CS 6320.002: Natural Language Processing

Fall 2019

Project Milestone 2 – 10 Points

Due 8:30am 11 Nov. 2019

Aniket Ashok – axa170068
Suraj Raghavendra Vadvadgi – srv180000
Vinit Kumar – vxk170008

1. What does the data look like? What is the input – document? sentence? word? And what is the gold standard output – class label? word embedding? real number?

Europarl corpus is a set of documents that consists of proceedings of the european parliament. The data are multiple sentences written in one of the European Languages and English in several files.
The input is going to be sentences that will be further tokenized.
The gold standard output will be translated language of the input file sentences.

2. Who collected the data? Was it you, or are you using someone else's dataset? If the latter, give the citation for the dataset.

The data was collected by a group of researchers led by Philip Koehn at University of Edinburgh for some research purposes in statistical /machine Translation and is available at — http://www.statmt.org/europarl/
It has the european parliament proceedings data from 1996 to present.

3. Where did the inputs come from? For example, the documents in the New York Times annotated corpus for summarization comes from archived NYT articles.

The data came from the website of the European Parliament Proceedings.

4. Where did the gold standard labels or annotations come from? For example, the NYT annotated corpus's gold standard summaries were written by the humans of the NYT Indexing Service for archival purposes.

There is a gold standard to assess the effectiveness of the data. They have been provided by Philip Koehn and are part of the standard dataset available at europarl page.

5. How many gold standard labels or annotations are there per document? If there are multiple labels or annotations per document, what is the inter-annotator agreement?

We have 2007723 number of sentences which have their french version and english translated version. These are the true labels that are available and will be helping us out with our model.

6. How large is the dataset? How many input/output pairs?

2,007,723 - sentences.
51, 388, 643 - French Words and
50, 196, 035 - English Words
Input/Output Pairs : (I/P: 2,007,723; O/P: 2,007,723)

7. What is the train/validate/test split? How many input/output pairs are in each set, and is there a standard split for this dataset? For example, the New York Times dataset has a standard split of 90%/5%/5% based on the dates that the articles were published.

If all the 21 languages are considered.

Train - 1.93 GB
Test - 15.6 Mb

Training Dataset —— 2, 007, 723 Input/Output Pairs for all words.
Test Dataset —— 10896 pairs for test dataset for the french to english words
The model is not using validation dataset.
No this dataset doesn't have a standard split.