



The Oxford Handbook of Computational Linguistics, Second Edition (2nd edn)

Ruslan Mitkov (ed.)

<https://doi.org/10.1093/oxfordhb/9780199573691.001.0001>

Published: 2014

Online ISBN: 9780191749643

Print ISBN: 9780199573691

CHAPTER

Text Summarization

Eduard Hovy

<https://doi.org/10.1093/oxfordhb/9780199573691.013.008>

Published: 05 October 2015

Abstract

Automated text summarization systems seek to provide the most important content contained in their (single or multiple document, and static or streaming over time) input. Extractive summarizers use various methods to assign an importance score to each fragment of the input and return the highest-scoring fragments, while abstractive summarizers attempt to compress and reformulate the extracted fragments and regenerate them in original and more elegant and coherent form. Numerous methods of scoring, combination, and compression have been developed. Evaluating the linguistic quality of a summary is much easier than evaluating the adequacy of its content. Various methods have been developed to compare system summary contents to the summaries of humans.

Keywords: automated text summarization, extract[ive] summary, abstract, maximal marginal relevance, ROUGE evaluation, Pyramid Method, summary content unit

Subject: Grammar, Syntax and Morphology, Linguistics

Series: Oxford Handbooks

1 Introduction

Automated text summarization systems are becoming increasingly desirable as the amount of online text increases. Experiments in the late 1950s and early 1960s suggested that text summarization by computer was feasible though not straightforward (Luhn 1959; Edmundson 1969). After a hiatus of some decades, progress in language processing, increases in computer memory and speed, and the growing presence of online text have renewed interest in automated text summarization research.

In this chapter we define a **summary** as follows:

Definition: A summary is a text that is produced from one or more texts that contains a significant portion of the information in the original text(s) and is no longer than half of the original text(s).

‘Text’ here includes single and multiple (possibly multimedia) documents, dialogues, hyperlinked texts, etc. Of the many types of summary that have been identified (Spärck Jones 1999; Hovy and Lin 1999), **indicative summaries** provide an idea of what the text is about without giving much content while **informative summaries** provide a shortened version of the content. **Extracts** are created by reusing portions (words, sentences, etc.) of the input text verbatim, while **abstracts** are created by regenerating the extracted content using new phrasing.

Section 2 outlines the principal approaches to automated text summarization. Problems unique to multi-document summarization are discussed in section 3. We review approaches to evaluation in section 4.

2 The Stages of Automated Text Summarization

Researchers in automated text summarization have identified three distinct stages (Spärck Jones 1999; Mani and Maybury 1999). The first stage, *topic selection or topic identification*, focuses on the selection of source material to include in the summary. Typically, topic identification is achieved by combining several methods that assign scores to individual (fragments of) sentences. The second stage, *topic interpretation and/or fusion*, focuses on the merging and/or compression of selected topics into a smaller number of encapsulating ones and/or a briefer formulation of them. The third stage, *summary generation*, focuses on producing the final summary in the desired form and format.

Most systems today include the first stage only, to produce pure extracts. Increasing degrees of sentence compression and topic fusion over selected fragments generally cause the need for generation, resulting in summaries that increasingly differ from their sources, approaching pure abstracts.

2.1 Stage 1: Topic identification

The overall approach is to assign a score to each unit (for example, each word, clause, or sentence) of the input text, and then to output the top-scoring N units, according to the summary length requested by the user (usually specified as a percentage of the original text length). Numerous methods have been developed to assign scores to fragments of text. Almost all systems employ several independent scoring modules, plus a combination module that integrates the scores for each unit.

The optimal size of the unit of text that is scored for extraction is a topic of research. Most systems focus on one sentence at a time. Fukushima et al. (1999) show that extracting subsentence-size units produces shorter summaries with more information. Strzalkowski et al. (1999) show that also including sentences immediately adjacent to important sentences increases coherence, by avoiding dangling pronoun references, etc.

The performance of topic identification modules working alone is usually measured using recall and precision scores (see section 4 and Chapter 15 on evaluation). Given an input text, a human’s extract, and a system’s extract, these scores quantify how closely the system’s extract corresponds to the human’s. For each unit, we let *correct* = the number of sentences extracted both by the system and the human; *wrong* = the number of sentences extracted by the system but not by the human; and *missed* = the number of sentences extracted by the human but not by the system. Then

$$\text{Precision} = \text{correct} / (\text{correct} + \text{wrong})$$

$$\text{Recall} = \text{correct} / (\text{correct} + \text{missed})$$

so that Precision reflects how many of the system's extracted sentences were good, and Recall reflects how many good sentences the system missed.

Topic identification methods can be grouped into families according to the information they consider when computing scores.

Positional criteria:

Thanks to regularities in the text structure of many genres, certain locations of the text (headings, titles, first paragraphs, etc.) tend to contain important information. The simple method of taking the lead (first paragraph) as summary often outperforms other methods, especially with newspaper articles (Brandow et al. 1995). Some variation of the **position method** appears in Baxendale (1958), Edmundson (1969), Donlan (1980), Kupiec et al. (1995), Teufel and Moens (1997), and Strzalkowski et al. (1999); Kupiec et al. and Teufel and Moens both consider this to be the single best method, scoring around 33%, for news, scientific, and technical articles.

In order to automatically determine the best positions, and to quantify their utility, Lin and Hovy (1997) define the genre- and domain-oriented Optimum Position Policy (OPP) as a ranked list of sentence positions that on average produce the highest yields for extracts, and describe an automated procedure to create OPPs, given texts and extracts.

Cue phrase indicator criteria:

Since in some genres certain words and phrases ('significant', 'in this paper we show') explicitly signal importance, sentences containing them should be extracted. Teufel and Moens (1997) report 54% joint recall and precision, using a manually built list of 1423 **cue phrases** in a genre of scientific texts. Each cue phrase has a (positive or negative) 'goodness score', also assigned manually. Teufel and Moens (1999) expand this method to argue that rather than single sentences, these cue phrases signal the nature of the multi-sentence rhetorical blocks of text in which they occur (such as Purpose/Problem, Background, Solution/Method, Conclusion/Claim).

Word and phrase frequency criteria:

Luhn (1959) uses Zipf's law of word distribution (a few words occur very often, fewer words occur somewhat often, and many words occur infrequently) to develop the following extraction criterion: if a text contains some words unusually frequently, then sentences containing these words are probably important.

The systems of Luhn (1959), Edmundson (1969), Kupiec et al. (1995), Teufel and Moens (1999), Hovy and Lin (1999), and others employ various frequency measures, and report performance of between 15% and 35% recall and precision (using word frequency alone). But both Kupiec et al. and Teufel and Moens show that word frequency in combination with other measures is not always better. Witbrock and Mittal (1999) compute a statistical model describing the likelihood that each individual word in the text will appear in the summary, in the context of certain features (part of speech tag, word length, neighbouring words, average sentence length, etc.). The generality of this method (also across languages) makes it attractive for further study.

