1. Choose your dataset: You can choose any public dataset of your choice (don't forget to cite them!). There are also a couple of useful databases that are available: Google Dataset Search and Kaggle. Explain why you chose this dataset. Furthermore, you can also look into creating your own custom dataset by scraping websites. If so, explain what kind of data you will be scraping.

Answer

It seems that there is no ready-made dataset for Shakespearean English vs modern English. Luckily this is a popular topic, especially with English students, and so websites like SparkNotes[1] have both Shakespearean and modern english versions side-by-side on their webpage. We will scrape these into two columns of data to prepare for preprocessing. The most likely candidate for scraping so far seems indeed to be SparkNotes.

It's important to note that most of this content, being from plays, has irrelevant text such as act and scene numbers, actor actions, and character names. These will not be part of the dataset and so will be filtered out.

[1] https://www.SparkNotes.com/shakespeare/

2. Methodology: Describe how you plan on approaching the project. This should be a high level overview of your plans, and this will allow us to judge the feasibility of your project. Be as thorough as you can, so we can give you critical feedback.

   (a) Data Preprocessing: Is the dataset you chose feasible? What information provided is/are the most useful? How are you planning on preprocessing the dataset to extract this information? You can take a look at these F2019 slides on data preprocessing.

   (b) Machine learning model: What do you want to predict/estimate from this dataset? Propose a machine learning model/algorithm for it, and explain your reasoning. Have you considered other alternative models? What are the pros and cons?

   (c) Evaluation Metric: BLEU score probably.

   (d) Final conceptualization: For demo purposes, we want you to be able to showcase your project!

Answers

(a) Yes, our chosen data should be feasible, as it is already neatly separated and formatted, and the translation is, at least at first glance, complete. As such, it will require minimal processing. Size-wise, we cannot tell how demanding the training will be in terms of hardware, so we will split the data into a smaller "experimental" set, and a full one. We will be able to run the small one ourselves, and should it turn out too slow, we'll turn to cloud computing. The host depends on our implementation and our needs, but for now we're thinking of Google, as they have initiatives for students.

Preprocessing will involve splitting the text into sentences pairs rather than lines. It will also involve replacing non-breaking whitespaces with spaces.

We might have to convert everything to lowercase, though we are unsure about that part yet. Moreover, some translations are imperfect or omit words. For instance, SparkNotes translated the dialog: "`What is the matter, my lord?`" to "`And what is the subject?`". This drop occurs for no reason, as 'lord' is not yet an archaic word. We will have to somehow ensure that there are no such deviations, or at least minimize them.

(b) We want to estimate what the equivalent of a given sentence would be in Shakespearean English if entered in modern English and vice versa. We will achieve our goal following a neural machine translation with attention model. The idea is to associate words from already translated sentences of Shakespeare's work with the original text using word-level tokenization.

There is no particular reason for this choice, rather, it is one of the few models capable of the given NLP task presented in MAIS202, so that's what we hope to apply.

(c) We will use the B.L.E.U. metric with brevity penalty to evaluate our model, as it is the one recommended by most references (including the MAIS202 textbook) for natural language processing.

(d) See q.3.

3. Application: We want you to integrate your model in a simple landing-page webapp. Give the general idea of your application, and the technologies you plan on using.

Answers

We intend to do a single page webpage with two text boxes, one for input and another for translated output. Also, a button or scroll-down menu allowing the user to switch between both modern and Shakespearean English. Basically something similar in appearance to the Google translation interface or to LingoJam. The application will be created with Flask.