A Study of Global Inference Algorithms in Multi-document Summarization

Ryan McDonald

Google Research
76 Ninth Avenue, New York, NY 10011
ryanmcd@google.com

Abstract. In this work we study the theoretical and empirical properties of various global inference algorithms for multi-document summarization. We start by defining a general framework for inference in summarization. We then present three algorithms: The first is a greedy approximate method, the second a dynamic programming approach based on solutions to the knapsack problem, and the third is an exact algorithm that uses an Integer Linear Programming formulation of the problem. We empirically evaluate all three algorithms and show that, relative to the exact solution, the dynamic programming algorithm provides near optimal results with preferable scaling properties.

1 Introduction

Automatically producing summaries from large sources of text is one of the oldest studied problems in both IR and NLP [7,13]. The expanding use of mobile devices and the proliferation of information across the electronic medium makes the need for such technology imperative. In this paper we study the specific problem of producing summaries from clusters of related documents – commonly known as multi-document summarization. In particular, we examine a standard paradigm where summaries are built by extracting relevant textual units from the documents [5,9,11,18].

When building summaries from multiple documents, systems generally attempt to optimize three properties,

- **Relevance:** Summaries should contain informative textual units that are relevant to the user.
- **Redundancy:** Summaries should not contain multiple textual units that convey the same information.
- **Length:** Summaries are bounded in length.

Optimizing all three properties jointly is a challenging task and is an example of a *global inference problem*. This is because the inclusion of relevant textual units relies not only on properties of the units themselves, but also properties of every other textual unit in the summary. Unlike single document summarization, redundancy is particularly important since it is likely that textual units from different documents will convey the same information. Forcing summaries to obey a length constraint is a common set-up in summarization as it allows for a fair empirical comparison between different possible

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 557–564, 2007.

outputs [1,12]. It also represents an important "real world" scenario where summaries are generated in order to be displayed on small screens, such as mobile devices.

The global inference problem is typically solved in two ways. The first is to optimize relevance and redundancy separately. For example, the work of McKeown et al. [14] presents a two-stage system in which textual units are initially clustered, and then representative units are chosen from each cluster to be included into the final summary. The second approach is to treat the problem truly as one of global inference and optimize all criteria in tandem. Goldstein et al. [9] presented one of the first global models through the use of the maximum marginal relevance (MMR) criteria, which scores sentences under consideration as a weighted combination of relevance plus redundancy with sentences already in the summary. Summaries are then created with an approximate greedy procedure that incrementally includes the sentence that maximizes this criteria. More recently, Filatova and Hatzivassiloglou [8] described a novel global model for their event-based summarization framework and showed that inference within it is equivalent to a known NP-hard problem, which led to a greedy approximate algorithm with proven theoretical guarantees. Daumé et al. [6] formulate the summarization problem in a supervised structured learning setting and present a new learning algorithm that sets model parameters relative to an approximate global inference algorithm.

In this work we start by defining a general summarization framework. We then present and briefly analyze three inference algorithms. The first is a greedy approximate method that is similar in nature to the MMR algorithm of Goldstein et al. [9]. The second algorithm is an approximate dynamic programming approach based on solutions to the knapsack problem. The third algorithm uses an Integer Linear Programming (ILP) formulation that is solved through a standard branch-and-bound algorithm to provide an exact solution. We empirically evaluate all three algorithms and show that, relative to the exact solution, the dynamic programming algorithm provides competitive results with preferable scaling properties.

2 Global Inference

As input we assume a document collection $\boldsymbol{D} = \{D_1, \ldots, D_k\}$. Each document contains a set of textual units $D = \{t_1, \ldots, t_m\}$, which can be words, sentences, paragraphs, etc. For simplicity, we represent the document collection as the set of all textual units from all the documents in the collection, i.e., $\boldsymbol{D} = \{t_1, \ldots, t_n\}$ where $t_i \in \boldsymbol{D}$ iff $\exists\, t_i \in D_j \in \boldsymbol{D}$. We let $S \subseteq \boldsymbol{D}$ be the set of textual units constituting a summary.

We define two primary scoring functions,

1. $\mathit{Rel}(i)$: The relevance of textual unit t_i participating in the summary.
2. $\mathit{Red}(i, j)$: The redundancy between textual units t_i and t_j. Higher values correspond to higher overlap in content.

These scoring functions are completely arbitrary and should be defined by domain experts. For instance, scores can include a term to indicate similarity to a specific query for query-focused summarization or include terms involving entities, coherence,