Project Milestone 2

Xiaojie Zhu XXZ180012

Bo Jin JXB180009

Hongzheng Wang HXW180004

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?***
   Answer:
   To achieve our goal of article polishing, we need two types of models: a model to detect errors and a model to correct the errors.

   We plan to use the The Louvain Corpus of Native English Essays (LOCNESS) and the International Corpus of Learner English as the training corpus for our models. The data is composed of sentences and articles.

   The gold standard outputs would be: the error detecting model would be maximum entropy classifiers for articles, and the meta classifier with tags as corrections such as CHANGE_, DELETE_, etc.

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.***
   Answer:
   The Louvain Corpus of Native English Essays (LOCNESS):
   Citation:
   Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) Learner English on Computer. Addison Wesley Longman : London & New York, 3-18.

   The International Corpus of Learner English was published under the collaboration with a wide range of partner universities internationally.
   Citation:
   Granger, Sylviane. International Corpus of Learner English. Version 2. 2009/01/01

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.***

Answer:

1. The Louvain Corpus of Native English Essays (LOCNESS):

The essays in LOCNESS are collected from British pupils, British university students and American university students.

2. The International Corpus of Learner English:

The International Corpus of Learner English contains argumentative essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish).

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.***

   Answer:

   1. The Louvain Corpus of Native English Essays (LOCNESS):

   The gold standard essays in LOCNESS are written by British pupils, British university students and American university students.

   2. The International Corpus of Learner English:

   The gold standard essays were written by higher intermediate to advanced learners of English from several mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish).

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

   Answer:

   In agreement with rules of English the errors were originally classified in the following way:

   – spelling,
   – syntax (grammar),
   – punctuation,
   – usage.

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

   Answer:

   1. The Louvain Corpus of Native English Essays (LOCNESS):

   British pupils’ A level essays: 60,209 words
   British university students essays: 95,695 words
   American university students’ essays: 168,400 words
   Total number of words: 324,304 words

   2. The International Corpus of Learner English:

   The International Corpus of Learner English consists of 6,085 essays and totals 3.7 million words of EFL writing from learners representing 16 mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana).

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.***
   Answer:
   For both data set, we plan to have a split of 90%/5%/5% based on the dates that the articles were published.