

# EXCERPT - a Within-Document Retrieval System Using Summarization Techniques

*Jürgen Reischer*

Information Science  
University of Regensburg  
93040 Regensburg  
juergen.reischer@sprachlit.uni-regensburg.de

## **Abstract**

A new approach to within-document retrieval is presented using methods of automatic summarizing. Segments of sentence to paragraph size can be retrieved, assembled to greater passages and ranked by informativeness. Parameters like coherence and cue words are used for scoring relevant passages, which are extracted according to user needs.

## **1 Introduction**

In view of the ever growing information jungle, techniques for the fast and accurate access to the most important contents of textual information are indispensable today. For that purpose, basically two methods are employed in current systems: (1) the automatic extraction of important information from texts in the sense of an extractive summary, or (2) the access of text passages using the standard 'find' or 'search' command, listing the corresponding text sections that contain the desired search terms. Method (1) mostly provides a representative summary ignoring the user's special interests or needs (expressed by a query); method (2) simply lists sections of text containing the search terms without rating and ranking of the passages found. Although the user does not only want to search in document collections but also in the documents themselves, surprisingly few genuine within-document retrieval systems and techniques exist today which exceed the two approaches mentioned

above. As a realistic scenario, imagine a WWW browser, a PDF reader, or a word processor providing a search interface similar to a search engine, where one or more search terms can be entered to retrieve relevant and ranked passages.

Although passage retrieval has gained some attention in the literature, it is not a fully explored subdiscipline of information retrieval despite its potential to drastically improve information access for the user. In many cases, passage retrieval is realized as an extension of the document retrieval process, neglecting the fact that it should be considered as a task in its own right: the user may want to scan *any* (stand-alone) text for the most relevant information, not only documents retrieved from a huge collection like the WWW. The benefit of a within-document retrieval tool is the possibility to access the most informative passages of a text, i.e. extract the most important or interesting information per se (in summary mode without search terms) or about some topic X (in retrieval mode given some search terms).

In the following sections, I will first give a short overview of existing tools and techniques in the discipline of within-document (passage) retrieval and summarization. After that, the EXCERPT<sup>1</sup> system is described in more detail, which implements the above ideas.

## 2 Overview

Two kinds of systems will be described in more detail below: passage retrieval and summarization systems.

### 2.1 Passage Retrieval Systems

Genuine systems for within-document retrieval are rare. Most applications in this area are passage retrieval systems serving one of the following purposes [cf. Stock 2007: 498 ff.]:

---

<sup>1</sup> Expert in Computational Evaluation and Retrieval of Passages of Text.

- retrieval of passages, which best match some search terms; mostly employed as a second step after document retrieval to improve precision (e.g. if only some passages are highly relevant to a query);
- ranking of passages of longer documents for fast scanning of the most relevant sections (e.g. in order to decide whether it is worth reading the entire document);
- retrieval of passages for question answering tasks where search terms are extracted from a question expressed in natural language (e.g. to extract a passage which contains the answer with high probability).

Within-document retrieval is most similar to the first and second point: search in an isolated document given some search terms to retrieve passages of any length (sentences, paragraphs, chapters etc.). Note that it is not necessary that the document was retrieved from a collection before.

A simple variant of passage retrieval is the 'Find' or 'Search' command available in almost every system that presents text, e.g. a browser, an editor, or a PDF reader. A more elaborated tool is described in [Harper & al. 2004] called 'ProfileSkim': This tool is designed to create a relevance profile of a document (according to a query), which is output as an interactive bar graph. With respect to passage retrieval, either within the context of document retrieval or question answering, several systems exist which search a document given a query (for an overview see [Lee & al. 2001]). Other systems for the extraction of passages to be mentioned here are [Salton & al. 1993], [Barker & al. 1998], [Dunlavy & al. 2007]; [Tellex & al. 2003] give a general overview of passage retrieval algorithms for question answering.

