Comp 550 Assignment 1

Jonathan Pearce, 260672004

September 25, 2018

Problem 1.

1. "Barry doesn't have a bat."

   This sentence is an example of lexical ambiguity. The ambiguity is that either Barry has no flying animal 'bat' or has no baseball 'bat'. The word 'bat' and its multiple meanings creates this ambiguity. In order for a natural language understanding system to disambiguate the passage it would need to recognize that the website (reddit.com/r/baseball) is all about baseball, or it could parse the discussion from the URL and pick up on the high frequency of words like 'pitcher' and 'walks' and infer that the topic of discussion is baseball, and therefore the 'bat' being discussed is a baseball bat.

   Source: Reddit; Baseball subreddit
   https://www.reddit.com/r/baseball/comments/64r890/what_if_barry_bonds_had_played_without_a_bat/

2. "eating noodles with chopsticks"

   This phrase is an example of syntactic ambiguity. The ambiguity is that there are either chopsticks in the noodles as part of the food or that chopsticks are being used to eat the noodles. The uncertainty of whether 'chopsticks' attaches to 'eating' (i.e. eating with chopsticks) or 'noodles' (i.e. noodles with chopsticks) creates the ambiguity. A natural language understanding system would need to have prior knowledge that chopsticks are traditionally used as utensils to consume noodles and not something that is added to noodles, in order to disambiguate the passage.

   Source: Today.com
   https://www.today.com/food/president-obama-anthony-bourdain-table-glass-case-vietna

3. "The boys are back in town"

   This is an example of pragmatic ambiguity. The ambiguity is whether this phrase is stating a fact, providing a warning, or displaying a sense of relief. Pragmatic ambiguity is caused by a lack of context, isolating this phrase from any scenario will create ambiguity, and it can only be cleared with a sense of what is going on in the scenario. A natural language understanding system would need to know whether the person speaking is happy (sense of relief), fearful (providing a warning), or speaking with neutral tone (stating a fact).

   Source: Genius.com
   https://genius.com/Thin-lizzy-the-boys-are-back-in-town-lyrics

4. "Why the turtle is slow"

   This is an example of semantic ambiguity. The ambiguity is whether the statement is about a particular turtle that is slow or whether it is discussing why the entire species is slow. Using the word 'the' creates the ambiguity. A natural language understanding system would need to assess whether a specific turtle is being discussed or whether the species as a whole are being analysed.

   Source: Story Jumper
   https://www.storyjumper.com/book/index/44780306/-Why-the-Turtle-is-Slow-

5. "Best days of my life!"

   This is an example of phonological ambiguity. The ambiguity arises in the case of someone saying this aloud and whether the sentence should be interpreted as 'Best days of my life!' or 'Best daze of my life'. The ambiguity is created by the fact that 'daze' and 'days' are pronounced the same way. A natural language understanding system would need to find out whether the person speaking is referring to a great state of confusion ('daze') or whether they are speaking about a time interval that they greatly enjoyed ('days')

   Source: Trip Advisor
   https://www.tripadvisor.ca/ShowUserReviews-g154942-d1809046-r601298580-Surf_Sister_Surf_School-Tofino_Clayoquot_Sound_Alberni_Clayoquot_Regional_Distri.html

Problem 2i.
Schematic transducer attached at the end of document.

Note 1: I mimicked the textbook format and placed input below edges and output above edges
Note 2: In places where there are square brackets (e.g. ˆ[a,e]s#). This simply means that the input must be ˆ(a or e)s#, I used this notation just to simplify this schematic transducer.

Problem 2ii. Note for table below: I copied the textbook format and so all cells follow the format: 'output' : 'input'