Query and title overlap criteria:

A simple but useful method is to score each sentence by the number of desirable words it contains. Desirable words are, for example, those contained in the text's title or headings (Kupiec et al. 1995; Teufel and Moens 1997; Hovy and Lin 1999), or in the user's query, for a **query-based summary** (Buckley and Cardie 1997; Strzalkowski et al. 1999; Hovy and Lin 1999). The query method is a direct descendant of IR techniques (for more on **information retrieval**, see Chapter 34).

Cohesion and lexical connectedness criteria:

Words can be connected in various ways, including repetition, coreference, synonymy, and semantic association as expressed in thesauri. Sentences and paragraphs can then be scored based on the degree of connectedness of their words; more-connected sentences are assumed to be more important. This method yields performances ranging from 30% (using a very strict measure of connectedness) to over 60%, with Buckley and Cardie's use of sophisticated IR technology and Barzilay and Elhadad's lexical chains (Salton et al. 1997; Mitra et al. 1997; Mani and Bloedorn 1997; Buckley and Cardie 1997; Barzilay and Elhadad 1999). Mani and Bloedorn represent the text as a graph in which words are nodes and arcs represent adjacency, coreference, and lexical similarity. In recent years, external resources like Wikipedia have increasingly been used to indicate semantic relatedness and thereby boost scores (see, for example, Bawakid and Oussalah 2010); Kennedy et al. (2010) implement an entropy-based score for sentence selection using *Roget's Thesaurus* and Copeck et al. (2009) use FrameNet as well.

Discourse structure criteria:

A sophisticated variant of connectedness involves producing the underlying discourse structure of the text and scoring sentences by their discourse centrality, as shown in (Marcu 1997, 1998). Using a GSAT-like algorithm to learn the optimal combination of scores from centrality, several of the abovementioned measures, and scores based on the shape and content of the discourse tree, Marcu's (1998) system does almost as well as people for *Scientific American* articles.

Combining the scores of modules:

In all cases, researchers have found that no single method of scoring performs as well as humans do to create extracts. However, since different methods rely on different kinds of evidence, combining them improves scores significantly. Various methods of automatically finding a combination function have been tried; all seem to work, and there is no obvious best strategy.

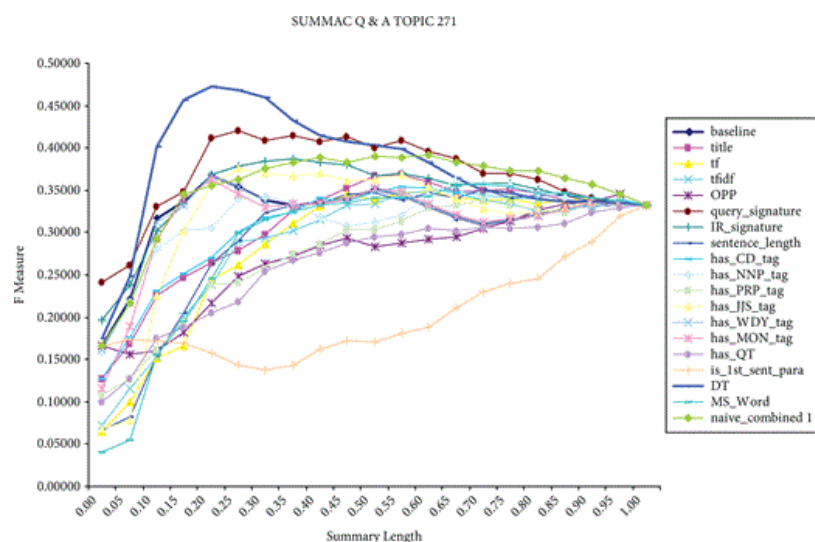
In a landmark paper, Kupiec et al. (1995) train a Bayesian classifier (see Chapter 11, 'Statistical Methods') by computing the probability that any sentence will be included in a summary, given the feature's paragraph position, cue phrase indicators, word frequency, upper-case words, and sentence length (since short sentences are generally not included in summaries). They find that, individually, the paragraph position feature gives 33% precision, the cue phrase indicators 29% (but when joined with the former, the two together give 42%), and so on, with individual scores decreasing to 20% and the combined five-feature score totalling 42%.

Also using a Bayesian classifier, Aone et al. (1999) find that even within the single genre, different newspapers require different features to achieve the same performance.

Using SUMMARIST, Lin (1999) compares 18 different features, plus straightforward and optimal combinations of them obtained using the machine learning algorithm C4.5 (Quinlan 1986; see also Chapter 13). These features include most of the abovementioned, as well as features signalling the presence in each

sentence of proper names, dates, quantities, pronouns, and quotes. The performances of the individual methods and the straightforward and learned combination functions are graphed in Figure 1, showing extract length against F-score (joint recall and precision). As expected, the top scorer is the learned combination function. The second-best score is achieved by query term overlap (though in other topics the query method did not do as well). The third-best score (up to the 20% length) is achieved equally by word frequency, the lead method, and the straightforward combination function. The curves in general indicate that to be most useful, summaries should not be longer than about 35% and not shorter than about 15%; no 5% summary achieved an F-score of over 0.25.

Figure 1



Summary length vs F-score for individual and combined methods of scoring sentences in SUMMARIST (Lin 1999).

Numerous recent studies on different methods of combining scores include an implementation of the pairwise algorithm of RankNet (Svore et al. 2007); linear regression to set weighting coefficients (Conroy et al. 2010); a learning to rank method using a three-layer (one hidden) neural network where the third layer contains a single node which is used as the ranking function (Jin et al. 2010). Genest et al. (2009) describe a system that performs its sentence selection in two steps: the first selects sentences and the second selects the best subset combination of them.

2.2 Stage 2: Topic interpretation or fusion

The stage of interpretation is what distinguishes abstractive summarization systems from extractive ones. During this stage, the topics selected are fused, represented in new terms, and/or otherwise compressed, using concepts or words not found in the original text. An extreme example would be to summarize a whole fable by Aesop into its core message, as in 'sour grapes' for the story of the fox and the grapes he cannot reach, or 'the race goes not always to the swift' for the race between the tortoise and the hare.

No system can perform interpretation without background knowledge about the domain; by definition, it must interpret the input in terms of something extraneous to the text. But acquiring enough (and deep enough) background knowledge, and performing this interpretation accurately, is so difficult that summarizers to date have only attempted it in a limited way.

