Assignment 5 Report

Marc-Andre Haley • Chu Duc Thang

Decision choices

Handling head and tail

During preprocessing we are separating each sentence from the data into individual words. However we are treating the head as one word and tail as one word. We decided to treat the head and tail this way because most if not all of them are proper nouns. In this case it makes more sense to treat these nouns as one entity because part of a proper noun doesn't contain as much information as the entire name itself. For example, let's say the head or tail is 'John Smith' and we add 'John' and 'Smith' to a bag of words during training. By doing this, we're increasing the probability of a test sentence with 'John' and or 'Smith' to be in the same class as the one with 'John Smith' seen in training. We don't want this because 'John' or 'Smith' might be referring to someone different and unrelated (i.e. John Cena or Will Smith). By adding 'John Smith' to the bag of words instead, we're only increasing the probability of a sentence to be in the same class if it also contains 'John Smith'.

Stopwords

Our classifier is using a set of stop words called *stopwords_en* from *nltk*. It contains words that give us no information on a sentence's class because they're so common across all sentences. In addition to this set, we also added words like " 's ", " 'm " and other words that were left in the training that did not provide any information on a sentence's class. We also added punctuation to this set since they don't care what class has more punctuation. We also treated dates as stop words. The data contained a lot of years (ie. 2006, 1999 ...) and given our task of entity relations, dates do not provide information on the relation classes.

Smoothing

We implemented Laplace smoothing in our model so that the bag of words of every class had the same words. At first our classifier did not have smoothing and our results were the opposite of expected. Without smoothing, if a sentence to be classified had no words in the bag of words of a certain class, that class would get a probability of 0. On the other hand, the class with a bag of words with the most words in common with that sentence would get a lower score because of the sum of log probabilities, so this class was not chosen by argmax. Adding smoothing fixed

this problem by giving low probability to unseen words by a class during training, instead of 0 probability.

Handling Unknown words

Our classifier ignores words in the test data that were not seen in training. Knowing which class has the most unknown words is not useful in our case.

Accuracy

Classifier accuracy on training and test sets

	Training (with 3-fold c-v)	Test
Accuracy	0.837	0.855

Confusion Matrix - Precision and Recall

Confusion matrix with results of classifier output on test data

	characters	performer	publisher	director	precision
characters	80	7	6	10	80/(80+13)=**0.8602**
performer	7	85	3	8	85/(85+14)=**0.8586**
publisher	2	3	94	1	94/(94+12)=**0.8868**
director	4	4	3	83	83/(83+19)=**0.8137**
recall	80/(80+23) =**0.7767**	85/(85+18) =**0.8252**	94/(94+6) =**0.9400**	83/(83+11) =**0.8830**	

*(row=original_label; column=predicted_label)

Macro-ave. precision: (0.8602+0.8586+0.8868+0.8137)/4 = **0.8548**

Micro-ave. precision: (80+85+94+83)/[(80+13)+(85+14)+(94+12)+(83+19)] = 342/400 = **0.855**

Macro-ave. recall: (0.7767+0.8252+0.94+0.883)/4 = **0.8437**

Micro-ave. recall: (80+85+94+83)/[(80+23)+(85+18)+(94+6)+(83+11)] = 342/400 = **0.855**

Error Analysis

As is shown in the confusion matrix, the most misclassified case is characters misclassified as director. This is probably because characters and directors are closely related and are both related to the same industries: film, TV, theater etc. This shared context results in similar words

being used in both of these sentence classes more than others which explains some of the misclassification. The class with highest accuracy, precision and recall is the publisher class. The difference in results between this class and the others can be explained by the fact that, unlike the other classes, a publisher is not an individual but a company or organization. This makes the context around this class different and therefore the words used in this context are more unique to this class. On the other hand directors, characters and performers are all individuals. This similar context results in more words being used in common across these classes and therefore more misclassifications. This is reflected in the results above.