A Machine Learning Approach to Abbreviation Word Sense Disambiguation

Homework 4 Report

Xiaohan Li

February 21, 2023

Contents

1 Literature Review

2 Data Exploration and Preprocessing

2.1 Data Exploration

Original data contains 75 different abbreviations, and each abbreviation comes with 500 samples with pre-annotations.

In order to pick an abbreviation that could yield the most meaningful result, the following steps were taken:

1. Get number of senses for each abbreviation.

   One abbreviation with only 1 sense is dropped.

2. Get standard deviation of the occurrence of each sense.

3. The results are ranked by standard deviation from low to high, and then by number of senses from high to low.

The final abbreviation chosen is `"CVA"`, which has 2 senses and the standard deviation of the occurrence of each sense is 39.6, resulting in a relatively even distribution of the senses.

2.2 Data Preprocessing

After careful investigation of the data. The following preprocessing steps were taken: