A Machine Learning Approach to Abbreviation Word Sense Disambiguation

Homework 4 Report

Xiaohan Li

February 25, 2023

Contents

1 Literature Review 1

2 Data Exploration and Preprocessing 1

2.1 Data Exploration 1

2.2 Data Preprocessing 2

3 Methods 3

3.1 Feature Extraction 3

3.2 Feature Set Selection 3

3.2.1 Feature Set Exploration 4

3.2.2 Best Performing Feature Set 5

3.3 Model Selection 5

4 Error Analysis 5

4.1 Classification Statistics 5

4.2 Error Analysis 6

5 Results and Discussions 7

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 rational is to pick the abbreviation with the most even distribution of senses, and the most number of senses. 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:

1. For each block of sample text provided, the text block are broken down into list of sentences using `'.'` as the delimiter. Sentences are then broken down into list of words using `' '` as the delimiter.

2. Special characters are removed from the sentences. e.g. `'('`, `')'`. If a word contains special characters, it is cut into 2 words, e.g. `'metacarpophalangeal(mp)'` would be cut into `'metacarpophalangeal'` and `'mp'`.

3. sentences that do not contain the abbreviation are removed.

4. If a text block contains 2 or more sentences that contain the abbreviation, the text block is split into corresponding number of text blocks, each containing only 1 sentence that contains the abbreviation.

5. De-identified date and time in different formats are unifed into `'_%#DATE#%_'`, similarly for `'_%#ZIP#%_'`.

6. After manual inspection, all 4 digits numbers in the data set are indeed year numbers, and are unifed into `'_%#DATE#%_'`.

A total number 518 samples are generated after the preprocessing steps.

3 Methods

3.1 Feature Extraction

Features explored in this project is based on *n-gram* where *n-1* is the window size of the *n-gram*. Features are generated by combining the window size and the following feature types:

1. word only e.g. `'[word1, word2...]'`.
2. word with direction e.g. `'[l-word1, l-word2..., r-word1, r-word2...]'`.
3. word with direction and distance to abbreviation. e.g. `'[l-n-word1, l-(n-1)-word2..., r-1-word1, r-2-word2...]'`.

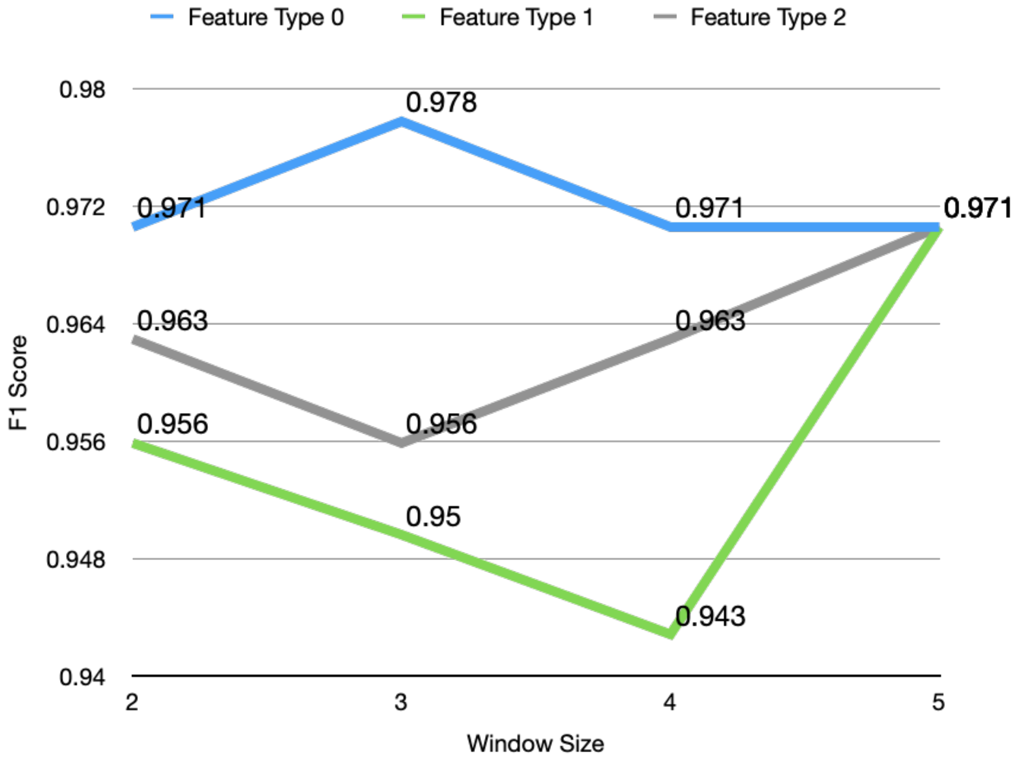
$n \in (2, 3, 4, 5)$ as window sized are used in this project. Resulting a total number of $4 \times 3 = 12$ types of features. The final feature generated is all words combined with their respective frequencies as values. The labels are senses mapped into integers.

3.2 Feature Set Selection

The generated features are fed into a random forest classifier with hyperparameter tuning by grid search to find the best combination of features. The results are ranked by the F1 score of the classifier.

