COMP-551: Applied Machine Learning
Project #1: A multilingual dialogue dataset
Due on September 27, 11:59pm.
(Lead TAs: Chris and Koustuv. CMT help: Herke)

Background:

One of the most important open problems in AI is to build conversational agents. This was first identified by Turing as being the ultimate test of intelligence. Research in dialogue systems has progressed steadily, with many new approaches proposed recently where the dialogue strategy is learned from a large corpus of human-human conversations using neural network architectures. We have recently released a survey of available datasets for learning dialogue models (https://arxiv.org/abs/1512.05742) As you will notice, these datasets are almost exclusively in English.

The goal of this project is to curate dialogue datasets in other languages. Each group should build a corpus containing several conversations in a language of their choice (1 language per corpus). The deliverables include the corpus, and a technical description of the corpus.

General instructions:

Prepare a corpus containing a large number of conversations in a language of your choice (other than English). The corpus size is up to you, but you should aim for several thousands of conversations. Good sources of such data include online forums, discussion boards. Make sure you have the right to download and use the data.

The format of the data you will submit is the following. Every conversation should have multiple utterances, or “turns”, and each conversation should start with the symbol `<s>` and end with the symbol `</s>`. Each utterance in a conversation should be wrapped by the symbols `<utt>` and `</utt>`. The speaker should be anonymized and identified with a “uid” attribute of the `<utt>` token. For clarity, separate each individual dialog with a new line. Wrap the full document in two tags `<dialog>` and `</dialog>`, and save the file in XML format with the following naming convention: `<groupname>_<language>.xml`, where `<language>` is the ISO 639-2 code for the language for your dataset (see http://www.loc.gov/standards/iso639-2/php/code_list.php for reference).

```
samplegroup_eng.xml

<dialog>
<s><utt uid=”1”>Hey, how are you?</utt><utt uid=”2”>I’m fine thank you!</utt><utt
uid=”1”>Nice!</utt></s>
<s><utt uid=”1”>Who’s around for lunch?</utt><utt uid=”2”>Me!</utt><utt uid=”3”>Me
too!</utt></s>
</dialog>
```

Team organization: The project must be completed in a group of (exactly) 3 students. You will be asked to change teams for each project. Please plan accordingly: If you want to do the final project with your best friend, don’t work with them for this first project! You can use the class discussion board on *myCourses* to find team members. Anyone auditing the course is welcome to participate in the submission and/or review process, however you should not work with people who are taking the course for credit, to avoid mismatched expectations.

Submission requirements (1 submission per team, not per individual):

- You must **submit the dialogue corpus** created for the project. Rather than upload the code, include a URL in the header of your technical report.
- You must **submit a written report** describing your dataset, how it was acquired, and key characteristics. The report should respect the following structure:
  - Project title. (Do not include a cover page.)
  - List of team members, including their full name, email and student number.
  - URL where your dataset can be accessed.
  - Introduction: 1-2 sentences giving an overview of your dataset.
  - Dataset description: Describe what it contains, how it was acquired, how it is represented. Don't hesitate to include graphs and statistics characterizing the data (e.g. number of words, number of turns per conversation, etc.)
  - Discussion: How does your dataset compare to existing dataset? What are some of the key characteristics of your dataset? You can use the survey listed above as the primary reference for this, though you are encouraged to support your discussion with other sources.
  - Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project. At the end of the Statement of Contributions, add the following statement: "We hereby state that all the work presented in this report is that of the authors." Make sure this statement is truthful!
  - References (optional). Use appropriate referencing style throughout the report, with the list of references given at the end of the report. References are optional, but should appear if you used any additional data, or adopt methods for feature encoding that were not presented in class, etc.

  The main text of the report should not exceed 4 pages. References and appendix can be in excess of the 4 pages. The format should be double-column, 10pt font, min. 1" margins. You can use the standard IEEE conference format, e.g. *ewh.ieee.org/soc/dei/ceidp/docs/CEIDPFormat.doc*.

Evaluation criteria:

Marks will be attributed based on: 40% for overall quality of the corpus curated, and 50% for the written report. Both these components will be assessed per team, not per individual (including late penalties). The remaining 10% is attributed following participation in the peer-review process (i.e. for assessing reports of other groups). This is assessed individually. The code will not be marked, but may be used to validate the other components.

For the dataset, the evaluation criteria include:

- Size of the data (15%)
- How interesting & useful it is for training dialogue systems, how distinct from existing datasets (15%)
- Respect of the prescribed file format. (10%)

For the written report, the evaluation criteria include:

- Clarity of description, plots and figures (don't forget captions, axes labels, etc.) (30%)
- Overall organization and writing (don't forget to spell-check!). (20%)

For the peer-review, the instructions and evaluation criteria will be given in class (and included in slides) later; this is not due on September 27.

Submission instructions:

We will be using an online conference management system to coordinate submission of project files and peer-reviews: https://cmt3.research.microsoft.com/APPLIEDML2017 (The site is now open.)
You should create an account (one per person) on this site before September 27. You will use your account both as an "author" and as "PC members" (=peer reviewer) throughout the course. Make sure to use the same account for all your activities and submissions. YOU MUST USE YOUR MCGILL EMAIL TO CREATE YOUR ACCOUNT.

When submitting your report, select "Create new submission". Create one submission per group (any of the authors can do this), and link your team members as co-authors. The written report should be submitted as the "Submission file". Only acceptable file format for the report is *.pdf*. Skip the page on "Edit Conflicts of Interest".

You can revise your submission anytime up to the deadline; do not create more than one submission. CMT uses pacific time as the default, therefore the deadline has been set to 9pm pacific time = midnight eastern time. Once the deadline expires, you will not be able to submit files. If you are submitting the project late (subject to automatic 30% penalty), send all files by email to the course instructor.

Final remarks:

As specified in the syllabus, you are expected to display **initiative, creativity, scientific rigour, critical thinking, and good communication skills**. You don't need to restrict yourself to the requirements listed above – feel free to go beyond, and explore further. It is not expected that all team members will contribute equally to all components. However every team member should make integral contributions to the project.

You can discuss methods and technical issues with members of other teams, but you cannot share any code or data with other teams. Any team found to cheat (e.g. use external information, use resources without proper references) on either the code, predictions or written report will receive a score of 0 for all components of the project.