6-2011

Automatic Summarization

Ani Nenkova
Univesity of Pennsylvania, nenkova@cis.upenn.edu

Kathleen McKeown
Columbia University

Nenkova, A. & McKeown, K., Automatic Summarization, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* , June 2011, doi: anthology/P11-5003

Automatic Summarization

Abstract
It has now been 50 years since the publication of Luhn's seminal paper on automatic summarization. During these years the practical need for automatic summarization has become increasingly urgent and numerous papers have been published on the topic. As a result, it has become harder to find a single reference that gives an overview of past efforts or a complete view of summarization tasks and necessary system components. This article attempts to fill this void by providing a comprehensive overview of research in summarization, including the more traditional efforts in sentence extraction as well as the most novel recent approaches for determining important content, for domain and genre specific summarization and for evaluation of summarization. We also discuss the challenges that remain open, in particular the need for language generation and deeper semantic understanding of language that would be necessary for future advances in the field.

Disciplines
Computer Sciences

Comments
Nenkova, A. & McKeown, K., Automatic Summarization, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* , June 2011, doi: anthology/P11-5003

Foundations and Trends® in
Information Retrieval
Vol. 5, Nos. 2–3 (2011) 103–233

DOI: 10.1561/1500000015

Automatic Summarization

By Ani Nenkova and Kathleen McKeown

Contents

Foundations and Trends® in
Information Retrieval
Vol. 5, Nos. 2–3 (2011) 103–233

DOI: 10.1561/1500000015

Automatic Summarization

Ani Nenkova[1] and Kathleen McKeown[2]

[1] *University of Pennsylvania, USA, nenkova@seas.upenn.edu*
[2] *Columbia University, USA, kathy@cs.columbia.edu*

Abstract

It has now been 50 years since the publication of Luhn's seminal paper on automatic summarization. During these years the practical need for automatic summarization has become increasingly urgent and numerous papers have been published on the topic. As a result, it has become harder to find a single reference that gives an overview of past efforts or a complete view of summarization tasks and necessary system components. This article attempts to fill this void by providing a comprehensive overview of research in summarization, including the more traditional efforts in sentence extraction as well as the most novel recent approaches for determining important content, for domain and genre specific summarization and for evaluation of summarization. We also discuss the challenges that remain open, in particular the need for language generation and deeper semantic understanding of language that would be necessary for future advances in the field.

We would like to thank the anonymous reviewers, our students and Noemie Elhadad, Hongyan Jing, Julia Hirschberg, Annie Louis, Smaranda Muresan and Dragomir Radev for their helpful feedback. This paper was supported in part by the U.S. National Science Foundation (NSF) under IIS-05-34871 and CAREER 09-53445. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

1

Introduction

Today's world is all about information, most of it online. The World Wide Web contains billions of documents and is growing at an exponential pace. Tools that provide timely access to, and digest of, various sources are necessary in order to alleviate the information overload people are facing. These concerns have sparked interest in the development of automatic summarization systems. Such systems are designed to take a single article, a cluster of news articles, a broadcast news show, or an email thread as input, and produce a concise and fluent summary of the most important information. Recent years have seen the development of numerous summarization applications for news, email threads, lay and professional medical information, scientific articles, spontaneous dialogues, voicemail, broadcast news and video, and meeting recordings. These systems, imperfect as they are, have already been shown to help users and to enhance other automatic applications and interfaces.

1.1 Types of Summaries

There are several distinctions typically made in summarization and here we define terminology that is often mentioned in the summarization literature.

Extractive summaries (extracts) are produced by concatenating several sentences taken exactly as they appear in the materials being summarized. *Abstractive summaries (abstracts)*, are written to convey the main information in the input and may reuse phrases or clauses from it, but the summaries are overall expressed in the words of the summary author.

Early work in summarization dealt with *single document summarization* where systems produced a summary of *one* document, whether a news story, scientific article, broadcast show, or lecture. As research progressed, a new type of summarization task emerged: *multi-document summarization*. Multi-document summarization was motivated by use cases on the web. Given the large amount of redundancy on the web, summarization was often more useful if it could provide a brief digest of many documents on the same topic or the same event. In the first deployed online systems, multi-document summarization was applied to clusters of news articles on the same event and used to produce online browsing pages of current events [130, 171].[1] A short one-paragraph summary is produced for each cluster of documents pertaining to a given news event, and links in the summary allow the user to directly inspect the original document where a given piece of information appeared. Other links provide access to all articles in the cluster, facilitating the browsing of news. User-driven clusters were also produced by collecting search engine results returned for a query or by finding articles similar to an example document the user has flagged as being of interest [173].

Summaries have also been distinguished by their content. A summary that enables the reader to determine about-ness has often been called an *indicative summary*, while one that can be read in place of the document has been called an *informative summary* [52]. An indicative summary may provide characteristics such as length, writing style, etc., while an informative summary will include facts that are reported in the input document(s).

[1] `http://lada.si.umich.edu:8080/clair/nie1/nie.cgi`, `http://newsblaster.columbia.edu`.

Most research in summarization deals with producing a short, paragraph-length summary. At the same time, a specific application or user need might call for a *keyword summary*, which consists of a set of indicative words or phrases mentioned in the input, or *headline summarization* in which the input document(s) is summarized by a single sentence.

Much of the work to date has been in the context of *generic summarization*. Generic summarization makes few assumptions about the audience or the goal for generating the summary. Typically, it is assumed that the audience is a general one: anyone may end up reading the summary. Furthermore, no assumptions are made about the genre or domain of the materials that need to be summarized. In this setting, importance of information is determined only with respect to the content of the input alone. It is further assumed that the summary will help the reader quickly determine what the document is about, possibly avoiding reading the document itself.

In contrast, in *query focused summarization*, the goal is to summarize only the information in the input document(s) that is relevant to a specific user query. For example, in the context of information retrieval, given a query issued by the user and a set of relevant documents retrieved by the search engine, a summary of each document could make it easier for the user to determine which document is relevant. To generate a useful summary in this context, an automatic summarizer needs to take the query into account as well as the document. The summarizer tries to find information within the document that is relevant to the query or in some cases, may indicate how much information in the document relates to the query. Producing snippets for search engines is a particularly useful query focused application [207, 213]. Researchers have also considered cases where the query is an open-ended question, with many different facts possibly being relevant as a response. A request for a biography is one example of an open-ended question as there are many different facts about a person that could be included, but are not necessarily required.

Update summarization addresses another goal that users may have. It is multi-document summarization that is sensitive to time; a

summary must convey the important development of an event beyond what the user has already seen.

The contrast between generic, query-focused, and update summarization is suggestive of other issues raised by Sparck Jones in her 1998 call to arms [194]. Sparck Jones argued that summarization should not be done in a vacuum, but rather should be viewed as part of a larger context where, at the least, considerations such as the purpose of summarization (or task which it is part of), the reader for which it is intended, and the genre which is being summarized, are taken into account. She argued that generic summarization was unnecessary and in fact, wrong-headed. Of course, if we look at both sides of the question, we see that those who write newspaper articles do so in much the same spirit in which generic summaries are produced: the audience is a general one and the task is always the same. Nonetheless, her arguments are good ones as they force the system developer to think about other constraints on the summarization process and they raise the possibility of a range of tasks other than to simply condense content.

1.2 How do Summarization Systems Work?

Summarization systems take one or more documents as input and attempt to produce a concise and fluent summary of the most important information in the input. Finding the most important information presupposes the ability to understand the semantics of written or spoken documents. Writing a concise and fluent summary requires the capability to reorganize, modify and merge information expressed in different sentences in the input. Full interpretation of documents and generation of abstracts is often difficult for people,[2] and is certainly beyond the state of the art for automatic summarization.

How then do current automatic summarizers get around this conundrum? Most current systems avoid full interpretation of the input and generation of fluent output. The current state of the art in the vast majority of the cases relies on sentence extraction. The extractive approach to summarization focuses research on one key question: how

[2] For discussion of professional summarization, see [114].

can a system determine which sentences are important? Over the years, the field has seen advances in the sophistication of language processing and machine learning techniques that determine importance.

At the same time, there have been recent advances in the field which move toward semantic interpretation and generation of summary language. Semantic interpretation tends to be done for specialized summarization. For example, systems that produce biographical summaries or summaries of medical documents tend to use extraction of information rather than extraction of sentences. Research on generation for summarization uses a new form of generation, *text-to-text generation* and focuses on editing input text to better fit the needs of the summary.

1.2.1 Early Methods for Sentence Extraction

Most traditional approaches to summarization deal exclusively with the task of identifying important content, usually at the sentence level. The very first work on automatic summarization, done by Luhn [111] in the 1950s, set the tradition for sentence extraction.

His approach was implemented to work on technical papers and magazine articles. Luhn put forward a simple idea that shaped much of later research, namely that some words in a document are descriptive of its content, and the sentences that convey the most important information in the document are the ones that contain many such descriptive words close to each other that. He also suggested using frequency of occurrence in order to find which words are descriptive of the topic of the document; words that occur often in the document are likely to be the main topic of this document. Luhn brought up two caveats: some of the *most* common words in a technical paper or a magazine article, and in fact in any type of document, are not at all descriptive of its content. Common function words such as determiners, prepositions and pronouns do not have much value in telling us what the document is about. So he used a predefined list, called a *stop word list*, consisting of such words to remove them from consideration. Another class of words that do not appear in the stop word list but still cannot be indicative of the topic of a document are words common for a particular domain. For example, the word "cell" in a scientific paper in

cell biology is not likely to give us much idea about what the paper is about. Finally, words that appear in the document only a few times are not informative either. Luhn used empirically determined high and low frequency thresholds for identifying descriptive words, with the high thresholds filtering out words that occur very frequently throughout the article and the low thresholds filtering out words that occur too infrequently. The remaining words are the descriptive words, indicative of the content that is important. Sentences characterized by high density of descriptive words, measured as clusters of five consecutive words by Luhn, are the most important ones and should be included in the summary.

In the next section we discuss how later work in sentence extraction adopted a similar view of finding important information but refined the ideas of using raw frequency by proposing weights for words, such as TF*IDF, in order to circumvent the need for coming up with arbitrary thresholds in determining which words are descriptive of a document. Later, statistical tests on word distributions were proposed to decide which words are topic words and which are not. Other approaches abandoned the idea of using words as the unit of operation, and used word frequency indirectly to model the similarity between sentences and derive measures of sentence importance from these relationships. We present these approaches in greater detail in Section 2, as they have proven to be highly effective, relatively robust to genre and domain, and are often referenced in work on automatic summarization.

There are some obvious problems with Luhn's approach. The same concept can be referred to using different words: consider for example "approach", "method" "algorithm", and "it". Different words may indicate a topic when they appear together; for example "hurricane", "damage", "casualties", "relief" evoke a natural disaster scenario. The same word can appear in different morphological variants — "show", "showing", "showed" — and counts of words as they appear in the text will not account for these types of repetition. In fact, Luhn was aware of these problems and he employed a rough approximation to morphological analysis, collapsing words that are similar except for the last six letters, to somewhat address this problem. After our presentation of word frequency-driven approaches in Section 2.1, we briefly discuss

work based on the use of coreference systems and knowledge sources that perform input analysis and interpretation. These methods can better address these challenges and are discussed in Section 3. In essence these are still frequency approaches, but counting is performed in a more intelligent manner. Such approaches incur more processing overhead, which is often undesirable for practical purposes, but comes closer to the ideal of developing systems that are in fact interpreting the input before producing a summary.

Edmundson's [52] work was the foundation of several other trends in summarization research which eventually led to machine learning approaches in summarization. He expanded on Luhn's approach by proposing that multiple features may indicate sentence importance. He used a linear combination of features to weight sentences in a scientific article. His features were: (1) number of times a word appears in the article, (2) the number of words in the sentence that also appear in the title of the article, or in section headings, (3) position of the sentence in the article and in the section, (4) the number of sentence words matching a pre-compiled list of cue words such as "In sum". A compelling aspect of Edmundson's work that foreshadows today's empirically based approaches, was the creation of a document/extractive summary corpus. He used the corpus both to determine weights on the four features and to do evaluation. His results interestingly suggest that word frequency is the least important of the four classes of features, for his specific task and corpus. His other features take advantage of knowledge of the domain and genre of the input to the summarizer. We discuss such domain dependent approaches, which make use of domain-dependent knowledge sources and of specific domain characteristics, for summarization of scientific articles, medical information and email in Section 5.

In other relatively early and seminal work, Paice [164, 165] shifted the research focus toward the need for language generation techniques in summarization. He focused on the problem in extractive summarization of accidentally selecting sentences that contain unresolved references to sentences not included in the summary or not explicitly included in the original document. The problem can arise not only because of the presence of a pronouns but also because of a wide variety

of other phrases (exophora) such as "Our investigations have shown this to be true." and "There are three distinct methods to be considered." Paice built an extractive summarizer which uses the presence of phrases from a list that he compiled, such as "The main goal of our paper . . . ", to determine an initial set of seed sentences that should be selected. Then an aggregation procedure adds sentences preceding or following the seed until all exophora are resolved. Paice also suggested modifying sentences to resolve exophora when the reference can be found but did not implement an actual system for doing this. Paice's research was the first to point out the problem of accidentally including exophora in extractive summaries, but the solution of simply adding more sentences until the antecedent is found is not satisfactory and much later research on using language generation for summarization has revisited the problem as we discuss in Section 4.

1.2.2 Non-extractive Approaches

The current state of the art in the vast majority of the cases completely ignores issues of language generation and relies on sentence extraction, producing *extractive summaries* composed of important sentences taken verbatim from the input. The sole emphasis in such systems is to identify the important sentences that should appear in the summary. Meanwhile, the development of automatic methods for language generation and text quality has become somewhat independent subfields of research motivated but not directly linked to the field of summarization. Below we briefly introduce some of the main areas of research that are needed for enhancing current summarization systems.

Sentence ordering. This is the problem of taking several sentences, such as those deemed to be important by an extractive summarizer, and presenting them in the most coherent order.

Below we reproduce an example from [9], that shows two different orderings of the same content. The first example is one rated as poor by readers, and the second is one rated as good. The examples make it clear that the order of presentation makes a big difference for the overall quality of the summary and that certain orderings may pose

problems for the reader trying to understand the gist of the presented information.

Summary 1; rated poor

P1 Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania.
P2 Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large.
P3 President Clinton said, The intended victims of this vicious crime stood for everything that is right about our country and the world.
P4 U.S. federal prosecutors have charged 17 people in the bombings.
P5 Albright said that the mourning continues.
P6 Kenyans are observing a national day of mourning in honor of the 215 people who died there.

Summary 2; rated good

P1 Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania. Kenyans are observing a national day of mourning in honor of the 215 people who died there.
P2 Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large. U.S. federal prosecutors have charged 17 people in the bombings.
P3 President Clinton said, "The intended victims of this vicious crime stood for everything that is right about our country and the world". Albright said that the mourning continues.

Sentence revision. Sentence revision was historically the first language generation task attempted in the context of summarization [89, 116, 146, 147, 162]. Sentence revision involves re-using text collected from the input to the summarizer, but parts of the final summary are automatically modified by substituting some expressions with other more appropriate expressions, given the context of the new summary. Types of revisions proposed by early researchers include ELIMINATION of unnecessary parts of the sentences, COMBINATION of information originally expressed in different sentences and SUBSTITUTION of a pronoun with a more descriptive noun phrase where the context of the summary requires this [116]. Given that implementation of these revision operations can be quite complex, researchers in the field eventually

established largely non-overlapping sub-fields of research, each concentrating on only one type of revision.

Sentence fusion. Sentence fusion is the task of taking two sentences that contain some overlapping information, but that also have fragments that are different. The goal is to produce a sentence that conveys the information that is common between the two sentences, or a single sentence that contains all information in the two sentences, but without redundancy.

Here we reproduce two examples of fusion from [96]. The first one conveys only the information that is common to two different sentences, A and B, in the input documents to be summarized (intersection), while the second combines all the information for the two sentences (union).

Sentence A Post-traumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.

Sentence B Post-traumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Fusion 1 Post-traumatic stress disorder (PTSD) is a psychological disorder.

Fusion 2 Post-traumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Sentence compression. Researchers interested in sentence compression were motivated by the observation that human summaries often contain parts of sentences from the original documents which are being summarized, but some portions of the sentence are removed to make it more concise.

Below we reproduce two examples from the Ziff-Davis corpus of sentence compression performed by a person, alongside the original sentence from the document. It is clear from the examples that compression not only shortens the original sentences but also makes them much easier to read.

Comp1 The Reverse Engineer Tool is priced from $8,000 for a single user to $90,000 for a multiuser project site.

Orig1 The Reverse Engineer Tool is available now and is priced on a site-licensing basis, ranging from $8,000 for a single user to $90,000 for a multiuser project site.

Comp2 Design recovery tools read existing code and translate it into definitions and structured diagrams.

Orig2 Essentially, design recovery tools read existing code and translate it into the language in which CASE is conversant — definitions and structured diagrams.

We discuss sentence ordering and language generation approaches in Section 4. The examples above clearly demonstrate the need for such approaches in order to build realistic, human-like, summarization systems. Yet the majority of current systems rely on sentence extraction for selecting content and do not use any of the text-to-text generation techniques, leaving the opportunity for significant improvements with further progress in language generation.

1.3 Evaluation Issues

The tension between generic and query-focused summarization, sentence-extraction and more sophisticated methods was also apparent in the context of the DUC (Document Understanding Conference) Evaluation Workshops [163]. Despite its name, DUC was initially formed in 2001 to evaluate work on summarization and was open to any group interested in participating. Its independent advisory board was charged with identifying tasks for evaluation. Generic summarization was the initial focus, but in later years it branched out to cover various task-based efforts, including a variation on query-focused summarization, topic-based summarization. The first of the topic-based tasks was to provide a summary of information about a person (similar to a biography), given a set of input documents on that person, while later on, a system was provided with a paragraph-length topic and a set of documents and was to use information within the documents to create a summary that addressed the topic.

Generic single document summarization of news was discontinued as a task at DUC after the first two years of the evaluations because no automatic summarizer could outperform the simple baseline consisting of the beginning of the news article, when using manual evaluation metrics. Similarly, for the task of headline generation — creating a 10-word summary of a single news article — no automatic approach

outperformed the baseline of using the original headline of the article. For both tasks, human performance was significantly higher than that of the baselines, showing that while not yet attainable, better performance for automatic systems is possible [148].

DUC, which was superseded by the Text Analysis Conference (TAC) in 2007, provided much needed data to the research community, allowing the development of empirical approaches. Given the difficulty of evaluation, DUC fostered much research on evaluation. However, because the metrics emphasized content selection, research on the linguistic quality of a summary was not necessary. Furthermore, given the short time-frame within which tasks were introduced, summarization researchers who participated in DUC were forced to come up with a solution that was quick to implement. Exacerbating this more, given that people like to win, researchers were more likely to try incremental, safe approaches that were likely to come out on top. Thus, DUC in part encouraged the continuation of the "safe" approach, sentence extraction, even while it encouraged research on summarization and evaluation.

1.4 Where Does Summarization Help?

While evaluation forums such as DUC and TAC enable experimental setups through comparison to a gold standard, the ultimate goal in development of a summarization system is to help the end user perform a task better. Numerous task-based evaluations have been performed to establish that summarization systems are indeed effective in a variety of tasks. In the TIPSTER Text Summarization Evaluation (SUMMAC), single-document summarization systems were evaluated in a task-based scenario developed around the tasks of real intelligence analysts [113]. This large-scale study compared the performance of a human in judging if a particular document is relevant to a topic of interest, by reading either the full document or a summary thereof. It established that automatic text summarization is very effective in relevance assessment tasks on news articles. Summaries as short as 17% of the full text length sped up decision-making by almost a factor of two, with no statistically significant degradation in accuracy. Query-focused summaries are also

very helpful in making relevance judgments about retrieved documents. They enable users to find more relevant documents more accurately, with less need to consult the full text of the document [203].

Multi-document summarization is key for organizing and presenting search results in order to reduce search time, especially when the goal of the user is to find as much information as possible about a given query [112, 131, 181]. In McKeown et al. [131], users were given a task of writing reports on specified topics, with an interface containing news articles, some relevant to the topic and some not. When articles were clustered and summaries for the related articles were provided, people tended to write better reports, but moreover, they reported higher satisfaction when using the information access interface augmented with summaries; they felt they had more time to complete the task. Similarly, in the work of Mana-López et al. [112], users had to find as many aspects as possible about a given topic. Clustering similar articles returned from a search engine together proved to be more advantageous than traditional ranked list presentation, and considerably improved user accuracy in finding relevant information. Providing a summary of the articles in each cluster that conveys the similarities between them, and single-document summaries highlighting the information specific to each document, also helped users in finding information, but in addition considerably reduced time as users read fewer full documents.

In summarization of scientific articles, the user goal is not only to find articles relevant to their interest, but also to understand in what respect a scientific paper relates to the previous work it describes and cites. In a study to test the utility of scientific paper summarization for determining which of the approaches mentioned in the paper are criticized and which approaches are supported and extended, automatic summaries were found to be almost as helpful as human-written ones, and significantly more useful than the original article abstract [199].

Voicemail summaries are helpful for recognizing the priority of the message, the call-back number, or the caller [95]; summaries of threads in help forums are useful in deciding if the thread is relevant [151], and summaries of meetings are a necessary part of interfaces for meeting browsing and search [205].

Numerous studies have also been performed to investigate and confirm the usefulness of single document summaries for improvement of other automated tasks. For example, Sakai and Sparck Jones [182] present the most recent and extensive study (others include [22] and several studies conducted in Japan and published in Japanese) on the usefulness of generic summaries for indexing in information retrieval. They show that, indeed, indexing for retrieval based on automatic summaries rather than full document text helps in certain scenarios for precision-oriented search. Similarly, query expansion in information retrieval is much more effective when potential expansion terms are selected from a summary of relevant documents instead of the full document [100].

Another unexpectedly successful application of summarization for improvement of an automatic task has been reported by [23]. They examined the impact of summarization on the automatic topic classification module that is part of a system for automatic scoring of student GMAT essays. Their results show that summarization of the student essay significantly improves the performance of the topical analysis component. The conjectured reason for the improvement is that the students write these essays under time constraints and do not have sufficient time for revision and thus their writing contains some digressions and repetitions, which are removed by the summarization module, allowing for better assessment of the overall topic of the essay.

The potential uses and applications of summarization are incredibly diverse as we have seen in this section. But how do these systems work and what are the open problems not currently handled by systems? We turn to this discussion next.

1.5 Article Overview

We begin our summarization overview with a presentation of research on sentence extraction in Section 2. In that section, we first present earlier research on summarization that experimented with methods for determining sentence importance that are based on variants of frequency. Machine learning soon became the method of choice for determining pertinent features for selecting sentences. From there, we move

to graph-based approaches that select sentences based on their relations to other sentences. Finally, we close Section 2 by looking at the use of sentence extraction for query-focused summarization. One case of query-focused summarization is the generation of biographical summaries and we see that when the task is restricted (here to one class of queries), researchers begin to develop approaches that differ substantially from the typical generic extraction based approach.

In Section 3, we continue to survey extractive approaches, but move to methods that do more sophisticated analysis to determine importance. We begin with approaches that construct lexical chains which represent sentence relatedness through word and synonym overlap across sentences. The hypothesis is that each chain represents a topic and that topics that are pursued for greater lengths are likely to be more salient. We then turn to approaches that represent or compute concepts and select sentences that refer to salient concepts. Finally, we turn to methods that make use of discourse information, either in the form of rhetorical relations between sentences, or to augment graph-based approaches.

In Section 4, we examine the different sub-fields that have grown up around various forms of sentence revision. We look at methods that compress sentences by removing unnecessary detail. We then turn to methods that combine sentences by fusing together repeated and salient information from different sentences in the input. Next, we turn to work that edits summary sentences, taking into account the new context of the summary. We close with research on ordering of summary sentences.

In the final section on approaches, Section 5, we survey research that has been carried out for specific genres and domains. We find that often documents within a specific genre have an expected structure and that structure can be exploited during summary generation. This is the case, for example, with journal article summarization. At other times, we find that while the form of the genre creates problems (e.g., speech has disfluencies and errors resulting from recognition that cause difficulties), information beyond the words themselves may be available to help improve summarization results. In speech summarization, acoustic and prosodic clues can be used to identify important information, while in very recent work on web summarization, the structure of the

web can be used to determine importance. In some domains, we find that domain dependant semantic resources are available and the nature of the text is more regular so that semantic interpretation followed by generation can be used to produce the summary; this is the case in the medical domain.

Before concluding, we provide an overview in Section 6 on research in summarization evaluation. Much of this work was initiated with DUC as the conference made evaluation data available to the community for the first time. Methodology for evaluation is a research issue in itself. When done incorrectly, evaluation does not accurately reveal which system performs better. In Section 6, we review *intrinsic methods* for evaluation. Intrinsic refers to methods that evaluate the quality of the summary produced, usually through comparison to a gold standard. This is in contrast to *extrinsic evaluation* where the evaluation measures the impact of the summary on task performance such as the task-based evaluations that we just discussed. We review metrics used for comparison against a gold standard as well as both manual and automatic methods for comparison. We discuss the difference between evaluation of summary content and evaluation of the linguistic quality of the summary.

2

Sentence Extraction: Determining Importance

In this section, we survey methods for determining sentence importance for extractive summarization.

We begin with a detailed discussion of robust, easily computable features for determining the importance of a sentence; these features do not rely on any sources of information beyond the input to the summarizer. The approaches in this class are completely unsupervised, and do not require any sort of human judgements about what sentences need to be present in the summary. The features vary in terms of expressive power and sophistication but are all ultimately directly related to tracking some form of repetition, redundancy or frequency in the input. Frequency can be counted on the word level and we present three approaches with increasing complexity and descriptive power for doing so: word probability, TF*IDF word weight, and the log-likelihood ratio (LLR) test for determining if a word is indicative of the topic of the input. We discuss these in detail because variants of TF*IDF weights are used in some form in most extractive summarizers, while a system that relies on the LLR test achieved the best performance in determining content in the official DUC evaluations for multi-document summarization of news [40]. We also introduce methods in which repetition

of information is tracked on the sentence level, through sentence clustering, or via a combination of lexical and sentence information, as in graph-based models.

The repetition of the same entity or topic can be done using different lexical items: either synonyms or related words, or pronouns. In order to make possible a more accurate account of frequency, researchers have made use of coreference resolution systems and of existing lexical resources which are either manually or automatically constructed and which contain information about related words. In these methods, importance is determined for concepts, lexical chains or entities instead of words. Latent semantic analysis, which has been extensively used for summarization of meetings, combines aspects of the lexical frequency approaches and those informed by knowledge on lexical relationships, but the knowledge is implicit, derived from word co-occurrences in a large number of different documents.

Use of machine learning approaches allows researchers the freedom to use numerous features. We overview only the main machine learning approaches that had the most significant impact on summarization research at the time they were proposed. Numerous later papers discuss machine learning either as a standard tool or in the context of comparing different machine learning approaches and techniques, without direct bearing on summarization itself.

Finally, we discuss a distinction in sentence extraction that is rarely emphasized: are sentences selected one by one, or jointly. In greedy approaches in the tradition of maximal marginal relevance, sentences are selected in order, one after the other. In global optimization approaches, the best overall summary is selected, consisting of several sentences that jointly satisfy some conditions.

2.1 Unsupervised Data-driven Methods

Unsupervised methods for sentence extraction are the predominant paradigm in extractive summarization. They do not require human annotation for training statistical methods for importance. They are "data-driven" because they do not rely on any external knowledge sources, models or on linguistic processing and interpretation.

2.1.1 Methods Based on Word Frequency

As we discussed in our historical overview of automatic summarization in Section 1, the earliest work in the area, conducted by Luhn, explored the use of lexical frequency as an indicator of importance. One way to capture his intuitions is to calculate *word probability* for content words from the input to the summarizer. But as Luhn pointed out, some of the most frequent words in the document might not be indicative of what is important in a particular document because they occur often in many documents. TF*IDF *weights* have been proposed in information retrieval to deal with exactly this problem; in addition to frequency in the input, this method incorporates evidence from a background corpus to adjust the weights of individual words so that the words with highest weights are those most likely to be descriptive of the topic of the particular document. *Log-likelihood ratio* approaches not only use a background corpus, but also allow for the definition of topic signature words that are the sole words in the input that determine importance of sentences while other words are entirely ignored in the calculation of importance. Of these three approaches, the log-likelihood one leads to best results for greedy sentence-by-sentence multi-document summarization of news [40, 70].

Word probability is the simplest form of using frequency in the input as an indicator of importance.[1] The probability of a word w, $p(w)$ is computed from the input, which can be a cluster of related documents or a single document. It is calculated as the number of occurrences of a word, $c(w)$ divided by the number of all words in the input, N:

$$p(w) = \frac{c(w)}{N} \tag{2.1}$$

Given this probability distribution over words, the likelihood of a summary can be computed based on a multinomial distribution:

$$L[\text{sum}] = \frac{M!}{n_1! \ldots n_r!} p(w_1)^{n_1} \cdot \cdots \cdot p(w_r)^{n_r} \tag{2.2}$$

[1] Raw frequency would be even simpler, but this measure is too strongly influenced by document length. A word appearing twice in a 10 word document may be important, but not necessarily so in a 1000 word document. Computing word probability makes an adjustment for document length.

where M is the number of words in the summary, $n_1 + \cdots + n_r = M$ and for each i, n_i is the number of times word w_i appears in the summary and $p(w_i)$ is the probability of w_i appearing in the summary estimated from the input documents.

Nenkova et al. [155] analyzed 30 DUC inputs consisting of multiple news articles on the same topic, along with four human abstracts for each, and found that for this data, the likelihood of human summaries is higher than that of automatically produced summaries. These findings indicate that when writing abstracts for multi-document inputs, people do tend to be guided by frequency in their selections of topics to be included in their summaries.

SumBasic is one system developed to operationalize the idea of using frequency for sentence selection. It relies only on word probability to calculate importance [212]. For each sentence S_j in the input it assigns a weight equal to the average probability $p(w_i)$ of the content words in the sentence, estimated from the input for summarization:

$$\text{Weight}(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|\{w_i | w_i \in S_j\}|} \tag{2.3}$$

Then, in a greedy fashion, SumBasic picks the best scoring sentence that contains the word that currently has the highest probability. This selection strategy assumes that at each point when a sentence is selected, a single word — that with highest probability — represents the most important topic in the document and the goal is to select the best sentence that covers this word. After the best sentence is selected, the probability of each word that appears in the chosen sentence is adjusted. It is set to a smaller value, equal to the square of the probability of the word at the beginning of the current selection step, to reflect the fact that the probability of a word occurring twice in a summary is lower than the probability of the word occurring only once. This selection loop is repeated until the desired summary length is achieved.

An approach that optimizes the occurrence of important words globally over the entire summary instead of greedy selection outperforms SumBasic [198].

TF*IDF weighting (Term Frequency*Inverse Document Frequency) One problem with frequency as an indicator of importance is the fact that in any general text the occurrence of words follows a Zipfian distribution [3], with a few words appearing very often and many words appearing infrequently. The most frequent words are known in information retrieval as stop words. Stop words include determiners, prepositions and auxiliary verbs, or common domain words. Obviously, such words are not indicative of topicality. Determining a useful cut-off threshold that works for a variety of documents and document lengths is not trivial. Instead, researchers have opted for use of a stop word list consisting of the most frequent words in a language and a particular domain — such words would not play any role in determining sentence importance. Deciding what words should be included in the stop word list is not straightforward though and some most recent methods have attempted to dynamically model which words should be considered representative of general English rather than of the topic of a given document [45].

Alternatively, the TF*IDF weighting of words [184, 193], as traditionally used in information retrieval, can be employed. This weighting exploits counts from a large background corpus, which is a large collection of documents, normally from the same genre as the document that is to be summarized; the background corpus serves as indication of how often a word may be expected to appear in an arbitrary text.

The only additional information besides the term frequency $c(w)$ that we need in order to compute the weight of a word w which appears $c(w)$ times in the input for summarization is the number of documents, $d(w)$, in a background corpus of D documents that contain the word. This allows us to compute the inverse document frequency:

$$\text{TF}^*\text{IDF}_w = c(w) \times \log \frac{D}{d(w)} \tag{2.4}$$

In many cases $c(w)$ is divided by the maximum number of occurrences of any word in the document, which normalizes for document length. Descriptive topic words are those that appear often in a document, but are not very common in other documents. In contrast, stopwords appear in practically all documents, so their IDF weight will

be close to zero. The TF*IDF weights of words are good indicators of importance, and they are easy and fast to compute. These properties explain why TF*IDF is one of the most commonly used features for extractive summarization: it is incorporated in one form or another in most current systems [56, 58, 61, 62, 63, 81].

Log-likelihood ratio test for topic signatures An even more powerful use of frequency is the application of the log-likelihood ratio test [51] for identification of words that are highly descriptive of the input. Such words have been traditionally called "topic signatures" in the summarization literature [106].

Topic signatures are words that occur often in the input but are rare in other texts, similarly to words with high TF*IDF weight. Unlike TF*IDF, the log-likelihood ratio test provides a way of setting a threshold to divide all words in the input into either descriptive or not.

Information about the frequency of occurrence of words in a large background corpus is necessary to compute the statistic on the basis of which topic signature words are determined. The likelihood of the input I and the background corpus is computed under two assumptions: (H1) that the probability of a word in the input is the same as in the background B or (H2) that the word has a different, higher probability, in the input than in the background.