The best examples of interpretation come from information extraction (Chapter 35), whose frame template representations serve as the interpretative structures of the input. By generating novel output text from an

instantiated frame, one obtains an abstract-type summary (DeJong 1978; Rau and Jacobs 1991; Kosseim et al. 2001; White et al. 2001).

Taking a more formal approach, Hahn and Reimer (1999) develop operators that condense knowledge representation structures in a terminological logic through conceptual abstraction (for more on knowledge representation see Chapter 5). To date, no parser has been built to produce the knowledge structures from text, and no generator to produce language from the results.

As proxy for background knowledge frames or abstractions, various researchers have used variants of topic models. Hovy and Lin (1999) use topic signatures—sets of words and relative strengths of association, each set related to a single headword—to perform topic fusion. They automatically construct these signatures from 30,000 *Wall Street Journal* texts, using *tf.idf* to identify for each topic the set of words most relevant to it. They use these topic signatures both during topic identification (to score sentences by signature overlap) and during topic interpretation (to substitute the signature head for the sentence(s) containing enough of its words). Similarly, Allan et al. (2001) use topic models to recognize the occurrence of new events in ongoing news streams and summarize them, while Wang et al. (2009) produce summaries using a Bayesian sentence-based topic model that uses both term-document and term-sentence associations.

Given the canonical structure of most news articles—all the most important material appears in paragraph 1—it is very difficult to build a summarizer that outperforms a baseline extractor. Recently, the summarization community has taken up the challenge of ‘guided summarization’, which tries to encourage a deeper linguistic (semantic) analysis of the source documents instead of relying only on position and word frequencies to select important concepts. The task is to produce a 100-word summary of a set of ten news articles for a given topic, focusing on specific semantic facets of the topic. As guidance, systems (and human summarizers) are given a list of aspects central to each category, and a summary must include all aspects found for each category. For details, see TAC (2010). The 2014 Biomedical Summarization Task discussed in section 4 is a more recent version of this task.

Deep interpretation, for example to automatically derive the theme(s) of a novel, the moral of a fable, or the plot of a film, remains beyond the scope of current capabilities, and hence powerful abstractive summarization is not yet a reality.

2.3 Stage 3: Summary generation

The third major stage of summarization is generation. When the summary content has been created through abstracting and/or information extraction, it exists within the computer in an internal notation that must be rendered into natural language using a generator (see Chapter 29).

Although in general extractive summarizers do not require generation, various disfluencies tend to result when sentences (or other extracted units) are simply extracted and printed—whether they are printed in order of importance score or in text order. A process of ‘smoothing’ can be used to identify and repair typical disfluencies, as first proposed in Hirst et al. (1997). **The most typical disfluencies that arise include repetition of clauses or NPs (where the repair is to aggregate the material into a conjunction), repetition of named entities (where the repair is to pronominalize),** and inclusion of less important material such as parentheticals and discourse markers (where the repair is to eliminate them). In the context of summarization, Mani et al. (1999) describe a summary revision program that takes as input simple extracts and produces shorter and more readable summaries.

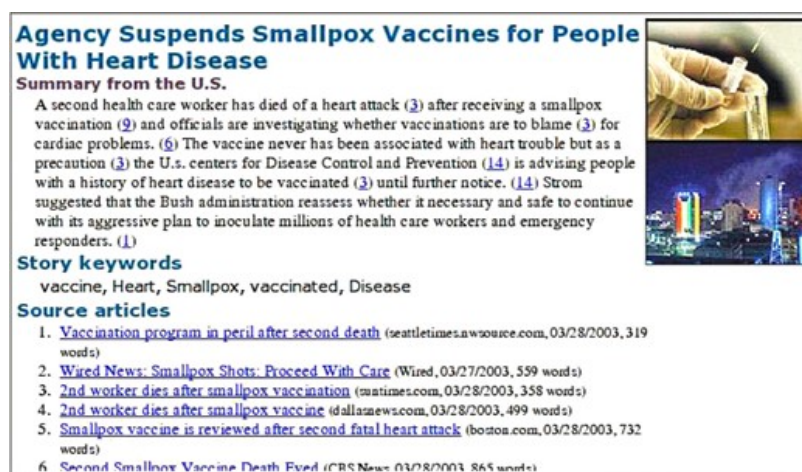
Text compression is another promising approach. Knight and Marcu’s (2000) prize-winning paper describes using the EM algorithm to train a system to compress the syntactic parse tree of a sentence in order to produce a shorter one, with the idea of eventually shortening two sentences into one, three into two

(or one), and so on. Banko et al. (2000) train statistical models to create headlines for texts by extracting individual words and ordering them appropriately.

Jing and McKeown (1999) argue that summaries are often constructed from a source document by a process of cut and paste—fragments of document sentences are combined into summary sentences—and hence that a summarizer need only identify the major fragments of sentences to include and then weave them together grammatically. In an extreme case of cut and paste, Witbrock and Mittal (1999) extract a set of words from the input document and then order the words into sentences using a bigram language model. Taking a more sophisticated approach, Jing and McKeown (1999) train a hidden Markov model to identify where in the document each (fragment of each) summary sentence resides. Testing with 300 human-written abstracts of newspaper articles, Jing and McKeown determine that only 19% of summary sentences do not have matching sentences in the document.

In important work, Barzilay and McKeown (2005) develop a method to align sequences of words carrying the same meaning across various input documents in order to identify important content, and then show how to weave these sentence fragments together to form coherent sentences and fluent summaries. Parsing source sentences into dependency trees, they pack the words into a lattice that records frequencies. Operations are applied to excise low-frequency fragments and merge high-frequency ones from other sentences as long as they fit syntactically and semantically. Linearizing the resulting fusion lattice produces long and syntactically correct sentences, which include for each fragment pointers back into the sections of the source documents. Figure 2 shows an example.

Figure 2



Agency Suspends Smallpox Vaccines for People With Heart Disease

Summary from the U.S.

A second health care worker has died of a heart attack (3) after receiving a smallpox vaccination (9) and officials are investigating whether vaccinations are to blame (3) for cardiac problems. (6) The vaccine never has been associated with heart trouble but as a precaution (3) the U.S. centers for Disease Control and Prevention (14) is advising people with a history of heart disease to be vaccinated (3) until further notice. (14) Strom suggested that the Bush administration reassess whether it necessary and safe to continue with its aggressive plan to inoculate millions of health care workers and emergency responders. (1)

Story keywords

vaccine, Heart, Smallpox, vaccinated, Disease

Source articles

1. [Vaccination program in peril after second death](#) (seattletimes.nwsource.com, 03/28/2003, 319 words)
2. [Wired News: Smallpox Shots Proceed With Care](#) (Wired, 03/27/2003, 559 words)
3. [2nd worker dies after smallpox vaccination](#) (mstimes.com, 03/28/2003, 358 words)
4. [2nd worker dies after smallpox vaccine](#) (dallasnews.com, 03/28/2003, 499 words)
5. [Smallpox vaccine is reviewed after second fatal heart attack](#) (boston.com, 03/28/2003, 732 words)
6. [Second Smallpox Vaccine Death Fied](#) (CBS News, 03/28/2003, 865 words)

Example output from Barzilay and McKeown's (2005) summarization system that aligns and fuses sentence fragments that occur commonly in the source documents.

