COMP 551 - Project 1: French Language dialogue dataset

Ryan Knight
Bioresource Engineering
McGill University
Montreal, Canada
ryan.knight@mail.mcgill.ca
ID: 260531961

Xin Tong Wang
School of Computer Science
McGill University
Montreal, Canada
xin.t.wang@mail.mcgill.ca
ID: 260640319

Eric Zhang
School of Computer Science
McGill University
Montreal, Canada
eric.zhang2@mail.mcgill.ca
ID: 260551466

***Abstract*—The data set can be accessed at: https://github.com/ExTee/COMP551A1/ Under the file french_reddit_corpus.xml**

I. Introduction

Our dataset is a French language corpus which comprises of 7125 dialogue entries taken from Reddit.com.

II. Dataset description

A. Dataset contents

Reddit.com, a social news aggregation, web content rating, and discussion website is currently the 9th most visited website in the world, with 542 million monthly visitors. Social interaction on the website relies on threads, which a user posts, and subsequently, other users may respond to said thread by posting comments. These users may reply to each other's comments as well, thus creating a social platform akin to a discussion room.

The dataset contains 7125 dialogues and 27145 utterances from a varied set of conversation topics. Our team took advantage of the fact that Reddit classifies its threads into broad topics, known as subreddits. These subreddits may have their common interests or language. For our dataset, we looked at subreddits which pertain to French language, as well as interests associated to francophones.

Starting from a suggested list of french subreddits, we limited our search to the following 11 subreddits: r/banalites, r/besoindeparler, r/forumlibre, r/france, r/francophonie, r/geographie, r/jeuxvideo, r/le_pen, r/philosophie, r/quebec and r/rance.

B. Dataset acquisition

The dataset was acquired by initially scraping manually determined subreddits on Reddit using the third party software, redditDataExtractor [1], a Python-based web-scraper. Reddit threads were sorted by "hot" and up to 1000 threads per subreddit were extracted. The output of the application was a set of JSON files where each JSON file represented an individual thread within any of the selected subreddits.

A Python script was then created to automatically convert the JSON files into a single file that followed the specified XML format. Due to the tree-like structure of Reddit threads, multiple replies could be given to a single comment. A decision was made to view each unique path as a separate conversation from other paths. As such, a single thread could give rise to many different conversations. The Python script performed a DFS search over each threads reply tree and separated each path into its own conversation. After extracting all conversations across all threads, a basic culling step was made to reduce the number conversations in the corpus. Specifically, all threads that contained no comments or contained no text component to the root post were removed from the corpus.

Next, a passover of the corpus was performed to remove most non-text components of the text bodies and provide human readability to the raw dataset. Using regular expressions, excessive newlines, erratic characters and reddit specific formatting were removed. Additionally, urls were replaced with special strings denoting that a url had been referenced that that location, [[url]].

Finally, human readability was introduced into the corpus with the insertion of newline and tab characters.

C. Dataset representation

The corpus itself is represented by a human-readable XML format. A `<dialog>` tag envelopes the entirety of the corpus. Within the corpus is a list of `<s>` tags that denote that a single conversation is contained between the open and close tag. Finally, within each converstation contains a list of `utt` tags representing the utterances carried out in the conversation in seqential order. Each `utt` tag also carries a "uid" attribute containing an integer representing the person who performed the utterance. Each unique speaker is given their own unique integer within a conversation.

III. Discussion

Many corpora available for data-driven dialogue systems which are primarily in English use dialogue either spoken in conversation between two people transcribed into writing or scripted dialogues from movie and theater plays [2]. Other corpora use dialogues from chat services or web forums where two or more users sequentially correspond to each other in written form [2]. However, this corpus differs from

	Total	Average per Subreddit	Average per Thread	Average per Conversation	Average per Comment
Subreddits	11	-	-	-	-
Threads	858	78	-	-	-
Conversations	7125	647.73	8.37	-	-
Comments	27145	2467.73	31.87	3.81	-
Words	2460737	223703.36	2868	342.65	90.65
Users	75126	-	-	2.77	-

TABLE I
TABLE CONTAINING DESCRIPTIVE STATISTICS OF OUR HUMAN-HUMAN WRITTEN DIALOGUE DATASET

those corpora since the dialogues are taken from the web forum Reddit.com where one person begins by posting a thread and people will comment on that original post or reply to comments of others which leads threads to have tree like structure. It was decided to make every path down the tree of possibilities a conversation, so this corpus as many instances of discussion originating from the same questions or statements but can greatly differ depending on which path is chosen. Therefore, this corpus specializes in Mob-chat/Group-discussion.

When comparing the statistics of our human-human written dialogues corpus to those of human-human spoken dialogues corpora, the main differences arise in the average number of speakers, utterances per conversation and length of the utterances [2]. When comparing to other human-human written dialogues corpora which are chat-based there are more similarities but not has much as when comparing to forum-based corpora [2].

It is also worth noting that, because Reddit is an internet blog/forum, many posts contain neologisms, internet memes and statements that may appear non-sensical to people who are unaware of a particular subculture. Moreover, because there is no proofreading available on the website, orthographic and grammatical mistakes are unfiltered. Thus, we believe that our corpus provides an interesting account of human interaction online, as opposed to a perfectly grammatically correct corpus such as a movie script corpus for example.

IV. STATEMENT OF CONTRIBUTIONS

Preliminary thread scraping and cleanup was done by Xin Tong Wang. Conversion from JSON to XML and dataset description was done by Eric Zhang. Corpus analysis and comparison was done by Ryan Knight.

Eric Zhang, Ryan Knight and Xin Tong Wang declare: We hereby state that all the work presented in this report is that of the authors.

REFERENCES

[1] NSchrading, "Reddit Data Extractor", Github repository, 2016. https://github.com/NSchrading/redditDataExtractor

[2] Vlad Serban, Iulian; Lowe, Ryan; Henderson, Peter; Charlin, Laurent; Pineau, Joelle, "A Survey of Available Corpora for Building Data-Driven Dialogue Systems," CoRR. arXiv:1512.05742, Dec 2015.