H1: $P(w|I) = P(w|B) = p$ (w is not descriptive)

H2: $P(w|I) = p_I$ and $P(w|B) = p_B$ and $p_I > p_B$ (w is descriptive)

The likelihood of a text with respect to a given word of interest, w, is computed via the binomial distribution formula. The input and the background corpus are treated as a sequence of words w_i: $w_1 w_2 \ldots w_N$. The occurrence of each word is a Bernoulli trial with probability p of success, which occurs when $w_i = w$. The overall probability of observing the word w appearing k times in the N trials is given by the binomial distribution

$$b(k, N, p) = \binom{N}{k} p^k (1-p)^{N-k} \tag{2.5}$$

For H1, the probability p is computed from the input and the background collection taken together. For H2, p_1 is computed from the input, p_2 from the background, and the likelihood of the entire data

is equal to the product of the binomial for the input and that for the background. More specifically, the likelihood ratio is defined as

$$\lambda = \frac{b(k, N, p)}{b(k_I, N_I, p_I).b(k_B, N_B, p_B)} \tag{2.6}$$

where the counts with subscript I are computed only from the input to the summarizer and those with index B are computed over the background corpus.

The statistic equal to -2λ has a known statistical distribution (χ^2), which can be used to determine which words are topic signatures. Topic signature words are those that have a likelihood statistic greater than what one would expect by chance. The probability of obtaining a given value of the statistic purely by chance can be looked up in a χ^2 distribution table; for instance a value of 10.83 can be obtained by chance with probability of 0.001.

The importance of a sentence is computed as *the number of topic signatures* it contains, or the proportion of topic signatures in the sentence; words are no longer assigned weights in this approach.

An example summary produced using topic signature words as indicators of importance [38], as well as a human abstract and a baseline summary consisting of the beginning of the latest article in the input are shown below (set D30003 from DUC 2004). The example shows that the baseline, as expected, focuses on only a single aspect of the story, while the topic words summarizer is able to select sentences from the input that cover most of the information chosen by the person writing the model abstract.

Baseline The Swiss government has ordered no investigation of possible bank accounts belonging to former Chilean Dictator Augusto Pinochet, a spokesman said Wednesday. Weekend newspaper reports in Spain said a Spanish judge who ordered Pinochet's arrest has issued a petition aimed at freezing any accounts the 82-year-old general might have in Luxembourg and Switzerland. But government spokesman Achille Casanova said no accounts have so far been frozen in Switzerland and no investigation order has been given to federal banking authorities. Pinochet has been held at a London clinic since his arrest earlier this month.

Topic signatures As his lawyers in London tried to quash a Spanish arrest warrant for Gen. Augusto Pinochet, the former Chilean Dictator, efforts began in Geneva and Paris to have him extradited. Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean Dictator has diplomatic immunity is ridiculous. Margaret Thatcher entertained former Chilean Dictator Gen. Augusto Pinochet at her home two weeks before he was arrested in his bed in a London hospital, the ex-prime minister's office said Tuesday, amid growing diplomatic and domestic controversy over the move.

Human abstract Former Chilean dictator Augusto Pinochet has been arrested in London at the request of the Spanish government. Pinochet, in London for back surgery, was arrested in his hospital room. Spain is seeking extradition of Pinochet from London to Spain to face charges of murder in the deaths of Spanish citizens in Chile under Pinochet's rule in the 1970s and 1980s. The arrest raised confusion in the international community as the legality of the move is debated. Pinochet supporters say that Pinochet's arrest is illegal, claiming he has diplomatic immunity. The final outcome of the extradition request lies with the Spanish courts.

The idea of topic signature terms for summarization was introduced by Lin and Hovy [106] in the context of single document summarization, as part of the SUMMARIST [81] system. Later systems also used topic signature features for the task of multi-document summarization of news [38, 40, 98]. Topic signatures are more powerful than direct use of frequency because they give a natural cut-off for deciding which words should be considered topical, based on an actual probability distribution [70].

2.1.2 Sentence Clustering

In multi-document summarization of news, the input by definition consists of several articles, possibly from different sources, on the same topic. In this setting, it is very likely that across the different articles there will be sentences that contain similar information. Information that occurs in many of the input documents is considered important and worth selecting in a summary. Of course, verbatim repetition on

the sentence level is not that common across sources. Rather, *similar* sentences can be *clustered* together [133, 76, 190]. Clusters with many sentences represent important topic themes in the input. Selecting one representative sentence from each main cluster is one way to produce an extractive summary using this approach, while minimizing possible redundancy in the summary.

The sentence clustering approach to multi-document summarization again exploits repetition, but at the sentence rather than the word level. The more sentences there are in a cluster, the more important the information in the cluster is considered. Below is an example of a sentence cluster from different documents in the input to a multi-document summarizer. All four sentences share common content, which is considered important.

S1 PAL was devastated by a pilots' strike in June and by the region's currency crisis.
S2 In June, PAL was embroiled in a crippling three-week pilots' strike.
S3 Tan wants to retain the 200 pilots because they stood by him when the majority of PAL's pilots staged a devastating strike in June.
S4 In June, PAL was embroiled in a crippling three-week pilots' strike.

A drawback of the clustering approach is that each sentence is assigned to only one cluster, which is a restrictive requirement for typical sentences that express several facts. The graph-based approaches discussed next allow for much more flexibility in representation and lead to better overall results in content selection for multi-document summarization.

2.1.3 Graph-based Methods for Sentence Ranking

Graph-based methods for sentence ranking productively exploit repetition in the input, both on the word and sentence level. Sentence similarity is measured as a function of word overlap, so frequently occurring words would link many sentences, and similar sentences give support for each other's importance. In this way, graph approaches combine advantages from word frequency and sentence clustering methods [169]. Moreover, such methods provide a formal model for computing sentence importance.

In the graph-based models that have been most popular recently and have been evaluated on DUC data [56, 135], the input is represented as a highly connected graph. Vertices represent sentences and edges between sentences are assigned weights equal to the similarity between the two sentences. The method most often used to compute similarity is cosine similarity with TF*IDF weights for words. The weights of edges connecting sentences that share many lexical items will be higher than for those connecting sentences that have fewer lexical items in common. In this way, word frequency plays a direct role in determining the structure of the graph. Sometimes, instead of assigning weights to edges, the connections between vertices can be determined in a binary fashion: the vertices are connected only if the similarity between the two sentences exceeds a pre-defined threshold. In addition, this graph representation is more flexible than sentence clustering — tightly connected cliques in the graph could be seen as representing a cluster, but sentences are no longer constrained to belong to exactly one cluster.

Vertex importance or centrality can be computed using general graph algorithms, such as PageRank [56, 135]. When the weights of the edges are normalized to form a probability distribution so that the weight of all outgoing edges from a given vertex sum up to one, the graph becomes a Markov chain and the edge weights correspond to the probability of transitioning from one state to another. Standard algorithms for stochastic processes can be used to compute the probability of being in each vertex of the graph at time t while making consecutive transitions from one vertex to next. As more and more transitions are made, the probability of each vertex converges, giving the *stationary distribution* of the chain. The stationary distribution simply gives the probability of (being at) a given vertex and can be computed using iterative approximation. Vertices with higher probabilities correspond to more important sentences that should be included in the summary.

A summary produced using this graph-based approach for a DUC 2004 input set is shown below [55]:

Graph-based summary Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean Dictator Augusto Pinochet calling it a case of international meddling. Pinochet, 82, was

> placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. The Chilean government has protested Pinochet's arrest insisting that as a senator he was traveling on a diplomatic passport and had immunity from arrest. Castro, Latin America's only remaining authoritarian leader, said he lacked details on the case against Pinochet but said he thought it placed the government of Chile and President Eduardo Frei in an uncomfortable position.

In the official DUC 2004 manual evaluation where both the topic words summarizer discussed in Section 2.1 [40] and the graph-based summarizer [55] were evaluated, the topic signature summarizer performed better in content selection. Both summarizers outperform by a wide margin the two traditional baselines for multi-document summarization: the beginning of the latest article in the cluster or a collection of the first sentences of most recent articles. Both topic word and graph-based approaches are concerned only with pinpointing important content and do not make any special decisions about sentence ordering, for example, even though the topic words system does perform some linguistic processing of the input text.

Graph-based approaches have been shown to work well for both single-document and multi-document summarization [56, 135]. Since the approach does not require language-specific linguistic processing beyond identifying sentence and word boundaries, it can also be applied to other languages, for example, Brazilian Portuguese [136]. At the same time, incorporating syntactic and semantic role information in the building of the text graph leads to superior results over plain TF*IDF cosine similarity [29].

Using different weighting schemes for links between sentences that belong to the same article and sentences from different articles can help separate the notions of topicality within a document and recurrent topics across documents. This distinction can be easily integrated in the graph-based models for summarization [216].

Graph representations for summarization had been explored even before the PageRank models became popular. For example, the purpose of a graph-based system for multi-document summarization developed by Mani and Bloedorn [115] is to identify salient regions of each

story related to a topic given by a user, and compare the stories by summarizing similarities and differences. The vertices in the graph are *words*, *phrases* and *named entities* rather than sentences and their initial weight is assigned using TF*IDF. Edges between vertices are defined using synonym and hypernym links in WordNet, as well as coreference links. Spreading activation is used to assign weights to non-query terms as a function of the weight of their neighbors in the graph and the type of relation connecting the nodes.

In order to avoid problems with coherence that may arise with the selection of single sentences, Salton et al. [183] argue that a summarizer should select full paragraphs to provide adequate context. Their algorithm constructs a text graph for a document using cosine similarity between each pair of paragraphs in the document. The shape of the text graph determines which paragraphs to extract. In their experiments, they show that two strategies, selecting paragraphs that are well connected to other paragraphs or first paragraphs of topical text segments within the graph, both produce good summaries.

A combination of the subsentential granularity of analysis where nodes are words and phrases rather than sentences and edges are syntactic dependencies has also been explored [103]. Using machine learning techniques, Leskovec et al. [103] attempt to learn what portions of the input graph would be included in a summary. In their experiments on single document summarization of news articles, properties of the graph such as incoming and outgoing links, connectivity and PageRank weights are identified as the best class of features that can be used for content selection.

2.2 Machine Learning for Summarization

As researchers proposed more and more indicators for sentence importance, it became necessary to come up with ways in which the different indicators can be combined. To address these problems, Kupiec et al. [97] introduced their proposal for using machine learning techniques on a corpus of summary/document pairs. Kupiec et al. claimed that statistical analysis of the corpus would both reveal what features should be used and how they should be weighted relative to

each other. For their experiments they choose three features following earlier work by Paice [165], Luhn [111], and Edmundson [52]: fixed (cue) phrases, paragraph features related to the position features of earlier work distinguishing sentence-initial and sentence-final sentences within a paragraph as well as paragraph-initial and paragraph-final positions, and word frequency features. They also introduce two new features: uppercase word feature to account for the importance of proper names and acronyms and sentence length cut-off to avoid including sentences below a length threshold.

For learning, they used a Naive Bayes classifier, making the assumption that the employed features are independent of each other given the class. Their training corpus consisted of technical articles drawn from a collection provided by Engineering Information. In addition, the corpus contained abstracts for each article created by professional abstractors. It consisted of a total of 188 document/summary pairs which were drawn from 21 publications in the scientific and technical domains. Abstract sentences were matched against input article sentences using a variety of techniques, such as *exact match*, *join* of two article sentences for the abstract sentences, or *incomplete* partial matches. Seventy-nine percent of the abstract sentences in their corpus were exact matches.

The results from the machine learning experiments for content selection show that a combination of location of the sentence, fixed-phrase and sentence length gave the best results. When frequency features were added for this single-document summarization task, performance decreased slightly.

In later work, Conroy and O'Leary [39] show how a Hidden Markov Model (HMM), which has fewer independence assumptions than Naive Bayes learners, can be used for summarization. They experiment with three features similar to those described in earlier work: position of sentence in the document, number of terms in a sentence, and the probability of a term estimated from the input. Conroy and O'Leary also deliberately exploit Markov dependencies: they suggest that the probability that the next sentence in a document is included in a summary will depend on whether the current document sentence is already part of the summary.

To obtain training data, they asked a single person to produce extractive summaries.

In later versions of the HMM summarizer, the number of topic signature words in the sentence was added as a feature. Follow up experiments showed that a data-driven summarizer using the topic word features has a similarly high performance to that of the supervised HMM system.

Since this seminal work on summarization by learning, quite a few others have also looked at the use of learning for summarization [60, 74, 81, 103, 160, 220, 232]. This later work focuses primarily on the comparison of different learning techniques and feature classes. For generic multi-document summarization of news, these methods have not been shown to directly outperform competitive unsupervised methods based on a single feature such as the presence of topic words and graph methods. Machine learning approaches have proved to be much more successful in domain or genre specific summarization, where classifiers can be trained to identify specific types of information such as sentences describing literature background in scientific article summarization or utterances expressing agreement or disagreement in meetings.

An additional problem inherent in the supervised learning paradigm is the necessity of labeled data on which classifiers can be trained. Asking annotators to select summary-worthy sentences [209] is time consuming, and thus, many researchers have concentrated their efforts on developing methods for automatic alignment of human abstracts and the input [8, 44, 88, 121, 232] in order to provide labeled data of summary and non-summary sentences for machine learning. This approach could be potentially problematic because, as we discuss in Section 6, different writers can choose different content for their abstracts and therefore summary-worthy information may not be identifiable based on a single abstract. To mitigate this issue, some researchers have proposed to leverage the information from manual evaluation of content selection in summarization in which multiple sentences can be marked as expressing the same fact that should be in the summary [41, 60]. Alternatively, one could compute similarity between sentences in human abstracts and those in the input in order to find very similar sentences, not necessarily doing full alignment [28].

Overall, given our current knowledge about sentence extraction approaches and their evaluation, the introduction of supervised methods has not lead to measurable improvements in content selection for generic summarization of news. Extraction approaches using a single feature, such as topic words or centrality measure weights in the graph methods, are the ones that have the best reported results. Currently, there are no good ways to combine such features in a joint measure of importance, except through supervised methods. Future work will need to establish if indeed the combination of strong data-driven indicators of importance will outperform the individual unsupervised methods.

2.3 Sentence Selection vs. Summary Selection

Most summarization approaches choose content sentence by sentence: they first include the most informative sentence, and then if space constraints permit, the next most informative sentence is included in the summary and so on. Some process of checking for similarity between the chosen sentences is also usually employed in order to avoid the inclusion of repetitive sentences.

One of the early summarization approaches for both generic and query focused summarization that has been widely adopted is *Maximal Marginal Relevance* (MMR) [25]. In this approach, summaries are created using greedy, sentence-by-sentence selection. At each selection step, the greedy algorithm is constrained to select the sentence that is maximally relevant to the user query and minimally redundant with sentences already included in the summary. MMR measures relevance and novelty separately and then uses a linear combination of the two to produce a single score for the importance of a sentence in a given stage of the selection process. To quantify both properties of a sentence, Carbonell and Goldstein use cosine similarity [118]. For relevance, similarity is measured to the query, while for novelty, similarity is measured against sentences selected so far. The MMR approach was originally proposed for query-focused summarization in the context of information retrieval, but could easily be adapted for generic summarization, for example as using the entire input as a user query as proposed by Gong and Liu [69]. Many have adopted this seminal approach, mostly

in its generic version, sometimes using different measures of novelty to select new sentences [142, 202, 221].

This greedy approach of sequential sentence selection might not be that effective for optimal content selection of the entire summary. One typical problematic scenario for greedy sentence selection (discussed in McDonald [127]) is when a very long and highly relevant sentence happens to be evaluated as the most informative. Such a sentence may contain several pieces of relevant information, alongside some not so relevant facts which could be considered noise. Including such a sentence in the summary will help maximize content relevance at the time of selection, but at the cost of limiting the amount of space in the summary remaining for other sentences. In such cases it is often more desirable to include several shorter sentences, which are individually less informative than the long one, but which taken together do not express any unnecessary information.

General global optimization algorithms can be used to solve the new formulation of the summarization task, in which the best overall summary is selected. Given some constraints imposed on the summary, such as maximizing informativeness, minimizing repetition, and conforming to required summary length, the task would be to select the best summary. Finding an exact solution to this problem is NP-hard [58], but approximate solutions can be found using a dynamic programming algorithm [127, 224, 225].

Even in global optimization methods, informativeness is still defined and measured using features well-explored in the sentence selection literature (see Section 2). These include word frequency and position in the document [225], TF*IDF [58, 127], and concept frequency [224]. Global optimization approaches to content selection have been shown to outperform greedy selection algorithms in several evaluations using news data as input, and have proved to be especially effective for extractive summarization of meetings [67, 179].

In a detailed study of global inference algorithms [127], it has been demonstrated that it is possible to find an exact solution for the optimization problem for content selection using Integer Linear Programming. The performance of the approximate algorithm based on dynamic programming was lower, but comparable to that of the exact solutions.

In terms of running time, the greedy algorithm is very efficient, almost constant in the size of the input. The approximate algorithm scales linearly with the size of the input and is thus indeed practical to use. The running time for the exact algorithm grows steeply with the size of the input and is unlikely to be useful in practice [127].

2.4 Sentence Selection for Query-focused Summarization

Query-focused summarization became a task in the Document Understanding Conference in 2004, reviving interest in this type of news summarization which was considered standard in earlier SUMMAC evaluations [113]. Given a topic expressed as a short paragraph statement, the idea was to generate a summary that addresses the topic. The creation of this task was in reaction to claims that generic summarization was too unconstrained and that human generated summaries were typically produced in the context of a task. The first efforts on query-focused summarization addressed the problem of generating biographical summaries. Later, the task was extended to more open-ended topics.

