Laura B. Bell
February 10, 2019
Dissertation Excerpt: Regression and random forest analyses and results

Introduction: Direct speech act category

The direct speech act category of an utterance is determined by its primary communicative purpose. The speech act taxonomy used in this study is as follows:

1) **Assertive:** The primary purpose is to assert belief in the truth of a proposition.
2) **Directive**: The primary purpose is to give an order or command.
3) **Commissive:** The primary purpose is to commit oneself to a future action.
4) **Expressive:** The primary purpose is to express a psychological state.
5) **Interrogative**: The primary purpose is to ask a question.
6) **Exercitive:** The primary purpose is to effect a change in the real world.

In search of which, if any, overt linguistic features prove themselves to be good predictors of direct speech act category in the corpus, logistic regression and random forest techniques were performed on the annotated Twitter corpus, with direct speech act category as the response variable.

Summary of logistic regression analysis and results

The logistic regression analysis was performed in RStudio, using the **nnet** package. All except two of the explanatory variables were coded via data manipulation in R. The two exceptions, “indirect act choice” and “serious yes/no”, were the result of questions that the corpus annotators were asked at the same time that they classified the direct speech act of the segment.

All together, the analysis tested 114 explanatory variables, for a total of 230 features when interaction terms were added. This was far too many variables for them all to be meaningful, especially for the amount of data in the corpus, so model selection was a crucial aspect of the analysis. Stepwise feature dropping was used to whittle the 230 features down to only those features whose absence caused a significant change in the model. A change was considered significant if it resulted in a p-value of <0.05 in the analysis of variance (ANOVA) between the two models. A significant change meant that the feature was kept, while an insignificant change meant that the feature remained dropped.

After model selection, 8 individual terms were determined by the model to be the most relevant to direct act category classification. These terms, and brief descriptions when necessary, are listed here:

1) **Indirect act choice:** This variable indicates the indirect act choice for the segment, including “None” as an option.

2) **Serious yes/no:** This variable indicates whether the direct speech act was annotated as being meant “seriously” or not.[1]
3) **Wh-word count**
4) **“Tell” count:** This variable indicates the number of times “tell” appears, as a verb, in the segment.
5) **Proper noun count**
6) **Punctuation:** This variable is categorical, giving the final character of the segment out of period, comma, exclamation point, question mark, colon, semi-colon, or “other” (anything else).
7) **Directive bigram, “gotta stop”:** This variable indicates the number of times the bigram “gotta stop” appears in the segment, and it is one of 30 bigram features that were included in the analysis. The bigrams consisted of the two word sequences from each direct speech act category with the highest tf-idf scores in the training portion of the corpus (the top 5 for each category were included).
8) **Expressive bigram, “i’m so”**

Summary of random forest analysis and results

The random forest analysis was performed in RStudio, using the **randomForest** package. The same explanatory and response variables that were used in the logistic regression analysis were used for the random forest, but model selection took a different form. The random forest output includes the mean decrease in accuracy (MDA) of the model if the feature were to be left out. The lower the value for MDA, the less important the feature. The random forest output also includes the mean decrease in Gini coefficient (MDG) for each feature. A higher mean decrease in Gini coefficient means that a feature contributes more to the purity of the tree nodes.

To arrive at a list of only the important variables in the random forest model, both MDA and MDG were used. First, any feature with an MDA value of zero (or less) was eliminated. After that, features were eliminated in a stepwise fashion by establishing a threshold MDG value. Features with MDG values below the threshold were eliminated. If this elimination led to an increase in model accuracy, those features remained eliminated, and the threshold was raised. The threshold was gradually raised until the increase in threshold led to a decrease in model accuracy.

After model selection, 4 individual terms were determined by the model to be the most relevant to direct speech act category classification. These terms, and brief descriptions when necessary, are listed here:

1) **Punctuation**
2) **Segment length:** This variable indicates the length of the segment, in number of tokens.

[1] “Indirect act choice” and “Serious yes/no” are actually not overt, but are inferred (which is why these variables were the result of annotation, as opposed to simple data manipulation). They were included primarily because if they appear significant for direct act identification, this would support a conclusion that inferred information affects classification (alongside, potentially, overt information).

3) **Tweet length:** This variable indicates the length of the tweet that a given segment is a part of, in number of segments.
4) **Segment position:** This variable indicates the position of the segment within its respective tweet.

