COMP-551: Applied Machine Learning

Project #1: A multilingual dialogue dataset

Rajveer Gandhi
rajveer.gandhi@mail.mcgill.ca
260719747

Sunyam Bagga
sunyam.bagga@mail.mcgill.ca
260777459

Yanis Hattab
yanis.hattab@mail.mcgill.ca
260535922

Dataset URL: https://drive.google.com/open?id=0B1ItK6JlQ6ImRXAzMm1jSU9aOTA

I. Introduction

In order to provide conversational training data in other languages than English we propose parsing openly available theatre plays in French. For this purpose, we will be curating dialog datasets in French, obtained by crawling through websites that aggregate openly available theatre works in a consistent and parseable format. In addition, we will parse sample interviews, released by authors through free sources on the web as well as language tutorials that feature conversations in French.

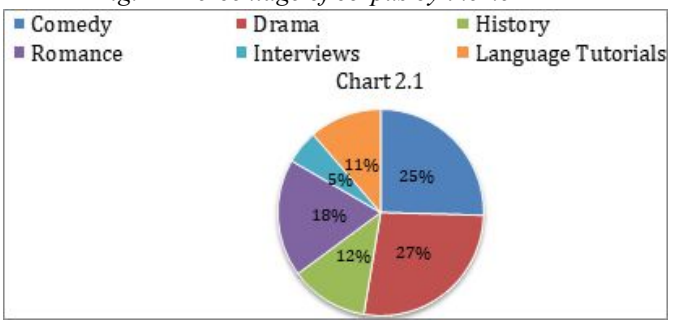
II. Dataset description

The corpus we built consists of utterances from a large number of both fictional and real conversations that are extracted from original theatre plays and sample interview scripts found on French websites. Theatre plays were chosen because they provide a source of high quality sophisticated speech about varied themes from varied time periods. Our corpus contains 5,211 conversations which include a total of 50,091 utterances averaging at about 10 utterances per conversation. (Table 2.1)

The plays have been obtained from multiple authors, still alive or from prior centuries whose plays have been released for free on the two websites we picked; *Comediatheque* [4] and *WikiSource* [5]. The sample interview scripts and the sample educational conversations have been obtained from French language learning websites that showcase different French accents from all around the world like Europe, Quebec and Africa. [2][3]

We have also put emphasis on targeting different genres of plays in order to provide diverse data (Chart 2.1) that can be used to train conversational agents with overall themes such as comedy, drama, romance or history.

Fig. 1 - Percentage of corpus by theme



Our approach to extract the dialogs from the websites was to implement a web crawler in Python (2.7) based on our investigation of the HTML structures of the webpages that hosted the plays. We used several Python libraries to gather this dialog dataset more easily from the diverse resources. Namely, we crawled the web to access and retrieve data from these websites using *urllib,* which provides a high-level interface for fetching data through HTTP or HTTPS links from across the World Wide Web. Once the website data was fetched using *urllib*, we used *Beautiful Soup* which is a Python library for pulling tags and markup data out of HTML and XML files as a tree for easy traversal. It provides different ways to navigate, search, and modify the parse tree. Furthermore, we used Python's *re* module (regular expressions) to clean the raw data extracted from these websites, for example- removing the parentheses. We also encountered encoding issues while parsing French accents, which we were able to solve with the *unicodedata* library that allows normalization of unicode characters.

There were some complexities involved in this task. Some webpages of the same website had different formatting of the data, so, we had to accommodate all of these differences in the script. A lot of time was spent proof checking the results of the parsing and tuning the scripts to get rid of non dialog text and irrelevant web page elements. We also had to come

Table 2.1

Name	Number of Conversations	Utterances	Average Length of Conversation	Number of Speakers
un succes de librairie	120	1258	10.48333333	17
apero tragique a beaucon les deux chateaux	96	1104	11.5	20
les copines davant et les copains dapres	117	1005	8.58974359	5
le pire village de france comedie de jean pierre martinez	118	996	8.440677966	12
breves de trottoirs	85	965	11.35294118	15
hors jeux interdits	102	931	9.12745098	9
erreur des pompes funebres en votre faveur	117	922	7.88034188	13
heritages a tous les etages	118	904	7.661016949	18
nos pires amis 2	90	891	9.9	6
miracle au couvent de sainte marie jeanne	99	835	8.434343434	15
La Tour de Nestle	37	1112	30.0540540541	16
GABRIEL	44	1059	24.0681818182	16
COSIMA	66	864	13.0909090909	16
Henry V	69	883	12.79710145	71
Macbeth	78	692	8.871794872	53
Henry VI	89	675	7.584269663	66
Le Fils naturel	41	608	14.8292682927	16

up with a logical way to break down the plays into different short yet coherent conversations. For that, we settled on splitting the play dialogs whenever there was a stage direction, which is a description of changes happening on the theatre stage. Alongside with scene changes they were displayed in italics on the web pages and could be detected programmatically. In the code for the *comediatheque* website, we could do that by just detecting <*em*> tags as a cue for starting a new conversation. On the other website stage directions were embedded in a <*i*> tag. Other post processing tasks involved replacing the names of the characters in the play with the properly generated *uids*. Combining the *Beautiful Soup* traversal functionalities with the powerful Python string processing operations, we were able to easily crawl the web, post process the plays and present it in the desired *xml* format.

