Hypertext Summary Extraction for Fast Document Browsing

Kavi Mahesh

Computing Research Laboratory (3CRL)
New Mexico State University
Las Cruces, NM 88003-8001 USA
mahesh@crl.nmsu.edu
http://crl.nmsu.edu/users/mahesh/papers/hyper-summary.html

Abstract

This article describes an application of Natural Language Processing (NLP) techniques to enable fast browsing of on-line documents by automatically generating Hypertext summaries of one or more documents. Unlike previous work on summarization, the system described here, *HyperGen*, does not produce plain-text summaries and does not throw away parts of the document that weren't included in the summary. HyperGen is based on the view that summarization is essentially the task of synthesizing Hypertext structure in a document so that parts of the document "important" to the user are accessible up front while other parts are *hidden* in multiple *layers of increasing detail*. In fact, HyperGen generates short descriptions of the contents and rhetorical purposes of the hidden parts to label the Hypertext links between the summary and the different layers of detail that it generates. A prototype HyperGen system has been implemented to illustrate the techniques and demonstrate its usefulness in browsing World Wide Web documents.

1. Summarization as Hypertext Generation

There is no correct algorithm for summarization. What belongs in a summary extracted from a document is determined by the needs, likes, and stylistic preferences of users and the tasks they are performing using the summary. Why then should a summarization system decide what to include in a summary and what to throw away? On the other hand, not throwing away any part of the document defeats the very purpose of summarization, namely, fast browsing of documents. It seems that in general the best way to summarize a document is to assign to each piece of the document a level of interestingness (or importance). The most interesting pieces should be most easily accessible to the user and can be considered to constitute the summary of the document. Less interesting (or more detailed) pieces should still be accessible but hidden behind more interesting ones in multiple layers of detail. In this view of summarization, each layer can be considered a summary of the more detailed layer immediately behind it.

Linking the different pieces and layers together in a Hypertext structure enables the user to navigate to the different parts of the document as desired. Labeling such Hypertext links with a short phrase or two indicates to the user what lies behind a link. In HyperGen, a prototype system that generates Hypertext summaries of English and Spanish texts, these labels are in fact very short summaries of the pieces of text that lie behind them. Each label summarizes the content and/or the perceived rhetorical purpose (e.g., illustration, historical background, comparison, explanation, etc.) of one or more pieces of text. Users can "open" a hidden piece and "zoom into" any part of the document if its label suggests matter of interest to them.

Hypertext links in a HyperGen summary need not be limited to other parts of the same document. Such links can in fact function as cross-links to related parts of other documents, even those in other languages. Such cross-links are generated by HyperGen using keywords that it recognizes while summarizing the documents involved.

HyperGen has been developed under the belief that a summarization system that (i) runs much faster than reading the full document, (ii) is robust, (iii) portable, (iv) richly customizable, and (v) provides excellent aids for the user to navigate through the difference parts of a document, will be useful in practice even if the summaries are less than ideal. HyperGen is intended for document browsing scenarios over any large collection of docu-

ments, such as an on-line library or the World Wide Web, whether the documents themselves are in plain or Hypertext. Current Web browsers and search engines merely follow predefined Hypertext links and retrieve entire documents by searching for keywords in pre-computed index databases. HyperGen goes beyond this technology and dynamically constructs Hypertext presentations of documents at multiple levels of detail, from brief summaries to entire documents. This view of summarization as a means for document visualization through a mere manipulation of Hypertext links between different parts of documents promises to deliver the full potential of Hypertext document collections such as the WWW to the large community of Internet users.

2. Example

Figure 1 shows a Hypertext summary extracted by HyperGen. The Web browser screen on the left shows the main summary along with automatically generated keywords in the document and several labels in between paragraphs that have Hypertext links to parts of the document not included in the summary. The paragraphs should in fact be considered as "callouts" ("highlights" or "pullquotes") extracted from the document. The labels, Hypertext links, and the separate files for the hidden parts are all automatically generated by HyperGen. The part of the document hidden behind the first label "The Dealer Network Option" is shown on the right half of Figure 1. It can be seen that this label is in fact a section heading in the document. Other labels, such as "Rank Xerox's expertise...the chaos..." are automatically extracted from one of the "most interesting" sentences in the hidden part.

