

Recent Developments in Text Summarization

Inderjeet Mani
The MITRE Corporation
11493 Sunset Hills Road, W640
Reston, VA 20190, USA
+1-703-883-6149
imani@mitre.org

ABSTRACT

With the explosion in the quantity of on-line text and multimedia information in recent years, demand for text summarization technology is growing. Increased pressure for technology advances is coming from users of the web, on-line information sources, and new mobile devices, as well as from the need for corporate knowledge management. Commercial companies are increasingly starting to offer text summarization capabilities, often bundled with information retrieval tools. In this paper, I will discuss the significance of some recent developments in summarization technology.

Categories and Subject Descriptors

H.3.1. [Content Analysis and Indexing]: Abstracting Methods.
I.2.7. [Natural Language Processing]: Text Analysis.

General Terms

Algorithms, Management, Experimentation.

Keywords

Text summarization, information reduction, evaluation, extracts, abstracts, information and knowledge management.

1. INTRODUCTION

The explosion of the World Wide Web has brought with it a vast hoard of information, most of it relatively unstructured. This has created a demand for new ways of managing this rather unwieldy body of dynamically changing information. Some form of automatic summarization seems indispensable in this environment. Increased pressure for technology advances in summarization is coming from users of the web, on-line information sources, and new mobile devices, as well as from the need for corporate knowledge management. Commercial companies are increasingly starting to offer text summarization capabilities, often bundled with information retrieval tools. In this paper, I will discuss recent developments in summarization.

The goal of text summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs [11]. There are a variety of different kinds of summaries. Summaries can be user-focused (or topic-focused, or query-focused), i.e., tailored to the requirements of a particular user or group of users, or else, they can be 'generic', i.e., aimed at a particular - usually broad - readership community. Traditionally, generic summaries written by authors or professional abstractors served as surrogates for full-text in information access environments. However, as our computing environments have continued to accommodate full-text searching, browsing, and personalized information filtering, user-focused summaries have assumed increasing importance. A summary can take the form of an extract, i.e., a summary consisting entirely of material copied from the input, or an abstract, i.e., a summary at least some of whose material is not present in the input (see [8] [9] [11] for a detailed introduction to the field).

Automatic summarization in some form has been in existence since the 1950's. Two main influences have dominated research in this area. Work in library science, office automation, and information retrieval has resulted in a focus on methods for producing extracts from scientific papers, including the use of "shallow" linguistic analysis and the use of term statistics. The other influence has been research in artificial intelligence, which has explored "deeper" knowledge-based methods for condensing information. While there are a number of problems remaining to be solved, the field has seen quite a lot of progress, especially in the last decade, on extraction-based methods. This progress has been greatly accelerated by the rather spectacular advances in shallow natural language processing, and the use of machine learning methods which train summarization systems from text corpora consisting of source documents and their summaries. Research systems are now able to summarize meetings [16], TV broadcast news [12], and the medical literature [4], and also generate biographies [15]. A number of commercial vendors such as InXight, IBM, SRA, etc., now offer summarization products.

In this paper, I will focus on three recent developments. The deluge of on-line data has given rise to an interest in characterizing content in collections of documents rather than just single-documents, thus paving the way for *multi-document summarization*. The ability to deliver information anytime, anywhere has created an interest in *summarization for hand-held devices*. Last, but not least, there has been increasing activity in *summarization evaluation*, with several large-scale evaluations being carried out.

2. MULTI-DOCUMENT SUMMARIZATION

Multi-Document Summarization (MDS) is, by definition, the extension of single-document summarization to collections of related documents. It is primarily concerned with summarizing collections of related documents so as to remove redundancy while taking into account similarities and differences in information content. Most MDS systems involve a degree of pre-filtering using clustering methods to allow the system to focus on subsets of closely-related documents, as well as presentation methods which allow a multi-document summary to be displayed along with supporting information from the source documents.

While identifying differences requires deeper, domain-specific methods, the shallow approaches work quite well in finding similar passages across documents. One particular shallow approach [6] explores the relationship between relevance and redundancy. Consider a search engine scenario where the top 100 hits returned by a search engine are such that the first 20 are about the same event, but hits 36, 41, and 68 are very different, although marginally less relevant. A user who scans through the first 10 or 20 hits may get tired and miss the different, marginally less-relevant information lurking further down the hit list. The topic-focused MDS system of [6] addresses this problem by offering a ranking parameter that trades off relevance of a passage (or document) to a query against diversity from (and non-redundancy with) the passages (or hits) seen so far.