## **2.2 Summarization Systems**

Summarizing systems are more established than within-document retrieval tools and approaches (see [Mani & Maybury 1999] and [Endres-Niggemeyer 1998] for an overview). Basically, the summary of a text may be an extract or abstract, the latter restating the most important points of the text contents in own words, the former using original material from the text. Another basic distinction concerns the kind of content presentation: an informative summary should be a stand-alone abbreviation of the most relevant or important/interesting information contained in the full text, an indicative summary should

give the user just enough information to decide whether it is worth reading the original article (cf. [Borko & Bernier 1975] and [Spärck Jones 2007] for an overview and further distinctions).

Several strategies are employed for automatic summarization of texts: discourse-oriented approaches exploiting the structure of a text (e.g. [Barzilay & Elhadad 1999]), corpus-based systems using statistical or linguistic information from a text (e.g. [Hovy & Lin 1999]), or knowledge-based approaches using linguistic and world (domain) knowledge for the compression of documents (e.g. [Hahn & Reimer 1999]). For the purpose of summarization by extraction of text passages, several parameters are used: keywords based on frequency count using a variant of tf•idf and/or appearance within a heading or the title, cue words defined by the system or user (e.g. bonus or malus/stigma words), sentence/text coherence structure where terms or sentences are the more central or important the more semantic links they share, sentence position within a paragraph or a text where initial sentences are to be rated higher, and some other parameters (for a collection see [Edmundson 1969], [Kupiec & al. 1995], [Teufel & Moens 2002], [Hovy 2004]).

### **3 The EXCERPT System**

The EXCERPT system introduced below is a conceptual successor of the Ival system described in [Reischer 2007]. While Ival was a concept-based system suffering from the word sense disambiguation problem for lexical chaining, EXCERPT is a purely term-based system exploiting the text's coherence structure instead. Furthermore, linguistic analysis was improved, e.g. with respect to synonym detection (see below).

#### **3.1 Text Analysis**

Before retrieval operations can be executed by the user, the text has to be analyzed and transformed into an internal representation. Several analysis modules are implemented for this task in EXCERPT:

- *Term normalization:* After sentence detection, the terms of every sentence are normalized by stripping off inflectional morphemes as provid-

ed by WordNet [Fellbaum 1998]. If a normal form (stem) is not present in WordNet, it is decomposed; if it is not decomposable, it is considered as a neologism. Function words as well as frequent and 'void' terms ("Mr." etc.) are excluded. From the remaining forms, a term-sentence index is created with information about the occurrence(s) of every term in the sentences of the text.

- *Multiword term detection*: The system assembles single terms belonging to a multiword term which then replaces the sequence of original terms. For example, the expression "Internet Explorer" is *one* term in WordNet which is more specific than the single terms.
- *Synonym detection*: In a final step, the system collects potential synonyms to further improve precision of linguistic analysis. For example, a text by [Pournelle 2005] uses the synonyms "Internet Explorer", "Explorer", and "IE" interchangeably. If the alternative forms "Explorer" and "IE" would not be matched with their full form "Internet Explorer", the accuracy of term frequency count would be decreased.

After this preprocessing stage, the text is analyzed for several linguistic parameters, which are the basis of sentence scoring for within-document retrieval. Each sentence is scored for its informativeness (importance, interestingness) irrespective of a certain query. Below, several parameters used so far in the scoring process are described in more detail:

- *Coherence*: The thematic connectedness of a text proves itself in the coherence structure of sentences and paragraphs. For a sentence or paragraph to be coherent, the terms must be semantically linked to each other (cf. [Halliday & Hasan 1976], [Hoey 1991], [Barzilay & Elhadad 1999], [Vechtomova & al. 2006]). For example, if one sentence contains the term "browser" and another the term "Internet Explorer", both are connected by a hypernymy/hyponymy relation. The more semantic relations exist between two passages, the more thematically coherent they are. The automatic detection of text coherence uses the WordNet thesaurus, which contains relations like synonymy, antonymy, hypernymy/hyponymy, holonymy/meronymy, etc. The coherence structure of a passage can be used as a measure of its *importance*: the more semantic links it has to other passages, the more important or central it is (cf. [Mihalcea 2004], [Mihalcea & Tarau 2004]; [Brin & Page 1998]). This implies that a passage containing many topical (frequent)

terms also shows many connections.