In this section we look at two classes of approaches. The first adapts techniques for generic summarization of news, following the reasonable intuition that in query-focused summarization the importance of each sentence will be determined by a combination of two factors: how relevant is that sentence to the user question and how important is the sentence in the context of the input in which it appears. The second class of approaches develops techniques that are particularly appropriate given the question type. We illustrate this class of approaches through a discussion of methods for generating biographical summaries.

2.4.1 Adapting Generic Approaches

Conroy et al. [40] propose adaptation of the use of topic signature words, where sentences that contain more such words are considered more suitable for inclusion in a generic summary and the weight of a sentence is equal to the proportion of topic signature words it contains. To extend this model for query-focused summarization, they

assume the words that should appear in a summary have the following probability: a word has probability zero of appearing in a summary for a user defined topic if it neither appears in the user query nor is a topic signature word for the input; the probability of the word to appear in the summary is 0.5 if it either appears in the user query or is a topic signature, but not both; and the probability of a word to appear in a summary is 1 if it is both in the user query and in the list of topic signature words for the input. These probabilities are arbitrarily chosen, but in fact work well when used to assign weights to sentences equal to the average probability of words in the sentence [40].

The graph-based approach of Erkan and Radev [56] has also been adapted for query-focused summarization with minor modifications. For query-focused summarization, cosine similarity is computed between each sentence and the user query as a measure of the relevance of the sentence to the user need. The similarity between sentences in the input is computed as in the generic summarization setting. Rather than picking one concrete value for the parameter that specifies the relative importance of the user topic and the input as done by Conroy et al. [40] who assigned equal weight to them, Otterbacher et al. [161] chose to find experimentally what parameters would give best results. Their experiments indicate that emphasizing the user query leads to better overall performance of the PageRank model for sentence weighting. The graph formulation of the query-focused task significantly outperforms a competitive baseline which chooses sentences based on their word overlap similarity with the user topic.

The intuitions behind the two approaches described above for adapting generic summarization approaches for the query-focused task can be incorporated in a fully formal Bayesian model for summarization and passage retrieval [45]. In this framework, a sophisticated inference procedure is used to derive a query model: the probability of a given word given the user information need. This model is based not only on the words that actually appear in the user defined topic, but also incorporates knowledge about what distinguished the input for summarization, assumed to be relevant to the user request, from any other documents that are not relevant to the user query.

2.4.2 Biographical and Definition Summarization

Early research on biographical summarization was carried out simultaneously by two different groups of researchers who both proposed hybrid solutions for the problem. Both groups developed systems combining approaches which look for particular types of information in the input and approaches which are data driven, using summarization techniques to sift through lots of information and pulling out the most prominent facts. The systems handle questions asking for a biography of a specified person ("Who is X?") as well as requests for definitions ("What is X?").

Blair-Goldensohn et al. [16] developed a system, DefScriber, which uses rhetorical predicates to specify information that should be included in a definition or biography. For a definition, they use three predicates: *genus*, *species* and *non-specific definitional*. A sentence containing a genus predicate specifies the conceptual category of the person/entity to be described, as for example in the sentence "The Hajj is a type of ritual"; sentences expressing a species predicate convey unique characteristics of the entity, as in "The Hajj begins in the 12th month of the Islamic year". Pattern-based matches were used to find good sentences expressing these kinds of target rhetorical predicates, where patterns were represented as partially lexicalized trees. For example, a pattern for *genus* looked for the query term X as subject, followed by a formative verb where the object represents the genus and modifiers of the object represent the *species*. When a pattern matched a sentence in an input document, the sentence was included in the summary. DefScriber blended this top-down search for information with a bottom-up, data-driven summarization which used search over the web to identify all non-specific definitional sentences related to the term X. For example, the sentence "Pilgrims pay substantial tariffs to the rulers of the lands they pass through" is a non-specific definition for the *What is Hajj?* question. DefScriber applied a classifier to select non-specific definitional sentences from the set of documents returned from a search, then ranked them based on distance from the centroid of all sentences. DefScriber also used coherence constraints when selecting which sentence should come next, in some cases overriding the rank ordering.

While DefScriber was originally developed to generate definitions of queried objects, it also was adapted to generate biographies.

Weischedel et al. [217] also developed a hybrid approach to generating summaries that answer definitional questions. They use linguistic features to target information that should be included in a definitional summary and evaluated the contribution of appositives, copulas, relations, and propositions. If one of these constructions is found modifying a question term, it is extracted for the answer (e.g., in "Tony Blair, the British Prime Minister", where the appositive serves as a good answer in a biography). They also experimented with manual and learned patterns (e.g., "X also known as Y" is a pattern that yields good alternative terms for the query term). They also used a question profile which represents typical responses to similar questions; the question profile was used to rank information extracted using the patterns. They have implemented their system for both biographies and definitional questions and have tested it for both English and Chinese [167]. In their evaluation, they investigated the contribution of targeting particular linguistic features for extraction as an answer versus using the patterns to extract answers, measuring the contribution of manual and learned patterns separately and together. They found that they got the best results with a combination of linguistic features and patterns; since each type performs equally well alone, they conclude that linguistic features and patterns are complementary and both are helpful. They note that the improvement is greater for biographical questions than for definitional questions, which overall seem harder. Patterns are more effective for biographical questions than for definitional questions. Of the linguistic features, they find that copulas are the most effective, with appositives and relations a close second.

A similar combination of goal-driven and data-driven approaches was used in systems developed specifically for generating biographical summaries alone. Schiffman et al. [185] target certain types of syntactic phrases that they predict will more likely provide descriptions of a person. They extract appositions and relative clauses which have the person name as a head, as well as main clauses which have the person as deep semantic subject. They then used information about the verbs involved in the clauses to further filter the content that should

be included in the biography. They carried out a corpus analysis to gather semantic information about the verbs, determining whether a verb carries important information by virtue of being a verb specific to describing people of a particular occupation; clauses with descriptive verbs are retained. In addition, they used typical summarization features such as TF*IDF to filter out clauses. The resulting biographies resemble a profile of a person (who is often mentioned in the news), listing descriptions of the person and his/her typical activities.

Zhou et al. [234] developed a system for multi-document summarization of biographical information. Their system was entered in the DUC 2004 evaluation of biographical summarization and was one of the top performers. They identified nine types of facts that should appear in a biography (e.g., fame, nationality, scandal, work) and use text classification to classify sentences in potentially relevant documents into one of these nine classes. Their approach also included a redundancy detection module. Their approach, therefore, was more targeted than any of the biographical response generators we have already described as it sought information of particular types.

In quite recent work, Biadsy et al. [15] used a combined top-down and bottom-up approach to generate summaries. Like Zhou et al., they are aiming at the generation of more conventional biographies as opposed to summaries that describe what a person has been up to lately. They also developed a classifier that can determine whether a sentence is biographical or not, but rather than specifying particular categories of information, they combine two novel techniques. First, they use Wikipedia biographies as models and train an information extraction system to find similar types of information by using unsupervised learning to mine extraction patterns. The top–down module of the system uses the resulting patterns to find biographical information in the input documents. In addition, they build two language models, one based on the Wikipedia biographies and the other trained on a non-biographical news corpus. The biography can then be generated by selecting sentences which match the biographical language model more than they match the general news language model.

Other approaches to biographical summarization have also used learning to identify biographical type sentences. Duboue et al. [50] used

a database of celebrity biographies and genetic algorithms to identify biographical sentences. Feng and Hovy [57] use a bootstrapping method to learn patterns to find the answers for five fields that typically occur in a biography: birth date, death date, birth place, death place, and spouse. They use seed answer pairs, querying the web to find a set of possible patterns and then prune the patterns using a precision-based method (which requires additional training data) and a re-ranking method using a variant of TF*IDF. They get quite good accuracy for dates, with lower accuracy for the other fields.

2.5 Discussion

In this section, we overviewed the main approaches for content selection in extractive summarization, including basic features that are likely to be mentioned in many publications on the topic, as well as competing paradigms for selection such as data-driven vs. supervised, greedy selection vs. global summary optimization, and generic vs. query-focused summaries. The emphasis on frequency-related features will prepare a novice in the research area to better understand research articles in which these methods are often assumed as common knowledge.

The majority of published work deals with unsupervised greedy sentence selection for sentence extraction for generic summarization. The use of supervised methods has not been well justified in generic summarization of news, but as we show in the case of biographical summarization, is the standard when moving to a specific genre of summaries, in which precise types of information have to be identified. What works best, i.e., features or modeling approaches, is highly dependent on the intended use. Similarly, current work has not clearly demonstrated if greedy selection or global optimization is to be preferred. Global selection is algorithmically and theoretically much more appealing, but greedy selection is generally faster and much more likely to cognitively mirror the human process of summarization. As for query focused summarization of general news, most approaches seem to be simple adaptations of existing methods for generic summarization. When dealing with a specific query type (e.g., biographies), more sophisticated, targeted approaches can be used. These approaches

typically integrate methods that seek specific types of information (e.g., through the use of learned patterns) with data-driven, generic methods. Better capabilities for input interpretation are necessary in order to develop even more sophisticated approaches.

We devote the remaining sections to methods using semantics and discourse information and on domain and genre specific methods, which utilize knowledge of the domain and additional resources to facilitate the summarization process.

3

Methods Using Semantics and Discourse

All methods overviewed in the previous section compute sentence importance on the basis of repeated occurrence of the same word in different places in the input. Even early researchers acknowledged that better understanding of the input would be achieved via methods that instead track references to the same entity, or the same topic in the document. These methods either rely on existing manually constructed semantic resources (lexical chains, concepts), on coreference tools, or on knowledge about lexical items induced from large collections of unannotated text (Latent Semantic Analysis, verb specificity).

3.1 Lexical Chains and Related Approaches

Lexical chains [7, 64, 192] attempt to represent topics that are discussed throughout a text or text segment. They capture semantic similarity between noun phrases to determine the importance of sentences. The lexical chains approach exploits the intuition that topics are expressed using not a single word but instead different related words. For example, the occurrence of the words "car", "wheel", "seat", "passenger" indicates a clear topic, even if each of the words is not by itself very frequent. The approach heavily relies on WordNet [137], a manually

compiled thesaurus which lists the different sense of each word, as well as word relationships such as synonymy, antonymy, part-whole and general-specific. In addition, the lexical chains approach requires some degree of linguistic preprocessing, including part of speech tagging and division into topically related segments of the input to the summarizer.

Barzilay and Elhadad [7] present a summarizer that segments an input document, identifies lexical chains first within segments and then across segments, identifies and scores lexical chains, and finally selects one sentence for each of the most highly scored chains.

A large part of Barzilay and Elhadad's work is on new methods for constructing good lexical chains, with emphasis on word sense disambiguation of words with multiple meaning: for example the word "bank" can mean a financial institution or the land near a river or lake. They develop an algorithm that improves on previous work by waiting to disambiguate polysemous words until all possible chains for a text have been constructed; word senses are disambiguated by selecting the interpretations (i.e., chains) with the most connections in the text. Later research further improved both the run-time of the algorithms for building of lexical chains, and the accuracy of word sense disambiguation [64, 192].

Barzilay and Elhadad claim that the most prevalent discourse topic will play an important role in the summary and argue that lexical chains provide a better indication of discourse topic than does word frequency simply because different words may refer to the same topic. They define the strength of a lexical chain by its length, defined as the number of words found to be members of the same chain, and its homogeneity, where homogeneity captures the number of distinct lexical items in the chain divided by its length. They build the summary by extracting a sentence for each strong chain, choosing the first sentence in the document containing a representative word for the chain.

In later work, researchers chose to avoid the problem of word sense disambiguation altogether but still used WordNet to track the frequency of all members of a concept set. In the robust multi-document summarization system DEMS [186], *concepts* were derived using WordNet synonyms, hypernyms and hyponyms relations. Rather than attempting to disambiguate polysemous words and only then find semantically related words, as was done in the lexical chains approach,

in the DEMS system, words with more than five senses ("matter", "issue", etc.) are excluded from consideration. Given that many common words are polysemous, this policy of exclusion can be viewed as too restrictive. In order to compensate for the loss of information, highly polysemous words were replaced by other nouns that were strongly associated with the same verb. For example if the word "officer" is excluded from consideration because it has many senses, "policeman" would be added in, because both nouns are strongly associated with the verb "arrest".

After concepts are formed, frequency information can be collected much more accurately, counting the occurrence of a concept rather than a specific word. Sample concepts for one article consisted of $C_1 = \{$war, campaign, warfare, effort, cause, operation, conflict$\}$, $C_2 = \{$concern, carrier, worry, fear, scare$\}$, $C_3 = \{$home, base, source, support, backing$\}$. Each of the individual words in the concept could appear only once or twice in the input, but the concept itself appeared in the document frequently.

Shallow semantic interpretation on the level of concepts was also employed by Ye et al. [224]. They also used WordNet to derive the concepts, but to find semantically related words they employ a measure of the content overlap of the WordNet definitions, called glosses, of two words rather than the WordNet relations. The intuition is that the more content is shared in the definitions, the more related two words are. Example concepts derived using their approach are {British, Britain, UK}, {war, fought, conflict, military}, {area, zone}.

The heavy reliance on WordNet is clearly a bottleneck for the approaches above, because success is constrained by the coverage of WordNet and the sense granularity annotated there. Because of this, robust methods that do not use a specific static hand-crafted resource have much appeal, explaining the adoption of Latent Semantic Analysis as an approximation for semantic interpretation of the input.

3.2 Latent Semantic Analysis

Latent semantic analysis (LSA) [46] is a robust unsupervised technique for deriving an implicit representation of text semantics based on observed co-occurrence of words. Gong and Liu [69] proposed the use

of LSA for single and multi-document generic summarization of news, as a way of identifying important topics in documents without the use of lexical resources such as WordNet.

At the heart of the approach is the representation of the input documents as a word by sentence matrix A: each row corresponds to a word that appears in the input and each column corresponds to a sentence in the input. Each entry a_{ij} of the matrix corresponds to the weight of word i in sentence j. If the sentence does not contain the word, the weight is zero, otherwise the weight is equal to the TF*IDF weight of the word. Standard techniques for singular value decomposition (SVD) from linear algebra are applied to the matrix A, to represent it as the product of three matrices: $A = U\Sigma V^{\mathrm{T}}$. Gong and Liu suggested that the rows of V^{T} can be regarded as mutually independent topics discussed in the input, while each column represents a sentence from the document. In order to produce an extractive summary, they consecutively consider each row of V^{T}, and select the sentence with the highest value, until the desired summary length is reached. Steinberger et al. [195] later provided an analysis of several variations of Gong and Liu's method, improving over the original method. Neither method has been directly compared with any of the approaches that rely on WordNet for semantic analysis, or with TF*IDF or topic word summarizers.

An alternative way of using the singular value decomposition approach was put forward by Hachey et al. [73]. They followed more directly the original LSA approach, and build the initial matrix A based on the information of word occurrence in a large collection of documents instead of based on the input documents to be summarized. They compared the performance of their approach with and without SVD, and with a TF*IDF summarizer. SVD helped improve sentence selection results over a general co-occurrence method but did not significantly outperform the TF*IDF summarizer.

3.3 Coreference Information

Yet another way of tracking lexically different references to the same semantic entity is the use of coreference resolution. Coreference resolution is the process of finding all references to the same entity in a

document, regardless of the syntactic form of the reference: full noun phrase or pronoun.

Initial use of coreference information exclusively to determine sentence importance for summarization [4, 18] did not lead to substantial improvements in content selection compared to shallower methods. However, later work has demonstrated that coreference resolution can be incorporated in and substantially improve summarization systems that rely on word frequency features. A case in point is a study on generic single document summarization of news carried out by Steinberger et al. [195]. The output of an automatic system for anaphora resolution was used to augment an LSA-driven summarizer [69]. In one experiment, all references to the same entity, including those when pronouns were used, were replaced by the first mention of that entity and the resulting text was given as an input to the traditional LSA summarizer. In another experiment, the presence of an entity in a sentence was used as an *additional* feature to consider when determining the importance of the sentence, and the references themselves remained unchanged. The first approach led to a decrease in performance compared to the traditional LSA summarizer, while the second gave significant improvements. An oracle study with gold-standard manual coreference resolution showed that there is further potential for improvement as the performance of coreference systems gets better.

3.4 Rhetorical Structure Theory

Other research uses analysis of the discourse structure of the input document to produce single document summaries. Rhetorical Structure Theory (RST) [117], which requires the overall structure of a text to be represented by a *tree*, a special type of graph (see Section 2.1.3), is one such approach that has been applied to summarization. In RST, the smallest units of text analysis are *elementary discourse units* (EDUs), which are in most cases sub-sentential clauses. Adjacent EDUs are combined through rhetorical relations into larger spans. The larger units recursively participate in relations, yielding a hierarchical tree structure covering the entire text. The discourse units participating in a relation are assigned nucleus or satellite status; a nucleus is considered

to be more central in the text than a satellite. Relations characterized by the presence of a nucleus and a satellite are called mononuclear relations. Relations can also be multinuclear, when the information in both participating EDUs is considered equally important. Properties of the RST tree used in summarization include the nucleus–satellite distinction, notions of salience and the level of an EDU in the tree.