3 Multi-Document Summarization

Summarizing a collection of thematically related documents poses several challenges beyond single documents (Goldstein et al. 2000; Fukumoto and Suzuki 2000; Kubota Ando et al. 2000). In order to avoid repetitions, one has to identify and locate thematic overlaps. One also has to decide what to include of the remainder, to deal with potential inconsistencies between documents, and, when necessary, to arrange events from various sources along a single timeline. For these reasons, multi-document summarization has received more attention than its single-document cousin.

An important study (Marcu and Gerber 2001) shows that for the newspaper article genre, even some very simple procedures provide essentially perfect results. For example, taking the first two or three paragraphs of the most recent text of a series of texts about the same event provides a summary as coherent and complete as those produced by human abstracters. In the same vein, the straightforward algorithm of Lin and Hovy (2002) that extracts non-overlapping sentences and pairs them with the first sentence of their respective documents (to set context) performed surprisingly well in the first multi-document summarization evaluation (DUC 2001).

More complex genres, such as biographies of people or descriptions of objects, require more sophisticated methods. Various techniques have been proposed to identify cross-document overlaps. SUMMONS (Radev 1999), a system that covers most aspects of multi-document summarization, takes an information extraction approach. Assuming that all input documents are parsed into templates (whose standardization makes comparison easier), SUMMONS clusters the templates according to their contents, and then applies rules to extract items of major import. SUMMONS deals with cross-document overlaps and inconsistencies using a series of rules to order templates as the story unfolds, identify information updates (e.g. increasing death tolls), identify cross-template inconsistencies (decreasing death tolls), and finally produce appropriate phrases or data structures for the language generator.

To determine what additional material should be included, first identify the units most relevant to the user's query, and then estimate the 'marginal relevance' of all remaining units.

Carbonell et al. (1997) introduce a measure called **Maximum Marginal Relevance** (MMR) to determine which sentence should be included from the sources as the pool of selected summary sentences grows. The MMR formula

$$\text{MMR} = \operatorname{argmax}_{D_i} [\lambda \operatorname{Sim}_1(D_i, Q) - (1 - \lambda) \max_{D_j} [\operatorname{Sim}_2(D_i, D_j)]]$$

balances the relevance of each candidate sentence (D_i) to the user's topic of interest (Q), measured by Sim_1 , against the similarity Sim_2 of the candidate to all other sentences already selected (D_j). By moving λ closer to 1, the user obtains summaries that concentrate on the desired topic with minimal content diversity; while by moving λ toward zero diversity is maximized.

An important line of research focuses on so-called **update summaries**: descriptions of only the important material that has occurred since the previous update, given a temporal stream of input material. This task extends multi-document summarization to consider a growing pool of known material. Methods to perform this include topic models (Allan et al. 2001).

An impressive system is Columbia University's Newsblaster (McKeown et al. 2002) at <<http://newsblaster.cs.columbia.edu/>> that summarizes online news 24 hours a day.

4 Evaluating Summaries

Many NLP evaluators distinguish between black-box and glass-box evaluation (for more on evaluation see Chapter 15). Taking a similar approach for summarization systems, Spärck Jones and Galliers (1996) define **intrinsic evaluations** as measuring output quality (only) and **extrinsic evaluations** as measuring user assistance in task performance.

More completely, one can differentiate three major aspects of summaries to measure: *form*, *content*, and *utility*.

Form is measured by considering linguistic considerations such as lexical aptness, sentence grammaticality, text coherence, and overall fluency. Standard text quality metrics have been developed and applied for example in machine translation (<http://www.issco.unige.ch/en/research/projects/isle/femti/>) and reading comprehension (Flesch 1948 *et seq.*) studies. For summarization, Brandow et al. (1995) performed one of the larger studies, in which evaluators rate systems' summaries according to some scale (readability; informativeness; fluency; and coverage).

Content is the most difficult to quantify. In general, to be a summary, the summary must obey two requirements:

- it must be shorter than the original input text;
- it must contain the important information of the original (where importance is defined by the user), and not other, totally new, information.

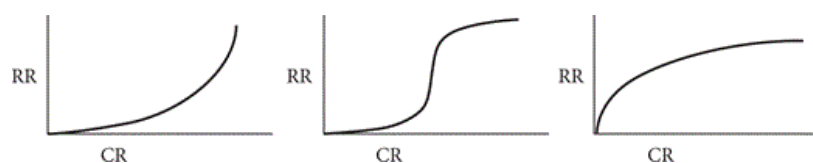
One can define two measures to capture the extent to which a summary S conforms to these requirements with regard to a text T :

$$\text{Compression Ratio : } CR = (\text{length } S) / (\text{length } T)$$

$$\text{Retention Ratio : } RR = (\text{info in } S) / (\text{info in } T)$$

However one chooses to measure the length and the information content, one can say that a good summary is one in which CR is small (tending to zero) while RR is large (tending to unity). One can characterize summarization systems by plotting the ratios of the summaries produced under varying conditions. For example, Figure 3) shows a fairly normal growth curve: as the summary gets longer (grows along the x axis), it includes more information (grows also along the y axis), until it equals the original. Figure 3) shows a more desirable situation: at some special point, the addition of just a little more text to the summary adds a disproportionately large amount of information. Figure 3) shows another: quite early, most of the important material is included in the summary; as the length grows, the added material is less novel or interesting. In both the latter cases, summarization is useful.

Figure 3



Compression Ratio (CR) vs Retention Ratio (RR).

For CR , measuring length is straightforward; one can count the number of words, letters, sentences, etc. For a given genre and register, there is a fairly good correlation among these metrics, in general.

For RR , measuring information content is difficult. Ideally, one wants to measure not information content, but *interesting* information content only. Although it is very difficult to define what constitutes *interestingness*, one can approximate measures of information content in several ways. The Categorization

and Ad Hoc tasks of the 1998 TIPSTER-SUMMAC study (Firmin Hand and Sundheim 1998; Firmin Hand and Chrzanowski 1999), described below, are examples. We discuss content measures in sections 4.2 and 4.3.

The growing body of literature on the interesting question of summary evaluation suggests that summaries are so task- and genre-specific and so user-oriented that no single measurement covers all cases.