- *Cue words*: Some types of words appear more often in summaries than in full texts. For example, comparatives and superlatives as well as adverbs of conclusion ("thus", "therefore") and indefinite articles seem to be good indicators of informative sentences (comparisons like "X is *better than* Y" express a proposition of interest, indefinite articles indicate the introduction of new entities in the discourse). Contrary to these 'bonus words', 'malus/stigma words' are more often found in sentences not to be included in a summary, e.g. indefinite and 1st/2nd person pronouns (cf. [Edmundson 1969], [Goldstein & al. 1999]).
- *Specificity*: Infrequent terms are more informative than frequent ones, as also expressed by  $tf \cdot idf$ . This includes rare and neologistic terms expressing specific and novel concepts. Such concepts are more interesting to the reader than every-day terms appearing in most texts. A further indication of specificity is the number of meanings assigned to a term: monosemous terms are more specific than polysemous ones, where the latter must be regarded as a disjunction of all its potential meanings; in addition to that, polysemous terms are generally more familiar (i.e. less novel) to the reader than monosemous (cf. [Jastrzembski 1981], [Tengi 1998]).
- *Novelty*: An additional parameter used here concerns the information structure of a text. Sentences containing a high ratio of newly introduced terms in the text, e.g. at the beginning of a new thematic section, are rated higher than sentences just repeating already mentioned terms. This scores down sentences with redundant (repeating) material.

The parameters described above are actually implemented in the EXCERPT system and already perform quite well in the sentence scoring and extraction process (see 3.3). In order to compute a total score involving all possible parameters, a separate sentence ranking list for every parameter set is created (e.g. for bonus and malus words alone). The ranking index positions of the sentences in every parameter list are regarded as quasi-normalized scores, which then can be computationally merged with other index scores.

## 3.2 Within-document Retrieval

The following sections describe the process of retrieving and ranking passages which are then assembled to greater textual units.

### 3.2.1 Retrieval and Ranking

After a text has been analyzed, the user may retrieve passages by entering one or more search terms or create an extractive summary instead by executing an 'empty query' (no search term). If search terms are provided, they are analyzed like input text, i.e. the terms are normalized with multiword term detection. Every search term is matched against the text by using the term-sentence index resulting in a list of all sentences where one or more search terms could be found. By use of the WordNet thesaurus, the matching process can be optimized for recall or precision: in the latter case, only exact matches (of normalized forms) select a sentence containing the term; in the former case, also semantically related terms create a match (e.g. "Internet Explorer" and "IE", "IE" and "browser" etc.). Non-exact matches generally receive a lower score than exact matches depending on the semantic distance of the two terms in the WordNet thesaurus (e.g. identical terms have distance 0, synonyms distance 1, hypernyms/hyponyms distance 2).

The overall set of retrieved sentences is then sorted (ranked) in three steps:

- for every different number  $N$  of search terms found in any sentence, a separate list  $L_N$  of sentences containing  $N$  search terms is created;
- within every such list  $L_N$ , the sentences are sorted for their matching distance score (exact matches with distance 0 are favoured against inexact matches), where sentences of the same score  $S$  are again assembled to a separate list  $M_S$  containing all sentences with score  $S$ ;
- finally, within every  $M_S$ , the sentences are ranked for informativeness as calculated by the parameters described above.

In other words: sentences are sorted (ranked) primarily for the number of matching search terms with direct matches preferred over indirect matches; secondarily and finally, sentences are ranked for importance and interestingness scores as calculated from coherence structure, cue words, specificity, and novelty (see 3.1).

The advantage of this incremental sorting/ranking process is twofold: First, sentences matching best with a query, i.e. have the most (exact) search terms in common, are scored highest and presented first; second, if many sentences are scoring equal (are within the same  $M_s$ ), especially if a summary is requested where all sentences have no primary score, sentences are exclusively ranked by their informativeness. With this procedure, both summary and search requests may be handled effectively by one and the same algorithm.

### *3.2.2 Assembling the Results*

The set of retrieved and ranked sentences, both by topical relevance and informativeness, may be assembled to passages of various size according to the needs of the user. Three main parameters for passage construction are used:

- passage size: 1 to N with N being the absolute or relative number of words, sentences, or passages desired from the text;
- passage measure: type of unit counted by passage size, i.e. words, sentences, or paragraphs;
- passage span: maximum distance between two adjacent sentences or paragraphs, i.e. continuous or discontinuous units in passage.

