LÉLU: A French Dialogue Corpus from Reddit

Amir El Bawab
amir.elbawab@mail.mcgill.ca
260645260

Breandan Considine
breandan.considine@mail.mcgill.ca
260815673

Zeyad Saleh
zeyad.saleh@mail.mcgill.ca
260556530

I. Introduction

In this paper, we present LÉLU, a French dialog corpus that contains a rich collection of human-human, spontaneous written conversations, extracted from Reddit's public dataset available through Google BigQuery. Our corpus is composed of 556,621 conversations between 2,035,268 Reddit users, with 1,583,083 utterances in total. This dataset can be accessed at the following URL: http://bit.ly/2k4GKoq

II. Dataset description

In this section, we present the source of our data and discuss its contents, as well as its acquisition and representation. Additionally, we discuss the pipeline implemented to automate the data creation process.

A. Data Source

Our corpus is extracted from a database of public Reddit comments. Reddit is a social news and content aggregation website that has become a huge repository of valuable knowledge on a diverse range of topics [1]. It allows users to start discussion threads and leave comments, who engage in a wide variety of digressive conversations. We found Reddit to be particularly interesting due to the ease of access to its public dataset as well as its significant size. However, handling this large amount of data presented several challenges that we will be discussing throughout this paper.

1) Structure: In order to understand the structure of the available data, we used Google's BigQuery platform to examine the schema of the public Reddit dataset. We considered all comments submitted to the website Reddit between December 12th, 2005 and August 31st, 2017. The raw dataset has a size of 832 GB and is spread across several tables, partitioned by month. Each row corresponds to a single comment and each column represent the data associated with that comment, such as the comment body, author, subreddit, a globally unique comment ID, the comment's parent ID as well as several other properties, including a comment's reader-voted score.

2) Acquisition: After structuring the dataset, it was important to find the right query as well as a data format that could be easily parsed and processed. Our query used the most relevant attributes from the table, mainly those that best explain the data and allow for a hierarchical reconstruction of the discussion threads. This query also serves to filter the Reddit database for French discussions. This was done by filtering comments within Francophone subreddits, more specifically /r/france, /r/FrancaisCanadien, /r/truefrance, /r/paslegorafi, and /r/rance. These comments were further refined to ensure all contained some French text. Google BigQuery allows for exporting the data in two formats, CSV and JSON. The CSV format proved to be problematic due to the frequence comments which caused the CSV parser to fail. So we used JSON instead.

3) Content: The contents of the Reddit comments can be described as spontaneous written conversations. These conversations share some similarities with spoken dialogue, with less emphasis on grammatical or syntactical correctness. As a result, the raw dataset contains a large amount of typos, slang, abbreviations as well as special characters such as "&gt;" or "&amp;" that are encoding-specific and serve to format the post, or encode special characters. Additionally, comments are frequently written in a language besides French and thus should be discarded. The contents of the dataset and its format presented an interesting challenge that required extensive preprocessing before any dialogue reconstruction took place.

B. Pipeline

1) Pre-Processing: The pre-processing phase consists of applying a filter to each comment in the dataset to remove any outliers. Our goal was to make the dataset as general as possible. We filter to remove the quoted content within the comment body as well as empty comments, comments containing the text "[deleted]" or "[removed]", and comments with empty text. Furthermore, comments containing the text "I am a bot", were discarded.[1] Finally, our pre-processing phase filters out non-French comments using the Python library langdetect, with a minimum threshold of 0.8 French to non-French words. Comments below this threshold were discarded.

2) JSON to XML: The dataset that Google BigQuery provides is a flat structured JSON file. However, the attributes attached to each utterance can be used in order to reconstruct a forest of threads where each tree recursively defines a comment and its child replies. Note that each path from root to leaf represents a complete conversation. Storing every possible path through the dialog tree would result in a large number of duplicate utterances across conversations. To avoid this redundancy, we propose two candidate algorithms: longest path first (LPF) and shortest path first (SPF). See Algorithms 1 and 2 for a detailed implementation of these two approaches. We selected SPF to generate our final dataset as it maximizes the length of utterances per conversation over LPF as illustrated in Fig. 1.

C. Statistics

Along with our corpus, we provide a conversion tool that converts conversational data represented in JSON format to

[1] This heuristic not guarantee removal of all bot-generated comments.

Algorithm 1 Longest Path First

```
function LPF($forest$)
    $queue$
    $conversations$
    for each $rootNode$ in $forest$ do
        push $rootNode$ to $queue$
    end for
    while $queue$ is not empty do
        $top \leftarrow$ front of $queue$
        remove front of $queue$
        $path \leftarrow$ longest path from $top$
        add $path$ to $conversations$
        for each $node$ in $path$, except last do
            for each $child$ in $node$ do
                if $child$ is different from next $node$ then
                    push $child$ to $queue$
                end if
            end for
        end for
    end while
    return $conversations$
end function
```

Algorithm 2 Shortest Path First

```
function SPF($forest$)
    $queue$
    $conversations$
    for each $rootNode$ in $forest$ do
        push $rootNode$ to $queue$
    end for
    while $queue$ is not empty do
        $top \leftarrow$ front of $queue$
        remove front of $queue$
        $path \leftarrow$ shortest path from $top$
        add $path$ to $conversations$
        for each $node$ in $path$, except last do
            for each $child$ in $node$ do
                if $child$ is different from next $node$ then
                    push $child$ to $queue$
                end if
            end for
        end for
    end while
    return $conversations$
end function
```