4.1 Utility: Extrinsic evaluation studies

Utility is measured by extrinsic evaluations, and different tasks suggest their own appropriate metrics. For extrinsic (task-driven) evaluation, the major problem is to ensure that the metric applied correlates well with task performance efficiency. Examples of extrinsic evaluation can be found in Morris et al. (1992) for GMAT testing, Miike et al. (1994) for news analysis, and Mani and Bloedorn (1997) for information retrieval.

A large extrinsic evaluation is the TIPSTER-SUMMAC study (Firmin Hand and Sundheim 1998; Firmin Hand and Chrzanowski 1999), involving some 18 systems (research and commercial), in three tests. In the Categorization Task, testers classified a set of texts and also classified their summaries, created by various systems. The agreement between the classifications of texts and their corresponding summaries was measured; the greater the agreement, the better the summary was deemed to capture its content. In the Ad Hoc Task, testers classified query-based summaries as Relevant or Not Relevant to the query that generated each source document. The degree of relevance was deemed to reflect the quality of the summary. Space constraints prohibit full discussion of the results; a noteworthy finding is that, for newspaper texts, all extraction systems performed equally well (and no better than the lead method) for generic summarization, a result that is still generally true today.

This early study showed the dependence of summarization systems on the specifics of the evaluation metric and method applied. In a fine paper, Donaway et al. (2000) show how summaries receive different scores with different measures, or when compared to different (but presumably equivalent) ideal summaries created by humans. Jing et al. (1998) compare several evaluation methods, intrinsic and extrinsic, on the same extracts. With regard to inter-human agreement, they find fairly high consistency in the news genre, as long as the summary (extract) length is fixed relatively short (there is some evidence that other genres will deliver less consistency; Marcu 1997). With regard to summary length, they find great variation. Comparing three systems, and comparing five humans, they show that the humans' ratings of systems, and the perceived ideal summary length, fluctuate as summaries become longer.

4.2 Content: Intrinsic evaluation studies

Most existing evaluations of summarization systems are intrinsic. Typically, the evaluators create a set of ideal summaries, one for each test text, and then compare the summarizer's output to it, measuring content overlap (often by sentence or phrase recall and precision, but sometimes by simple word overlap). Since there is no 'correct' summary, some evaluators use more than one ideal per test text, and average the score of the system across the set of ideals. Comparing system output to some ideal was performed by, for example, Edmundson (1969), Paice (1990), Ono et al. (1994), Kupiec et al. (1995), Marcu (1997), and Salton et al. (1997). To simplify evaluation of extracts, Marcu (1999) and Goldstein et al. (1999) independently developed an automated method to create extracts corresponding to abstracts.

As mentioned above, not all material is equally informative or relevant. A popular method to measure the interestingness of content is to ask humans to decompose the texts (both system summaries and human gold standards) into semantically coherent units and then to compare the overlap of the units. The more popular a unit with several judges, the more informative or important it is considered to be. In a careful study, Teufel and van Halteren (2004) experimented with various unit lengths and found that (1) ranking

against a single gold-standard summary is insufficient, since rankings based on any two randomly chosen summaries are very dissimilar (average correlation $\rho = 0.20$); (2) a stable consensus summary can only be expected when at least 30–40 summaries are used; and (3) similarity measurements using unigrams show a similarly low ranking correlation.

In order to measure the importance of each unit, Nenkova and Passonneau (2004) developed the **Pyramid Method**: first, several humans independently identified from the text whatever units (here called **Summary Content Units** or SCUs) they considered relevant, after which each unit was assigned an importance score the number of times it was defined. Using DUC evaluation materials (see below), Nenkova and Passonneau demonstrated reasonable correlation with human ratings of the summaries, despite difficulties between judges to obtain exact agreement in unit length/content, and differences in subjective judgements about unit equivalence across texts (two units may be quite substantially different and still be considered equivalent by some judges). As a ‘unit popularity’ measure, the Pyramid Method has been the basis of human evaluation in recent TAC evaluation workshops.

4.3 Automated summarization evaluation

The cost and difficulty of evaluating summaries suggested the need for a single, standard, easy-to-apply even if imperfect, evaluation method to facilitate ongoing research. Following the lead of the successful automated BLEU evaluation metric used in machine translation, Lin and Hovy (2003) introduced **ROUGE**, a variant of BLEU, that compares a system summary to a human summary using a variety of (user-selected) metrics: single-word overlap, bigram (two-word) overlap, skip-bigram (bigram with omitted words) overlap, etc. As with BLEU, the assumption is that the quality of a summary will be proportional to the number of unit overlaps between system summary and gold-standard summary/ies, though of course the correspondence increases with more gold-standard examples. In contrast to BLEU, which measures precision, ROUGE measures recall—it counts not the correctness of system output units but rather the number of system units that are included in the gold standard. A series of subsequent studies showed acceptably strong correlation between human evaluations of summaries and ROUGE scores for them, especially using ROUGE-2. For example, in single-document summarization tasks (summaries of 100 words or ten words), ROUGE achieved Pearson’s ρ correlations of over 85% with humans.

ROUGE has been used in the NIST evaluation series and is the current most common evaluation standard. The ROUGE software package is described at and downloadable from <http://www.berouge.com/Pages/default.aspx>.

Since the creation of ROUGE, several other variations and similar metrics have been developed, including Pourpre (Lin and Demner-Fushman 2005) and Nuggeteer (Marton and Radul 2006), that score short units (‘nuggets’) against gold standards. Louis and Nenkova (2008) note that it is reasonable to expect that the distribution of terms in the source and a good summary are similar to each other. To compare the term distributions, they apply KL and Jensen-Shannon divergence, cosine similarity, and unigram and multinomial models of text. They find good correlations with human judgements, with Jensen-Shannon divergence giving a correlation as high as 0.9.

The BEwT-E package (Tratz and Hovy 2008) generalized ROUGE by using not ngrams but instead Basic Elements, minimal syntactic units obtained from the parse trees of sentences in the system output and gold-standard summaries. In contrast to SCUs, Basic Elements are automatically produced using about two dozen transformations to widen matching coverage, performing, for example, abbreviation expansion (‘USA’ and ‘US’ and ‘United States’ all match), active–passive voice transformation, and proper name expansion. As with the Pyramid Method, individual BEs are also assigned scores reflecting their popularity

in the gold standard. BEwT-E can be downloaded from <<http://www.isi.edu/publications/licensed-sw/BE/index.html>>.

4.4 Conference series: TAC, DUC, and NTCIR

To guide and focus research, the National Institute of Standards and Technology (NIST) in the USA and the Japan Society for Promotion of Science have held annual meetings, at which papers are presented and the results of challenge tasks they organize and evaluate are announced. NIST's DUC (<http://duc.nist.gov/>) and later TAC (<http://www.nist.gov/tac>) workshops have met annually since 2001, and NTCIR's workshops (<http://research.nii.ac.jp/ntcir/outline/prop-en.html>) since 1999. These have become the principal venues for discussing progress in automated summarization and system evaluation, and for developing new tasks and corpora.