Infinitive	1 Sg	2 Sg	3 Sg	1 Pl	2 Pl	3 Pl
Regular verbs						
andar	and a:o r:ϵ	and a:a r:s	and a:a r:ϵ	and a:a r:m ϵ:o ϵ:s	and a:á r:i ϵ:s	and a:a r:n
contestar	contest a:o r:ϵ	contest a:a r:s	contest a:a r:ϵ	contest a:a r:m ϵ:o ϵ:s	contest a:á r:i ϵ:s	contest a:a r:n
beber	beb e:o r:ϵ	beb e:e r:s	beb e:e r:ϵ	beb e:e r:m ϵ:o ϵ:s	beb e:é r:i ϵ:s	beb e:e r:n
correr	corr e:o r:ϵ	corr e:e r:s	corr e:e r:ϵ	corr e:e r:m ϵ:o ϵ:s	corr e:é r:i ϵ:s	corr e:e r:n
vivir	viv i:o r:ϵ	viv i:e r:s	viv i:e r:ϵ	viv i:i r:m ϵ:o ϵ:s	viv i:í r:s	viv i:e r:n
recibir	recib i:o r:ϵ	recib i:e r:s	recib i:e r:ϵ	recib i:i r:m ϵ:o ϵ:s	recib i:í r:s	recib i:e r:n
Irregular verbs						
ser	s e:o r:y	s:e e:r r:e ϵ:s	s:e e:s r:ϵ	s e:o r:m ϵ:o ϵ:s	s e:o r:i ϵ:s	s e:o r:n
haber	h a:e b:ϵ e:ϵ r:ϵ	ha b:s e:ϵ r:ϵ	ha b:ϵ e:ϵ r:ϵ	h a:e b:m e:o r:s	hab e:é r:i ϵ:s	ha b:n e:ϵ r:ϵ

Problem 2iii.
FST attached at the end of document.

Note: I mimicked the textbook format and placed input below edges and output above edges

Problem 3.

Problem and Setup: In this part of the assignment we were tasked with finding a machine learning classifier that could accurately classify segments from movie reviews as either positive or negative sentiments. We were provided three possible classifiers to experiment with and evaluate; logistic regression, support vector machine (with a linear kernel), and the Naive Bayes algorithm. We were also supplied a data set that has been previously used in Natural Language Processing research. The dataset was labelled and contained both positive and negative movie review sentiments, 5331 of each for a total data set size of 10662. Since there were equal number of positive and negative sentiments our dataset was balanced, which made model evaluation more straightforward. The first task of this problem was to read in the data and pre-process all the sentiments. The only preprocessing that was done was removing all punctuation from the segments.

Experimental Procedure: The first step in evaluating the models was to divide the data into a test and training set. I choose to divide the data in the following way, 80% of the data went into the training set, and 20% into the test set. The training data was used to find the optimal parameters for each of the 3 classifiers. In order to find the best parameter settings for a classifier, a 5-fold cross validation technique was utilized and the evaluation metric was the average accuracy of the classifier across all 5 folds. Thus we were able to directly compare all the different parameter settings and uncover the best combination of parameters for each classifier.

This process was done independently for each of the three types of classifiers. Once we found the optimal parameter settings for each classifier we proceeded to evaluate which classifier was best suited for this classification task of distinguishing positive and negative sentiments from movie reviews. First, all 3 optimal models were re-trained on the entire training set. Next, the test set was used to evaluate the 3 optimal classifiers with data they had never seen before. The classifier with the best accuracy was chosen since it appeared to be the most successful at generalizing to unseen data.

Parameter Settings

Stop Words Allowed	$\{true, false\}$
N-grams	$\{unigrams, unigrams + bigrams\}$
Min Frequency	$\{0.0, 2\}$
Max Frequency	$\{0.5, 1.0\}$

Note: for min frequency the parameter 2 is equivalent to "ignore terms that appear in less than 2 segments". All other parameters in min/max frequency are true frequency values.

Results and Conclusions: After training the three machine learning classifiers with the 16 different variations of parameters and evaluating each with 5-fold cross validation. These were the parameter settings that each classifier had the highest average accuracy with:

Classifier	Parameters:$\{stopwordsallowed, n-grams, minfreq., maxfreq.\}$
Support Vector Machine	$\{true, unigrams + bigrams, 0.0, 1.0\}$
Logistic Regression	$\{true, unigrams + bigrams, 2, 0.5\}$
Naive Bayes	$\{true, unigrams + bigrams, 0.0, 0.5\}$

With these 3 final models I re-trained all of them on the entire training set and then predicted the class (positive or negative) of each sentiment in the unused test set. I also ran a random classifier as a benchmark. The results are below.