Some of the labels are in fact links to large chunks of the document which often contain several subsections. These chunks are in fact summarized again by HyperGen to create multiple layers of intermediate detail. HyperGen is currently being extended to produce labels describing the rhetorical purpose of a piece of text. For example, rhetorical categories are identified and presented using phrases such as "comments," "analysis," and "example." Optionally, any multimedia elements in the document will also be retained by HyperGen in the appropriate parts of the Hypertext summary.

3. Generating Hypertext Summaries

The core summarization problem is taking a single text and producing a shorter text in the same language that contains all the main points in the input text. Figure 2 shows the basic architecture of the HyperGen summarization engine.

HyperGen has adopted a robust, graded approach to building the core engine by incorporating statistical, syntactic, and document-structure analyses among other techniques. This approach is less expensive and more robust than a summarization technique based entirely on a single method. HyperGen is designed in such a way that as additional resources, such as lexical and other knowledge bases or text processing and machine translation engines, become available, they can be incorporated into HyperGen's architecture to incrementally enhance its capabilities and improve the quality of summaries. For example, there is no morphological analyzer (or stemmer) in the current implementation of HyperGen. If available, it can be added easily as shown in Figure 2.

Some of the main modules in HyperGen are (i) document structure analysis, (ii) sentence selection, (iii) sentence simplification, (iv) summary construction, and (v) user customization.

3.1. Document Structure Analysis

Document structure analysis is important for extracting the topic of a text (Paice and Jones, 1993; Salton and Singhal, 1994; Salton, et al, 1995). In such an analysis, for example, titles and sub-titles would be given a more important weight than the body of the text. Similarly, the introduction and conclusion for the text itself and for each section are more important than other paragraphs, and the first and last sentence in each paragraph are more important than others. The applicability of these depends, of course, on the style adopted in a particular domain, and on the language: the stylistic structure and the presentation of arguments vary significantly across genres and languages. Structure analysis must be tailored to a particular type of text in a particular language. Document structure analysis in HyperGen involves the following subtasks:

- **Document Structure Parsing:** HyperGen assumes that documents have HTML markers for headers, section headings, and paragraph breaks. It separates the title, sections and subsection headings, and other data and graphics from paragraphs of text.
- **Sentence Segmentation:** HyperGen breaks each paragraph into sentences by looking for sentence boundaries. It uses a stop list to deal with special cases such as certain uses of '.' in abbreviations.

0069: A joint venture between the Xerox Corporation....

Date: December 27, 1990, Thursday, Late Edition – Final
Source: Copyright (c) 1990 The New York Times Company

Keywords in Document: state owned Rank Xerox Europe Eastern

The Dealer Network Option

A joint venture between the Xerox Corporation and the Rank Organization of Britain, Rank Xerox sells Xerox photocopiers outside the Americas and has been involved in Eastern Europe since 1964.

Rank Xerox's expertise ...the chaos ...now

In the year ended Oct. Rank Xerox sold about $50 million in equipment to Eastern Europe and about $50 million to the Soviet Union.

Czechoslovakia, ...where ...there ...a mere

Rank Xerox estimates that it has about 70 percent of the market in the Soviet Union and about one-third in Eastern Europe, where its share ranges from 10 percent in Hungary to about 50 percent in Romania.

Plant in Soviet Union

With big state-owned companies being broken up throughout Eastern Europe and with small businesses expected to multiply rapidly, Rank Xerox -- like other Western companies -- is also trying to create a network of dealers and service organizations to serve the newly decentralized market.

Room for Improvement

"With the secret police gone," Mr. Leimgruber said, "contact with Western companies is no longer on the no-no list."

