COMP550 Natural Language Processing
Assignment 1

Jonathan Guymont

September 24, 2018

Question 1

Case 1: "That's Wong on so many levels."

This is an *orthographic* ambiguity. The sentence could be interpreted has *there is a lot of peoples call Wong on all the floors* or has the common expression *that is wrong on so many levels*. The word *Wong* which could be mistaken for *wrong* is the cause of ambiguity. Knowing that Names are starting by a capital letter could prevent a system from making a mistake. For instance, a spell checker would probably change wong for wrong otherwise. Having access to the context (the scene) would also disambiguate the sentence.

Source: https://www.pinterest.ca/pin/360076932688489937/?lp=true

Case 2: "But I know what I am and I'm glad I'm a man, and so is Lola"

This is an *Syntactic* ambiguity. The sentence could be interpreted has *I am glad to be a man and Lola is glad to be a man* or has *I am glad to be a man and Lola is glad I am a man*. The way the sentence is writen: *I am glad I am [something], and so is [another person]* cause ambiguity. The adjacent sentences would disambiguate the sentence; for instance, reference to Lola as a her etc. Knowing that Lola is more a woman name would disambiguate the passage.

Source: https://www.reddit.com/r/Music/comments/2izue5/serious_is_lola_a_transvestite/

Case 3: "GOP Lawmakers Grill IRS Chief Over Lost Emails"

This is a *lexical* ambiguity. The sentence could be interpreted has *GOP lawmakers are grilling the chief of IRS (has in burning over a fire) for losing emails* or has *GOP lawmakers are giving a hard time (without any violence) to the chief of IRS*. The word *Grill* is the cause of ambiguity. Knowing that US government employees do not get tortured by the government disambiguate the passage.

Source: https://www.wsj.com/articles/gop-lawmakers-grill-irs-chief-over-loss-emails-1403278518?mg=id-wsj

Case 4: "you look like I need a drink"

This is a *pragmatic* ambiguity. The sentence could be interpreted has *I have a biological need to drink* or has *I would like a drink*. The expression *I need a drink* meaning *I would like to have a drink* is the cause of ambiguity. Knowing the expression would disambiguate the passage.

Source: https://genius.com/Justin-moore-you-look-like-i-need-a-drink-lyrics

Case 5: "this is a big plane"

This is a *phonological* ambiguity. The sentence could be interpreted has *this is a big plain* or has *this is a big plane*. The word *plane* cause of ambiguity, because we use the same phonome to say *plane* or *plain*. Knowing the english level of the writer and the context would help determining the chance he or she meant *plain*. Adjacent passage would help determine wether we are talking of a plane or a plain.

Source: http://livestly.com/new-worlds-largest-plane/

Question 2

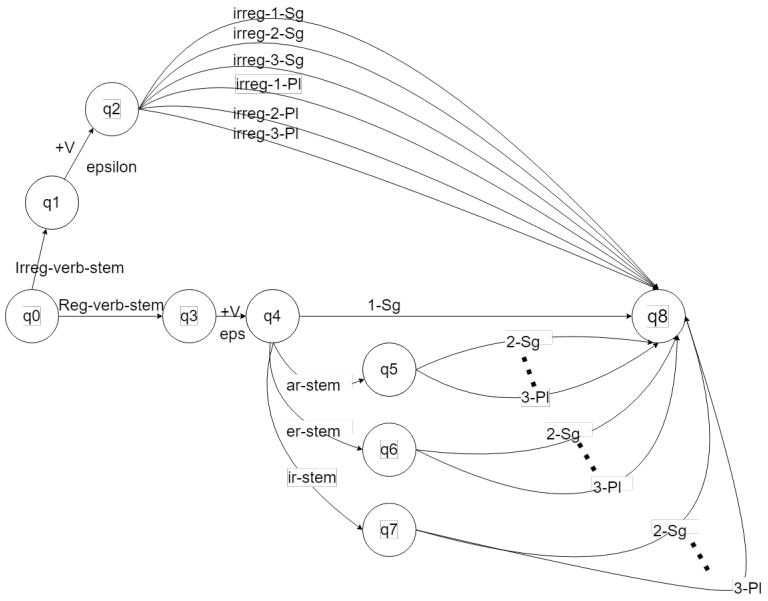

Figure 1: Schematic Transducer

infinitive	**1-sg-reg**	**2-sg-reg**	**3-sg-reg**	**1-pl-reg**	**2-pl-reg**	**3-pl-reg**
andar	and a:o e: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
infinitive	**1-sg-irreg**	**2-sg-irreg**	**3-sg-irreg**	**1-pl-irreg**	**2-pl-irreg**	**3-pl-irreg**
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

