing the actual conversation data mining:

A. Scrapy

Scrapy is a python framework designed for crawling websites and extracting structured data. It is widely used for mining text data and was the primary method of getting comments from reddit.

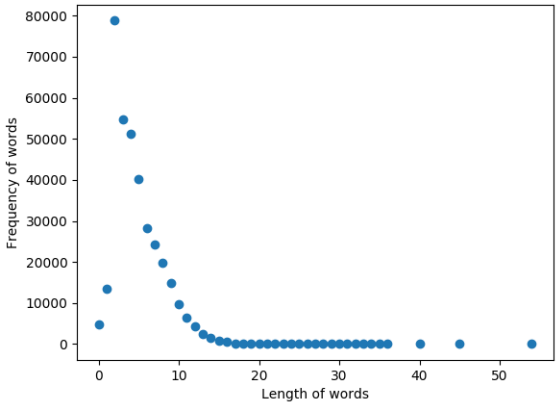

Fig. 2. Frequency of words as a function of word length

It allows users to create and instantiate Spiders, agents that crawl one or more website. More pertinent to our application, it also allows users to utilize Xpath to navigate the structure of the website as well as look at the text content [2].

B. PRAW

PRAW, or Python Reddit API Wrapper is a python package built specifically for accessing Reddit. It is a popular choice for creating bots and writing scripts to perform different actions on subreddits [3].

This package simplifies the process of extracting data from Reddit and obeys all the rules of the site regarding data mining. Moreover, it provides details about each submission including all the comments and replies in a hyarchial parent-child format as well user names, time of posting and other vital information as conveniently accessible methods.

C. Choosing Scrapy

Although PRAW is a solution specifically catering to our data source, Reddit, we chose Scrapy due to its better adaptability. It could have potentially allowed us to mine any website given the ease access to structure that Xpath provides. Moreover, it gave us all the data in a hierarchical format similar to PRAW. It has the convenience of being able to format output data directly in Xml format which was the requirement for this project. Lastly, PRAW does not deal well diacritic marks on words and replaces them with their equivalent unicode which we would have had to locate and replace from our conversations.

IV. Representation

To outline the method in which conversations were represented, we need to first outline how subreddits are structured.

Each subreddit contains multiple submissions with a title. Subscribers can then comment under the submission and vote to promote more popular comments. Users can reply to the submission itself as well as comments by other users.

We had a variety of options when it came to structuring the conversations. In our final implementation, we chose to traverse each submission and choose the ”root” comment, i.e. the comment that replies to the submission, and that top reply to the root comment to create conversations.

This structuring prevented two potential problems with other structuring methods we considered. Given the hierchial structure of the data, we considered storing entire conversations with all replies in a tree. However, extracting conversations either using preorder or postorder traversal of the tree lost the hierarchy of the comments and therefore the context of the conversation.

Another strucutures we considered but chose not to implement was to create a new conversation for each reply to a comment. However, to retain context, this would have created a lot of repetition of parent utterances with a lot of replies, potentially biasing the datasource while training a conversational agent.

V. Discussion

Given the easy access to social media sites like Reddit and Twitter as well as easy tools to extract conversations, there is no dearth of datasets with conversations from these sources.

Our dataset would be classified as a spontaneous written corpora as per [4] as the topic of the conversation is not pre-specified.

The Reddit Corpus is a similar but far more extensive dataset containing 1.7 billion comments from Reddit. However, given that English is the primary mode of communication on Reddit, this dataset would not be particularly useful creating coversational agents in other languages. Our selective use of French subreddits as well as checking that users primarily use French while posting puts our dataset in contrast to this.

The Twitter Corpus [6] also carries some similarities to our dataset given the similar nature of comments and posts on social media sites. As with Twitter, comments by Redditers in our dataset also involve use of colloquial terms as well as references to events outside of the topic of the conversation [**?**]. However, the 140 character limit imposed on Tweets makes them less useful for training formal chat agents due to the high volume of abbreviations used. With Reddit’s lack of such a restrictive limit, the conversations are more varied.

VI. Conclusion

In conclusion, our dataset provides a large collection of French conversations. The choice to use subreddits from France and Quebec creates some additional variation. Although there a number large datasets from Reddit, we could find none catering to a specific language which makes this dataset invaluable for training conversational agents in French. The high variation between conversations and the large number of training examples can potentially train a conversational agent that sounds converses more naturally than traditional machine-human conversation bots.

VII. Statement of Contributions

Samual Whaite: Implemented the Scrapy web crawler that extracted majority of the data for the dataset.

Salman Memon: Implemented the PRAW script to extract comments and conversations directly from Reddit.

Yawar Khalid: Worked on an alternate approach using podcast transcripts to add to the dataset using MATLAB which was not complemeted on time.

We hereby state that all the work presented in this report is that of the authors.

REFERENCES

[1] Loerzel, Christy. "What is Reddit and why should you care?", symantec.com, Symantec Official Blog, 11 April 2013, Web, 26 September 2017

[2] Developers, Scrapy. "Scrapy Tutorial" docs.scrapy.org, Scrapy Docs, Web, 26 September 2017

[3] Boe, Bryce et al. "Quick Start" praw.readthedocs.io, PRAW Documentation, Web, 26 September 2017

[4] Serban, Iulian Vlad, et al. "A survey of available corpora for building data-driven dialogue systems." arXiv preprint arXiv:1512.05742 (2015).

[5] Stuck_In_The_Matrix, "I have every publicly available Reddit comment for research. 1.7 billion comments @ 250 GB compressed. Any interest in this?", Reddit, https://files.pushshift.io/, Web, 26 September 2017

[6] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In North American Chapter of the Association for Computational Linguistics (NAACL 2010), 2010