EVALUATION: A PERSPECTIVE ON AUTOMATIC ABSTRACTING RESEARCH

Frances.C Johnson
Manchester Metropolitan University

This short article presents a broad overview of automatic abstracting research (and its close relations of text extraction and summarisation)[1]. The motivation is to bring to the fore some of the issues of evaluation prompted by the announcement of the Summarization Conference (SUMMAC) sponsored by DARPA TIPSTER Text Program which represents a step towards contextual user centered evaluations of such technologies. The key issues highlighted are discussed elsewhere[2] and the reader is referred to further readings. Evaluation is an important research agenda item in the advancement of new systems and it is shown that this area of research stands to benefit from the developments in evaluation metrics which are set to take place.

The goal of automatic summarisation or abstracting, in the broadest sense, is to produce a concise representation which effectively conveys the central message of a text to its reader. This non trivial task (seemingly requiring both text understanding and generation) presents many challenges around which researchers from different fields and perspectives have coalesced. From its early days, back in the late 50s, researchers from an Information Science background have sought to determine the extent to which techniques successfully applied to document indexing could be applied to the task of abstract generation. Based on sentence extraction, the approach employs a statistical and/or pattern-matching analysis of text for cues identifying content-indicative sentences to form the basis of an abstract. At a more sophisticated level of processing, some form of [partial] parsing may be employed to deal with text cohesion and coherence, in particular to identify rhetorical relations, thematic progression in text and the resolution or elimination of referring expressions[3]. In general, these methods for summarisation achieve robustness in their ability to handle texts of unrestricted domains, although the readability of the resulting abstract or extract can not be guaranteed. A distinct approach comes from the field of AI and related disciplines which employ computational techniques for discourse understanding and language generation. Typically, these attempt to derive general statements of content using forms of domain knowledge representation such as semantic networks[4]. As an application for the techniques of Natural Language Understanding, the task is seen to provide a demonstration of the system's understanding of a given text based on a model of human text comprehension.

[1] Although distinctions are made between the use of the terms abstracting, summarising and extracting these are not spelt out in detail here since the intention is to present a broad picture of techniques used to identify the central themes of a text.

[2] F.C Johnson. "A critical review of system-centred to user-centred evaluation of automatic abstracting research" paper in preparation.

[3] Comprehensive overviews of the techniques can be found in Paice 1990 , Johnson et al 1993, and Kupiec 1995.

[4] A range of techniques and applications are discussed in the proceedings of workshops on text summarization, 1993 [4 and 1997 [5]

Evaluation of automatic abstracting or text summarisation, in terms of obtaining some measure of the systems' success, is potentially problematic. It is interesting to note, however, that the methods of sentence extraction are prevalent in the recent commercial systems such as Oracle's *ConText*, Mead Data Central's *Searchable Lead* and British Telecom's *Text Summariser*. Furthermore, previous evaluations of these methods indicate that the simple location based heuristics, extracting from particular sections of a text, consistently provides the optimum results (see for example Kupiec, 1995). The central goal of these systems, and thus the remit or motivation for evaluation, is to identify the most important or relevant information in a text. This goal is shared by technologies of automatic abstracting, text summarisation, and information extraction and it is not therefore surprising to find some overlap in the techniques employed. However subtle distinctions can be made which affect the approach taken for evaluation. Information extraction analyses in depth only document sections which may contain information relevant to the slots of templates representing information categories to be extracted. The metrics for comparative evaluation are fairly well established based on the measures for retrieval system performance of recall and precision:where recall equals the fraction of relevant data actually extracted and precision the fraction of information extracted accurately. These (and similar measures of overgeneration and fallout) may also be used with slight modifications for the evaluation of text summarisation where the measures are calculated based on sentence coselection or cointension with some target representation, or summary sentence set. However in the absence of defined extraction criteria, a significant difficulty lies with determining which information is relevant, and equally which detail is irrelevant. For the purpose of abstracting this is exacerbated by the fact that an abstract representing a document's content can be expressed by many valid subsets of sentences and, furthermore, that different pieces of information will be considered relevant by different people with varying individual interests and information needs. Thus, it could be said that this benchmarking exercise provides little indication of the functionality or utility of the summaries produced. That is how effectively the summary conveys the message of the text and assists the reader in their information requirements for the completion of a specific task. Summaries of text content are key aids to resource discovery and access: abstracts are particularly useful when a search returns a number of potentially relevant documents from which the user has to select the most appropriate. Given the developments in communications technology, making widely accessible increasing amounts of information, and its associated tools of retrieval and filtering, to assist the user in finding specific pieces of information, it can be argued that such an oversight is no longer acceptable.

A shift towards evaluation of the functionality of summary forms will require a shift towards investigations of a higher level which aim to obtain a better understanding of the precise contribution of the features and characteristics of a summary which, so to speak, increase the odds that their intended function is effectively served. On the one hand, abstracts function as determinants of retrieval performance: their success defined in terms of identifying relevant texts when used to match against a query. On the other hand, abstracts function as determinants of users' search performance: their success defined in terms of the support provided in allowing relevancy judgements with respect to the users' information need. In general then what is required is some measure of their indicativity and/or informativeness with respect to the individual user and task and some indication of the impact of other characteristics, such as presentational form.

The SUMMAC Conference sponsored by the DARPA TIPSTER Text Program will provide researchers in the field with evaluation metrics to compare the performance of summary types. The proposed approach does not assume a single correct summary but rather provides a measure based on the time taken to make a relevance judgement (does the summary capture the information sought by a user as given in a query) and categorisation decisions (are the predetermined key text concepts captured in the summary). A measure of the value or informativeness of the surrogates is thus determined as a fraction of the relevance judgements made that were the same as those made on seeing the full text. Additional qualitative measures of user preference will also be collected according to a range of acceptability criteria. This is a much welcomed development towards an user-centered approach to evaluation. It is, however, imperative that individual contributions to the development of evaluation metrics continue. In particular, a programme of research is proposed which aims to establish which summary features assist these judgements, and by varying users, tasks and the evaluation dimensions seek to determine under which circumstances a system of abstract generation performs best and why. Only then will the real value of automatic summary or abstract generation, possibly varied in content and format, be realised at a time when, more than ever, users stand to benefit from the summary presentations of information.

References

1. Johnson, F.C., Paice, C.D., Black, W.J., and Neal, A.P. (1993) The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management*, 1(3): 215-242. (Reprinted *In*: Karen Spark Jones and Peter Willett (Eds). *Readings in Information Retrieval*. San Francisco: Morgan Kauffman Publishers, 1997, 538-553)

2. Kupiec, Julian., Pedersen, Jan. and Chen, Francine. (1995) *A trainable document summarizer*. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle July 9-13, 68-73

3. Paice, C.D. (1990) Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1): 171-186.

4. Summarizing text for intelligent communication, Seminar Program, December 1993 http://www.phil.uni-sb.de/FR/Infowiss/Abstract/program.html

5 Intelligent scalable text summarization. Proceedings of a workhop sponsored by the Association for Computational Linguistics. Madrid, Spain, 11 July, 1997.

6 DARPA TIPSTER Text Program. http://www.tipster.org