

Abstract

Summaries are used in daily life to condense information in a manner suitable for the intended recipient's use and ideally suit the recipient's information seeking goals. In the case of text, examples of such summaries include newswire articles, headlines, and the information snippets returned by Google. Previous research has focused on summarizing newswire articles or clusters of newswire documents, scientific articles, books, and extracting opinion (sentiment) sentences from reviews.

Our research addresses how to create short multi-sentence summaries meeting user's goals within specific genres. The methodology is first to determine the genre of the document and then, based on the genre, present applicable summaries designed to address the user's information seeking goals. For example, in the movie review genre, a goal-focused summary for a review could be an overview summary, a plot summary, or an opinion summary, each of which has a different focus and hence a different summary composition.

We describe our experiments with genre identification using different sets of features and varying numbers of training documents and show that genre tagging using classifiers (Support Vector Machines and Random Forests) is probably at a sufficient level of accuracy to inform a summarization system. We discuss the creation of goal focused single document summaries for seven genres (newswire articles, editorials, interviews, biographies, movie reviews, product reviews, and product press releases). Our results indicate that genre oriented goal-focused summarization algorithms perform better than our two baselines, lead sentence and the newswire summarization algorithm.

We also examine email summarization and, based on previous research in speech acts, present categories of the communicative intent of the sender. We discuss our experiments in identifying these email speech acts using a small annotated corpus of personal emails. In addition, for textual summaries of a sender's email, we analyze a human annotated subset of the Enron corpus and based on a user study, suggest that the subject line and one sentence extracted from the email text body may be an effective summary length.

We briefly explore multi-document summarization for the newswire genre and present results indicating that, by using maximal marginal relevance (MMR) to eliminate redundancy, there is more coverage of the subtopics in a cluster than our baseline - which uses our single document newswire summarization algorithm on the concatenation of all articles and no MMR. MMR based summaries were also preferred in a ranking produced by one unbiased human evaluator.