Trivial Classification: What features do humans use for classification?

Abstract

When humans classify of piece of text, what features do they use? To answer this question, we develop a technique that leverages an existing community that competitively classifies text in an online fashion: the collegiate trivia circuit. Individuals receive features (clues) in a continuous stream, halt the transmission of features once they have enough information to give a classification (answer), and then produce an answer. Because of the compeditive pressure to answer correctly as quickly as possible, answers happen immediately after useful information has been revealed (Figure 1).

By providing the community with a web-based system to practice, we collected tens of thousands of word-by-word ratings of feature usefulness for eliciting the correct association of questions to answers (Figure 2). We trained a classifier to produce text classifications (mapping question text to answers) with the weight vector mean either being zero, tf-idf based, or based on what clues humans used to answer questions (Table 1).

This shows that text classification can benefit not just from human provided labels, but also from the human process of classification. By embedding the process of human classification in a game, we were able to quickly collect a large data set.
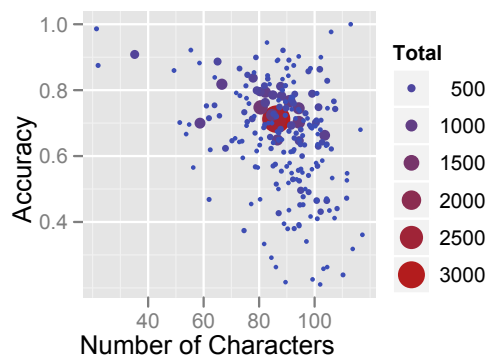
Figure 2: Each user is positioned on a plot of accuracy vs. the number of characters – on average – the user took to give an answer. The size of the point represents the number of questions answered. Users that answered questions later in the question (smaller number of characters before the buzz) in general had higher accuracy (proportion of correct answers). However, there were users that were able to answer questions relatively early without sacrificing accuracy.

Figure 1: A typical quiz bowl question on the answer William Jennings Bryan. Words are shaded based on the number of times the word triggered a buzz (darker is more). Diamonds ($\diamond$) indicate buzz positions.

Weighting	Error
zero	0.37
tf-idf	0.14
human	**0.09**

Table 1: Classification error for different prior weighting schemes. Incorporating human knowledge improves classification error.