# Arabic/English Multi-Document Summarization with CLASSY—The Past and the Future

Judith D. Schlesinger[1], Dianne P. O'Leary[2], and John M. Conroy[1]

[1] IDA/Center for Computing Sciences, Bowie MD 20715, USA,
[judith,conroy]@super.org
[2] University of Maryland, CS Dept. and Inst. for Advanced Computer Studies,
College Park MD 20742, USA,
oleary@cs.umd.edu

**Abstract.** Automatic document summarization has become increasingly important due to the quantity of written material generated world-wide. Generating good quality summaries enables users to cope with larger amounts of information.

English-document summarization is a difficult task. Yet it is not sufficient. Environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events.

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic(/query)-driven summaries for single documents or clusters of documents. CLASSY has performed well in the Document Understanding Conference (DUC) evaluations and the Multilingual (Arabic/English) Summarization Evaluations (MSE).

We present a description of CLASSY. We follow this with experiments and results from the MSE evaluations and conclude with a discussion of on-going work to improve the quality of the summaries–both English-only and multi-lingual–that CLASSY generates.

## 1 Introduction

Automatic multi-document summarization poses interesting challenges to the Natural Language Processing (NLP) community. In addition to addressing single document summarization issues such as determining the relevant information, pronoun resolution, and coherency of the generated summary, multi-document summary-generating systems must be capable of drawing the "best" information from a set of documents.

Automatic single document text summarization [11] has long been a field of interest, beginning in the 1950s, with a recent renaissance of activity beginning in the 1990s. System generated single document summaries for English are generally of good quality. Therefore, NIST ended single document summarization evaluation after the 2002 Document Understanding Conference (DUC). See [17] for DUC research papers and results over the years.

In contrast to the single document task, summarization of multiple documents written in English remains an ongoing research effort. A wide range of strategies to analyze documents in a collection and then synthesize/condense information to produce a multi-document summary have been explored by various research groups. System performance has improved but still lags behind human performance.

Nevertheless, environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events. Thus it is vital that English speakers be able to access documents in a variety of languages.