COMP 551: Applied Machine Learning Project 1

UWM8 Corpus: A collection of sarcastic and satirical conversations in French collected from Reddit

Dataset URL: `https://raw.githubusercontent.com/vicrep/551-P1/master/uwm8_fre.xml`

Victor Repkow
260569786, victor.repkow@mail.mcgill.ca

Remi Martin
260814461, remi.martin@mail.mcgill.ca

Parker King-Fournier
260556983, parker.king-fournier@mail.mcgill.ca

September 28, 2017

1 Introduction

The corpus collected contains written human to human conversations in French. Each conversation in the corpus features some form of satire or sarcasm and was collected from francophone subreddits using the native Reddit application program interfaces (APIs). The title of the dataset, UWM8 is an abbreviation of the words "u wot m8?", a common meme satirizing a Northern English accent.

2 Dataset

The website www.reddit.com (referred to as Reddit) is a popular forum-style website. Users on Reddit can anonymously post links publicly on the site and can upvote for, downvote against, or comment on any post. Each user is allowed one up/downvote per post; these votes are then used to sort the relevancy of posts. When commenting on a post, one can comment as a direct reply to the post, or as a reply to a comment. This gives a tree structure for each post in which the original post, P_i, is the root and each node is a comment, C_j, having children that are replies, R_k (fig. 1).

Comment-Trees on Reddit

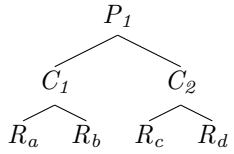

fig. 1

In order to find a sufficient amount of exclusively French conversations, it was necessary to identify multiple francophone subreddits. A subreddit is a subforum of the website Reddit with the URL form www.reddit.com/r/subreddit_name (referred to as r/subreddit_name, e.g.

r/france) focusing around a particular topic. Using a directory provided by r/france returned many francophone only subreddits [1].

The quality of posts, comments and replies were taken into consideration when choosing conversations. In order to find comments that were satirical or sarcastic, post flairs, which categorize posts (e.g. political, religious, etc.) were used, as well as searching for the symbol "/s" which is used to denote sarcasm on Reddit. Reddit's native voting system, described above also aided in the determination of conversation quality. Posts, comments or replies that had high ratios of upvotes to downvotes were considered favourable, as well as posts that had a large number of votes and comments with a large number of replies.

Reddit's native APIs were used on Comment-Trees containing posts, comments and replies that met the above criteria to extract the text which was then outputted in Markdown, preserving the relationship between individual replies, comments or posts.

The data was processed from Markdown and stripped of unnecessary features. Any text that was bold, italicized, strike-through, or otherwise stylized was removed. Titles and subtitles, if found, added little information and were also removed. Posts or comments that contained images had them removed before extracting only the text. Lastly, any comments in English or other languages were ignored along with all their replies.

While Reddit users have anonymous usernames, the username can be used to track down the user through post content and personal history. For this reason, a unique identifier was generated for each distinct Reddit username.

This dataset defines a conversation to be the path from the roof of a Comment-Tree to any leaf node. Thus it was necessary to partially flatten the data returned from the Reddit APIs into a tree which has an arbitrary node called *dialogue* whose children represent all suitable posts. Each post has one branch for each conversation in its comments, as a conversation is defined above. Comment-Trees were transformed using a variation on Depth First Search and returning the path to the root as a full conversation each time the algorithm reached a leaf node. Figure 2 shows how this transformation changes the Comment-Tree shown in Figure 1.

Transformation of Comment-Tree

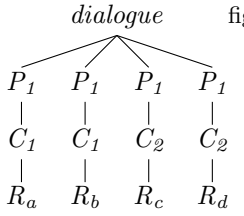

fig. 2

3 Discussion

The design and use of Reddit ensure that this dataset has specific characteristics which should be mentioned. It is possible, and no doubt probable, that many of the conversations curated have a very high similarity. Looking at Figure 2, one can see that the conversation-lists, L_i, corresponding to leaf nodes R_a and R_b are highly similar, differing only in the last reply (Figure 3).

[1] For the collection of this dataset, subreddits r/france, r/paslegorafi, r/ivrevirgule and r/jememarre were used

Conversation-Lists

$$L_1 = dialogue \rightarrow P_1 \rightarrow C_1 \rightarrow R_a$$
$$L_2 = dialogue \rightarrow P_1 \rightarrow C_1 \rightarrow R_b$$

fig. 3

This high similarity between conversations should be kept closely in mind if using this dataset to train any sort of intelligent program. Treating two similar conversations, such as L_1 and L_2, as independent could lead to a biased learning pattern which more heavily favours posts or comments that are close to the beginning of a conversation and thus appear in many highly similar conversations. As a toy example, note that an assumption of independence on Figure 2 would show that 50% of sarcastic or satirical conversations contain comment C_1, and that 100% of sarcastic or satirical conversations contain post P_1: an obvious mistake. This property is a result of Comment-Tree flattening which removes some information about the relations of posts, comments and replies. It is important to remember the dependence and relationships between posts, comments and replies when using the data. This could be achieved by weighting each message in a conversation such that messages with a low depth in the conversation tree (original posts, first comments, etc.) carry less weight, as they would appear in many separate conversations.

