Medical Text Simplification

Team members

Team number : 13

Shanmukh Alle (201531085)
Ralla Akhil (201502200)
Sri Keshav Kothapalli (20161023)
Shivam Agrawal (20172074)

GitHub Repo : https://github.com/shanmukh98/Medical-Text-Simplification.

Description

- Normal medical text is difficult for average human to understand, It is essential for a simplified version of this medical text to exist to spread basic awareness about the advances in the subject.
- Manually simplifying all of medical research for this cause is extremely difficult as the time of people with the necessary knowledge can be better spent elsewhere.
- Medical texts are generally difficult to understand due to the frequent occurrence of complex medical jargon which generally has a simple easy to understand representation in the common language for example leukemia can replaced with blood cancer which is an easily understandable replacement phrase.
- One can see that we can search for such complex words patterns and replace them with their simplified counterparts. This requires us to write specific rules to accomplish the task. Which is not efficient and fade in performance in comparison to neural machine translation models.
- But neural machine translation models require a large amount of labelled data to work properly. Data in a specific domain is rare due to lack of resources and large public interest in the topic.
- Hence in the last part of the project we shall explore unsupervised and semi supervised models for text simplification using online blogs which in a way can be considered to be a simplified version of a corresponding complex medical report.

Dataset

- **Dataset 1 :** 167,000 sentences mapped to simpler sentences. The sentences are not medical specific.
- **Dataset 2:** This dataset has 60,000 document pairs one from wikipedia and another from simple wikipedia. The data is not medical specific.
- **Dataset 3:** Pairs of research paper tile and corresponding blog title pairs. The data is not specific.

Challenges

1. Get labelled data for supervised models and to evaluate the unsupervised models.
    a. Medical text simplification datasets are largely non existent, but a good number of text simplification datasets are available.
    b. Separate out medical sentence pairs from the general text simplification datasets using medical entity recognizers.
2. Identification of synonyms of medical terms using ULMS sets and choosing the best replacement.
3. Build a neural text simplification model and test and evaluate it on the medical data generated.
4. Fine Tuning the neural text simplification model for medical text simplification.
5. Exploring unsupervised / semi supervised models for medical text simplification.

First Deliverable

1. Separate medical specific data from the collect text simplification corpus using medical entity recognizers.
2. Simplify text using synonym replacement models and evaluate their performance.

Second Deliverable

1. Build a neural machine translation model and train it on generic simplification set and test it to simplify medical text and evaluate its performance.
2. Fine tune the model to medical text simplification.

Tools

- Cliner (Medical entity recognizer)
- Unified medical Language System (ULMS Metamap)
- Python(Programming language)
- Pytorch
- Keras + Tensorflow (Machine learning library)
- NLTK
- Scikit learn

References

- Wikipedia Datasets
- Medical Text simplification using synonym replacement
- A survey of research on text simplification
- Exploring Neural Text Simplification Models