NLP Final Project
Due Friday, May 14, 2009

For this final assignment, there are two types of projects to choose from, but you may also propose your own.

I. Annotation and Analysis of Data

For this task, you will annotate data in a corpus, and there are two corpora in particular to work with. Choose Corpus A or Corpus B:

A. The Tweet corpus

For another research project, a number of tweets have been downloaded from Twitter that are on the topic of the Obama Health Care plan. For each tweet, we would like to have a label that says whether the opinion of the tweet is positive, negative or neutral towards the plan. For purposes of this class, we have separated the data randomly into groups of approximately110 tweets, each in its own sheet of an excel file. For people who choose this topic, each person should get a different set of tweets to annotate, so that we get as much annotated data as possible. Furthermore, we would like the annotated data to be made available for people who choose the classification task. Coordination of who is annotating what spreadsheet and making the annotated data available will be achieved via a discussion list in the iLMS. Each person should annotate 1 sheet.

For analysis purposes, in addition to the three labels positive, negative or neutral, you should assign an extra label of “hard”, meaning hard to decide, for any tweet that is difficult to tell what the opinion is. (We will had two columns to each sheet, one for the class labels and one for the possibly hard labels.)

B. The Movie Review corpus sentence level

For this annotation, you will investigate the movie review corpus by analyzing the polarity of the opinions in the reviews at the sentence level. First, pick about 5 reviews (for each person) from the Movie Review corpus that you are comfortable working with. Produce a file that assigns an opinion level to each sentence. The opinion must be that of the reviewer towards the movie itself or towards aspects of the movie, such as actors, director, or particular scenes. Disregard opinions about anything else, like other movies, or factual statements. So each sentence can be labeled with +, - or nothing, indicating no opinion in that sentence.

For analysis purposes, in addition to the three labels +, - , or nothing, you should assign an extra label of “hard”, meaning hard to decide, for any sentence that is difficult to tell what the opinion is.

Semantic analysis of the Annotated data

The results of your project will be the annotated data file and a report that contains the following.

i. From the annotation, give a few examples each of tweets or sentences that typify each of the three labels. Then give a few examples of the ones which were marked hard and give any comments that you have about why they were difficult.

ii. Analyze the sentences that you annotated and look for short phrase patterns or words that indicated to you the polarity of the opinion in that sentence. For each word or pattern, look them up in WordNet and see if the word or pattern can be generalized by adding hypernyms, synonyms, etc. Discuss whether you think that these patterns or words could be used to predict the opinions in future movie reviews.

iii. Design positive and negative sentiment word lists. For this, you can use your annotation, semantic resources such as WordNet to get hypernyms, synonyms, etc. that you used in part ii, or anything else that you can think of.

II. Classification of Data

For this task, you should choose to either work further on classification of the Movie Review data at the document level as we did in lab week 11, or to work on classifying the tweet data.

i. Using the data, define features in the NLTK and classify using Naïve Bayes. If you are using the tweets, you will need to read in the data as a set of documents from csv files produced from one or more of the spreadsheets. (There is one annotated spread sheet to start with.) Try getting better accuracy with some experiments. Here are some ideas of things to try:

Use the top frequency all words, but apply a stop word list to filter out stopwords.
Improve the sentiment/significant word lists, either from a team member's annotation
of part I, reading or inspection of the data.
Apply POS tagging and use only adjectives and/or adverbs.
Use a sentiment/significant word list and look up hypernyms in WordNet and add those
to the feature set as well (probably only useful for non-adjectives in WordNet)

ii. Similar to i, but use the Stanford Parser to process the documents, where the POS tagger should be more accurate than in the NLTK. Use some functions from Assignment 2 to pick out all the adjectives and or adverbs. Or use functions to pick out noun phrases and see if they improve the classification.

iii. Use the NLTK to prepare feature sets, using one or more of the techniques above. Output the feature sets as arff files and use Weka to do the classification. (There will be an additional tutorial session on using weka.)

What to Hand In

If you are working in a group, you should choose a task for each person. If you do annotation, hand in the annotation data, submitting the tweet documents to the iLMS discussion as soon as they are done. Every group should hand in a report with the description of all that you did and the discussion of the results. As usual, submit these document to the iLMS system by the end of the day on the due date.