III. Discussion

Our goal with this dataset was to provide a sundry and professional corpora in order to build more knowledgeable models for dialog datasets using free and publicly available data.

- *Comparison with Existing Corpora:*

In order to bring this diversity, we have focused on obtaining data from different sources such as plays, interviews, and french language tutorials. Furthermore our theatre data is eclectic and involves plays of various genres like comedy, drama, romance which exhibit varying number of utterances per dialog (See Fig 2). We also have some historic plays in the dataset by famous writers such as William Shakespeare. The advantage of taking professionally written plays is that we get high quality dialogs without grammar or spelling mistakes and a sustained high level of language. It allows training of conversational agents for professional and business settings were the familiar language that can be found in usual existing datasets taken from social media or forums might be inappropriate. The data also exhibits conversations between people often of adult age which avoids idioms and expressions particular to youth speech which is common in other corpora taken from contemporary sources. In addition, we have added conversations which involve more than 2 participants. Our plays consist of many characters which speak in different ways and enhance the variety of our corpus. We have a total of about 110 plays out of which 25 plays involve two participants and 35 plays involve three to eight participants. There are 60 plays that involve nine or more participants which can be very useful for training chatbots that aim to converse with a group of users, instead of a one-to-one conversation.

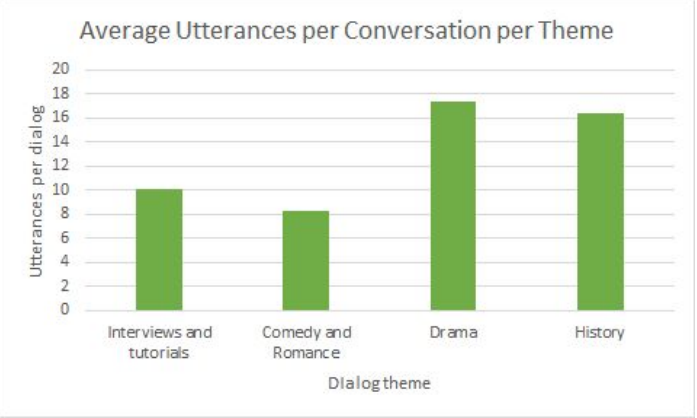
Most of the existing French dialog datasets involve recent data from sources like online forums or public social media posts. In contrast, we have mostly covered theatre plays that gave us more well-written and coherent conversations. Using theatre plays, we wanted to achieve our goal to provide datasets which were both spontaneous as well as well-written [1].

With the addition of interview dialogs, we wanted to complement our corpus with a new dimension by providing conversations which include personal or open ended questions which may help us obtain more depth while training our conversation agents and cover areas of speech not addressed in theatre plays. This information could help us model algorithms to discover common interests, developing authentic connections and fostering mutual empathy and understanding.

- *Key Characteristics:*

Harnessing Playwrights' speech writing talents allows us to get dialogs which describe the characters' circumstances, thoughts and their inner states. With these dialogs we aim to find the true character of the speaker, providing us with information on the character's cultural and social background, emotional and psychological state. Within this information, we may have different dialects with vocabulary ranging from old-historic grammar to slang dialogs. Using these dialogs we can determine what a character is feeling and attain other objectives.

Fig. 2 - Number of utterances per dialog by theme



Our non theatre data involves conversations in different social and political contexts, since some of the interviews involved are with the Senator and Member of Parliament. We also have dialogs that are samples of spoken french from Europe, Quebec, and Africa. These are under different settings such as two young women preparing for a job interview, or a courtroom scene with the lawyer questioning a witness etc.

In conclusion, our aim with this corpus has been to provide professional business-ready and yet thematically varied datasets to be used to train models with elevated language. Link to the code repository: https://github.com/Dzinator/AMLproject1

IV. Statement of Contributions

Yanis Hattab & Sunyam Bagga worked on investigating and crawling the websites to access and retrieve data as well as outputting it in the required *XML* format in the end with all the underlying implementation. Rajveer Gandhi worked on cleaning the raw data that was extracted from the parsing.
The report was written in conjunction by all 3 members.

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

References

1. I. Vlad Serban, R. Lowe, P.Henderson, L.Charlin, J.Pineau, “A survey of Available Corpora for Building Data-Driven Dialogue Systems”, 2017
2. "Apprendre Le Français C'est Facile ! - Français Avec Pierre". 2017. Français Avec Pierre. Accessed September 27 2017. https://www.francaisavecpierre.com/.
3. Stanley Aléong. 2017. "Speak French Fluently - Learning To Speak French Fluently, Accurately And Idiomatically". Fluentfrenchnow.Com. Accessed September 27 2017. http://www.fluentfrenchnow.com/.
4. "Théâtre Pièce Télécharger Textes Gratuit Comédiathèque". 2017. La Comédiathèque. Accessed September 27 2017. http://comediatheque.net/theatre/.
5. "Catégorie:Théâtre - Wikisource". 2017. Fr.Wikisource.Org. Accessed September 27 2017. https://fr.wikisource.org/wiki/Cat%C3%A9gorie:Th%C3%A9%C3%A2tre.