The DUC and TAC evaluations generated a wealth of evaluation resources for summarizing news. But far less material is available for other domains and genres. In 2014 TAC hosted the Biomedical Summarization Task (<http://www.nist.gov/tac/2014/BiomedSumm/>) on the genre of research papers. The task identified two kinds of summary—the abstract produced by the author at the start of a paper, and the collection of sentences in other papers that cite the given target paper—and asked systems to identify the specific text fragment in the target paper each citation refers to, and what facet (chosen from Goal, Method, Result/Data, and Conclusion) each citation highlights. This information enables systems to produce summaries structured by the facets that can be compared to the author's own abstract and/or to other summaries of the paper produced by humans for different purposes.

Further Reading and Relevant Resources

Mani (2001) provides a thorough though now somewhat dated overview of the field, and Mani and Maybury (1999) include a useful collection of 26 early papers about summarization, including many of the most influential. The overviews in Lin (2009) and Radev (2004) are helpful.

A list of useful URLs can be found at <<http://www.summarization.com/>>.

Text summarization systems can be downloaded from:

- Open Text Summarizer: <<http://libots.sourceforge.net/>>
- Radev's MEAD summarizer: <<http://www.summarization.com/mead/>>
- QuickJist summarizer: <http://download.cnet.com/QuickJist-summarizer/3000-12512_4-10882271.html>

Evaluation has been very much investigated. Measures of readability (fluency, comprehensibility) are often taken from the machine translation community (see Chapter 32). Measures of content, however, require specialized treatment. The following packages are available:

- ROUGE package: <<http://www.berouge.com/Pages/default.aspx>>
- BEwT-E package: <<http://www.isi.edu/publications/licensed-sw/BE/index.html>>
- Pyramid Method (instructions for 2006): <<http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html>>

Useful workshop material can be found at:

- TAC Proceedings (2008–): <<http://www.nist.gov/tac/publications/index.html>>
- DUC Proceedings (2001–2007): <<http://www-nlpir.nist.gov/projects/duc/pubs.html>>
- NTCIR Proceedings: <<http://research.nii.ac.jp/ntcir/publication1-en.html>>
- Older workshop proceedings: Goldstein and Lin (2001); Hahn et al. (2000); Hovy and Radev (1998)

Acknowledgements

Thanks to Chin-Yew Lin, Daniel Marcu, Inderjeet Mani, Hoa Trang Dang, Hao Liu, Mike Junk, Louke van Wensveen, Thérèse Firmin Hand, Sara Shelton, and Beth Sundheim.

References

Allan, James, Rahul Gupta, and Vikas Khandelwal (2001). 'Topic Models for Summarizing Novelty'. In *Proceedings of the ACM SIGIR 01 Conference*, 9–12 September, New Orleans, 10–18. New York: ACM.

Aone, Chinatsu, Mary-Ellen Okurowski, James Gorlinsky, and Bjornar Larsen (1999). 'A Scalable Summarization System using Robust NLP'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 71–80. Cambridge, MA: MIT Press.

Banko, Michelle, Vibhu Mittal, and Michael Witbrock (2000). 'Headline Generation Based on Statistical Translation'. In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL)*, October, Hong Kong, 318–325. Stroudsburg, PA: Association for Computational Linguistics.

Barzilay, Regina and Michael Elhadad (1999). 'Using Lexical Chains for Text Summarization'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 111–121. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Barzilay, Regina and Kathleen McKeown (2005). 'Sentence Fusion for Multidocument News Summarization', *Computational Linguistics* 31(3): 297–328.

[Google Scholar](#) [WorldCat](#)

Bawakid Abdullah and Mourad Oussalah (2010). 'Summarizing with Wikipedia'. In *Proceedings of the Text Analytics Conference (TAC-10)*, 15–16 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Baxendale, Phyllis (1958). 'Machine-Made Index for Technical Literature—An Experiment', *IBM Journal* 2(4): 354–361.

[Google Scholar](#) [WorldCat](#)

Brandow, Ronald, Karl Mitze, and Lisa Rau (1995). 'Automatic Condensation of Electronic Publications by Sentence Selection', *Information Processing and Management* 31(5): 675–685.

[Google Scholar](#) [WorldCat](#)

Buckley, Chris and Claire Cardie (1997). 'Using EMPIRE and SMART for High-Precision IR and Summarization'. In *Proceedings of the TIPSTER Text Phase III 12-Month Workshop*, October, San Diego, CA.

Carbonell, Jaime, Yibing Geng, and Jade Goldstein (1997). 'Automated Query-Relevant Summarization and Diversity-Based Reranking'. In *Proceedings of the IJCAI-97 Workshop on AI in Digital Libraries*, 24 August, Nagoya, Japan, 12–19. San Mateo: Morgan Kaufmann.

Conroy, John, Judith Schlesinger, Peter Rankel, and Dianne O'Leary (2010). 'Guiding CLASSY Toward More Responsive Summaries'. In *Proceedings of the Text Analytics Conference (TAC-10)*, 15–16 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Copeck, Terry, Alistair Kennedy, Martin Scaiano, Diana Inkpen, and Stan Szpakowicz (2009). Summarizing with Roget's and with FrameNet. In *Proceedings of the Text Analytics Conference (TAC-09)*, 16–17 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

DeJong, Gerald (1978). 'Fast Skimming of News Stories: The FRUMP System'. PhD dissertation, Yale University.

Donaway, Robert, Kevin Drummey, and Laura Mather (2000). 'A Comparison of Rankings Produced by Summarization Evaluation Measures'. In *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA, 69–78. Stroudsburg, PA: Association for Computational Linguistics.

Donlan, Dan (1980). 'Locating Main Ideas in History Textbooks', *Journal of Reading* 24: 135–140.

[Google Scholar](#) [WorldCat](#)

DUC (2001). *Proceedings of the Document Understanding Conference (DUC) Workshop on Multi-Document Summarization Evaluation*, at the SIGIR-01 Conference, 13–14 September, New Orleans. New York: ACM. See <<http://duc.nist.gov/>>.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Edmundson, Harold (1969). 'New Methods in Automatic Extraction', *Journal of the ACM* 16(2): 264–285.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Firmin Hand, Thérèse and Michael Chrzanowski (1999). 'An Evaluation of Text Summarization Systems'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 325–335. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Firmin Hand, Thérèse and Beth Sundheim (1998). 'The TIPSTER-SUMMAC Summarization Evaluation'. In *Proceedings of the TIPSTER Text Phase III 18-Month Workshop*, Fairfax, VA (no page numbers in binder).