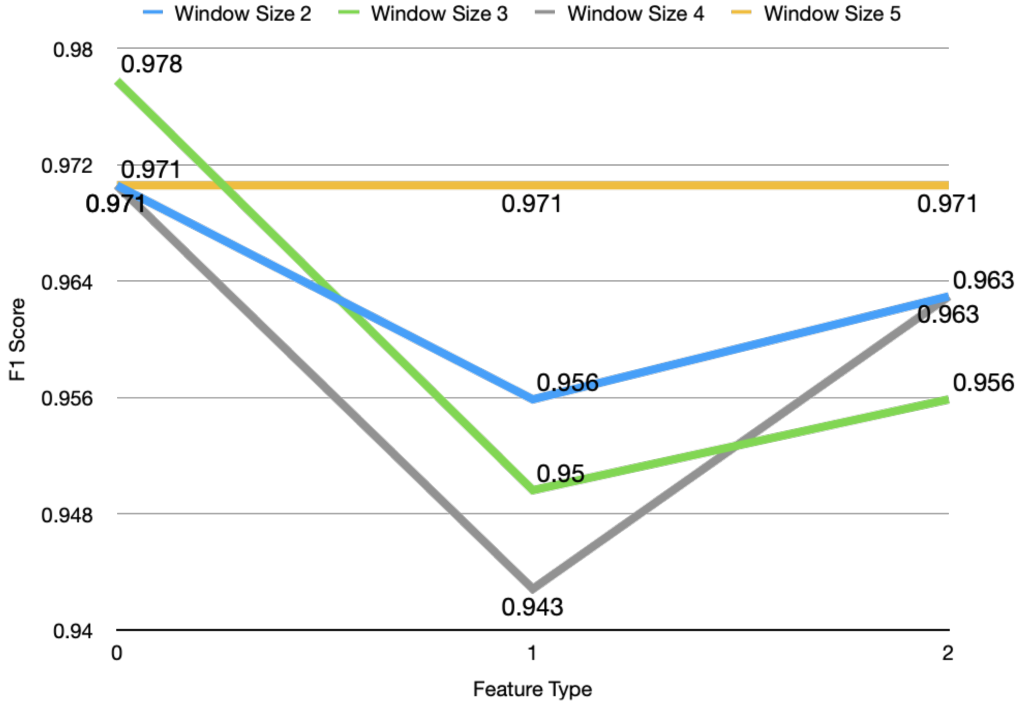
3.2.1 Feature Set Exploration

The following graphs shows the impact of the window size given the feature type:



Impact of Window Size by Feature Type

The following graphs shows the impact of the feature type given the window size:



Impact of Feature Type by Window Size

3.2.2 Best Performing Feature Set

The best performing set of featrues a 3-gram with basic word features, with:

- Accuracy: 99.22 (0.94%)
- F1 score: 0.98
- ROC AUC Score: 0.97

3.3 Model Selection

Using the best performing feature set established in the previous section, the following models are explored and compared with 5 fold stratified cross validation and grid search hyperparameter tuning (20% out of bag data as test set):

Classifier	Accuracy & STD	F1 Score	ROC AUC
Random Forest	99.22% (0.88%)	0.98	0.97
Gradient Boosting	99.21% (0.94%)	0.98	0.97
Ada Boost	97.07% (1.99%)	0.98	0.97
Gaussian Naive Bayes	88.89% (3.00%)	0.93	0.92

As the table suggested, random forest classifier is the best performing model with a slight edge over gradient boosting classifier.

4 Error Analysis

4.1 Classification Statistics

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.95	0.96	37
1	0.97	0.99	0.98	67
accuracy			0.97	104
macro avg	0.97	0.97	0.97	104
weighted avg	0.97	0.97	0.97	104

The confusion matrix:

	Actual 0	Actual 1
Predicted 0	35	2
Predicted 1	1	66

Since the data set is somewhat imbalanced, the performance when predicting the minority class is expected to be lower than the majority class.

The following table shows the data ID, actual and predicted labels of the 3 misclassified samples and the original n-gram:

Data ID	Actual	Predicted	Original 3-gram
249	1	0	['No', 'CVA', 'tenderness']
42	0	1	['the', 'left', 'CVA', 'costal', 'margin']
69	0	1	['Spine', 'and', 'CVA', 'did', 'not']

4.2 Error Analysis

- ID_249: The sentence containing the abbreviation is only 3 word long including the abbreviation itself, and the sentence appears evenly distributed across different senses. Thus the model is unable to make a good Classification.

- ID_42: After examining the data and doing some preliminary research, we found that the phrase `'CVA costal margin'` is not a commonly used combination of words since the *CVA* is formed by the 12th rib and the spine while the *costal margin* is an arch formed by the medical margin of the seventh rib to the tenth rib. Thus the text does not match the usual pattern of the abbreviation well enough to be classified correctly.

- ID_69: The sentence containing the abbreviation is in a negative context, while 3 out of 4 words in the 3-gram are `'and'`, `'did'` and `'not'`. which are common words that are not specific to any sense. Thus the model is unable to make a good classification.

5 Results and Discussions