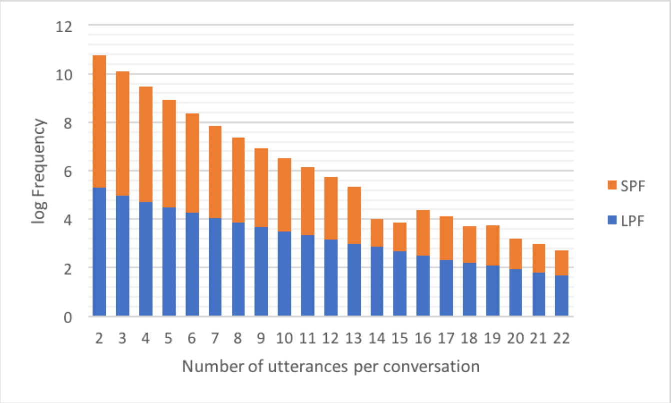

Fig. 1. Log frequency of conversations by length.

XML, as well as a command-line analysis tool, that runs an N-gram analysis on the data.

III. Discussion

This dataset is intended to provide a large and rich corpus of French human-human conversations. The main motivation is that the majority of the available datasets are in English, and only very few are available in other languages. However, there exists a number of interesting French datasets that are publicly available. The OTG and ECOLE_MASSY datasets are part of the "Parole Publique" project that aims to provide a large corpus of spoken French dialogue [2]. The project aims to gather various types of dialogues through human-human, Wizard of Oz, man-machine interactions, and is primarily intended for research on man-machine communication. It is based on recorded conversations in different settings that are restricted to the topic of tourism. Media is another French corpus that aims to define a protocol for the evaluation of speech understanding modules for dialog systems [3]. It is based on real spoken dialogues related to hotel reservation and tourist information. These conversations are also recorded, transcribed and semantically annotated. The DECODA corpus is a a call-center, human-human, spoken conversation corpus [4]. It aims to propose robust speech data mining tools in the framework of call-center monitoring and evaluation. It is based on the recorded conversations between operators and customers provided by the call-center of the Paris transport authority (RATP).

These datasets are extracted from the existing French corpora and their methodologies share multiple similarities, resulting in a corpora that contains fewer examples as well as significant grammatical and syntactical differences from Reddit vernacular [5].

The dataset we are presenting was built with the missing key features of the existing French corpora in mind. It is based on the spontaneous written conversations extracted from Reddit's posts and comments. It is based on human-human turn-taking which, in contrast to human-machine interactions, is very rich as it reflects natural dialogue interactions [5]. The range of the topics that can be found in this corpus is very wide, and can vary from gaming recommendations to discussions about French politics. These can be further filtered by topic for specific applications if needed. However, the conversations may depend on external unobserved events such as images, links or videos, which cannot be expressed in text and thus can make data-driven learning more difficult [5].

IV. Statement of Contributions

Amir worked on the Python module that serializes data to XML, conversations and dialogues and converts them to an XML file as well as the bash file that produces the Ngrams. Amir also built the C++ module that reconstructs the tree representing the hierarchical threads from the JSON file, as well as building Linux containers and Bash scripts. Breandan worked on dataset collection, the original Python script that built the dialog tree and filtering the JSON file. Amir and

Breandan worked on the processing pipeline and tree building algorithms together. Zeyad worked on the pre-processing filter for the CSV file, he also did and extensive research to find the available datasets and their format, and wrote the report. We hereby state that all the work presented in this report is that of the authors.

All our source code and tools for generating the dataset may be found at the following URL: https://github.com/amirbawab/corpus-tools

V. Conclusion

In this paper, we have presented a French human-human dialogue corpus extracted from the Reddit's public dataset available on Google BigQuery. The data is based on human-human spontaneous written conversations that reflect natural dialogue interactions between Reddit users. We acquired the data through Google BigQuery in CSV and JSON formats and found the JSON more amenable to deserialization. Since the data presented multiple sources of inconsistency, we applied a pre-processing filter to remove the most common outliers. Then we reconstructed the hierarchical discussion tree and extracted conversations from the comments, while collecting statistical data. Along with our corpus, we provide the tools and scripts that collect data, convert JSON files with a specific format to XML, as well as command line tools to view relevant statistics about the data.

References

[1] Weninger, Tim, Xihao Avi Zhu, and Jiawei Han. "An exploration of discussion threads in social news sites: A case study of the reddit community." In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 579-583. ACM, 2013.

[2] Nicolas, Pascale, Sabine Letellier-Zarshenas, Igor Schadle, Jean-Yves Antoine, and Jean Caelen. "Towards a large corpus of spoken dialogue in French that will be freely available: the" Parole Publique" project and its first realisations." In LREC. 2002.

[3] Bonneau-Maynard, Hlene, Matthieu Quignard, and Alexandre Denis. "MEDIA: a semantically annotated corpus of task oriented dialogs in French." Language Resources and Evaluation 43, no. 4 (2009): 329.

[4] Bechet, Frederic, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. "DECODA: a call-centre human-human spoken conversation corpus." In LREC, pp. 1343-1347. 2012.

[5] Serban, Iulian Vlad, Ryan Lowe, Laurent Charlin, and Joelle Pineau. "A survey of available corpora for building data-driven dialogue systems." arXiv preprint arXiv:1512.05742 (2015).