Flesch, Rudolf (1948). 'A New Readability Yardstick', *Journal of Applied Psychology* 32: 221–233.

[Google Scholar](#) [WorldCat](#)

Fukumoto, Fumiyo and Yoshimi Suzuki (2000). 'Extracting Key Paragraph [sic] Based on Topic and Event Detection: Towards Multi-Document Summarization'. In *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA, 31–39. Stroudsburg, PA: Association for Computational Linguistics.

Fukushima, Takahiro, Terumasa Ehara, and Katsuhiko Shirai (1999). 'Partitioning Long Sentences for Text Summarization' [in Japanese], *Journal of the Society of Natural Language Processing of Japan* 6(6): 131–147.

[Google Scholar](#) [WorldCat](#)

Genest, Pierre-Etienne, Guy Lapalme, Luka Nerima, and Eric Wehrli (2009). 'A Symbolic Summarizer with 2 Steps of Sentence Selection for TAC 2009'. In *Proceedings of the Text Analytics Conference (TAC-09)*, 16–17 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Goldstein, Jade, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell (1999). 'Summarizing Text Documents: Sentence Selection and Evaluation Metrics'. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, Berkeley, CA, 121–128. New York: ACM.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz (2000). 'Multi-Document Summarization by Sentence Extraction'. In *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA, 40–48. Stroudsburg, PA: Association for Computational Linguistics.

Goldstein, Jade and Chin-Yew Lin (eds) (2001). *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA. Stroudsburg, PA: Association for Computational Linguistics.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Hahn, Udo and Ulrich Reimer (1999). 'Knowledge-Based Text Summarization: Saliency and Generalization Operators for Knowledge Base Abstraction'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 215–232. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Hahn, Udo, Chin-Yew Lin, Inderjeet Mani, and Dragomir Radev (eds) (2000). *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA. Stroudsburg, PA: Association for Computational Linguistics.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Hirst, Graeme, Chrysanne DiMarco, Eduard Hovy, and Kim Parsons (1997). 'Authoring and Generating Health-Education Documents that Are Tailored to the Needs of the Individual Patient'. In Anthony Jameson, Cecile Paris, and C. Tasso (eds), *Proceedings of the Sixth International Conference on User Modeling (UM97)*, Sardinia, Italy, 107–118. Vienna and New York: Springer-Verlag. See <<http://um.org>>.

Hovy, Eduard and Chin-Yew Lin (1999). 'Automating Text Summarization in SUMMARIST'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 81–97. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Hovy, Eduard and Dragomir Radev (eds) (1998). *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*. Stanford University, CA: AAAI Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Jin, Feng, Minlie Huang and Xiaoyan Zhu (2010). 'The THU Summarization Systems at TAC 2010'. In *Proceedings of the Text Analytics Conference (TAC-10)*, 15–16 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Jing, Hongyan, Regina Barzilay, Kathleen McKeown, and Michael Elhadad (1998). 'Summarization Evaluation Methods: Experiments and Results'. In Eduard Hovy and Dragomir Radev (eds), *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, 60–68. Stanford University, CA: AAAI Press.

Jing, Hongyan and Kathleen McKeown (1999). 'The Decomposition of Human-Written Summary Sentences'. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, Berkeley, CA, 129–136. New York: ACM.

Kennedy, Alistair, Terry Copeck, Diana Inkpen, and Stan Szpakowicz (2010). 'Entropy-Based Sentence Selection with Roget's Thesaurus'. In *Proceedings of the Text Analytics Conference (TAC-10)*, 15–16 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Knight, Kevin and Daniel Marcu (2000). 'Statistics-Based Summarization—Step One: Sentence Compression'. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence (AAAI)*, 30 July–3 August, Austin, TX, 703–710. Stanford University, CA: AAAI Press.

Kosseim, Leila, Stéphane Beauregard, and Guy Lapalme (2001). 'Using Information Extraction and Natural Language Generation to Answer E-mail', *Data and Knowledge Engineering* 38: 85–100.

[Google Scholar](#) [WorldCat](#)

Kubota Ando, Rie, Branimir Boguraev, Roy Byrd, and Mary Neff (2000). 'Multi-Document Summarization by Visualizing Topical Content'. In *Proceedings of the NAACL Workshop on Text Summarization*, Seattle, WA, 79–88. Stroudsburg, PA: Association for Computational Linguistics.

Kupiec, Julian, Jan Pedersen, and Francine Chen (1995). 'A Trainable Document Summarizer'. In *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR-95)*, Berkeley, CA, 68–73. New York: ACM.

Lin, Chin-Yew (1999). 'Training a Selection Function for Extraction'. In Lorna Uden, Leon S. L. Wang, Juan Manuel Corchado Rodríguez, Hsin-Chang Yang, and I-Hsien Ting (eds), *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM)*, 2–6 November, Kansas City, MO, 1–8. Dordrecht: Springer.

Lin, Chin-Yew and Eduard Hovy (1997). 'Identifying Topics by Position'. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, 31 March–3 April, Washington, DC, 283–290. San Francisco, CA: Association for Computational Linguistics.

Lin, Chin-Yew and Eduard Hovy (2002). 'From Single to Multi-Document Summarization: A Prototype System and its Evaluation'. In *Proceedings of the 40th Conference of the Association of Computational Linguistics (ACL-02)*, July, Philadelphia, PA, 457–464. Stroudsburg, PA: Association for Computational Linguistics.

Lin, Chin-Yew and Eduard Hovy (2003). 'Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics'. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational*

Linguistics (HLT-NAACL 2003), 27 May–1 June, Edmonton, Canada, 71–78. Stroudsburg, PA: Association for Computational Linguistics.

Lin, Jimmy and Dina Demner-Fushman (2005). 'Automatically Evaluating Answers to Definition Questions'. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 6–8 October, Vancouver, Canada, 931–938. Stroudsburg, PA: Association for Computational Linguistics.

Lin, Jimmy (2009). 'Summarization'. In Ling Liu and M. Tamer Özsu (eds), *Encyclopedia of Database Systems*. New York: Springer.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Louis, Annie and Ani Nenkova (2008). 'Automatic Summary Evaluation without Human Models'. In *Proceedings of the Text Analytics Conference (TAC-08)*, 17–19 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Luhn, Hans Peter (1959). 'The Automatic Creation of Literature Abstracts', *IBM Journal of Research and Development* 2(2): 159–165.
[Google Scholar](#) [WorldCat](#)

Mani, Inderjeet (2001). *Automatic Summarization*. Amsterdam and Philadelphia: John Benjamins.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Mani, Inderjeet and Eric Bloedorn (1997). 'Multi-Document Summarization by Graph Search and Matching'. In *Proceedings of the 14th Conference of the American Association of Artificial Intelligence (AAAI-97)*, 27–31 July, Providence, RI, 622–628. Stanford University, CA: AAAI Press.