Hidden Subpart 1

Back to the Summary

Nobody has to tell Sepp Leimgruber how difficult it is to do business in a land that is in between Communism and capitalism. He has two choices for selling Xerox photocopiers in Czechoslovakia, and as far as he is concerned, both are rotten. The first is to continue doing business with the two crumbling state-owned companies that had held the monopoly for distributing office equipment under the Communists.

Given how bureaucratic and inefficient they have always been, Mr. Leimgruber, the area director of Hungary and Czechoslovakia for Rank Xerox Ltd., would not ordinarily mourn their demise.

The Dealer Network Option

The second choice -- the one he would prefer -- is to create a network of independent dealers.

The problem, he said, is that the only people with the necessary capital are those who were corrupt Communist officials or those involved in money-changing or other black-market activities. "Nobody but the big crooks have the money," Mr. Leimgruber said, "and the question is whether you as a respectable company want to do business with the crooks."

Figure 1. A Hypertext summary generated by HyperGen is shown on the left. On the right is one of the parts of the document hidden under the label "The Dealer Network Option." The document is 200 lines long.

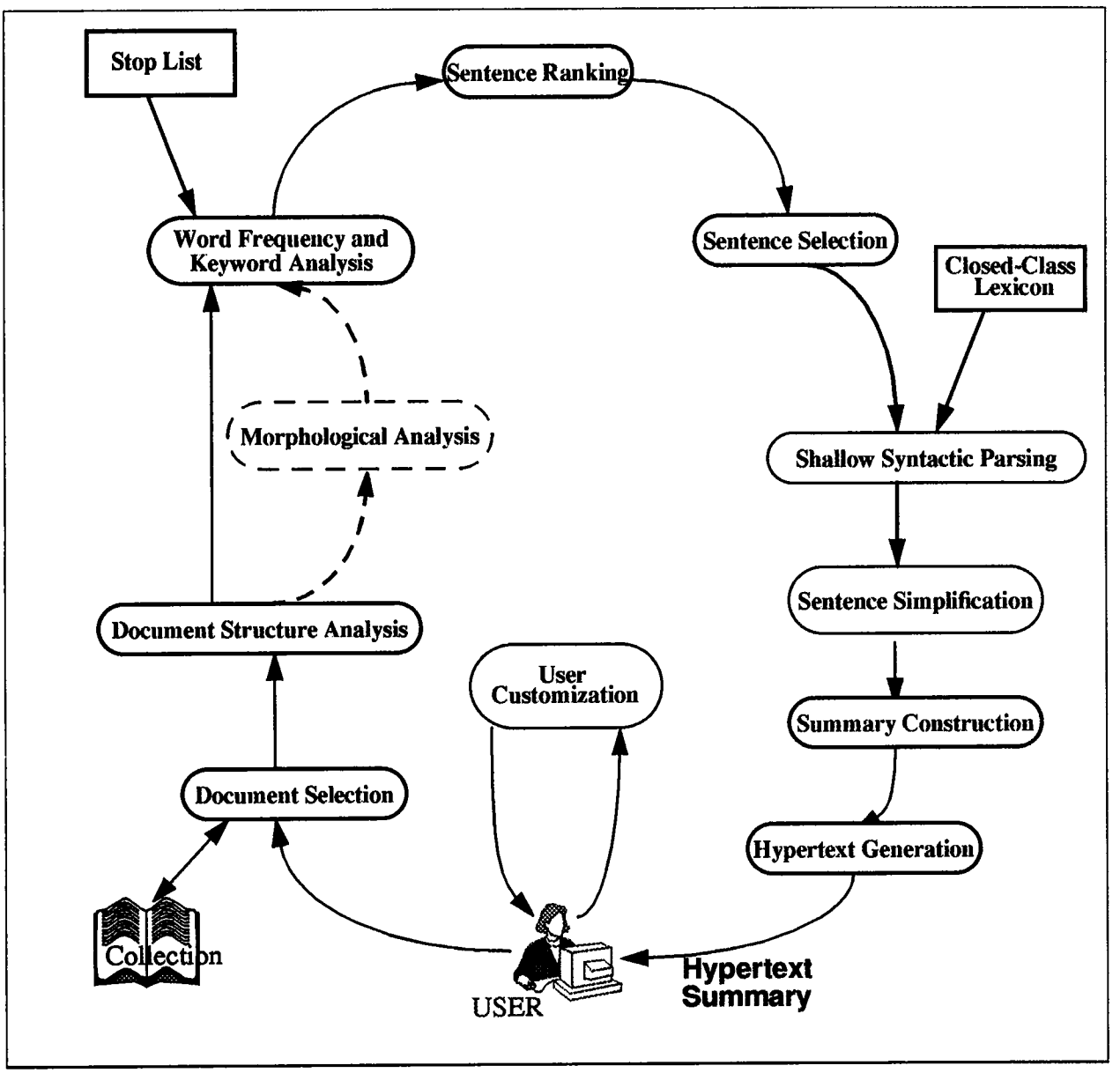

Figure 2. Architecture of HyperGen showing the basic steps in extracting the Hypertext summary of a document. Note that only selected sentences are parsed.

- **Text Structure Heuristics:** HyperGen uses heuristic rules based on document structure to rank the sentences in a document (see below). For example, different scores are assigned to first sentences of paragraphs, to single-sentence paragraphs, and so on.

3.2. Sentence Selection

In order to allow a multitude of techniques to contribute to sentence selection, HyperGen uses a flexible method of scoring the sentences in a document by each of the techniques and then ranking them by combining the different scores. Text-structure based heuristics provide one

way to rank and select sentences in a document. Additional methods included in HyperGen are described below.

3.2.1. Word Frequency Analysis

The basic technique is to sort the words in the document by frequency of occurrence within the document and select a few of the most frequent content words (i.e., words other than articles, prepositions, conjunctions and other closed-class words). Sentences containing those words get a score increment. Supporting processes needed for word frequency analysis include:

- **Morphological Analysis (Optional):** Statistical analysis works more reliably on a text which has been morphologically processed to recognize the same word with different inflections. The current implementation of HyperGen uses simple string matching for counting and comparing words.

3.2.2. Corpus Statistics

Word frequency analysis can be enhanced by counting word frequencies across an entire corpus instead of just within a document. Words that have significantly higher frequencies within the document relative to the whole corpus are likely to indicate significant parts of the document. Such techniques will be integrated into future enhancements of HyperGen.

3.2.3. Keyword and Keyword Pattern Analysis

Preset or user-specified keywords from either the domain or to focus on particular style can be used to introduce elements of targeted summarization. Keyword-based ranking and selection has already been implemented in HyperGen. This will be further developed in the future to accept more expressive keyword patterns. The new capability will enable other query-based IE/IR modules to be integrated with HyperGen.

Open-class words in document titles and section headings are also treated as keywords. Sentences in each section containing such keywords are given score increments to boost their chances of being selected for inclusion in the summary.

Finally, sentences in the document are ranked based on the scores returned by all of the above techniques. Highest ranking sentences are selected to constitute a summary of a given length (which is either a default or set by the user).

3.3. Sentence Simplification

Sentences selected for inclusion in the summary are often lengthy and can be simplified to further shorten the summary. HyperGen has implemented a novel "phrase dropping" (or parse-tree pruning) algorithm based on phrase-structure heuristics for English and Spanish. For example, it drops embedded clauses and right-branching prepositional adjuncts. Figure 3 shows an example of sentence simplification.

This method requires a robust, shallow parser, at least a partial one that identifies phrase boundaries, with good coverage. The simplification module is currently being integrated with the sentence selection module so as not do drop parts that were deemed important by the selection scores. The summary shown in Figure 1 was produced without sentence simplification.