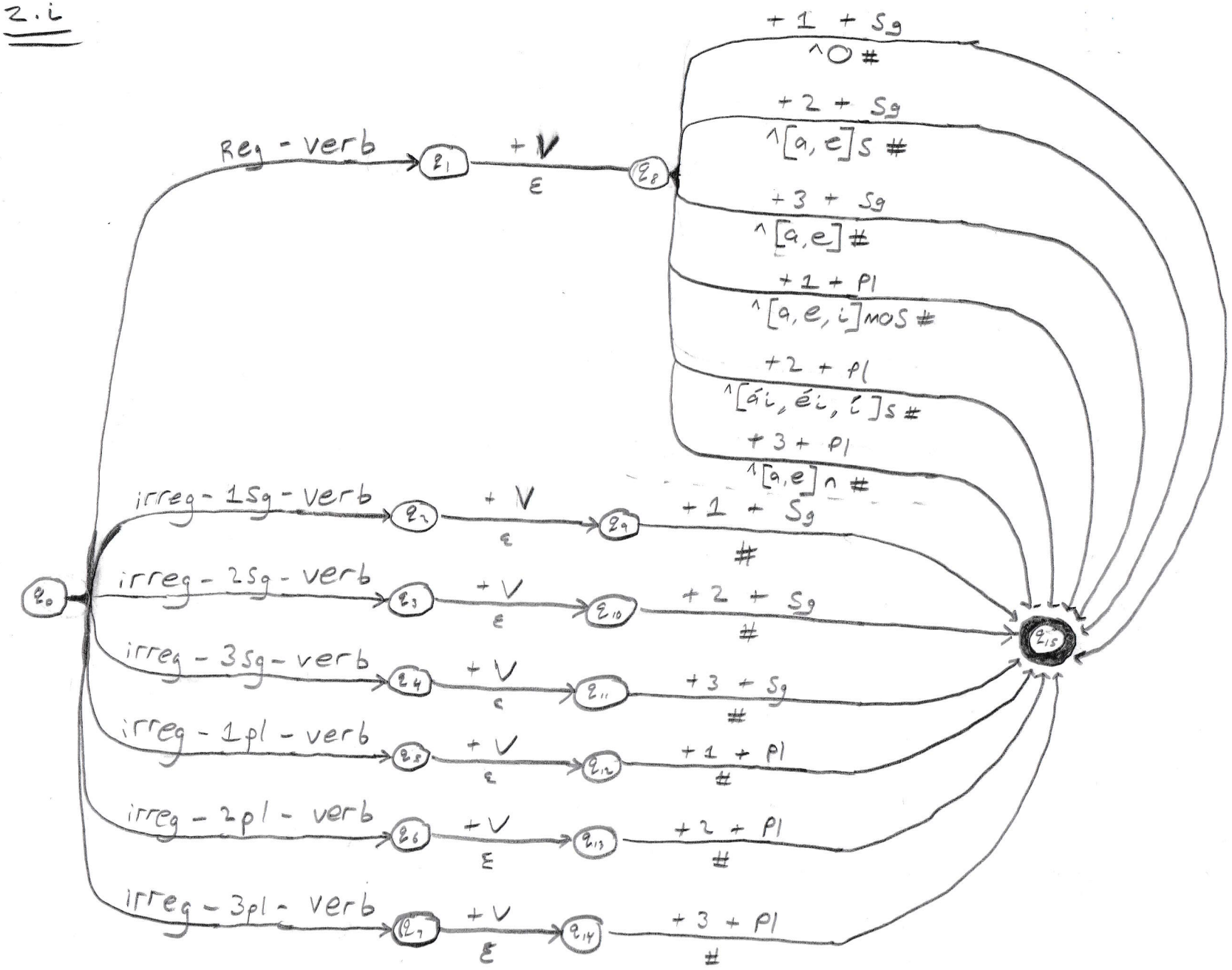
Classifier	Accuracy on Test Set
Random Classifier	0.5188
Support Vector Machine	0.7744
Logistic Regression	0.7767
Naive Bayes	0.7885