For example, a passage size of 25% of a text containing 60 sentences with a measure in sentences and a span of 1 creates a ranked list of summary-style passages of 15 continuous sentences per passage (starting at any arbitrary position in the text). With a span of 0, the system assembles the 15 best-scoring discontinuous sentences (from anywhere in the text) first, then the next 15 scoring 16<sup>th</sup> to 30<sup>th</sup> best and so on.

## **3.3 Evaluation**

To evaluate the performance of the EXCERPT system, we can employ several strategies: (i) we can compare it to other within-document retrieval systems like ProfileSkim which offer similar functionality; (ii) we can compare the quality of the system's output with the baseline output provided by the 'Find' command (with respect to a certain task) as described in [Harper & al. 2004]; (iii) we can create our own test set of queries and relevant answer passages rated by humans; (iv) we can make use of summarization data and

evaluation strategies (cf. [Jing & al. 1998] and [Mani 2001] for an overview; see also [TAC 2008]).<sup>2</sup> Evaluation scenarios (i) and (ii) are planned to be carried out within the next months; scenarios (iii) and (iv) are work in progress, so that at least some preliminary results can be presented here.

Scenario (iv) is primarily intended to optimize the performance of the passage selection process, not to outperform other summarization systems. The idea is that if EXCERPT performs in summarizing mode on a level comparable to other summarizers, it is also good enough to create summary-style ranked passages in retrieval mode. For that purpose, one small collection to be found in [Zechner 1995] was used, where 13 persons rated 6 texts each: the task was to identify 5 to 7 sentences which are most central or relevant to the content of the text. The results of the human ratings were surprisingly homogenous, which makes this corpus highly interesting.

The best scoring 6 or 7 sentences (about 1/3 of the text) were used as the set of relevant items for which precision values could be computed: precision was calculated as proportion of retrieved relevant items (sentences) within the set of 6 or 7 retrieved items which would constitute the optimal extract according to the user ratings. For that purpose, all 6 texts were extracted from the electronic version of [Zechner 1995] and converted to Unicode resulting in 6 files which could be analyzed by EXCERPT. The results of the automatic extraction by EXCERPT are given in table 1 below; as a baseline for comparison, both the highly optimized results of Zechner's system and the first 6/7 initial sentences – which is a good baseline for news texts that is often hard to beat [cf. Ledeneva & al. 2008] – of the texts were used:

Text A			Text B			Text C			Text D			Text E			Text F		
B	Z	E	B	Z	E	B	Z	E	B	Z	E	B	Z	E	B	Z	E
2/6	3/6	4/6	3/6	2/6	4/6	3/7	4/7	5/7	2/6	4/6	4/6	2/7	3/7	3/7	4/6	4/6	3/6

Table 1: Precision of baseline (B), Zechner's system (Z), and EXCERPT (E)

---

<sup>2</sup> It should be noted here that passage retrieval for question answering and pure summarization are not genuine fields of application for EXCERPT.

The average precision values for the baseline are 0.42, for Zechner 0.53, for EXCERPT 0.61. The results show that EXCERPT retrieves – with the exception of text E – at least half of the relevant sentences which humans rated as most important; in only one case of text F, EXCERPT is worse than the baseline and Zechner's system. It must be stated that *no optimizations* of any kind were used, especially no positional parameter for sentences or weighting of parameters to enhance performance (as in the original system of Zechner overweighting initial and final sentences). The only parameters used for scoring were coherence structure, specificity, novelty, and cue words containing comparatives, superlatives, and indefinite articles (bonus) as well as 1st/2nd person and indefinite pronouns (malus).

One further pre-evaluation was carried out by the author: 30 persons were asked to indicate all *informative* sentences of a text (extracted from [Pournelle 2005]). The text consists of 57 sentences and deals with Pournelle's opinion about and work with the browser Firefox. The results were not as inhomogenous as expected despite the unspecific criterion of informativeness: about 30% of the sentences (18 of 57) were rated by at least 50% of the persons as informative, the best sentence selected by 23 persons (77%). There seems to be a quite good intuition about which passages of a text are informative to the user, where 'informative' was often equated with 'important' or 'interesting' by the raters. The precision of EXCERPT was 0.50 if the best 18 sentences are used as the set of relevant items, with the baseline performing very poorly at 0.17. This result also indicates that precision will certainly decrease with text length; the above performance of 61% can only be achieved for quite short texts of about 20–25 sentences.