Comparison and summary

Of the individual terms tested, punctuation proved to be the most important overt element for direct speech act classification, as it appeared in the final results of both the logistic regression and random forest analyses. Additionally, a chi-square test of independence showed punctuation and direct act category to have a significant relationship (p<2.2e-16). Figure 1 provides the association plot for that variable crossed with direct act category. On the "Direct Act" axis, "E" stands for "Exercitive," and "X" stands for "Expressive."

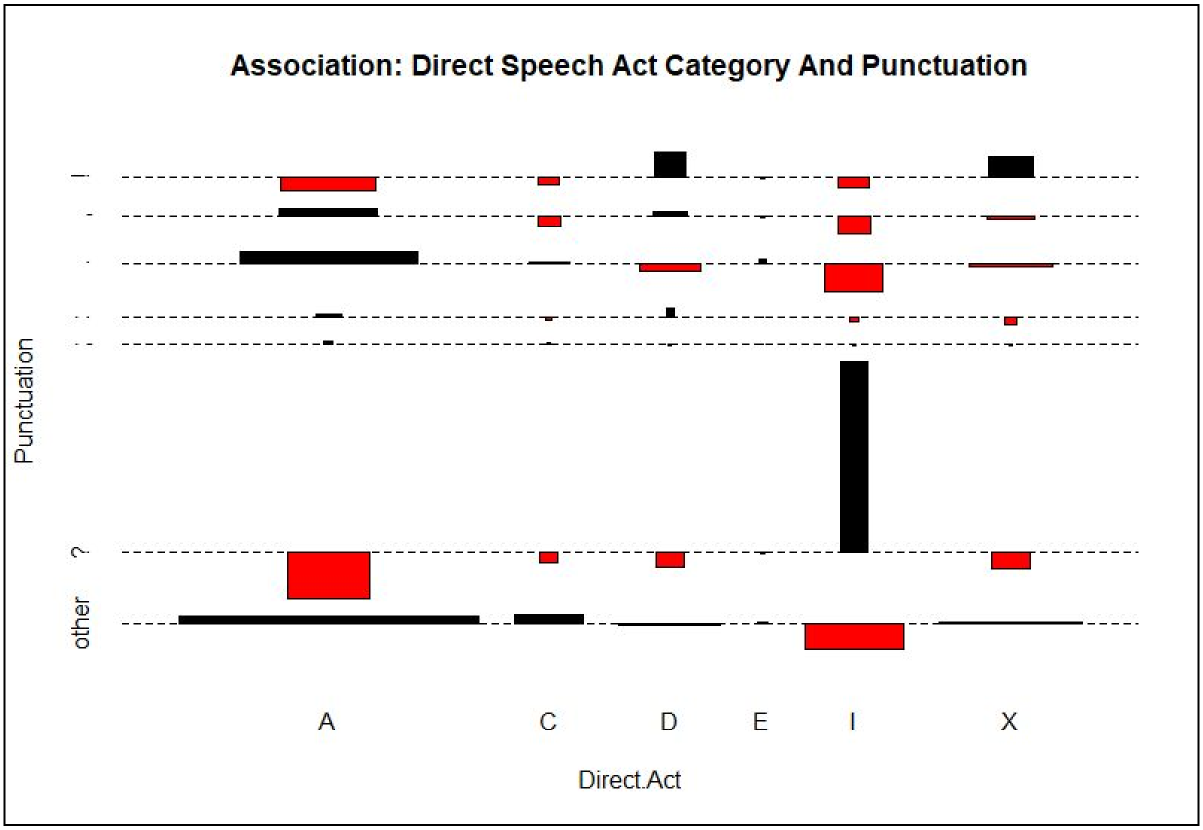

Figure 1: Direct speech act category and punctuation

Focusing on the most significant associations, figure 1 indicates a strong positive association between interrogative direct act and the presence of a question mark as the final character of a segment. The presence of an exclamation point (the top row) as the final character shows a positive association with both directive and expressive direct act categories, and an assertive direct act shows positive associations with both final periods and final commas.

While the conclusion that punctuation is one of the most important overt elements for Twitter speech act classification is not particularly surprising, the ramifications of this discovery as pertains to spoken dialogue is not immediately obvious. Spoken dialogue, of course, is not punctuated. To translate the results of this analysis into something relevant to the spoken mode, correlates to punctuation that are available in the spoken mode would need to be confirmed. Likely candidates include intonation, facial expression, and gaze. Any or all of these elements may prove to perform the same functions as punctuation.[2]

[2] More accurately, due to the antecedence of spoken and signed languages to written ones, punctuation would be a written correlate for these potential spoken speech act category indicators.