In early work, Ono et al. [159] suggested a penalty score for every EDU based on the nucleus–satellite structure of the RST tree. Satellite spans are considered less essential than spans containing the nucleus of a relation. With the Ono penalty, spans that appear with satellite status are assigned a lower score than spans which mostly take nucleus status. The penalty is defined as the number of satellite nodes found on the path from the root of the discourse tree to the EDU.

Marcu [120] proposes another method to utilize the nucleus–satellite distinction, rewarding nucleus status instead of penalizing satellite. He put forward the idea of a promotion set, consisting of salient/important units of a text span. The nucleus is considered as the more salient unit in the full span of a mononuclear relation. In a multinuclear relation, all the nuclei become salient units of the larger span. At the leaves, salient units are the EDUs themselves. Under this framework, a relation between two spans is defined to hold between the salient units in their respective promotion sets. Units in the promotion sets of nodes close to the root are hypothesized to be more important than those appearing at lower levels. The highest promotion of an EDU occurs at the node closest to the root which contains that EDU in its promotion set. The depth of the tree from this node gives the importance for that EDU. The closer to the root an EDU is promoted, the better its score.

A further modification of the idea of a promotion set [120] takes into account the length of the path from an EDU to the highest promotion set it appears in. An EDU promoted successively over multiple levels should be more important than one which is promoted fewer times. The depth score fails to make this distinction; all EDUs in the promotion sets of nodes at the same level receive the same scores. In order to overcome this, a promotion score was introduced which is a measure of the number of levels over which an EDU is promoted.

The RST approach for content selection has been shown to give good results for single document summarization of news and Scientific American articles [119, 120, 122].

3.5 Discourse-motivated Graph Representations of Text

In the RST based approaches, the importance of a discourse segment is calculated on the basis of the depth of the discourse tree and the position of the segment in it, relation importance, and nuclearity and satellite status. Marcu's work on using RST for single document summarization has been the most comprehensive study of tree-based text representations for summarization [119, 120, 122] but suggestions for using RST for summarization were proposed even earlier [159].

Graph-based summarization methods are very flexible and allow for the smooth incorporation of discourse and semantic information. For example, graph representations of a text that are more linguistically-informed than simply using sentence similarity can be created using information about the discourse relations that hold between sentences. Wolf and Gibson [219] have demonstrated that such discourse-driven graph representations are more powerful for summarization than word- or sentence level frequency for single document summarization. In their work, sentences again are represented by vertices in a graph, but the edges between vertices are defined by the presence of discourse coherence relation between the sentences. For example, there is a *cause–effect* relation between the sentences "There was bad weather at the airport. So our flight got delayed." Other discourse relations included *violated expectation*, *condition*, *similarity*, *elaboration*, *attribution* and *temporal sequence*. After the construction of the graph representation of text, the importance of each sentence is computed as the stationary distribution of the Markov chain, as in the sentence-similarity graph methods. Wolf and Gibson claim that their method outperformed summarization approaches using more restricted discourse representations such as RST.

Both the RST approach and the GraphBank work rely on *the structure* of the text, be it a tree or a general graph, to define importance of sentences. In recent work [110], the RST and GraphBank

methods were again compared directly with each other, as well as against discourse information that also included the *semantic type* of the discourse relation and non-discourse features including topic words, word probabilities and sentence position. The summarizers were tested on single-document summarization of news. Of the three classes of features — structural, semantic and non-discourse — the structural features proposed by Marcu lead to the best summaries in terms of content. The three classes of features are complimentary to each other, and their combination results in even better summaries. Such results indicate that the development of robust discourse parsers has the potential of contributing to more meaningful input interpretation, and overall better summarization performance.

3.6 Discussion

The discourse-informed summarization approaches described in this section are appealing because they offer perspectives for more semantically and linguistically rich treatment of text for summarization. At the same time, these methods require additional processing time for coreference resolution, lexical chain disambiguation or discourse structure. Because of this, they would often not be used for applications in which speed is of great importance. Furthermore, the above approaches have been tested only on single document summarization and have not been extended for multi-document summarization or for genre specific summarization. Discourse-informed summarization approaches will likely have a comeback as recent DUC/TAC evaluations have shown that systems have gotten impressively good at selecting important content but lack in linguistic quality and organization. Addressing these harder problems will require linguistic processing anyway and some of the discourse approaches could be used as the basis for techniques for improving the linguistic quality of the summaries rather than the content selection capabilities of summarizers (see Section 4).

In concluding this section, we would like to point out that the use of semantic interpretation in summarization seems intuitively necessary: the majority of summarization systems still process text at the word level, with minor pre-processing such as tokenization and stemming.

Current systems are complex, relying on multiple representations of the input and features and algorithms to compute importance. Virtually no recent work has attempted to analyze the direct benefits of using semantic representations, either based on WordNet or derived from large corpora. In this sense, the development and assessment of semantic modules for summarization remains much of an open problem. No clear answers can be given to questions such as how much run-time overhead is incurred by using such methods. This can be considerable, for example, if word sense disambiguation for full lexical chains is performed, or a coreference resolution system is run as a pre-processing step. Also unclear is by how much content selection is improved compared to simpler methods that do not use semantics at all? Future research is likely to address these questions, because the need for semantic interpretation will grow as summarization approaches move toward the goal of producing abstractive rather than extractive summaries, which will most likely require semantics.

4

Generation for Summarization

Determining which sentences in the input documents are important and summary-worthy can be done very successfully in the extractive summarization framework. Sentence extraction, however, falls short of producing optimal summaries both in terms of content and linguistic quality. In contrast to most systems, people tend to produce abstractive summaries, rewriting unclear phrases and paraphrasing to produce a concise version of the content found in the input. They do also re-use portions of the input document, but they often cut and paste pieces of input sentences instead of using the full sentence in the summary. While extensive research has been done on extractive summarization, very little work has been carried out in the abstractive framework. It is clearly time to make more headway on this alternative approach.

Content of an extractive summary may inadvertently include unnecessary detail along with salient information. Once an extractive approach determines that a sentence in an input document contains important information, all information in the sentence will be included regardless of its relevance. While methods do attempt to extract sentences that contain a large amount of salient material, this nonetheless remains an issue. In this section, we discuss two approaches that

address this problem: *compression* and *sentence fusion*. Compression is a method to remove unnecessary detail from selected sentences and has been used primarily in single document summarization. Sentence fusion is a method that cuts and pastes phrases from different sentences in the input documents, thereby removing unnecessary information and avoiding the possibility of short, choppy sentences in the summary. It has been used primarily for multi-document summarization. These approaches are discussed in Sections 4.1 and 4.2.

The linguistic quality of automatic summaries is also far from optimal, even when content is well chosen. In DUC 2005, for example, more than half of the summaries were perceived as not having good referential clarity, focus, structure and coherence. Addressing these problems requires the use of text generation techniques in order to improve information ordering and to help in making context dependent revisions to improve the flow of the summary. Approaches that use editing of extracted sentences to improve referring expressions are discussed in Section 4.3 and approaches that dynamically determine the order of sentences using constraints from discourse and from the input articles are discussed in Section 4.4.

4.1 Sentence Compression

In many cases, sentences in summaries contain unnecessary information as well as useful facts. This is likely to happen when summarizing documents that contain long sentences, particularly the very long sentences that occur in news articles. For example, in the sentence

(1) As U.S. Court of Appeals Judge Sonia Sotomayor made her Senate debut with a series of private meetings, Republicans said they would prefer holding hearings on her nomination in September, which could cloud the speedy summertime confirmation Obama wants.

two topics are mentioned: one about the series of meetings and the other about the timing of Senate hearings. Depending on the set of articles, one of these topics will not be relevant to the other sentences in the summary and should be dropped. Often the irrelevant information will be presented in a clause modifying the main sentence or in a modifier of one of the a clause argument (e.g., in this case, in the clause beginning

with "as"), but if the set of input documents focuses on the meetings that Sotomayor had, then it would be better to keep the first clause, dropping just the word "as" as well as the main clause of the sentence. A summarizer that can "compress" a sentence, removing unnecessary pieces, is able to produce a more concise and focused summary.

There is evidence that humans do exactly this when writing summaries. In a corpus analysis of human single document summaries, Jing and McKeown [86] observed that sentence reduction was often used by professional summarizers. For their analysis of human summarizing behavior, Jing and McKeown worked with the Ziff-Davis corpus,[1] which is a collection of newspaper articles about computer products along with a (human) summary of each article. They automatically aligned the sentences in the abstracts to sentences in the original article and were able to study the sentence-level transformations employed by the writers of the abstracts. They found that 78% of the summary sentences were written by editing the input document and of those, more than half of the edits were done using compression alone, removing information from a sentence extracted from the input document. The remaining edits used compression in addition to combining information from one or more other sentences. Thus, compression is an important component of summarization which, even now, is infrequently addressed.

In this section, we describe two general approaches to sentence compression, one using primarily linguistic techniques and the other, statistical techniques. These two approaches have been used both for single document summarization where useful corpora of paired human abstracts and documents exist, as well as for the task of headline generation, where the key points of a document must be compressed down to a few words. Additional approaches to compression have been developed for summarization of speech where it is important to remove disfluencies.

4.1.1 Rule-based Approaches to Compression

Rule-based approaches to sentence compression [87, 226] use both syntactic and discourse knowledge to determine how to compress a sentence. Syntactic knowledge includes the syntactic structure of an

[1] Available from the Language Data Consortium, catalog number LDC93T3A.

extracted sentence and knowledge about which constituents are less likely to be needed, while discourse knowledge includes information about how each constituent is connected with the rest of the summary.

Jing [87] developed a system for automatic sentence compression, or reduction, that uses multiple sources of knowledge to decide which phrases in an extracted sentence can be removed, including syntactic knowledge, contextual information, and statistics computed from a corpus of professionally written summaries. Compression can be applied at any granularity, including a word, a prepositional phrase, a gerund, a to-infinitive or a clause. In the first step of reduction, the candidate sentences are parsed and all nodes in the tree that are necessary in order to preserve the grammaticality of the sentence, such as the main verb or head of a noun phrase, are marked. Such nodes can not be removed by the reduction module. Obligatory arguments of verbs are also marked. Contextual information is used in order to decide which parts of a sentence are most closely linked to the overall topic of the article and these parts will not be deleted even if it is syntactically possible to do so. The context weight for each word is computed as the number of links to the rest of the article in terms of repetitions, occurrence of morphological variants or of semantically related words identified using the WordNet database. Finally, the likelihood of a human deleting a particular type of constituent is looked up in a table of precomputed corpus statistics. These three factors are weighted and a constituent below a certain weight will be removed from the sentence. This approach seems effective as evaluation demonstrated that 81% of the sentence reductions proposed by the system agreed with the reduction decisions made by a professional human abstractor.

Zajic et al. [226] also developed a set of rules to compress sentences. There are two key differences between their work and that of compressions for each document sentence *before* doing sentence extraction and they base the removal of constituents Intuitively, this should improve selecting information can operate on system they originally developed for headline generation, called *HedgeTrimmer*, which was intended to compress a sentence down to 10 words. Compression is done iteratively, removing one constituent at a time. The resulting compression at each round is provided as input to the sentence extraction module, which

uses six features to score sentences for extraction. Most of these features have been used in earlier sentence extraction approaches: sentence position, relevance to query for query-focused summarization or to document centroid for generic summarization, as well as a feature unique to their approach, the number of trims applied to a sentence. Sentence compression uses linguistic heuristics, described in Section 4.1.3, which remove specific syntactic constituents. *HedgeTrimmer* does not use information about discourse connections.

Zajic et al. entered their system in DUC06 and noted that they did not score well on fluency measures such as grammaticality, concluding that this indicates that a good number of trimmed sentences were included as these typically would not be as grammatical as the original human written sentences.

These results highlight why sentence extraction is by far the more commonly used method for summarization, even though it is also less innovative. Novel approaches to summarization, such as compression, are clearly needed as demonstrated by the amount of extraneous information that is otherwise included. Nonetheless, it often happens that systems will perform worse on some standard evaluation scores when difficult tasks such as compression are attempted. It is important to recognize that research progress can only be made if such decreases in scores are tolerated and even welcomed, as research continues to find the best way to produce compressions that are fluent. This line of research will also become more attractive when metrics that reward brevity and the inclusion of higher number of important facts are developed and adopted.

Others have also used a combination of syntactic and lexical heuristics to compress sentences. CLASSY [37] used shallow parsing combined with lexical triggers to remove unnecessary constituents. Siddharthan et al. [190] found that sentence simplification improves the quality of sentence clusters in multi-document summarization and hence, significantly improves the resulting summary, as measured by ROUGE scores.

4.1.2 Statistical Approaches to Compression

Statistical approaches to sentence compression have also been explored. In this paradigm, rules about which syntactic constituents can be

deleted are learned by the program. Thus, no linguistic rules need to be provided as input. Knight and Marcu [94] experiment with two different statistical approaches. Like Jing, they also use the Ziff-Davis corpus for their work. They develop and evaluate an approach based on the noisy channel model, as well as one using decision trees. The idea behind the noisy channel model is to assume that the input to the model was a short, compressed sentence, that somehow got distorted with additional words yielding a long sentence. Given the long sentence, the task is to infer the "original" compressed version. They adapt the noisy channel model so that it operates on trees and thus, they work with probabilities over trees. Given a long sentence, their program first produces a (large) parse tree and then hypothesizes different ways it can be pruned to produce smaller ones. They measure the goodness of the smaller ones by computing the goodness of the tree using standard probabilistic context free grammar (PCFG) scores and the probability of the sentence according to a bigram language model. They measure the likelihood of the transformation from the large parse tree to the small one using statistics collected over 1037 summary/document sentence pairs from the Ziff-Davis corpus. A feature of this approach is that it can produce many possible compressions, ranked by their score, and depending on the required summary length, a compressed version that fits can be chosen. It does, however, have drawbacks as well. The word order in the sentence cannot be changed and it does not allow reorganization of the existing tree structure. For example the tree cannot be revised by turning a prepositional phrase into an adjective.

Knight and Marcu compare this approach with a decision-based model for compression. They use a shift–reduce–drop parsing paradigm, where any of the rules in the grammar are applied to the long string, attempting to build the tree of the short string. At any point, the rules can *shift* a word from the input to the output string; *reduce* operations manipulate the existing tree by combining pieces of trees produced by the algorithm so far to make a new tree; *drop* is used to delete a portion of the long input corresponding to a constituent, and a fourth operation *assignType* allows the processor to re-assign part of speech tags to words. A decision tree learner is applied to learn the kinds of rules applicable in different contexts. This type of learning has more

flexibility than the noisy channel model as it does allow for changing words, part of speech and for reorganization of the tree.

To evaluate the two models, they had assessors separately score each compression for grammaticality and importance of retained information. Their evaluation shows that the decision tree approach generally produces more highly compressed sentences and that both approaches significantly outperform the baseline. However, the performances of the decision tree and the noisy channel model are not significantly different from each other. In determining which approach to use, the need to adjust the level of compression is probably most differentiating.

Later work [65, 206] addressed some of the shortcomings of the Knight and Marcu approach. Galley and McKeown [65] make two significant improvements. They move to a lexicalized head-driven approach that allows them to take the lexical heads of constituents into account before making deletions. This helps them to avoid deleting negation words such as "no" and "not", which would completely change the meaning of the sentence, a problem which Knight and Marcu's approach cannot handle. They add markovization, a standard technique applied in statistical parsing which deals with data sparsity issues by conditioning the probability of a node only on a small number of other nodes rather than on the full context. Finally, they develop techniques for tree alignment which allow for insertions and substitutions, not only deletion. By doing this, increases the amount of training data they can use from the Ziff-Davis corpus by seven-fold over Knight and Marcu's approach.

Turner and Charaniak [206] improve on Knight and Marcu's approach by developing a method for acquiring training data in an unsupervised fashion. They demonstrate how to compute probabilities of possible deletions by counting instances in the Penn TreeBank which match pairs of rules where the right-hand side of one CFG rule is a subset of the right-hand side of another rule. For example, one rule might specify NP $\rightarrow$ DT JJ NN while another may specify NP $\rightarrow$ DT NN. This method is applied to generate additional training data for the original noisy channel model algorithm. Turner and Charniak also propose an improvement on the model for measuring the goodness of the

resulting compression, replacing the bigram language model with the syntax-based language model developed for the Charniak parser.

Since these seminal approaches, the field of sentence compression has taken off on its own. There has been considerable research on the problem in isolation, considering it as a general purpose tool that could be used in a variety of applications in addition to summarization, such as generation of subtitles for television programs, generation for small mobile devices, and generation for devices that allow the blind to browse and search collections using audio [33, 36, 126, 157, 180, 206].

We focus here on just one of these approaches to represent this burgeoning new area. While primarily statistical in nature, it does draw on additional linguistic information encoded as hard constraints. Clarke and Lapate [33] present an approach that uses discourse information as well as a statistical approach to compression using Integer Linear Programming (ILP). Note that ILP is also explored by others [36, 126]. ILP is an optimization approach which finds the output that maximizes an objective function subject to certain constraints. The constraints and the objective can be formulated in a way that does not require labeled data, which is in fact what Clarke and Lapata do. So their approach has an advantage over noisy channel model approaches [94, 65] in that it is unsupervised and thus the problem of obtaining adequate training data is not an issue.