Shallow Syntactic Parsing. A shallow syntactic analysis helps not only sentence simplification but also the recognition of important sentences and extraction of semantically relevant parts of these sentences. For example, labels for Hypertext links are generated by extracting several noun phrases from the highest ranking sentence in the hidden part attached to the link. Syntactic analysis is also used in HyperGen to recognize the rhetorical purposes of texts. A very simple parser has been implemented for English using only a closed-class English lexicon. This lexicon merely specifies the part of speech of closed-class words and a few stop words. The parser can correctly tag many noun, verb, and prepositional phrases in texts. HyperGen is very fast in spite of using a parser, since the parser runs only on selected sentences.

On Thursday, **A group of bankers will assemble in Hong Kong to sign a @400 million 12-year loan agreement to help fund local tycoon Gordon Wu's dream of building a 120-kilometre, six-lane "super highway."** from the colony to Guangdong, capital of the neighbouring Chinese province of Guangzhou, formerly Canton.

Figure 3. Sentence simplification example showing the retained parts in boldface and the dropped parts in a smaller font.

3.4. Summary Construction

Using the simplified parts of selected sentences, a summary is constructed by extracting the corresponding parts of the source document. A number of issues such as capitalization and other punctuation must be addressed to render the summary and make it readable, especially when sentences have been simplified by dropping some of their parts. In addition, the document title (if any) and information about the source and date of the document are presented at the beginning of a summary. Several keywords (frequent words or user specified keywords) are also presented at the top of the summary. Future versions of HyperGen will highlight occurrences of the keywords in the body of the summary or the hidden parts and provide cross-links to other documents containing the same keywords.

3.4.1. Hypertext Generation

Each contiguous piece of unselected text is written into a separate HTML file with Hypertext links to the appropriate point in the summary (or the next layer in front of it, in general). Labels are generated for these links as follows: if there is a section heading for the hidden part, use it as the label; otherwise, select the highest scoring sentence(s) in the hidden part, parse them, simplify them by dropping phrases and if they are longer than a preset maximum length, pick any keywords or other noun phrases in them to construct a label of length less than the maximum length. These labels are in a sense summaries of the hidden parts and are intended to indicate the main topic (or theme) of the hidden part of the document.

If a hidden part is longer than a (user customizable) threshold or has multiple sections, then the part is summarized again by Hypergen to create an intermediate layer of summary with hidden subparts behind it. Thus the document is ultimately broken into individual paragraphs where the contents of each unit is summarized by a label that links it to other pieces and subpieces of the document in the Hypertext summary.

Current work on HyperGen is developing algorithms for identifying the rhetorical purpose of a paragraph by matching its sentences against a glossary of syntactic patterns that act as keys for the rhetorical purposes. The glossary contains syntactic patterns for phrases such as "consider x," "illustrated by x,"

Apart from Hypertext links between a summary and different parts of a document, future versions of HyperGen will generate additional navigational aids by matching keywords and key topic areas across documents in a collection. Results will be presented in the Hypertext summary by providing an overall set of links to related documents as well as links to related documents based on each keyword. Some such cross-links may in fact be between documents in different languages (if the keywords can be translated by a machine translation module).

3.5. User Customization

Users can customize Hypergen in the following ways:

- setting the length of the summary: relative to the length of the document or in absolute terms (number of characters or sentences)
- specifying keywords to use: e.g., "joint venture"
- specifying the number of frequent or title words to find
- controlling sentence ranking heuristics: by adjusting the levels of preference for frequent words, keywords, title words, first sentence in a paragraph, etc.
- extending the stop list: by specifying words to be ignored

In essence, all aspects of HyperGen's summarization behavior can be controlled and customized by a user through a friendly graphical interface. Future versions of

HyperGen will allow users to customize rhetorical purpose identification by adding new phrase patterns to look for or by telling HyperGen to focus on particular rhetorical types.

4. HyperGen: The System

HyperGen has been implemented entirely in Java and can be used in conjunction with a Web browser. It is fast, robust, modular, portable, essentially multilingual, and simple (e.g., does not require complete lexicons, full parsers, etc.). It has been tested on several English and Spanish news articles. No usability study has been conducted so far to determine if HyperGen summaries are in fact useful for Web browsing or other document filtering applications. A new version of HyperGen with rhetorical-purpose labels will be developed and demonstrated during the symposium. Results from testing HyperGen on larger collections of documents and preliminary usability studies will be reported during the symposium if possible.

