COMP 550 Assignment 1

Jingyuan Wang 260860682

September 24, 2019

Question 1

1. *Every student took a course.*

- Ambiguity: There is a scope ambiguity in the sentence which can be interpreted as ”Each student took an arbitrary course. The choice of course need not to be the exact same one as long as they all followed the ’take course’ action.”, or ”Every student needed to take a particular course.” and also confuse reader with what course it is talking about.
- Causes: The sentence is following the ”Everyone does something” structure where both ”Everyone” and ”something” can take the scope over each other.
- Domain involved: Semantics ambiguity.
- Complementary Knowledge: Additional explanation about the noun ”course”. For instance, adding ”arbitrary” or the phrase ”due to their own choice” to complement the word ”course” should indicate that ”Every student” took the larger scope whereas explicitly pointing out which specific course the students should take will lead to the latter interpretation.

2. *John was upset at Kevin but he didn’t care.*

- Ambiguity: Deictic ambiguity with deixis ”he”. It could be John was upset at Kevin but ”John didn’t care about the upset” or ”Kevin didn’t care about being upset at”.
- Causes: The ambiguous usage of deixis: using ”he” to regard a man while there are two male characters in the sentence.
- Domain involved: Anaphoric and Pragmatics.
- Complementary Knowledge: Distinction of deixis. Replace ”he” in the sentence with ”John” or ”Kevin”.

3. *Sara owns the newspaper.*

- Ambiguity: Uncertain meaning of the word ”newspaper”. It could be ”Sara owns a specific set of printed paper sheets” or ”Sara has control over the organization which publishes the newspaper”.
- Causes: Multiple meanings of the word ”newspaper”.
- Domain involved: Lexical.
- Complementary Knowledge: Complementary information like ’which set of paper belongs to Sara’ or ’name of the particular newspaper company’.

4. *He is my ex-father-in-law-to-be.*

- Ambiguity: There are different ways to interpret the compound word with hyphen. The first way is that ”the man” used to be my future father-in-law but not anymore probably because the speaker and his/her partner got married or they broke off the engagement. And the second possible meaning is that ”he” is going to be ”my ex-father-in-law”, which indicates ”he” is supposed to be ”my father-in-law” at the present time, but is going to be ”an ex” due to a divorce or something with his/her partner.
- Causes: The unclear sequence about hyphen usage. Whether ”ex-” in ”ex-father-in-law” or ”-to-be” in ”father-in-law-to-be” comes first decide the meaning of this sentence.
- Domain involved: Morphology
- Complementary Knowledge: Replace either part of the compound word with another adjective as ”further ex-father-in-law” or ”former father-in-law-to-be”.

5. *ttyl ;) [text message]*

- Ambiguity: First what the abbreviation is from. Also different implied meanings as ”goodbye” and want to cease the conversation or as literally ”pause the discussion and will be back to talk to the other person in some time” due to some emergency, etc.
- Causes: Lack of face expression and acoustic information like tones. So it is hard to know the meaning behind the word.
- Domain involved: Orthography, pragmatics and Lexical.
- Complementary Knowledge: First of all a machine system needs the corpus with shortened text messages and emoticons. Moreover, the reason for ”talk to you later” or an acoustic sound of this sentence should be needed.

Question 2

For a Naive Bayes classifier, we have:

$$\begin{aligned} P(y=1|x) &= \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1)+P(x|y=0)P(y=0)} \\ &= \frac{1}{1+\frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)}} \\ &= \frac{1}{1+\exp^{-\log\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)}}} \\ &= \sigma(\sum_j \log\frac{P(x_j|y=1)}{P(x_j|y=0)} + \log\frac{P(y=1)}{P(y=0)}) \\ &= \sigma(\sum_j (\theta_{j,1}-\theta_{j,0})\log x_j + \sum_j \log\theta_{j,0} + \log\frac{P(y=1)}{P(y=0)}) \\ &= \sigma(\sum_j \mathbf{w}_j^T\phi_j(x_j)+b) \end{aligned} \tag{1}$$

where ϕ denotes the feature space of x, and

$$b = \sum_j \log\theta_{j,0} + \log\frac{P(y=1)}{P(y=0)}, \tag{2}$$

A logistic function is,

$$P(y=1|x) = \sigma(\mathbf{w}^T\mathbf{x}+b) \tag{3}$$

Hence, the Naive Bayes classifier is equivalent to a logistic regression classifier on certain feature space and thus is a linear classifier.

Question 3

In this experiment, we will implement a sentiment analysis pipeline and investigate with different feature processing procedures and classifiers. The experiment is completed using *NLTK* and *scikit-learn*. The dataset we use is a movie review corpus where two list of sentences are given in a negative and a positive file respectively.

The pipeline of the system is to firstly read in two corpus files, assemble data and label matrix. We will tokenize each sentence and use *CountVectorizer* in *sklearn* to count each unigram occurrence which is finally trained through classifiers and model is evaluated through 10-fold cross-validation. After finishing the basic pipeline, token features were used with stopwords removal, stemming, lemmatization as well as low-frequency words removal. Different parameters and regularization methods are then tested on basis of the best performing features. By running such experiment on three classifiers: Naive Bayes, Support Vector Machine and logistic regression, we achieve our best model through comparison.
Following that, we will report experimental results on *logistic regression classifier* as an instance and look into what improves the model in this sentiment analysis task.

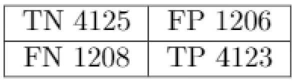
	Stemming		No steming	
Remove stopwords	Lemmatize	No lemmatize	Lemmatize	No lemmatize
	76.5051%	76.4863%	76.1580%	76.1019%
Not remove stopwords	Lemmatize	No lemmatize	Lemmatize	No lemmatize
	76.9837%	**77.0493%**	76.7679%	76.8616%

As shown in the table above, stemming can evidently improve the model accuracy by 0.2% to 0.5% due to better revealing simpler form of tokens while stopwords removing is not helpful since the usage of those common words might implicitly include some sentimental information. In contrast, lemmatization does not have a outstanding impact on model performance. The result with different low frequency thresholds *t*(remove unigrams that only exist in less than t documents) are shown in table below where by removing tokens that only appears in a single sentence, the accuracy is slightly increased. As t get larger, the accuracy drops, probably because we are running with a small dataset and by removing too many words we will lose information. Subsequently, I tried to imply *l1 penalty* to the model, which however, reduced the accuracy to 75.3986%. Finally, the stopping tolerance *tol* and regularization strength *C* were tuned and reached the highest accuracy of 77.1056% with *tol=1e-5, C=0.8.*

	t=1	t=2	t=3	t=5	t=10
accuracy	77.04935%	**77.0587%**	76.8149%	76.7774%	75.6237%

	Stopwords	Lemma	Stem	t	tol	C	10-fold accuracy
logistic regression	Keep	No	Yes	2	10e-5	0.8	77.1056%
svm	Keep	No	Yes	1	10e-4	0.8	75.6423%
nb	Keep	No	Yes	2	N/A	N/A	**78.5498%**
random	N/A	N/A	N/A	N/A	N/A	N/A	50.4783%

After running the same process on Naive Bayes and Support Vector Machine, the best set of parameter with each model is acquired with best performances in the table above. It is shown that Naive Bayes model with certain processing and parameters can generalize the best through this task with unigram occurrence. The confusion matrix is shown as follow which indicates our classifier have equal performance on positive and negative cases.



TN 4125	FP 1206
FN 1208	TP 4123