On the other hand, their approach operates on the level of words, deciding which words should be kept in the compression and which should be deleted, and does not manipulate syntactic trees as done in the noisy channel models. Thus, they are likely to have less control over the grammaticality of their output. The ILP approach used by Clarke and Lapata uses optimization to determine which words can be removed from a long sentence given as input to produce a short sentence. A scoring function is computed to determine the best word to remove. In earlier work [32], they used a language model to score the sentence resulting from each word deletion, almost the same as the model Hori and Furui [78] use for speech summarization. In the new research, they augment the constraints using information from discourse. They include constraints from centering theory, weighting named entities and noun phrases higher when they serve as backward-looking centers

and weighting forward-looking centers, next highest, essentially weighting more heavily the entities mentioned in adjacent sentences and entities in syntactically prominent positions. They enforce hard constraints that these higher weighted entities should never be deleted. They also incorporate lexical chains, weighting entities higher when they occur in longer lexical chains across the documents. Similarly to Jing's rule-based approach, they enforce hard constraints that such entities should not be deleted. They find it relatively easy to incorporate these discourse constraints into the ILP framework for scoring a deletion.

4.1.3 Headline Generation

A key difference between the document summarization problem and headline generation is in the length of the summary. Headlines are typically even shorter than a sentence, which makes the usual, generic summarization approach of sentence extraction inappropriate for the task. Another critical difference is in the amount of data available for training. Since a headline typically appears with every news article published, there are large amounts of training data that do not require any special modification. The two primary approaches to compression, one based on linguistic rules and one statistical, have also been applied to headline generation, and are surveyed here.

The main linguistic rule-based approach to headline generation [49] was developed prior to the rule-based approach to compression in multi-document summarization that we overviewed in the last section. It was used to generate multiple possible compressed sentences that the multi-document summmarizer could then opt to select as the headline for the summary using a set of features. In their work on headline generation, Dorr and her colleagues developed a system called *HedgeTrimmer* which uses linguistic heuristics to determine which syntactic constituents can be dropped. The development of heuristics was done through manual analysis of three different sets of headline–article pairs containing a total of almost 400 such pairs. Through this analysis, they identified the types of constituents that are typically dropped. They postulate first a set of heuristics that are always applied. For these, they identify the

lowest S node in the constituent parse of the sentence and its immediate arguments and drop the remaining constituents. They remove low-content words such as determiners and function words and they also delete time expressions. Following the application of these heuristics, they iteratively apply shortening rules until the sentence has reached the desired length, which they set at 10 words based on observation, though this number is parameterized and could easily be changed.

Their iterative shortening rules remove prepositional phrases and preposed adjuncts. They also have a general rule which they call XP-over-XP which basically looks for constituents such as NP or VP which contain an embedded NP or VP and remove all constituents except for the embedded XP. For example, given a sentence such as "A fire killed a firefighter who was fatally injured as he searched the house.", *HedgeTrimmer*'s parser produces an analysis of the sentence as containing an NP "a firefighter who was fatally injured as he searched the house" which contains an embedded NP ("a firefighter"). Using the XP-over-XP rule the relative clause is deleted. Using the determiner rule, "a" is deleted and the headline "Fire killed firefighter" is produced.

Statistical approaches to headline generation exploit the large amounts of training data. Witbrock and Mittal [218] and Banko et al. [5] use 8000 article–headline pairs, with 44,000 unique tokens in the articles and 15,000 in the headlines. They use a much simpler approach to compression than do Knight and Marcu. Their approach uses two models: one for content selection and the other for realization. For content selection, they compute the conditional probability of a word appearing in a headline, given its appearance in the article. For realization, they use a bigram model, trained on the headlines. They evaluated the system using unseen article–summary pairs, comparing it against both the original headlines and the top ranked summary sentence for the article. They generated headlines of multiple lengths since their system was not able to predict the ideal length. Overlap between the system headline and the original article headline ranged from 0.87 to 0.90 and overlap with the top summary sentence ranged from 0.85 to 0.90 across the different system headline lengths. They did not measure the quality of the headline in terms of word order or structure although this is clearly another factor that should influence evaluation.

4.2 Information Fusion

While research on compression does result in summary sentences that are not identical to sentences in the input document, those differences are limited; each summary sentence can differ at most by being a subset of a sentence in the original document. In addition to removing phrases from document sentences, human abstractors sometimes substitute one word for another and often combine information from two sentences to create a novel sentence. This was noticed early on by Jing who coined the term "cut and paste" to refer to the kind of operations performed [90]. She did extensive analysis of the types of edits that are typically carried out by human abstractors, but implemented just one operator other than compression. This operator created a new sentence using conjunction between two reduced sentences from the input document [89] under certain constraints such as having the same subject.

A more general approach to fusion of information from several sentences was introduced in Barzilay and McKeown [11] for the multi-document summarizer MultiGen. A common approach to multi-document summarization is to find similarities across the input documents and extract the similarities to form the summary. Often similarities are identified using sentence clustering; each cluster is considered to represent a theme in the input. While many systems simply extract a sentence from each cluster, Barzilay and McKeown introduce a text-to-text generation technique, which takes as input a set of similar sentences and produces a new sentence containing information common to most of them. Their approach addresses two challenges: the identification of phrases across sentences that convey common information and the combination of those phrases into a grammatical sentence.

The identification of common information is done using pairwise alignment of all sentences in a cluster of similar sentences that the algorithm receives as input. Pairwise alignment for multi-document summarization is quite different from the alignment that is commonly used for machine translation. First, only a subset of the subtrees in one sentence will align with a subset of the subtrees in the other sentences. Second, the order of the matching trees will be different in the different sentences and thus, there is no natural starting point for the

alignment. The authors develop a bottom-up local multisequence alignment algorithm for this problem. It uses words and phrases as anchors and operates on the dependency trees of the cluster sentences, considering a pair of sentences at a time. Alignment is driven by both similarity between the structure of the dependency trees and similarity between lexical items, considering paraphrases as well as identical lexemes. Once identified, the common phrases are combined by first building a fusion lattice which represents the alignments and then linearizing the lattice as a sentence using a language model.

MultiGen first selects a sentence which is most similar to other sentences to serve as a basis sentence and then transforms it into a fusion lattice by removing information that does not appear in the majority of other sentences and by adding in phrases that do occur in the majority of other sentences, but are missing from this one. Finally, the fusion lattice is linearized using a tree traversal performing lexical choice among alternative verbalizations available at each node and placing function words. The linearization component outputs all possible sentences that can be derived from the lattice and MultiGen uses a language model to choose the best one among these different possible fusion sentences.

Recent work has extended the sentence fusion algorithm to question-answering. In this context, it makes sense to combine all information that is relevant, regardless of whether it is repeated across sentences in a cluster. Thus, Marsi and Krahmer [123] introduce the notion of *union* of two input sentences, producing an output sentence that conveys all information expressed in the two input sentences without repetition. They note that this can enable generation of a sentence that would provide complete information in response to a query. They first explore different possibilities for alignment using a manual annotation of different translations of *The Little Prince*. They develop an algorithm for dependency tree alignment that is very similar to Barzilay and McKeown [11], but they experiment with merging and generation on the annotations thus allowing generation from aligned nodes where one is either identical, a paraphrase, a generalization, or more specific than the other. As in [11], a language model is used to select the best sentence. The resulting generated sentences can be either restatements of the input, generalizations or specifications. Their evaluation reveals

that roughly 50% of the generated sentences are judged as "perfect" (evaluation done by the authors), while the remainder are divided into "acceptable" and "nonsense." While not yet viable, this is an interesting approach that clearly identifies new avenues for further research.

Filippova and Strube [59] also explore a form of union fusion, accepting a cluster of similar sentences as input and using dependency tree alignment between pairs of sentences in the cluster. They use a much simpler form of alignment than earlier work, however, combining any nodes which share the same words. They then use integer linear programming (ILP) to transform the resulting graph into a tree and to remove unimportant nodes, using a sophisticated set of structural, syntactic and semantic constraints. For example, since their focus is on union fusion, they use semantic constraints to determine that a conjunction can be used to combine two phrases when they are semantically similar and fill the same roles of similar sentences (e.g., both are objects of identical verbs in the two different input sentences). Rather than overgenerating and using a language model to select from the resulting sentences, they use a syntactically constrained linearization approach. Their evaluation shows that their approach produces significantly more readable sentences than Barzilay and McKeown's, but that there are no significant differences in informativeness.

Despite the interesting research problems, few researchers in summarization today attempt this line of work and most persist with the safe, yet uninspired, extraction-based approach. The tasks selected for large-scale, shared evaluation and the metrics used tend to encourage extraction. Daume III and Marcu [43] present an empirical study whose results also discourage research on sentence fusion, claiming that it is an ill-defined task. They use the Ziff-Davis corpus as data, extracting summary sentences that are aligned with two different document sentences. The summary sentences are used as the references. To obtain examples of sentence fusion, annotators were presented with the two document sentences both of which aligned to the same summary sentence, and were asked to create a single sentence from the two that preserves important information and is 50% shorter than the two sentences combined. The human subjects were to do this without consulting the documents and thus, had no context in which the

sentences appeared. They measure agreement between humans by manually identifying the factoids in each summary sentence elicited in the experiment and comparing it with the corresponding reference. They also compare the factoids from each elicited summary sentence against every other elicited summary sentence from the same document sentences. The Kappa when comparing against the reference is low — less than 0.251 — and the highest Kappa between two elicited summary sentences is 0.47, still relatively low. They also perform two other evaluations, which they claim support this view.

It should be noted their definition of "sentence fusion" is more in line with "union" as defined by Marsi and Krahmer, since it requires extracting two semantically different phrases from the input sentences and combining them in one sentence. "Intersection" requires identifying semantically equivalent phrases from the different sentences in a cluster of similar sentences and expressing only such shared information. Thus, for intersection, the criteria for selecting phrases to combine is semantic equivalence, while in the Daume and Marcu study, the criteria is "importance" which is a much more vague and less restrictive criteria. It is not surprising that "importance", particularly with no context, does not yield agreement. Other work that deals with union uses more concrete criteria for combination (e.g., connection to previous discourse [87] or relevance to query [123]). In fact, recent work [96] shows that sentence fusion is better defined when performed in the context of a task like question answering.

4.3 Context Dependent Revisions

Revision is another device for producing non-extractive summaries. Different revision algorithms have been developed to merge phrases from different articles to revise references to people and to correct errors arising from machine translation in multilingual summarization.

In a preliminary study of how summary revisions could be used to improve cohesion in multi-document summarization [162], automatic summaries were manually revised and the revisions classified as pertaining to discourse (34% of all revisions), entities (26%), temporal ordering (22%) and grammar (12%). This study further supports findings from

early research that shows that unclear references in summaries pose serious problems for users [165].

Early work investigated generation of references. In research on multi-document summarization, Radev and McKeown [175] built a prototype system called PROFILE which extracts references to people from news, merging and recording information about people mentioned in various news articles. The idea behind the system was that the rich profiles collected for people could be used in summaries of later news in order to generate informative descriptions. However, the collection of information about entities from different contexts and different points in time leads to complications in description generation that are not faced in a pure multi-document environment — for example, past news can refer to Bill Clinton as "Clinton, an Arkansas native", "the democratic presidential candidate Bill Clinton", "U.S. President Clinton", or "former president Clinton". Which of these descriptions would be appropriate to use in a summary of a novel news item? In later work, Radev developed an approach to learn correlations between linguistic indicators and semantic constraints which could eventually be used to address such problems [170].

Single document summarization can also benefit from techniques for improving summaries through revision [116]. Mani et al. define three types of revision rules — *elimination* (removing parentheticals, sentence initial prepositional phrases and adverbial phrases), *aggregation* (combining constituents from two sentences) and *smoothing*. The smoothing operators covers some reference editing operations. They include substitution of a proper name with a name alias if the name is mentioned earlier, expansion of a pronoun with coreferential proper name in a parenthetical and replacement of a definite NP with a coreferential indefinite if the definite occurs without a prior indefinite.

While the rules and the overall approach are based on reasonable intuitions, in practice, entity rewrite for summarization does introduce errors, some due to the rewrite rules themselves, others due to problems with coreference resolution and parsing. Readers are very sensitive to such errors and notice them easily [149]. In the entity-driven rewrite for multi-document summarization of news, Nenkova [149] proposed incorporating noun phrase editing in a greedy sentence selection method for

summarization. References to the same entity from all input documents were collected automatically using the heuristic that noun phrases with the same head refer to the same entity (for example "the angry dog" and "the black dog" would be considered co-referential). After a sentence is selected for inclusion in the summary, all noun phrases co-referential with noun phrases contained in the sentence are evaluated for their informativeness, dependent on the information already included in the summary, using word probability. The most informative noun phrase is chosen and incorporated in the sentence. This method of noun phrase rewrite leads to a natural selection of more informative and descriptive references for the first mention of an entity, and short references for subsequent mentions. The rewrite method produces summaries that differ in content from the summaries produced by the purely extractive version of the same algorithm by 20–50% as measured by lexical overlap. In manual evaluation of linguistic quality, however, readers preferred the extractive summaries which consisted of sentences written by journalists.

In contrast, when editing is restricted to references to people alone, there are fewer edits per summary but the overall result is perceived by readers as better than the original [147]. Nenkova's work on reference to people distinguished two tasks: should the person be mentioned at all and if so, what is the impact on first mention? Nenkova et al. [154] showed that it is possible to automatically distinguish two characteristics of an entity: whether the entity is known to a general reader and whether the entity is a major participant in the event discussed in the input. If the entity is familiar to a general reader and is a major participant, a short reference by last name only is appropriate and is often used in human-written summaries. A rule-based system for deciding whether to include attributes such as title or role words (e.g., *Prime Minister, Physicist*), temporal role modifying adjectives (e.g., *former*), and affiliations (e.g., country or organization name) depending on the expected familiarity of the reader with the entity was able to reproduce references used in human summaries with 80% accuracy. This research complements earlier research [147] where a rewrite system for references to people in multi-document summaries of news was developed based on a corpus analysis and human preference studies. For the

first mention of the person in the summary, a reference containing the full name of the person and the longest, in number of words, premodifier of the name (e.g., "Swiss tennis star Roger Federer") is chosen among the references to that person in the multiple input documents. If no reference with premodification is found in the input, a reference with an appositive is selected ("Roger Federer, winner of five US Open titles"). Automatic summaries rewritten using these rules were almost always prefered to the original extracts by human assessors.

Rewriting references to people can also improve multi-lingual summarization where the input for summarization is in a language different from the language in which the summary is produced. Siddharthan and McKeown [189] report their results on the task of summarizing in English, news articles written in Arabic. They use several automatic machine translation systems to translate the input. As a result, there are many errors in grammar and lexical choice in the input for summarization. From the multiple translations of the input articles, Siddharthan and McKeown [189] automatically extract the name, role and affiliation for each person mentioned in the summary. Frequency of occurrence in the different translations is used to choose the best value for each of these attributes of the reference and these are used in the final summary, correcting the errors in machine translations. Several automatic evaluation measures show significant improvement in the output after error correction through rewrite of references to people.

4.4 Information Ordering

The order in which information is presented also critically influences the quality of a text. Reading and understanding a text in which sentences are randomly permuted is difficult. In a single document, summary information can be presented by preserving the order in the original document [104, 133, 174]. However, even in single document summaries written by professional summarizers, extracted sentences do not always retain their precedence orders in the summary. In multi-document summarization, the problem is even harder because the input consists of several documents and not all topics appear in all documents. Sentences with different wording and varying content can express the topic in the

different articles, and topics are not discussed in the same order in all documents. For this setting, a generalization of the original text ordering for single document summarization works quite well. Similar sentences from the different documents, discussing the same themes or topics, can be clustered [76]. A graph representing local ordering patterns in the input can be constructed, with each vertex representing a topic cluster. Two vertices are connected by an edge if in some document a sentence from one topic immediately preceeded a sentence from the other topic. The weight of the edge is the number of articles in which that particular ordering pattern was observed. Inferring a global ordering from the graph is an NP-complete problem [35], but an approximation algorithm can be used to solve the problem with remarkably good results. This approach to ordering has been called Majority Ordering [9]. Variations of the Majority Ordering approach were later studied by several groups [19, 20, 84].

An alternative ordering approach studied in Barzilay et al. [9] is Chronological Ordering. Using chronological order in the summary to describe the main events helps the user understand what has happened. For this approach, it is necessary to associate a time stamp with each sentence or topic. The time when a topic occurred was approximated in Barzilay et al.'s work by its first publication time; that is, the publication date of the earliest article that contributed a sentence to the topic cluster. When two topics have the same publication time, it means that they both are reported for the first time in the same article and the order in the article can be used to decide presentation in a summary. The initial results from applying Chronological Ordering to automatic summaries were discouraging and human assessments of the resulting summaries were low. An error analysis revealed that problematic orderings were produced when the sentences that needed to be ordered conveyed background information, or states rather than events. Another problem in summaries rated as poor by human assessors was that they contained abrupt switches of topic.

