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 worldwide. 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 Multi-lingual (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.

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 568–581, 2008.

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.

The quantity of non-English documents makes it impossible to expect quality (or, even, *any*) human translation. Therefore, we have come to rely on machine translation (MT) systems for translation to English. While MT systems continue to improve, generated translations remain difficult to read and understand, with critical words often omitted, and inconsistent translations for the same word in a document [5,6]. Translation of Arabic documents is particularly challenging due to errors introduced by incorrect sentence-splitting, tokenization, and lemmatization.

Volumes of documents in one or more languages may be summarized by:

- creating summaries in the original language(s) which can then be translated by either humans or MT systems to determine "importance".
- creating summaries of the (MT-translated) documents which can be used to determine which documents are important and should be translated by humans.

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic summarization system, developed for summarizing English documents. CLASSY uses trimming rules to shorten sentences in the document, identifies sentences as being more or less likely to be included in a summary, generates a summary for each document, selects sentences for a multi-document summary for a cluster of related documents, and finally organizes the selected sentences for the final summary.

Our approach to multi-lingual summarization is based on the second approach listed above: we use CLASSY to generate single or multi-document (cluster) summaries of MT-translated documents. The experiments presented in Sect. 4. helped determine the best way to accomplish this.

We participated in the two Multilingual Summarization Evaluations (MSE)[1], which evaluated summaries of document sets containing a mix of both English and Arabic documents. Both the Arabic source and the MT output were

[1] MSE 2005: *Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization Workshop* at the Annual Meeting of the Association of Computational Linguistics (ACL 2005), Ann Arbor Michigan, 25-30 June 2005. MSE 2006: *Multilingual Summarization Evaluation* at the 21st International Conference on Computational Linguistics (ACL 2006)/44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17-21 July 2006.