5. Related Work

Previous work in summarization (also called automatic abstracting) has addressed primarily the simpler problem of producing a plain-text summary of a single document in the same language (Cohen, 1995; Luhn, 1958; Paice, 1990; Pinto Molina, 1995; Preston and Williams, 1994; Salton et al., 1994). Document retrieval and classification efforts, on the other hand, have produced a multitude of techniques for selecting a subset of an entire collection of documents. However, they simply present the retrieved documents in their entirety, providing little support for quickly digesting the contents of an individual document. Information extraction (Cowie and Lehnert, 1996) systems have assumed that what is of interest to users is known a priori in the form of templates. Most work in natural language generation has focused on generating summaries of data, not texts (Kalita, 1989; Kukich, et al, 1994; McKeown, et al, 1995; Robin, 1994). Work in machine translation has assumed for the most part that source texts must be translated in their entirety. A key drawback of the above research is that they assumed that a document must be either processed (e.g., translated, retrieved, etc.) in its entirety or that it must be summarized while throwing away the unselected parts. By bringing some of these areas together, HyperGen has attempted to exploit NLP techniques to improve the utility of Web browsers. Since there is no algorithm for summarization that guarantees a summary that always meets the needs of a task (i.e., never misses on a piece of information that is of interest to the user or is relevant to the task on hand), it is critical for such a system to indicate to the user the contents of the unselected parts of a document. HyperGen provides this new functionality.

6. Future Work: Further Uses of NLP

This section briefly outlines additional applications of NLP techniques to improve the quality of HyperGen's summaries. The core summarization method outlined above has the advantage of simplicity, does not require a language generator, and delivers grammatical and fairly readable summaries even in untargeted summarization situations. However, it suffers from several problems which lead to important research issues:

- Pronouns, reference, anaphora, etc.: When the selected sentences contain pronouns or other references, anaphora, or ellipsis, the required context may be missing in the summary, thereby hurting its readability and ease of understanding. HyperGen needs simple reference resolution techniques to overcome the problems and heuristics for avoiding such problems by not selecting sentences that are beyond the capabilities of the resolution techniques.
- Poor flow in summary: Since the summary is put together by conjoining different pieces of the source text, it is not likely to have a good flow. However, this may not be a critical problem for HyperGen since it does not produce a plain-text summary. HyperGen summaries are inherently Hypertexts and are interspersed with hyperlinks to other pieces of documents.

Further applications of NLP techniques may yield benefits in several areas:

Ontology-based content classification: Future versions of HyperGen will attempt to classify the topic (or "theme") of a piece of text using a broad coverage ontology, such as the Mikrokosmos Ontology (Carlson and Nirenburg, 1990; Mahesh and Nirenburg, 1995; Mahesh, 1996). Words in a text will be mapped to ontological concepts using a semi-automatically acquired mapping from WordNet entries to concepts in the Mikrokosmos Ontology. An appropriate clustering algorithm will be used to select the main concept from among the concepts to which words in a piece of text are mapped. Such a method is expected to yield better labels that summarize hidden parts than current ones often generated by simplifying a sentence in the hidden text.

Translation issues: A glossary-based (or other fast and simple) translation engine can be used to translate necessary parts of documents in other languages to English. The entire document should be translated only if demanded by the user. Different methods for interleaving summarization and translation operations can be explored to minimize translation efforts. For example, the source text may be summarized in the source language and the summary translated to English. Unselected parts may sometimes need to be translated to English before labels can be generated to identify and describe their topic areas and rhetorical purposes.

Self-evaluation: Every decision made by HyperGen is based on scores assigned according to a variety of user-customizable parameters. Self-evaluation can be added to this design by attaching simple evaluation procedures to each of the parameters so that a measure of the system's confidence is determined whenever a decision is made and a score is assigned along some parameter. A simple algorithm can be developed for combining individual measures of confidence so that an overall measure can be presented to the user. Such confidence measures can be integrated into the labels and summaries by generating English phrases such as "certainly about," "appears to be," or "not sure about" to express HyperGen's evaluation of its own summarization decisions.