To address this issue, Barzilay et al. [9] used a topic segmentation tool to identify blocks of text in each article from the input that were tightly related to each other as indicated by word distribution and coreference cues. Topics that appeared in the same segment in more

than half of the input articles for multi-document summarization were considered to be strongly related. Finding the transitive closure of this relation builds groups of related themes and, as a result, ensures that themes that do not appear together in any article but which are both related to a third theme will still be linked. The chronological ordering algorithm is used to order *blocks of topics* rather than directly the topics themselves as in the original algorithm. They show that this *Augmented Algorithm* produced the best results compared to the Majority and Chronological order approaches.

A more general approach to information ordering, applicable to any text rather than specifically for summarization, was proposed by Barzilay and Lapata [6]. The main insight in this approach is that continuity between adjacent sentences will determine the overall coherence of the text. Continuity is defined as the occurrence of the same entity in two adjacent sentences. To capture patterns of continuity, Barzilay and Lapata [6] define a matrix M, called an *entity grid* in which rows correspond to sentences in the text and the columns correspond to entities that appear in the text. An element m_{ij} of the matrix corresponding to the ith sentence and jth entity can take one of four values: **S**ubject, **O**bject or **Ot**her (depending on the grammatical function of entity j in sentence i), or **N**one if entity j is not mentioned at all in sentence i. The entire text is characterized by 16 features, equal to the percentage of each type of transition between adjacent sentences observed in the text (for example **SS**, **SO**, **ON**, etc.). An SVM ranking model incorporating these features achieved 84% accuracy in predicting which of two summaries is more coherent. The data for training and testing the classifier contained a mix of human and machine produced summaries.

The two models described in this section relate only to information ordering, not content selection. For domain specific summarization, Hidden Markov Models (HMM) combining selection and ordering have been proposed [10, 61]. These models capitalize on the fact that within a specific domain, information in different texts is presented following a common presentation flow. For example, news articles about earthquakes often first talk about where the earthquake happened, what its magnitude was, then mention human casualties or damage, and finally discuss rescue efforts. Such "story flow" can be learned from multiple

articles from the same domain. An HMM model is very appropriate for the task. States correspond to topics in the domain, which discovered via iterative clustering of similar sentences from many articles from the domain of interest. Transitions between states in the model correspond to topic transitions in typical texts. These HMM models do not require any labelled data for training and allow for both content selection and ordering in summarization.

4.5 Discussion

If we want to develop automatic systems that emulate human summarization, then it is clear from all collected corpora of *in vivo* human summarization that generation of abstractive summaries is necessary. The research presented in this section takes a step toward that goal. Of the different approaches to abstraction presented here, compression is probably the most mature. Since the early work introducing this topic in the context of summarization and drawing on the Ziff-Davis corpus, a sub-field on just this topic has grown. Part of the interest in this sub-field may come in part from its similarities to machine translation since alignment is required and in part from its reliance on parsing, thus drawing researchers outside of the summarization community. Indeed, recent work on this topic explores new methods for learning over parse trees that enlarges the range of alignments allowed [222]. Looking ahead in this monograph to Section 5.5, compression has also been heavily used in the area of speech summarization, in part to remove disfluencies. It is a concretely circumscribed research problem with application to other problems in addition to summarization. Continuing research on compression must still deal with issues such as scarcity of training data, appropriate integration of syntax even when the input data comes from a noisy genre, and compressions involving lexical substitution and paraphrase.

Sentence fusion and revision have also been growing as summarization sub-fields. By combining phrases from different sentences, fusion approaches push the envelope of multi-document summarization in a way that few other systems do. Revision provides a more general approach than compression to producing abstractive summaries. Like

many compression approaches, fusion relies on parsing and alignment of parse trees, and in this respect, its robustness is an issue. For example, in experiments at Columbia using MultiGen in NewsBlaster, MultiGen produced less than 10% of the summaries in part because of its need to parse all input sentences, thus slowing down the process; NewsBlaster runs in a limited period of time and will choose summaries from the summarizers that finish first. Future research on fusion and revision will need to further address grammaticality of summary sentences, something that is not an issue with extractive approaches. Given that many of the evaluation set-ups favor extraction, another area of research is investigation into evaluations that jointly measure quality and content.

Finally, if summaries are actually to be read by people, as opposed to evaluated by comparison in content to human summaries, then sentence ordering remains an important research issue.

5

Genre and Domain Specific Approaches

In the previous sections, we presented a variety of approaches for generic summarization, applicable when we know little about the audience or user need. We also described how the generic summarization approach can be adapted when summarization takes the user need, specified as a query, into account. Similarly, generic approaches may not be appropriate when the input for summarization comes from a specific domain or type of genre. When the input documents have a known, specified structure or other unique characteristics, summarization algorithms can take advantage of these characteristics to more accurately identify important information. Journal articles, for example, very often have a results or conclusion section that identifies the key contributions of the article and these sections will be the place where important information for the summary will be located. Specific domains, such as medicine or law, may have specific requirements on the kind of information that is needed in a summary. Such domains may also have resources that can help the summarization process.

In this section, we overview research on summarization that has been carried out for genres and domains where the structure of the underlying documents are markedly different from news documents.

We look at genres that have much more structure than news, in particular, journal articles. We also examine genres which are much more informal than news, such as email, blogs, and speech. These genres are based on exchanges and the dialog characteristics must be taken into account. We also describe summarization in the medical domain, for both lay users and health professionals as intended users. This is an interesting domain not only because of the structure found in the documents, but also because it has a variety of resources that can aid the summarization process. Semantic resources are available and thus, summarization techniques that exploit these resources operate at a deeper semantic level than do techniques for generic summarization.

We first present summarization approaches within the medical domain and then present genre-specific summarization approaches for journal articles, email, blogs and speech.

5.1 Medical Summarization

Summarization in the medical domain is a remarkable example of an application where generic summarization approaches are simply not applicable. In this domain, summarization algorithms are developed with precisely defined intended uses such as to help doctors decide on a course of treatment, to review the latest research relevant to a particular patient or to help patients and their relatives access information pertinent to a condition or disease. Medical articles have predictable structure that algorithms exploit, and furthermore, in the medical domain, there are large-scale knowledge resources available, providing semantic information for millions of concepts and concept names.

End users of summarization systems in the medical domain include healthcare providers and consumers, both of whom turn online to find information of interest. In the medical community, the number of journals relevant to even a single specialty is unmanageably large, making it difficult for physicians to keep abreast of all new results reported in their fields.[1] Similarly, patients and family members who need information about their specific illness can also be overwhelmed with the choice of

[1] There are five journals in the narrow specialty of cardiac anesthesiology but 35 different anesthesia journals; 100 journals in the closely related fields of cardiology (60) and cardiothoracic surgery (40); over 1000 journals in the more general field of internal medicine.

online information related to their interest. Summarization can help patients determine which articles in search results are relevant and can also help them to browse more efficiently.

A summary can be tailored to the type of user, whether a healthcare provider or a patient. The content of the summary can be further tailored using information from the patient record, a unique form of user model available in the medical domain. Depending on the patient's diagnosis and various test results, different information may be more or less relevant for the summary; these can influence what is relevant both for the patient's physician as well as the patient and family. Finally, end users may either be looking for information on a specific problem (essentially searching) or they may be browsing for information that may be relevant to a patient problem. These two different tasks may also influence the kind of summary that is provided.

Substantial resources are also available in the healthcare domain. An ontology of medical concepts, the Unified Medical Language System (UMLS) [210] is available and can be automatically linked to the terms in the input articles. The entry for each concept includes multiple alternative ways that can be used to refer to the concept. For example, the entry for Hodgkin Disease also lists Hodgkin's sarcoma, Hodgkin lymphoma and Lymphogranulomatosis, which are all terms that can be used to refer to the disease. Concepts are assigned semantic types such as ORGANISMS, CHEMICALS, EVENTS, and physical, spatial, temporal, functional and conceptual relationships are explicitly encoded. This enables more semantic processing than is currently possible in other domains and as a result, summarizers developed in the medical domain tend to do deeper analysis than the generic and query-focused summarizers described so far. They also tend to use more generation from semantic representations than pure sentence extraction.

In this section, we overview systems for summarization of journal articles, intended for physicians, and summarization of consumer health information.

5.1.1 Summarization of Healthcare Documents for Patients

Patients and family members search for information in ways rather different from that of caregivers. Often their initial search query is

not precise, although they may desire a specific type of information. Although this problem is not limited to seekers of medical knowledge, researchers [12] have noted that the problem of underspecification is particularly acute in medicine. In addition, Berland et al. [14], in a study of internet-accessible consumer healthcare information, concludes that identification of conflicting viewpoints needs to be clearly shown.

Centrifuser [53, 92] is a summarizer that aims to help consumers in their search for information. It generates three distinct summary components that form an overview plus details embedded within a textual user interface that specifically targets the cognitive needs of consumers. The three components consist of: (1) hyperlinks to topics related to the query to facilitate navigation and query reformulation, (2) a high-level overview of the commonalities in the documents, and (3) a description of differences between the retrieved documents to help searchers select relevant items. An example of a summary generated by Centrifuser, taken from Elhadad et al. [53], is shown in Figure 5.1.

Centrifuser's summaries can be classified as multi-document and query-focused. This summarizer selects relevant information from multiple documents using the query to access topic segments in the document that are most likely to contain pertinent information. These segments provide navigation links in the first part of the tripartite summary and thus we see that Centrifuser also differs from the majority of summarizers described in this overview in that it provides an *indicative* summary that allows users to browse relevant documents. Centrifuser also generates similarities and differences between the documents it is summarizing. It uses sentence extraction to provide a synopsis of commonalities across input documents. It computes quantitative metrics for the typicality or rare-ness of the different topics discussed in input documents by mapping these to a topic tree for the particular domain. The topic tree itself is constructed off-line in an unsupervised manner by clustering segments from a large number of documents from the domain. A document topic tree represents the document as a hierarchy of topics (e.g., treatment topic consisting of dieting, surgery and medication subtopics) and enables more precise retrieval of relevant sentences by localizing query relevance to specific document sections. Centrifuser uses typicality scores in the tree in order

Overview summary of **Atrial Fibrillation**

We found **9** documents about atrial fibrillation.

(1) **Here are some links to more specific subtopics of atrial fibrillation**: definition — symptoms — causes — diagnosis — treatment — complications — prevention — prognosis

(2) **Here's some general information on atrial fibrillation put together from the 9 documents**: Atrial fibrillation is a common type of palpitation, where you experience an irregular and often rapid beating of the upper chambers of the heart, known as the atria. There are a number of treatment options for AF. The first line of treatment usually involves medications, but there are other treatments which may be appropriate. Symptoms of AF may include one or more of the following: heart palpitations, lack of energy or feeling over-tired, etc. The causes of atrial fibrillation include: rheumatic heart disease, ischaemic heart disease, etc. Treatment varies from case to case and your general outlook will depend on the severity of your underlying heart condition. Medications are prescribed in the management of atrial fibrillation depending on the overall treatment goal: Heart surgery Patients at "low risk" may be given aspirin 325 mg/d to prevent stroke. See also: Anticoagulants, Arrhythmias, etc.

(3) **We've grouped the 9 documents by their topics to guide you in your search for information on atrial fibrillation. Click on a document's title to see it in another window**:

- Atrial Fibrillation Study, Atrial Fibrillation Foundation and The Atrial Fibrillation Page examine key topics. Some of their topics are "controlling heart s irregular rhythm provides no benefit over controlling heart rate", "what is atrial fibrillation af", "atrial fibrillation foundations" and "study findings suggest revised approach to therapy for atrial fibrillation" . The first 2 documents are longer than usual.
- There are 5 documents, including Atrial Fibrillation, that are generally related to your question. These documents discuss topics such as "news and upcoming events", "patient education event atrial fibrillation foundation", "atrial fibrillation foundations" and "boston atrial fibrillation symposium" and are longer than usual. Management of Patients With Atrial Fibrillation contains more figures and tables than usual.
- Primary atrial fibrillation - HUM-MOLGEN doesn't seem to be related to the main sense of your question and contains more material than average.

Fig. 5.1 Centrifuser summary of documents retrieved for the query *What is atrial fibrillation?*. Underlined portions represent links to related queries or documents.

to report differences between retrieved documents. For example, a user may be interested in a document that describes a topic that is rarely described in other documents, but is relevant to the user's query.

5.1.2 Medical Journal Article Summarization

Research on summarization of medical documents has also investigated the problem of journal article summarization. In this section, we present two journal article summarization systems, one of which uses more

traditional sentence extraction techniques, while the other uses semantic analysis and generation of summary sentences.

One system for journal summarization, TAS [54], was developed as part of PERSIVAL [129], a patient specific digital library. It generates a summary for a set of documents retrieved as relevant to a user query. Thus, TAS also falls into the class of query-focused, multi-document summarization. TAS is strikingly different from most other summarizers as it does not use sentence extraction, but instead extracts information from sub-sentential units to fill in pre-defined templates, orders it, and generates summary sentences. The summaries produced by TAS are briefings containing results reported in clinical studies. TAS uses information extracted from the patient record to filter findings from the article, thus creating a summary for the physician of relevant journal article results that are tailored for the patient under his/her care. The patient record, which serves as a user model for TAS, is available to PERSIVAL from the electronic patient record system WebCIS (Web-based Clinical Information System) [82]. Any information that is repeated across the input articles is identified and presented to the user, while any contradictory results are explicitly signaled to the reader.

To produce a summary, TAS first automatically classifies the article according to its primary clinical purpose. Typical article types are diagnosis, prognosis and treatment. At the content selection stage, a set of templates, each representing a parameter (e.g., chest pain), relation (e.g., association), and finding (e.g., unstable angina), is instantiated for each input article. In this stage, relevant results are extracted exploiting the structure of the medical articles, each of which always has a result section. The templates that are not specific to the patient are filtered out by matching parameters against information found in the patient record. For instance, the template representing the result "Chest pain is associated with unstable angina" will only be included in a summary for a patient with chest pain. During the content organization stage, the relevant templates are clustered into semantically related units, and ordered. Finally, TAS uses a phrase-based language generator to produce a fluent summary.

Yang et al. [223] also describe a query-focused multi-document summarization system. They use sentence extraction in much the same

style as the generic summarizers described earlier, but their system is critically different in its attempt to summarize *all* information related to a particular mouse gene, specified in the user's query. The search for relevant information is performed in all of PubMed, a comprehensive online repository of medical abstracts.[2] Given a query containing a mouse gene of interest, their system returns a ranked list of all sentences across all articles that are relevant and users can choose how many sentences they would like to view. The number of sentences for each gene varied from one to 30,216. Like Elhadad et al., Yang and his co-authors also use semantic information available in the medical domain, but they use it to extract features for clustering and for extracting relevant sentences. In the first stage of their system, they created a database of all sentences indexed by mouse gene using their gene/protein named entity recognition system [34]. They also stored fields such as the MESH headings (i.e., keywords) and publication date for each sentence. Their sentence extraction system uses a set of features to score sentences for inclusion in the summary, including several based on domain specific semantic information. They model their sentence extraction on Edmundson's paradigm and thus, one feature is clue phrases that can be indicative of importance (e.g., "in conclusion"). Another set of key features includes five descriptive features representing clusters of genes in the database (e.g., MESH terms, descriptive words representing where the cluster falls in a gene ontology). The gene clusters are computed using a standard clustering algorithm over gene vectors consisting of MESH headings, gene ontology terms associated with the gene, and the words in the sentence associated with each gene as well as the sentences immediately preceding and following the target sentence. Other features for sentence extraction include the number of genes a sentence mentions, domain specific keywords, sentnece length, and recency of the article in which the sentence occurs. The scoring yields a ranking of sentences for inclusion in the summary and it is left to the user to specify summary length.

[2] http://www.pubmedcentral.nih.gov.

5.2 Journal Article Summarization in Non-medical Domains

While summaries of medical journal articles are clearly important for physicians, summarization of journal articles from other fields is also a topic of current research. Given that journal articles are typically organized in a more predictable way than texts from other genres, proposed approaches often exploit that predictable structure to find salient information. It is expected, for example, that the paper will have an introduction, statement of goals, comparison with related work, approach description, results and conclusions. In this section, we survey an approach using rhetorical status and others that use information about citations.

5.2.1 Using Genre-specific Rhetorical Status

A summarization system for journal articles can take advantage of the *rhetorical status* of a sentence in producing a summary. Teufel and Moens [200] proposed an annotation scheme for rhetorical status of sentences with seven categories. The categories include :

AIM The specific research goal of the current paper.
OWN A neutral description of methodology, results and discussion.
CONTRAST Statements of comparison with other work or weaknesses of other work.
BASIS Statements of agreement with other work or continuation of other work.

Human annotators were able to follow the scheme reliably to annotate each sentence in scientific articles with one of these rhetorical status categories. The annotated corpus was used to train a Naive Bayes classifier using a variety of features, including sentence location in the article, the section and the paragraph, type of section (conclusion, experiments, etc.), sentence length, number of words in the sentence that also appear in the paper title, presence of words with high TF*IDF weight, verb tense, voice and presence of modal auxiliaries, presence and nature of citation, rhetorical context, and presence of 644 formulaic expressions. The classifier was able to predict the rhetorical status of novel sentences with accuracy of 73%.