In between the two extremes of shallow and deep approaches, one finds MDS systems which can fuse together similar phrasal descriptions into a single sentence. As an example, [15] describes a biographical summarizer that provides short summaries of the salient attributes and activities of people in news collections. Syntactic processing using cascaded finite-state transducers and semantic processing using WordNet [13] are used to merge together hundreds of descriptions of a person into just a few sentences. For example, in a collection of 1300 wire service news documents on the Clinton impeachment proceedings, there are 607 sentences mentioning Vernon Jordan by name, from which the system extracted 82 descriptions expressed as 78 appositive phrases (e.g., “presidential friend”) and relative clauses (“who helped Lewinsky find a job”). The relative clauses are duplicates of one another, while the 78 appositives are merged using WordNet into just 2 groups: “friend” (or equivalent descriptions, such as “confidant”), and “adviser” (or equivalent such as “lawyer”).

These MDS capabilities, in condensing information content, allow for the restructuring of information derived from multiple documents, as well as, in some cases, a fusion of information from document collections and on-line databases. MDS summarizers rely on components which construct structured or semi-structured representations derived from the source data. When used to characterize the content of large information spaces, they are able to provide potentially useful collection (or web-site) meta-data which can help in cataloging and accessing content.

3. SUMMARIZATION FOR HAND-HELD DEVICES

Hand-held devices such as personal digital assistants (PDAs) and cell-phones provide an interesting niche application for summarization technologies. There are many opportunities for offering tailored summaries for these mobile environments, for example, based on location-sensitive profiles. Summarization in these environments can take advantage of hierarchical displays, document keyword extraction, and short summaries with sentence truncation. In certain display environments, word compaction [3] has also been used. However, more powerful sentence compaction strategies based on syntactic and semantic information can also be leveraged [1] [7]. Recent work [2] shows that methods based on a combination of keywords and single-sentence summaries can provide significant improvements in access times in web browsing and number of pen actions, as compared to other schemes.

These efforts represent only the early stages of what promises to become an important area for summarization. However, many challenging problems remain. Spoken summaries are somewhat hampered by the lack of natural prosody in speech synthesis. When speech is the output modality, the summarizer also needs to be especially cognizant of its subject matter, including awareness of terms such as proper names or other fields which are hard to pronounce naturally.

4. SUMMARIZATION EVALUATION

In recent years, there has been a vigorous examination of the issues behind summarization evaluation, motivated by a desire for more cost-effective, user-centered, repeatable evaluations that would offer feedback to developers at any point in their developmental cycle. Two broad varieties of evaluation have been explored. Summaries can be evaluated in and of themselves (an intrinsic evaluation), or in relation to some task (an extrinsic evaluation).

In 1997, the US government conducted a large-scale extrinsic evaluation of summarization systems as part of its Tipster program [10]. In one typical task in this evaluation, each user saw either a source document or a user-focused summary and had to decide if it was relevant to a topic. The results from relevance assessment by 51 subjects of the output of 16 summarization systems showed that subjects could assess relevance from summaries which discarded 77 to 90 percent of the source text, as from full text in almost half the time, with no statistically significant degradation in accuracy. In 2000, the Japanese Text Summarization Challenge evaluation [5] found similar results, finding that subjects could carry out relevance assessment using summaries which discarded 77% of the source in about two-thirds of the time compared to the full-text, without loss of accuracy. They also carried out an intrinsic evaluation that had subjects compare and evaluate single-document summaries produced by systems against reference summaries produced by humans. Here they found that abstracts produced by humans were preferred to human-produced extracts, which in turn were preferred to system-produced summaries.

A more recent intrinsic evaluation has been conducted by the U.S. government under the aegis of the Document Understanding

Conference (DUC)¹. Here, subjects evaluated and compared both single- and multi-document system summaries against human-produced reference summaries. The results are not available at the time of writing, and the evaluation will continue for the next few years. Finally, the National Science Foundation recently sponsored a six-week workshop at Johns Hopkins University [14], where a variety of different evaluation measures were studied in a cross-lingual information retrieval setting. Among the interesting results was the discovery that summaries preserve relevance, so that using a search engine against summaries results in a relevance ranking that is strongly correlated with the ranking obtained by searching against the corresponding full-text documents. Both these evaluations have made available annotated corpora that can be used for training and testing other summarizers.

