Homework 2

Shubhankar Patankar

January 31, 2020

1 Baselines

The majority class baseline presumes that all words are complex.

The word length baseline classifies words longer than a certain threshold length to be complex. I searched between the values of 1 and 10 to determine a suitable threshold length. Based on Figure 1, precision and recall track each other inversely as expected over the range of threshold values tested. The F1 metric is maximized when the threshold length is 7 (Figure 2). As a result, all words longer than 7 letters are classified as being complex.

Under the word frequency baseline, words that appear infrequently in a large corpus are classified as being complex. This involves determining the threshold past which a word is classified as being simple. I searched over a range between 1000000 and 70000000 to determine a suitable value of the threshold. Precision and recall track each other inversely (Figure 3). F1 score is maximized when the threshold frequency is 19900000. All words appearing fewer than 19900000 times are classified as being complex.

	Training			Development		
	P	R	F1	P	R	F1
Majority Class	0.433	1.000	0.604	0.418	1.000	0.590
Word Length Thresholding ($l = 7$)	0.601	0.844	0.702	0.605	0.866	0.713
Word Frequency Thresholding ($f = 19900000$)	0.566	0.816	0.668	0.557	0.844	0.671

Table 1: Classification baselines

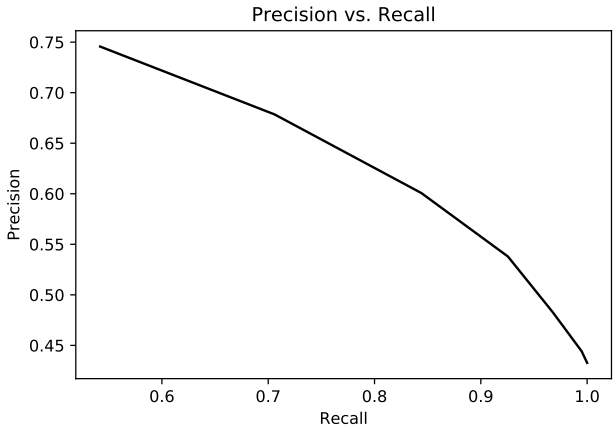

Figure 1: Precision-Recall Curve Length Thresholding

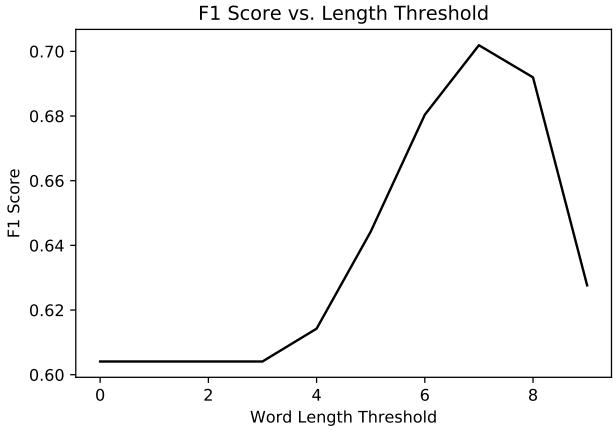

Figure 2: Length Threshold Selection

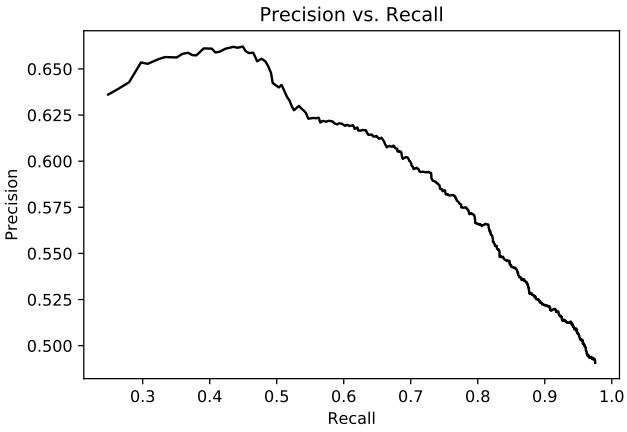

Figure 3: Precision-Recall Curve Frequency Thresholding

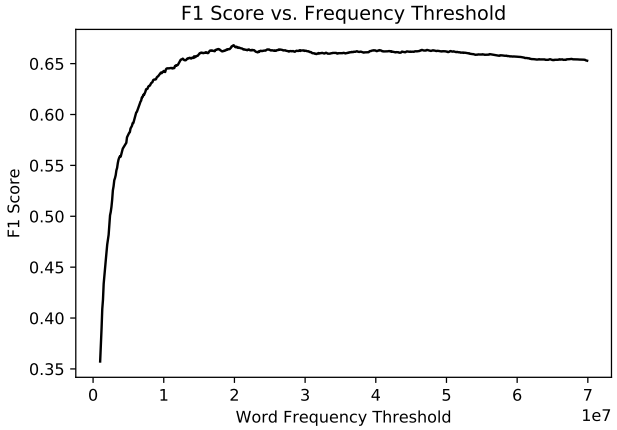

Figure 4: Frequency Threshold Selection

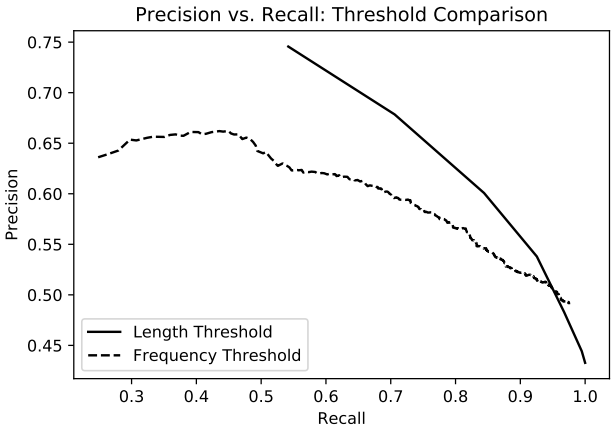

Figure 5: Threshold Comparison

Based on Figure 5, the length threshold consistently has a higher precision value than the frequency threshold. This implies that the length threshold, on average, retrieves more complex words than the frequency threshold suggesting that perhaps word length is a better predictor of complexity than word frequency. This is confirmed by the F1 scores for the two baselines with the word length baseline performing better than the word frequency baseline.

2 Two-feature Classification Models

The classification models described in Table 2 are comprised of two features only; namely word length and word frequency.

	Training			Development		
	P	R	F1	P	R	F1
Naive Bayes	0.495	0.980	0.658	0.469	0.969	0.632
Logistic Regression	0.725	0.658	0.690	0.727	0.694	0.710

Table 2: Two-feature Classification Models

Based on the F1 scores in Table 2, logistic regression outperforms naive Bayes. Note, however, that naive Bayes has a consistently higher recall than logistic regression, implying that it retrieves more complex words in general compared to logistic regression. A greater number of words that it classifies as complex are actually simple reflecting its higher sensitivity. Logistic regression, on the other hand, is correct more often when it classifies a word as being complex even though it may miss certain truly complex words. The reason for the naive Bayes classifier's underperformance relative to logistic regression may lie in its assumption of conditional independence of features. The length of a word and its frequency of usage are not uncorrelated features. A word that is longer is likely to be used less often.