The evaluations executed so far mainly serve to calibrate sentence scoring for informativeness. In the second phase, query-based evaluations measuring search term-based performance are of major interest, e.g. "Indicate the most informative sentences *about Firefox* (in comparison to *Internet Explorer*)". The purpose is to present the set of relevant retrieved passages such that the most informative ones are ranked first.

## 4 Conclusion and Prospect

The purpose of EXCERPT is to retrieve text passages of documents ranked by their topical relevance and informativeness according to several parameters also used in summarizing. The user may retrieve passages by entering 0 to N search terms, the former creating a representative summary passage satisfying general user interests, the latter assembling specific passages for fast scanning and incremental reading of texts according to the user needs. Under ideal circumstances, the system is able to handle all sorts of (informational) texts, like news, reports, articles, comments, etc. Although this is hard to achieve, a precision ranging from 30 to 60 percent is possible. Realistically, the more texts sorts and thematic domains the system must be able to handle, the worse performance will be on average. To optimize the system for general usage, further texts must be rated especially for informativeness, which is currently work in progress.

## References

- [Barker & al. 1998] Barker, K. & Chali, Y. & Copeck, T. & Matwin, S. & Szpakowicz, S. (1998): The Design of a Configurable Summarization System. <http://www.csi.uottawa.ca/tanka/uploadable/tr98-04.ps> (17.10.08)
- [Barzilay & Elhadad 1999] Barzilay & Elhadad (1999): Using Lexical Chains for Text Summarization. In [Mani & Maybury 1999], pp. 111–121.
- [Brin & Page 1998] Brin, S. & Page, L. (1998): The anatomy of a large-scale hypertextual Web search engine. Proceedings of the 7<sup>th</sup> international conference on World Wide Web 7, pp. 107–117.
- [Borko & Bernier 1975] Borko, H. & Bernier, C. L. (1975): *Abstracting Concepts and Methods*. New York & al.: Academic Press.
- [Dunlavy & al. 2007] Dunlavy, D. M. & O'Leary, D. P. & Conroy, J. M. & Schlesinger, J. D. (2007): QCS: A system for querying, clustering, and summarizing documents. *IPM*, 43, pp. 1588–1605.
- [Edmundson 1969] Edmundson, H. P. (1969): New Methods in Automatic Abstracting. *JACM*, 16(2), pp. 264–285.
- [Endres-Niggemeyer 1998] Endres-Niggemeyer, B. (1998): *Summarizing In-*

formation. Berlin & al.: Springer.

[Fellbaum 1998] Fellbaum, C. (1998; Ed.): *WordNet – An Electronic Lexical Database*. Cambridge & London: MIT Press.

[Goldstein 1999] Goldstein, J. & Kantrowitz, M. & Mittal, V. & Carbonell, J. (1999): Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of SIGIR'99*, pp. 121–128.

[Hahn & Reimer 1999] Hahn, U. & Reimer, U. (1999): Knowledge-based Text Summarization: Saliency and Generalization Operators for Knowledge Base Abstraction. In [Mani & Maybury 1999], pp. 215–232.

[Halliday & Hasan 1976] Halliday, M. A. K. & Hasan, R. (1976): *Cohesion in English*. London & N. Y.: Longman.

[Harper & al. 2004] Harper, D. J. & Koychev, I. & Sun, Y. & Pirie, Y. (2004): Within-Document Retrieval: A User-Centred Evaluation of Relevance Profiling. *IR*, 7, pp. 265–290.

[Hoey 1991] Hoey, M. (1991): *Patterns of Lexis in Text*. Oxford: University Press.

[Hovy & Lin 1999] Hovy, E. & Lin, C. (1999): Automatic Text Summarization in SUMMARIST. In [Mani & Maybury 1999], pp. 81–94.

[Hovy 2004] Hovy, E. (2004): Text Summarization. In Mitkov, R. (2004; Ed.): *The Oxford Handbook of Computational Linguistics*. Oxford: University Press, pp. 583–598.