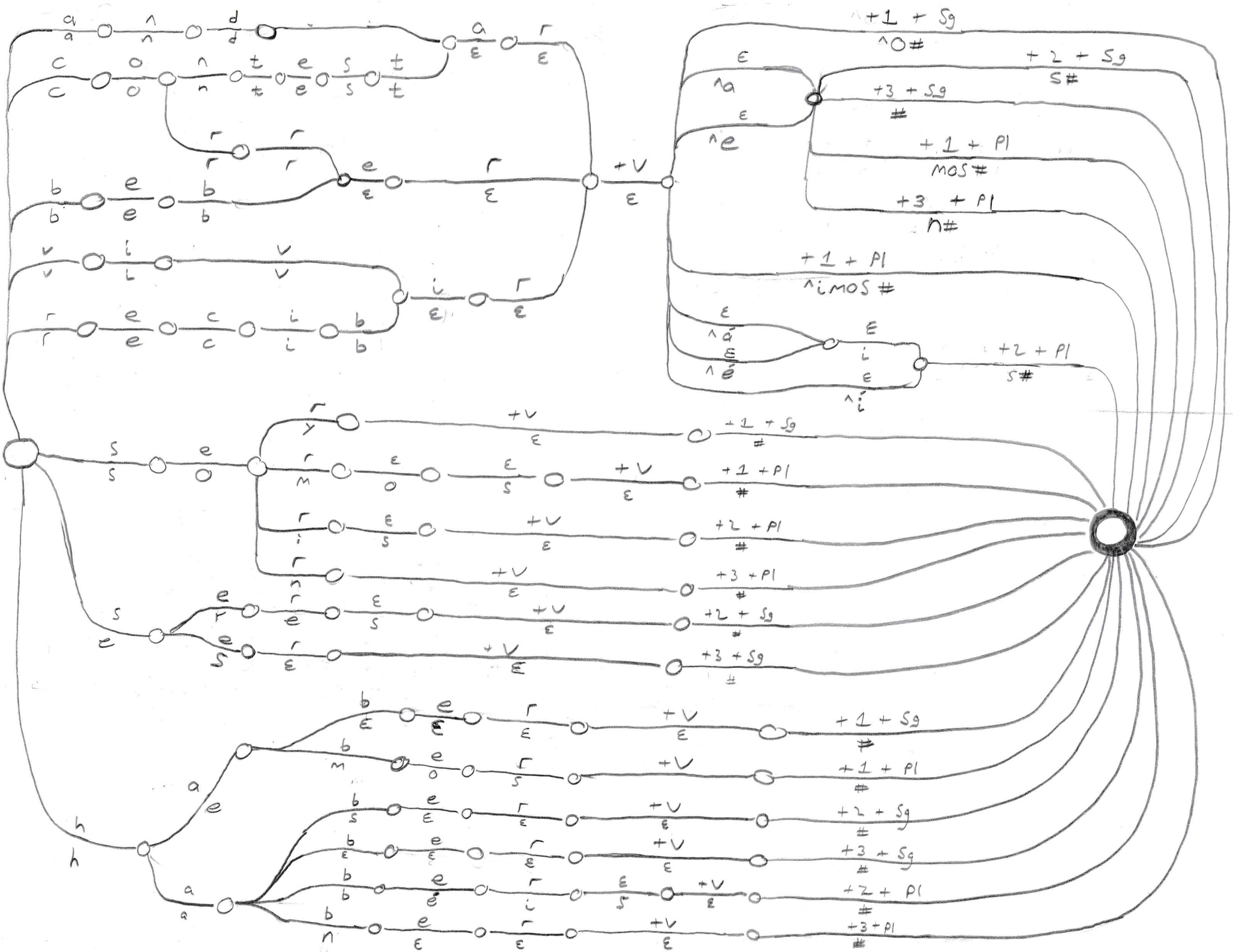
We conclude that the Naive Bayes classifier is the best algorithm for this classification task as it was able to predict the class of unseen data with the greatest accuracy. This result seems reasonable, as mentioned in class Naive Bayes usually does well with little training data (in our case 8000+ training segments). A confusion matrix of its predictions can give a little more insight into its performance.

	positive	negative
positive	838	228
negative	223	843

We see that the Naive Bayes classifier remained unbiased in classifying new data, and that it did not favour one class over the other in general. The number of false positive and false negatives were virtually equal which shows that classifying a negative sentiment is as difficult as classifying a positive sentiment.

2.i


Reg - verb
irreg - 1sg - verb
irreg - 2sg - verb
irreg - 3sg - verb
irreg - 1pl - verb
irreg - 2pl - verb
irreg - 3pl - verb
+ V
ε
+ 1 + Sg
^O #
+ 2 + Sg
^[a, e]s #
+ 3 + Sg
^[a, e] #
+ 1 + Pl
^[a, e, i]mos #
+ 2 + Pl
^[ái, éi, í]s #
+ 3 + Pl
^[a, e]n #
#


+1 +Sg
^O#
+2 +Sg
S#
+3 +Sg
#
+1 +Pl
mos#
+3 +Pl
n#
+1 +Pl
^imos#
+2 +Pl
S#
+V
ε
^a
^e
^á
^é
^í
^i
+1 +Sg
#
+1 +Pl
#
+2 +Pl
#
+3 +Pl
#
+2 +Sg
#
+3 +Sg
#