Table 1: Lexicon table

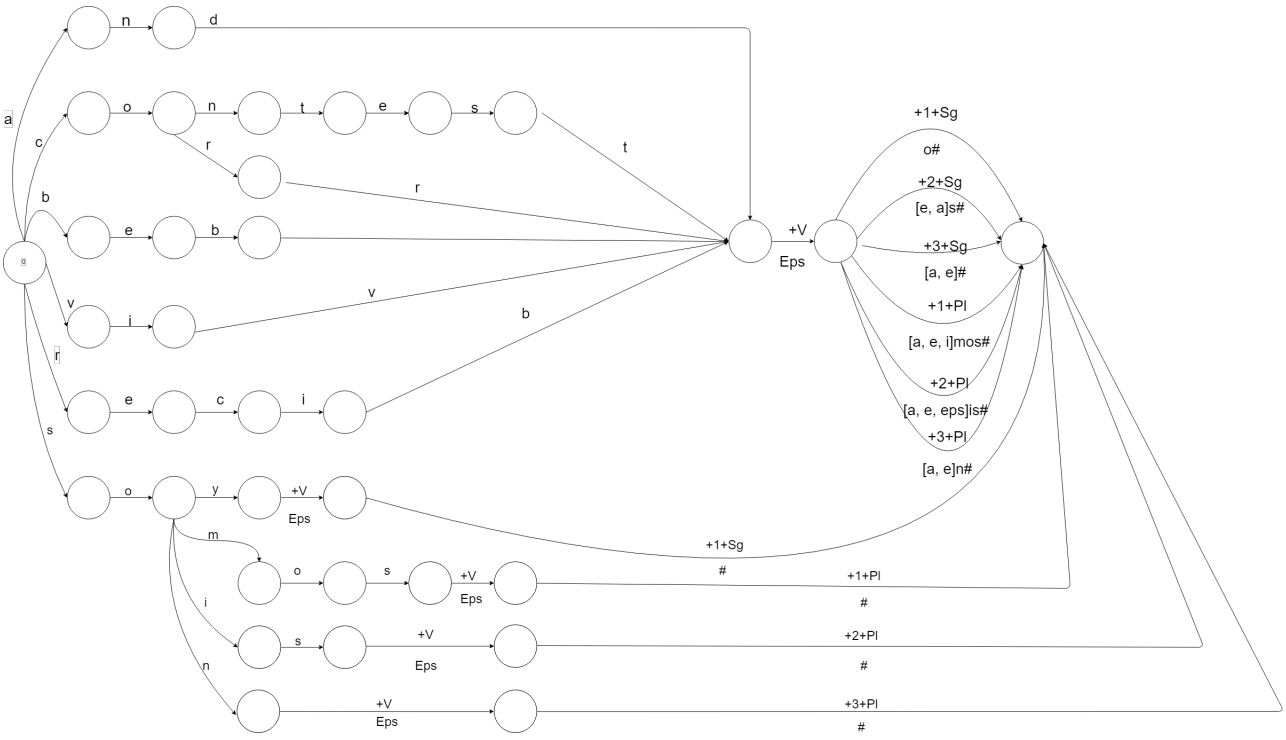

Figure 2: Fleshed-out FST

Question 3

Problem setup

The goal is to train a binary classifier to perform sentiment analysis on movie reviews. More predisely, the classifier should predict wether a review is positive or negative. The number of examples is equal to 10662. One half of the examples are positive review and the other half of are negative reeviews. The corpus size in 224067 and the lexicon size is 21425. Tree algorithm are compared during the experiment: Naive Bayes, Logistic Regression, and SVM.

Experimental procedure

The data was split into a training set and a test set. The size of the test set represent 15% of the data. The test examples are selected randomly among the all the examples. A 5-fold cross-validation procedure was used to select the hyperparameters for each models. During the cross-validation process, a grid search is perform over the predefined set of parameters and the cross-validation is performed on all the different combination of parameters. After the cross-validation process, the hyperparameters that gives the best average accuracy over the 5 folds is selected and refit over all the training set and test on the test set.

Range of parameter setting

Hyperparameters include parameter for the preprocessing and parameters of the models. All hyperparameters are selected simulaneously since a changing the preprocessing is likely to influence the model behavior. The preprocessor hyperparameters are the number of n-grams, the stopwords to remove, the threshold for which unfrequent words are removed (e.g. words that represent less then 0.01% of the corpus), and the threshold for which too frequent words are removed (e.g. word representing more then 10% of the copus). The classifier hyperparameters depend on the model. The models hyperparameters that were tuned during cross-validation and there range are shown in tables 2.

Preprocessing	Possible values	SVM	Possible values	Naive Bayes	Possible values	Logistic regression	Possible values
n-grams	{1-gram, 2-grams}	C	{0.25, 0.5, 0.75, 1, 2, 3, 4, 5}	α	{0.5, 0.75, 1., 1.5, 2}	C	{0.25, 0.5, 0.75, 1, 2, 3, 4, 5}
stopwords	{none, nltk english stopwords}	loss function	{squared hinge}				
Min frequency threshold	{none, 0.0001, 0.001}						
Max frequency threshold	{none, 0.2, 0.3, 0.4, 0.5}						

Table 2: Range of hyperparameters

Result and conclusion

The table 4 show the accuracy of the models that performed the best on the training set. The performance are comparable, but the logistic regression perform slightly better. In all case, the training set is overfited. The gap between the accuracy on the training set and the test set could be decrease with better regularisation. It is surprising that removing a set of stopwords do not improve the accuracy on any models. This indicates that the set of stopwords provided by the library nltk contains more relevant features then unrelevant ones. In particular, nltk list of stopwords include the negative form of some verbs. For instance: haven't, isn't, shouldn't,...These may be useful features to determine wether a review is positive. For example, in "This movie isn't good." the feature "isn't" is very important.

hyperparameters	SVM	Naive Bayes	Logistic Regressionm
n-grams	2-gram	2-gram	2-gram
stopwords	None	None	None
Min frequency threshold	0.	0.	0.
Max frequency threshold	0.2	1.	0.4
C	0.25	na	2.
α	na	0.75	na
train accuracy	1.0	0.995	1.0
test accuracy	0.791	0.787	**0.795**

Table 3: Models accuracy comparison (with there best choice of parameters). na means not applicable.

	0	1
0	0.8	0.2
1	0.21	0.79

Table 4: Confusion matrix: 0 for negative and 1 for positive.