[Jastrzembski 1981] Jastrzembski, J. E. (1981): Multiple Meanings, Number of Related Meanings, Frequency of Occurrence, and the Lexicon. *Cognitive Psychology*, 13, pp. 278–305.

[Jing & al. 1998] Jing, H. & Barzilay, R. & McKeown, K. & Elhadad, M. (1998): Summarization Evaluation Methods: Experiments and Analysis. *AAAI Symposium on Intelligent Summarization*, pp. 60–68.

[Kupiec & al. 1995] Kupiec, J. & Pederson, J. & Chen, F. (1995): A Trainable Document Summarizer. *Proceedings of SIGIR'95*, pp. 68–73.

[Ledeneva & al. 2008] Ledeneva, Y. & Gelbukh, A. & García-Hernández, R. A. (2008): Terms Derived from Frequent Sequences for Extractive Text Summarization. In Gelbukh, A. (2008; Eds.): *Computational Linguistics and Intelligent Text Processing. Proceedings of CICLing 2008*. Berlin & Heidelberg: Springer, pp. 593–604.

[Lee & al. 2001] Lee, S. S. & Shishibori, M. & Sumitomo, T. & Aoe, J.-I.

- (2001): Extraction of field-coherent passages. *IPM*, 38, pp. 173–207.
- [Mani & Maybury 1999] Mani, I. & Maybury, M. (1999; eds.): *Advances in Automatic Text Summarization*. Cambridge & London: MIT Press.
- [Mani 2001] Summarization Evaluation: An Overview. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf> (5.1.09)
- [Mihalcea & Tarau 2004] Mihalcea, R. & Tarau, P. (2004): TextRank – bringing order into texts. <http://www.cs.unt.edu/~rada/papers/mihalcea.emnlp04.pdf> (17.10.08)
- [Mihalcea 2004] Mihalcea, R. (2004): Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. <http://www.cs.unt.edu/~rada/papers/mihalcea.acl2004.pdf> (17.10.08)
- [Pournelle 2005] Firefox! [http://www.byte.com/documents/s=53/byt1135358505027/1205c\\_pournelle.htm](http://www.byte.com/documents/s=53/byt1135358505027/1205c_pournelle.htm) (15.10.08)
- [Reischer 2007] Reischer, J. (2007): Extracting Informative Content Units in Text Documents. In Oßwald, A. & Stempfhuber, M. & Wolff, C. (Eds.): *Open Innovation. Neue Perspektiven im Kontext von Information und Wissen*. Konstanz: UVK, pp. 285–301.
- [Salton & al. 1993] Salton, G. & Allan, J. & Buckley, C. (1993): Approaches to Passage Retrieval in Full Text Information Systems. *Proceedings of SIGIR'93*, pp. 49–58.
- [Spärck Jones 2007] Spärck Jones, K. (2007): Automatic Summarising: The state of the art. *IPM*, 43(6), pp. 1449–1481.
- [Stock 2007] Stock, W. (2007): *Information Retrieval*. München & Wien: Oldenbourg.
- [TAC 2008]: <http://www.nist.gov/tac/tracks/2008/summarization/index.html>. (29.10.08)
- [Tellex & al. 2003] Tellex, S. & Katz, B. & Lin, J. & Fernandes, A. & Marten, G. (2003): Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. *Proceedings of SIGIR'03*, pp. 41–47.
- [Tengi 1998] Tengi, R. I. (1998): Design and Implementation of the WordNet Lexical Database and Searching Software. In [Fellbaum 1998], pp. 105–127.
- [Teufel & Moens 2002] Teufel, S. & Moens, M. (2002): Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *CL*, 28(4), pp. 409–445.

[Vechtomova & al. 2006] Vechtomova, O. & Karamuftuoglu, M. & Robertson, S. E. (2006): On document relevance and lexical cohesion between query terms. *IPM*, 42, pp. 1230–1247.

[Zechner 1995] Zechner, K. (1995): Automatic Text Abstracting by Selecting Relevant Passages. Edinburgh: Master Thesis. <http://www.cs.cmu.edu/~zechner/abstr.pdf> (29.10.08)