5. CONCLUSION

These recent developments present many new challenges and opportunities for summarization. Further progress will depend on additional research in the natural language understanding and generation, and the availability of corpora for training summarizers, as well as more precise methods of evaluating progress in summarization on different types of tasks.

As we move into the 21st century, with very rapid, mobile communication and access to vast stores of information, we seem to be surrounded by more and more information, with less and less time or ability to digest it. Summarization offers the promise of helping humans harness the vast information resources of the future in a more efficient manner. Before this promise fully materializes, however, there is more research, in terms of both theory and practice, that must be carried out.

6. REFERENCES

- [1] B. Boguraev, R. Bellamy, and C. Swart. Summarisation Miniaturisation: Delivery of News to Hand-Helds. In Proceedings of the NAACL'2001 Workshop on Automatic Summarization. New Brunswick, New Jersey: Association for Computational Linguistics.
- [2] O. Buyukkokten, H. Garcia-Molina, A. Paepcke. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. The 10th International WWW Conference (WWW10). Hong Kong, China - May 1-5, 2001.
- [3] S. Corston-Oliver. Text compaction for display on very small screens. In Proceedings of the NAACL'2001 Workshop on Automatic Summarization. New Brunswick, New Jersey: Association for Computational Linguistics.
- [4] N. Elhadad and K. R. McKeown. Towards Generating Patient Specific Summaries of Medical Articles. In Proceedings of the NAACL'2001 Workshop on Automatic Summarization. New Brunswick, New Jersey: Association for Computational Linguistics.
- [5] Fukusima, T. and Okumura, M. Text Summarization Challenge: Text summarization evaluation in Japan. In

Proceedings of the NAACL'2001 Workshop on Automatic Summarization, 40-48. New Brunswick, New Jersey: Association for Computational Linguistics.

[6] Goldstein, J., Mittal, V. O., Carbonell, J. G., and Kantrowitz, M. Multi-Document Summarization by Sentence Extraction In Proceedings of the ANLP'2000 Workshop on Automatic Summarization, 40-48. New Brunswick, New Jersey: Association for Computational Linguistics.

[7] Grefenstette, G. Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the blind. In Working Notes of the 1998 Workshop on Intelligent Text Summarization, 111-117. Menlo Park, California: American Association for Artificial Intelligence Spring Symposium Series.

[8] Hahn, U. and Mani, I. The Challenges of Automatic Summarization. IEEE Computer 33, 11, 29-36.

[9] Mani, I. Automatic Summarization. Amsterdam: John Benjamins, 2001.

[10] Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Hirschman, L., Sundheim, B., Obrst, L. The TIPSTER SUMMAC Text Summarization Evaluation: Final Report. MITRE Technical Report MTR 98W0000138, 1998.

[11] Mani, I. and Maybury, M.T. (eds.) Advances in Automatic Text Summarization. Cambridge, Massachusetts: MIT Press.

[12] Merlino, A. and Maybury, M.T. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. In Advances in Automatic Text Summarization, I. Mani and M.T. Maybury (eds.), 391-401. Cambridge, Massachusetts: MIT Press.

[13] Miller, G. WordNet: A Lexical Database for English. Communications of the Association For Computing Machinery (CACM) 38, 11, 39-41.

[14] Radev, D., Blair-Goldensohn, S. and Zhang, Z. Experiments in Single and Multi-Document Summarization Using MEAD. In Proceedings of ACM SIGIR'01 Workshop on Text Summarization. New Orleans, Louisiana, September 13, 2001.

[15] Schiffman, B., Mani, I., and Concepcion, K. Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001), 450-457. New Brunswick, New Jersey: Association for Computational Linguistics.

[16] Waibel, A., Bett, M., Finke, M., and Stiefelhagen, R. Meeting Browser: Tracking and Summarising Meetings. In Proceedings of the 1998 DARPA Broadcast News Workshop.

¹ <http://www-nlpir.nist.gov/projects/duc/>