7. Conclusions

There is an overload of on-line documents on the users of the Internet. Current search engines and browsers are limited in their capabilities to ease this overload and to help users quickly find and digest the information they need from the large collection of arbitrarily irrelevant documents available to them. The HyperGen system described in this article attempts to remedy this situation by applying NLP techniques to enhance the capabilities of a Web browser. HyperGen automatically constructs Hypertext summaries of documents so that users can see interesting subparts of documents in a summary up front and, at the same time, see labels summarizing the parts that are hidden behind the main summary in layers of increasing detail. The author firmly believes that these summaries are more meaningful and useful than the first few lines of documents typically displayed in the results from a state-of-the-art Net search engine and that these capabilities cannot be achieved without the use of NLP techniques.

Acknowledgments

I would like to thank Sergei Nirenburg and Rémi Zajac for many insightful comments and discussions.

References

Carlson, L. and Nirenburg, S. (1990). "World Modeling for NLP." Center for Machine Translation, Carnegie Mellon University, Tech Report CMU-CMT-90-121.

Cohen, J.D. (1995). "Highlights: Language- and domain-independent automatic indexing terms for abstracting." Journal of the American Society for Information Science, 46(3):162-174.

Cowie, J. and Lehnert, W. (1996). "Information Extraction." Communications of the ACM, special issue on Natural Language Processing, January 1996

Kalita, J. (1989). "Automatically generating natural language reports." International Journal of Man-Machine Studies, 30:399-423.

Kukich, K., McKeown, K., Shaw, J., Robin, J., Morgan, N., and Phillips, J. (1994). "User-needs analysis and design methodology for an automated document generator." In Zampolli, A., Calzolari, N., and Palmer, M., editors, Current Issues in Computational Linguistics: In Honor of Don Walker. Kluwer Academic Press, Boston, MA.

Luhn, H.P. (1958). "The automatic creation of literary abstracts." I.B.M. J. of Res. Dev., 2(2):159-165.

Mahesh, K. and Nirenburg, S. (1995). "A situated ontology for practical NLP." In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, August 1995.

Mahesh, K. (1996). "Ontology development for machine translation: Ideology and methodology." Technical Report, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

McKeown, K., Robin, J., and Kukich, K. (1995). "Generating concise natural language summaries." Information Processing and Management (special issue on summarization).

Paice, C.D. and Jones, P.A. (1993). "The identification of important concepts in highly structured technical papers." Proc.of the 16th ACM SIGIR conference, Pittsburgh PA, June 27-July 1, 1993; pp.69-78.

Paice, C.D. (1990). "Constructing literature abstracts by computer: Techniques and prospects." Inf.Proc. & Management 26(1), 171-186.

Pinto Molina, M. (1995). "Documentary abstracting: toward a methodological model." Journal of the American Society for Information Science, 46(3):225-234.

Preston, K. and Williams, S. (1994). "Managing the information overload: New automatic summarization tools are good news for the hard-pressed executive." Physics in Business, Institute of Physics, June 1994.

Robin, J. (1994). "Revision-based generation of natural language summaries providing historical background: Corpus-based analysis, design, implementation and evaluation." Ph.D. Thesis, Computer Science Department, Columbia University, Technical Report CUCS-034-94.

Salton, G., Allen, J., Buckley, C. and Singhal, A. (1994). "Automatic analysis, theme generation, and summarization of machine-readable texts." Science, 264:1421-1426.

Salton, G., Singhal, A., Buckley, C., and Mitra, M. (1995). "Automatic text decomposition using text segments and text themes." Technical Report, Department of Computer Science, Cornell University, Ithaca, NY.

Salton, G. and Singhal, A. (1994). "Automatic text theme generation and the analysis of text structure." Technical Report TR94-1438, Department of Computer Science, Cornell University, Ithaca, NY.