To Be Equal or Not To Be Equal

That is the question

Dànae Canillas - danae.canillas@est.fib.upc.edu
Xavier Rubiés - xavier.rubies@est.fib.upc.edu

April 2020

1 Project Description

1.1 Objective

Quora is one of the world's most popular Q&A forums, where questions from many different topics are solved. Being that said, it is perfectly normal that multiple users post (almost) the same question without knowing it already exists on the website. So preventing that will be our task. This tackles some problems, as having repeated questions:

- adds useless load to the website system
- can cause seekers to spend more time finding the best answer to their question
- may make writers feel they need to answer multiple versions of the same question.

So, this system could be used when a user intends to post a question on the website. Word embedding could be used to find the k most similar questions to that one and then the model could be used with those words. If it was the case the question coincides with another one, it would not be posted and maybe the user would be given its best answer or would be redirected to the question page.

As our model will be trained with Quora data, maybe it will work better with test examples extracted from Quora. However, we would like to see whether it can have a generalised use. That is why we have the intention of testing it with examples from other places. With the increasing number of online community engagement, there are a lot of sites that work in a similar way, such as other forums like Stack Overflow or enterprises customers sections.

Problem Statement

- We are tasked with predicting whether a pair of questions are duplicates or not
- We will give a word embedding system and could think about additionally giving the similarity between the two questions in terms of it

1.2 Databases

The dataset will be taken from the Quora competition at Kaggle:
`https://www.kaggle.com/c/quora-question-pairs`

Contains 400,000 pairs of questions organized in the form of 6 columns as explained:

- *id*: Row ID
- *qid*1, *qid*2: The unique ID of each question in the pair
- *Question*1, *Question*2: The actual textual contents of the questions.
- *is_duplicate*: Label is 0 for questions which are semantically different and 1 for questions which essentially would have only one answer (duplicate questions).

63% of the questions pairs are semantically non-similar and 37% are duplicate questions pairs.

1.3 Initial work plan and tasks

1. Text cleaning and normalization: Tokenization, convert all tokens to lower case, remove punctuations, special tokens, etc. Restore abbreviations (e.g. "What's" to "What is", "We're" to "We are", etc.)
2. Build a vocabulary that contains all the words from dataset
3. Generate word embeddings
4. Create different models architectures and test it to get better accuracy.The models we want to test for this dataset are:
    - Siamese Deep Network: A network that has two identical sub-networks in them, with the peculiarity that the weights are shared on both sides and a dense layer to connect the two nets. We think it is interesting to process both questions at the same time adding also a layer of attention since the length of the expressions are long.
        - LSTM with Attention
        - Bi-LSTM with Attention
        - SiameseGRU model.
    - CBOW, RNN, BERT
5. Model tuning: Testing different hyperparameter values
6. Making predictions
7. Choosing the best model with the help of the metrics obtained in the evaluation

1.4 Reference paper and baseline

Siamese Recurrent Architectures for Learning Sentence Similarity:

- https://www.researchgate.net/publication/307558687_Siamese_Recurrent_Architectures_for_Learning_Sentence_Similarity

Natural Language Understanding within the context of Question Matching:

- https://cs.brown.edu/research/pubs/theses/ugrad/2019/li.michael.pdf

Natural Language Understanding with the Quora Question Pairs Dataset:

- https://arxiv.org/pdf/1907.01041.pdf

Detecting Duplicate Questions with Deep Learning:

- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2748045.pdf

1.5 Evaluation

In this last section, we will compare the accuracy, precision, recall and F1 score of each model used. We will also plot them to give us a graphical representation of how the different models compare to each other. We will also submit the answer to the test set from Kaggle in order to see the performance we get there. The measure which is given in the competition is the log loss.