Mani, Inderjeet, Barbara Gates, and Eric Bloedorn (1999). 'Improving Summaries by Revising Them'. In *Proceedings of the 37th Conference of the Association of Computational Linguistics (ACL-99)*, 558–565. Stroudsburg, PA: Association for Computational Linguistics.

Mani, Inderjeet and Mark Maybury (eds) (1999). *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press.
[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Marcu, Daniel (1997). 'The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts'. PhD dissertation, University of Toronto.

Marcu, Daniel (1998). 'Improving Summarization through Rhetorical Parsing Tuning'. In Eugene Charniak (ed.), *Proceedings of the Sixth Workshop on Very Large Corpora (COLING-ACL'98)*, 15–16 August, Montreal, 10–16. New Brunswick, NJ: Association for Computational Linguistics.

Marcu, Daniel (1999). 'The Automatic Construction of Large-scale Corpora for Summarization Research'. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, Berkeley, CA, 137–144. New York: ACM.

Marcu, Daniel and Laurie Gerber (2001). 'An Inquiry into the Nature of Multidocument Abstracts, Extracts, and their Evaluation'. In *Proceedings of the Workshop on Text Summarization at the 2nd Conference of the North American Association of Computational Linguistics*, Pittsburgh, 1–8. Stroudsburg, PA: Association for Computational Linguistics.

Marton, Gregory, and Alexey Radul (2006). 'Nuggeteer: Automatic Nugget-Based Evaluation Using Descriptions and Judgements'. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, June, New York, 375–382. Stroudsburg, PA: Association for Computational Linguistics.

McKeown, Kathleen, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman (2002). 'Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster'. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT-02)*, 280–285. San Francisco, CA: Morgan Kaufmann Publishers.

Miike, Seiji, Etsuo Itoh, Kenji Ono, and Kazuo Sumita (1994). 'A Full-Text Retrieval System with Dynamic Abstract Generation Function'. In *Proceedings of the 17th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR-94)*, 152–161. New York: ACM.

Mitra, Mandar, Amit Singhal, and Chris Buckley (1997). 'Automatic Text Summarization by Paragraph Extraction'. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, July, Madrid, Spain, 39–46. Stroudsburg, PA: Association for Computational Linguistics.

Morris, Andrew, George Kasper, and Dennis Adams (1992). 'The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance', *Information Systems Research* 3(1): 17–35.

[Google Scholar](#) [WorldCat](#)

Nenkova, Ani and Rebecca Passonneau (2004). 'Evaluating Content Selection in Summarization: The Pyramid Method'. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2–7 May, Boston, MA, 145–152. Stroudsburg, PA: Association for Computational Linguistics.

Ono, Kenji, Kazuo Sumita, and Seiji Miike (1994). 'Abstract Generation Based on Rhetorical Structure Extraction'. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 5–9 August, Kyoto, Japan, vol. 1, 344–384. Stroudsburg, PA: Association for Computational Linguistics.

Paice, Chris (1990). 'Constructing Literature Abstracts by Computer: Techniques and Prospects', *Information Processing and Management* 26(1): 171–186.

[Google Scholar](#) [WorldCat](#)

Quinlan, John Ross (1986). 'Induction of Decision Trees', *Machine Learning* 1: 81–106.

[Google Scholar](#) [WorldCat](#)

Radev, Dragomir (1999). 'Generating Natural Language Summaries from Multiple Online Sources: Language Reuse and Regeneration'. PhD dissertation, Columbia University.

Radev, Dragomir (2004). 'Text Summarization'. Tutorial at the 27th Annual International ACM SIGIR Conference, July, Sheffield University.

Rau, Lisa and Paul Jacobs (1991). 'Creating Segmented Databases from Free Text for Text Retrieval'. In *Proceedings of the 14th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR-91)*, 13–16 October, Chicago, 337–346. New York: ACM.

Salton, Gerald, Amit Singhal, Mandar Mitra, and Chris Buckley (1997). 'Automatic Text Structuring and Summarization', *Information Processing and Management* 33(2): 193–208.

[Google Scholar](#) [WorldCat](#)

Spärck Jones, Karen and Julia Galliers (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York: Springer.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Spärck Jones, Karen (1999). 'Automatic Summarizing: Factors and Directions'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 1–13. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Strzalkowski, Tomek, Gees Stein, Jin Wang, and Bowden Wise (1999). 'A Robust Practical Text Summarizer'. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 137–154. Cambridge, MA: MIT Press.

[Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Svore, Krysta Marie, Lucy Vanderwende, and Christopher Burges (2007). 'Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources'. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning (EMNLP-CONLL), June, Prague, 448–457. Stroudsburg, PA: Association for Computational Linguistics.

TAC (2010). *Proceedings of the Text Analysis Conference (TAC) on Multi-Document Summarization*, 15–16 November, National Institute of Standards and Technology, Gaithersburg, MD. See <<http://www.nist.gov/tac/2010/Summarization/index.html>>. [Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Teufel, Simone and Marc Moens (1997). ‘Sentence Extraction as a Classification Task’. In *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 11 July, Madrid, Spain, 58–65. Stroudsburg, PA: Association for Computational Linguistics.

Teufel, Simone and Marc Moens (1999). ‘Argumentative Classification of Extracted Sentences as a First Step toward Flexible Abstracting’. In Inderjeet Mani and Mark Maybury (eds), *Advances in Automated Text Summarization*, 155–175. Cambridge, MA: MIT Press. [Google Scholar](#) [Google Preview](#) [WorldCat](#) [COPAC](#)

Teufel, Simone and Hans van Halteren (2004). ‘Evaluating Information Content by Factoid Analysis: Human Annotation and Stability’. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Barcelona, 419–426. Stroudsburg, PA: Association for Computational Linguistics.

Tratz, Stephen and Eduard Hovy (2008). ‘Summarization Evaluation Using Transformed Basic Elements’. In *Proceedings of Text Analytics Conference (TAC-08)*, 17–19 November, National Institute of Standards and Technology, Gaithersburg, MD (no page number).

Wang, Dingding, Shenghuo Zhu, Tao Li, and Yihong Gong (2009). ‘Multi-Document Summarization Using Sentence-Based Topic Models’. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 4 August, Singapore, 297–300. Stroudsburg, PA: Association for Computational Linguistics.

White, Michael, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff (2001). *Multi-document Summarization via Information Extraction*. CoGenTex Technical Report.

Witbrock, Michael and Vibhu Mittal (1999). ‘Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries’. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, Berkeley, CA, 315–316. New York: ACM.