3 Multi-feature Classification Models

Models in Table 3 are built using more than just the word length and word frequency as features of a word. Additionally, they comprise of the following features:

- number of syllables in the word
- number of synonyms: determined using all WordNet senses for the word
- number of senses: number of word synsets from WordNet [not used in all models tried]

Number of syllables is a complementary feature to word length, and assumes that words with more syllables are used infrequently implying complexity. A word with lots of synonyms is also likely to be a simpler word, since it is easy to replace it with another word. A complex word is likely to be domain-specific and thus harder to replace. For instance, whereas 'onomatopoeia' has no synonyms, 'bottle' has 8. The number of senses a word has is a complementary feature to the number of senses.

	Training			Development			Testing
	P	R	F1	P	R	F1	F1
Naive Bayes	0.534	0.957	0.685	0.511	0.943	0.663	0.684
Logistic Regression	0.729	0.660	0.693	0.728	0.687	0.707	0.678
Support Vector Machine	0.707	0.694	0.701	0.711	0.713	0.712	0.684
Random Forest	0.978	0.953	0.965	0.716	0.670	0.692	0.707
AdaBoost	0.740	0.735	0.737	0.730	0.775	0.752	0.751
AdaBoost + Senses	0.744	0.722	0.733	0.748	0.773	0.760	0.727
AdaBoost + Senses - Length	0.735	0.683	0.708	0.743	0.725	0.734	0.710

Table 3: Comparison of Classification Models

AdaBoost (in yellow) outperforms all other classifiers tested. AdaBoost combines multiple classifiers in a successive manner such that the downstream classifiers fit to the residuals of the upstream ones. At each stage, the incorrectly classified examples of the current classifier are up-weighted for the next weak classifier to deal with. Whereas a random forest has high training accuracy compared to AdaBoost, it clearly fails to generalize well implying that it overfits to the training data. AdaBoost is not immune to this problem since it also has lower performance metrics for the development set compared to the training set. Overfitting may be addressed by testing a variety of weak learners that comprise the AdaBoost classifier, in addition to their number.

True Positives	chairwoman, crippling, misbehavior, proponents, unraveled, nutrition, objected, embedded, unprecedented
False Positives	laughter, cheeseburgers, delivers, briefing, manipulating, handprints, spontaneous, fireworks, proliferate
True Negatives	brother, suit, prices, thanked, truth, limit, strict, success, clean, speech, creative, sci-fi
False Negatives	lecture, eerie, yips, beige, embrace, basins, racism, performance, oath, felon, proper, seize

Table 4: AdaBoost Classifer Output

A cursory glance at Table 4 reveals that the classifier tends to classify longer words as being complex. A large number of false positives are words longer than 8 letters. At the same time, a large number of false negatives are words shorter than 5 letters. This indicates that perhaps the word frequency feature is being over-shadowed by the word length feature in the training data. The classifier also

tends to incorrectly classify specialty words as being complex. This includes domain-specific words like ‘seize’, ‘oath’ and ‘felon’. These words probably appear infrequently and have few synonyms causing incorrect classifications.