Table 1 shows a comparison of our dataset to other similar datasets. As many others have learned, online communities provide a wealth of written data. Multiple corpora have been curated from forums, some even from Reddit, but still have key differences from the UWM8 Corpus. Our dataset, apart from the others shown in Table 1, focuses on a class of conversation rather than a topic of the conversation. The Reddit Domestic Abuse Corpus revolves around the topic of Abuse Help, whereas sarcasm or satire may be found in any singular topic.

The availability of non-English dialogue corpora is low, and even lower for such corpora that utilize human to human written communication. This may be a results of internet communities, which have often adopted English as the language of choice due to a high number of English speakers. One such corpus is the Ubuntu-fr dataset which was collected from the Ubuntu platform's forums, mailing lists and IRC channels. While very similar to our dataset, Ubuntu-fr differs in both topic and scale. Ubuntu-fr features a larger variety of communication due to its use of platforms other than forums and again focuses on a particular topic rather than a class of conversation.

Using Table 1 to compare corpus statistics, one can see that the UWM8 dataset is comparable in number of total dialogues and average number of turns per conversation, but remains one of the smaller sets in terms of total number of words in the corpus. In addition to the information in Table 1, the average number of participants in a conversation was calculated to be 3.58 and the number of participants in the whole corpus was calculated to be 3175.

The dataset compares well to existing sets, but is not otherwise redundant. The unique choice of sarcasm and satire for the focus of this dataset distinguishes it from datasets pertaining to more concrete topics. The choice of a non-English language, albeit a popular one, is another robust feature due to the apparent lack of non-English human to human written datasets.

4 Contributions

The construction of the UWM8 Corpus was comprised of three stages. Each team member was responsible for working through one of the three components, allowing for communication between all members to ensure consistency and a variety of experience based knowledge, though all members assisted to some extent in each of the three components. The first step in making the corpus was to collect the data. Victor Repkow took responsibility for this aspect, writing scripts to query and fetch conversations through Reddit's APIs, providing curation at the higher level by filtering the data with search query parameters, as well as serializing the results into conversation trees. Remi Martin handled the flattening of aforementioned trees, text normalization, lower-level curation such as removing comments in English and image links, and serializing the results into the required XML format. After the corpus had been created in full, Parker King-Fournier summarized the dataset and methods with which is was obtained and curated in this report. This included researching existing data sets and analyzing the data of the UWM8 Corpus in order to compare it to similar preexisting corpora. We hereby state that all the work presented in this report is that of the authors.

Access to scripts Those interested in accessing or using the scripts written for this this project can visit the project's git repository at:
`https://github.com/vicrep/551-P1`.

5 References

J. N. Schrading. Analyzing domestic abuse using natural language processing on social media data. Master's thesis, Rochester Institute of Technology, 2015. http://scholarworks.rit.edu/theses.

J. Andreas, S. Rosenthal, and K. McKeown. Annotating agreement and disagreement in threaded discussion. In *LREC*, pages 818–822. Citeseer, 2012.

M. A Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *The International Conference on Language Resources and Evaluation (LREC)*, pages 812–817, 2012b

S. Rosenthal and K. McKeown. I couldnt agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, page 168, 2015.

N. Hernandez, S. Salim, and E. L. Clouet. Ubuntu-fr: a Large and Open Corpus for Supporting Multi-Modality and Online Written Conversation Studies. At *The Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. In *arXiv preprint:1512.05742*, 2015a.

Table 1: A comparison of similar datasets

Dataset	Topic	Total # of dialogues	Average # of turns	Total # of words	Description
UWM8 Corpus	Sarcasm/ Satire	8999	4.49	2.1M	Conversations from francophone subreddits containing sarcasm or satire.
Reddit Domestic Abuse Corpus (Schrading et al., 2015)*	Abuse Help	21,133	17.53	19M-103M	Reddit posts from either domestic abuse, or general chat.
Agreement in Wikipedia Talk Pages (Andreas et al., 2012)*	Unrestricted	822	2	110K	LiveJournal and Wikipedia Discussions forum threads. Agreement type and level annotated.
Internet Argument Corpus (Walker et al., 2012)*	Politics	11,000	35.45	73M	Debates about specific political or moral positions.
Agreement by Create Debaters (Rosenthal and McKeown, 2015)*	Unrestriced	10,000	2	1.4M	Create Debate forum conversations. Annotated what type of agreement (e.g. paraphrase) or disagreement.
Ubuntu-fr (Hernandez et al., 2016)	Ubuntu	27,400	4	30M	French corpus of online written conversations extracted from the Ubuntu platform's forums.

*Datasets labelled with an asterisk were taken from A Survey of Available Corpora for Building Data-Driven Dialogue Systems (Serban et al., 2017)