Listing the sentences that fulfill the AIM, CONTRAST, BASIS and BACKGROUND for each paper gives an excellent generic summary of a scientific article. The explicit indication of the rhetorical status is helpful for users not familiar with the paper that is being summarized. Usually, there are relatively few AIM, CONTRAST and BASIS sentences in the paper and all of them can be included in the summary. BACKGROUND sentences, on the other hand, are more numerous so displaying all of them might become problematic when stricter summary length restrictions are imposed. To solve this problem, one more classifier was trained to distinguish between summary-worthy sentences and other sentences. The classifier is used to select only the most appropriate BACKGROUND sentences and its use increases the overall performance of the summarizer.

5.2.2 Exploiting Citation Links between Papers

Genre-specific summarization of scientific articles can take advantage of a valuable and powerful source of information not available in most other genres: the references (citations) that link papers. The sentences and paragraphs in which references occur often contain concise summaries of the cited paper or other information relevant to that paper. Consequently, the analysis of related articles becomes an essential knowledge source for summarization.

Three types of extractive summaries based on citation link analysis have been proposed: *overview* of a research area (multi-document scientific paper summary) [145], *impact summary* (single document summary of a paper using sentences from the paper itself) [134] and *citation summary* (a mix of multi- and single document summarization, in which a single paper is summarized but the input to the summarizer are the sentences from other papers in which that paper is cited) [169].

Writing an overview of several related scientific papers is a difficult task, even for people. Nanba and Okumura [145] proposed a system for facilitating writing such overviews, which visualizes connections between papers and provides information on how the papers are related to each other. They develop a rule-based system to identify reference areas in a paper, in which other papers are discussed and cited.

In addition, each of these areas is classified as belonging to one of three types: (i) describing methods or approaches used in the paper, (ii) discussion of and comparison with related work, and (iii) other. Classification is performed using hundreds of manually coded cue words. The user can request to see each type of reference area for a paper of interest. With the ever increasing number of scientific publications, the need to further develop and improve such aids for paper browsing and access will only increase with time. In fact much progress has already been achieved in the automatic identification and classifications of citations and citation types [191].

Impact summarization is defined by Mei and Zhai [134] as the task of extracting sentences from a paper that represent the most influential content of that paper. They employ a language model approach to solve the task. For each paper to be summarized, they find other papers in a large collection that cite that paper and extract the areas in which the references occur. A language model is built using the collection of all reference areas to a paper, giving the probability of each word to occur in a reference area. This language model gives a way of scoring the importance of sentences in the original article: important sentences are those that convey information similar to that which later papers discussed when referring to the original paper. The measure of similarity between a sentence and the language model was measured by Kullback–Leibler (KL) divergence. In order to account for the importance of each sentence within the summarized article alone, they use word probabilities estimated from the article. The final score of a sentence is a linear combination of impact importance coming from KL divergence and intrinsic importance coming from the word probabilities in the input article. The method produces extractive summaries that are vastly superior, as measured by the ROUGE automatic measure, to baselines and to the generic summarization system MEAD [202]. One drawback of impact-based summaries is that, while they contain valuable content, they have low linguistic quality and are hard to read and understand, as can be seen from the example shown in Figure 5.2.

Citation summarization [169] is a different approach to summarizing a single article, based on the way other papers refer to it. Citation summarization does not use at all the text of the article being summarized.

1. Figure 5: Interpolation versus backoff for Jelinek-Mercer (top), Dirichlet smoothing (middle), and absolute discounting (bottom).
2. Second, one can decouple the two different roles of smoothing by adopting a two stage smoothing strategy in which Dirichlet smoothing is first applied to implement the estimation role and Jelinek–Mercer smoothing and Jelinek–Mercer smoothing is then applied to implement the role of query modeling.
3. We find that the backoff performance is more sensitive to the smoothing parameter than that of interpolation, especially in Jelinek–Mercer and Dirichlet prior.
4. We then examined three popular interpolation-based smoothing methods (Jelinek–Mercer method, Dirichlet priors, and absolute discounting), as well as their backoff versions, and evaluated them using several large and small TREC retrieval testing collections.
5. By rewriting the query-likelihood retrieval model using a smoothed document language model, we derived a general retrieval formula where the smoothing of the document language model can be interpreted in terms of several heuristics used in traditional models, including TF-IDF weighting and document length normalization.
6. We find that the retrieval performance is generally sensitive to the smoothing parameters, suggesting that an understanding and appropriate setting of smoothing parameters is very important in the language modeling approach.

Fig. 5.2 An example of an impact-based summary.

Instead, Qazvinian and Radev [169] propose to summarize the reference areas in other articles related to the target paper that has to be summarized. The input to the summarizer consists of the sentences or short paragraphs from other papers that discuss the target article. There is a high degree of repetition in the input because many papers refer to the same aspect of the target paper: an approach, result, main contribution, etc. Given this characteristic of the data, a clustering method for summarization seems highly appropriate (see Section 2.1.2). Similar sentences are clustered together and a representative sentence is chosen to convey the information in that cluster. The best way to find the most representative sentence in each cluster turned out to be applying the graph-based method discussed in Section 2.1.3 for summarization to each cluster. The evaluation was performed on 25 articles from five sub-areas of computational linguistics, using the manual Pyramid evaluation method.

1. The Czech parser of Collins et al. (1999) was run on a different data set and most other dependency parsers are evaluated using English.
2. More precisely, parsing accuracy is measured by the attachment score, which is a standard measure used in studies of dependency parsing (Eisner, 1996; Collins et al., 1999).
3. In an attempt to extend a constituency-based parsing model to train on dependency trees, Collins transforms the PDT dependency trees into constituency trees (Collins et al., 1999).
4. More specifically for PDT, Collins et al. (1999) relabel coordinated phrases after converting dependency structures to phrase structures, and Zeman (2004) uses a kind of pattern matching, based on frequencies of the parts-of-speech of conjuncts and conjunctions.
5. In particular, we used the method of Collins et al. (1999) to simplify part-of-speech tags since the rich tags used by Czech would have led to a large but rarely seen set of POS features.

Fig. 5.3 An example of a citation summary.

Citation summaries are even harder to read than impact summaries because they mix information about the work done in the paper which cites the target article with descriptions of the work described in the target article itself, as shown in Figure 5.3. Application of the approaches developed for categorization of types of reference areas should be helpful in overcoming this problem in future work.

All of these approaches reveal that there is much work that remains in the field of journal summarization. Very different approaches from generic summarization have been proposed that exploit the structure and specific characteristics of journal articles and this is appealing. The quality of the summaries that are produced, however, still leaves much to be desired. Further work is needed that takes these new approaches to the next level, perhaps by integrating work on fluency, cohesion and sentence ordering.

5.3 Email

Research in email summarization ranges from summarization of a single email message to summarization of a email collection, including both

summarization of a mailbox and summarization of a thread of related emails. The function of summarization varies in each of these tasks. For single email summarization, a short reminder of the message topic can suffice. Indications of topic can help users to prioritize incoming email, determining when an immediate reply is required and to quickly find older relevant messages. For a mailbox of emails, summarization can provide a browsing interface to help the user find emails of interest. Focused summarization is also a possibility as, for example, the focus may be to identify tasks that must be done. For an email thread, summarization must convey the essence of an extended conversation.

Summarization must be sensitive to the unique characteristics of email, a distinct linguistic genre that exhibits characteristics of both written text and spoken conversation. A thread or a mailbox contains one or more conversations between two or more participants over time. As in summarization of spoken dialog, therefore, summarization needs to take the interactive nature of dialog into account; a response is often only meaningful in relation to the utterance it addresses. Unlike spoken dialog, however, the summarizer need not concern itself with speech recognition errors, the impact of pronunciation, or the availability of speech features such as prosody. Furthermore, responses and reactions are not immediate and due to the asynchronous nature of email, they may explicitly mark the previous email passages to which they are relevant.

5.3.1 Summarization of Individual Email Messages

Summarization of a single email message can be accomplished by selecting noun phrases that are indicative of its topic. This is the approach taken in Gister [141, 208], a system that uses a combination of linguistic filtering and machine learning to select noun phrases for the summary.

Researchers at Microsoft Research [42] note that email often contains new tasks for the recipient. They suggest that it would be helpful if a system could automatically identify and summarize the tasks contained in emails. They have developed a system that uses machine learning to classify sentences within emails as task-directive or not. The system then reformulates each such sentence in the imperative to

make the task explicit. To collect data for their supervised approach, they had a group of annotators annotate a large body of email with speech acts specifying task and non-task acts. The features used for machine learning contained message-specific features (e.g., number of addressees, number of forwarded messages), superficial features (e.g., n-grams) and linguistic features (e.g., structural features that could be derived from part-of-speech tagging or parsing). They used Support Vector Machines for machine learning and performed feature ablation to determine the impact of different types of features. Their evaluation showed that the best results are achieved with all features, but there was little difference between the use of deep linguistic features and surface linguistic features, thus indicating that their system would still get good results without parsing.

Full document summarization has also been used to generate a summary of email messages. Lam et al. [99] use an existing IBM summarization system [17] to produce summaries. However, in order to make this feasible, they first must pre-process the message to remove pieces that would be problematic; their pre-processor removes certain headers, quoted text, forward information and email signatures. Summaries are generated by including the context of the email messages. Thus, an email message plus all its ancestor messages in the thread are sent to the document summarizer. The resulting summary is augmented with the results of named entity extraction and includes a list of important names and dates at the end of the textual summary. This summarization approach resembles some of the techniques that we describe in the next section, where summarization of full threads is discussed.

5.3.2 Summarization of an Email Thread

In early research on summarization of email threads, Nenkova and Bagga [151] developed a system to generate indicative summaries of email threads in an archived discussion. They used an extractive summarizer to generate a summary for the first two levels of the discussion thread tree, producing relatively short "overview summaries." They extracted a sentence for each of the two levels, using overlap with preceding context. For the root message, they extracted the

shortest sentence with the largest number of nouns which also occur in the email's subject header. For the follow-up email, they extracted the sentence with the most overlap of words with the root message. The threads in their corpus of email were relatively short, so this approach worked well.

Later work on summarization of email threads [177] also used extractive summarization. However, Rambow et al.'s work zeroed in on the dialogic nature of email. They experimented with two extractive summarizers based on machine learning. The first used features similar to any text summarizer and considered the input thread as simply a text document. The second summarizer relied on email specific features in addition to traditional features, including features related to the thread and features related to email structure such as the number of responders to a message, similarity of a sentence with the subject, etc. They used a rule-based classifier, RIPPER, and their results show that the full set of features yield the best summaries.

Email conversations are a natural means of getting answers to one's questions and the asynchronous nature of email makes it possible for one to pursue several questions in parallel. As a consequence, question–answer exchanges figure as one of the dominant uses of email conversations. These observations led to research on identification of question and answer pairs in email [140, 187] and the integration of such pairs in extractive summaries of email [132]. McKeown et al. experiment with different methods for integrating question and answer pairs into email summarization and show that all achieve significant improvement over an extractive approach alone, which may end up including either a question without its answer or an answer without its question.

Summarization of an email thread can also be used to allow a reader to catch up on an extended discussion that took place in her absence and make a decision about how to respond to an outstanding issue. Wan and McKeown [215] develop an approach that identifies the main issue within an email and the responses later in the thread that address that issue. Their approach makes the assumption that the issue will be found in the first email in the thread, and subsequently extracts all responses to the message. To determine the issue, they find the sentences that are most similar to the response. They find that using a centroid based

approach with singular value decomposition for similarity computation produces the best results. The intuition that motivates this method is very similar to that in impact summarization of scientific articles that we discussed previously in Section 5.2.1.

While the summarizers described here have sometimes used a techniques developed for another genre, overall it seems clear that attention to the characteristics that are specific to email result in better summarization systems. Those researchers that have experimentally compared both approaches have seen an improvement when they include email specific features. In general, machine learning results have shown that including the linguistic features that are used in generic summarization in addition to email specific features yields the best results.

5.3.3 Summarization of an Email Archive

The research we have reported on so far has generated summaries for an individual thread. Archives and mailboxes, however, consist of multiple threads and it can be difficult for an end-user to organize his mailbox in such a way that it is easy to find the email or thread of interest. Newman and Blitzer [156] present a system that can be used for browsing an email mailbox and that builds upon multi-document summarization techniques. They first cluster all email in topically related threads. Both an overview and a full-length summary are then generated for each cluster. The resulting set of summaries could be presented as a browsing page, much as is done with news summarization systems such as NewsBlaster [130] or NewsInEssence [173]. Newman and Blitzer rely on email specific features for generating both overviews and summaries. They also take quotes into account. Another key feature of overview and summary generation is the similarity of a sentence to a cluster centroid.

A more recent approach to summarization of email within a folder uses a novel graph-based analysis of quotations within email [27]. Using this analysis, Carenini et al.'s system computes a graph representing how each individual email directly mentions other emails, on the granularity of fragments and individual sentences. "Clue words" are defined as stems of words in an email message that are repeated in either the

parent or child node (fragment) from the quotation graph. The CWS (Clue Word Summarizer) scores sentences for inclusion based on frequency of the clue words they contain. Carenini et al. compare how well CWS works relative to MEAD [56], to a modified version of the statistical sentence extractor built by Rambow et al. [177], and to a combination of CWS with MEAD in which the CWS and MEAD scores for each sentence are normalized and combined. Their results show that the clue words summarizer does as well as or better than previous approaches. The evaluation was performed using sentence precision and recall on 20 email conversations from the Enron email dataset. Annotators were asked to produce extractive summaries with length 30% of the original text to obtain the gold-standard.

5.4 Web Summarization

The quantity of information appearing on the web has been the motivation for much research on summarization, yet research on summarization of web sources started much later than other work. As with the other genres and domains discussed in this section, systems developed for web summarization can take into account the characteristics of the web and the data being summarized to produce higher quality and more accurate summaries. To date, most research carried out for this genre has focused on summarization of web pages, with a smaller body of work looking at summarization of blogs.

5.4.1 Web Page Summarization

Research on web page summarization varies in terms of how much the summary is influenced by web page content or by web page context, that is, the text surrounding the links which point at that page. Most research in this area makes use of a large scale resource from the Open Directory, known as DMOZ,[3] consisting of summary/web page pairs all organized into a hierarchy with web pages on more general topics occurring near the top of the hierarchy and more specialized sub-topics occurring along the branches. The short human summaries

[3] http://www.dmoz.org/.

for each web page are a valuable resource and they can serve as model summaries against which automatically generated summaries can be compared.

Early research on web page summarization [13, 24] relied on page content alone to generate the summary. Berger and Mittal [13] take an approach modeled on statistical machine translation (MT). They use alignment from MT to align DMOZ summaries with the web page content. To generate the summary, they use two models, the first of which determines the words of the summary and the second of which determines their order. The novelty in their approach is that they generate a gist which may include words that did not occur in the web page at all. Their content model, which selects summary words, is based in part on word frequency within the document but they also generate semantically related words to replace one or more words in the document. They use traditional alignment developed for machine translation between DMOZ summaries and the corresponding web pages to generate the semantically related words. Their generation model is simply an n-gram model trained over the DMOZ summaries. Buyukkokten et al. [24], in contrast, focus on producing adjustable-length summaries that can be successively revealed on a small hand-held device. They use a relatively simple summarization method, selecting the most important sentence from a discourse segment within the web page, based entirely on Luhn's summarization method. The contribution of their approach is in their ability to expand and contract summaries by adding sentences within a discourse segment or across segments.

Later research on web page summarization takes an interesting turn and explores the use of context on the web. Discussion here centers on whether summaries can be generated from the context alone — as in the citation summarization approach for scientific articles — or whether algorithms should take into account web page content as well. One type of context to consider is the text in pages that link to the one that has to be summarized, in particular the text surrounded by the hyperlink tag pointing to the page. This text often provides a descriptive summary of a web page (e.g., "Access to papers published within the last year by members of the NLP group"). Proponents of using context to provide summary sentences argue that a web site includes

multimedia, may cover diverse topics, and it may be hard for a summarizer to distinguish good summary content from bad [47]. The earliest work on this approach was carried out to provide snippets for each result from a search engine [2]. To determine a summary, their system issued a search for a URL, selected all sentences containing a link to that URL and the best sentence was identified using heuristics. Delort et al. [47] use a very similar procedure to select context sentences. They extend the older approach through an algorithm that allows selection of a sentence that covers as many aspects of the web page as possible and that is on the same topic. For coverage, Delort et al. used word overlap, normalized by sentence length, to determine which sentences are entirely covered by others and thus can be removed from consideration for the summary. To ensure topicality, Delort's system selects a sentence that is a reference to the page (e.g., "CNN is a news site") as opposed to content (e.g., "The top story for today..."). He computes topicality by measuring overlap between each context sentence and the text within the web page, normalizing by the number of words in the web page. When the web page does not have many words, instead he clusters all sentences in the context and chooses the sentence that is most similar to all others using cosine distance. This algorithm thus uses some aspects of the web page in addition to context, an advance over earlier work, although it is clear that word overlap is a very rough approximation of measuring representativeness.

Sun et al. [197] propose another type of context: the use of clickthrough data. They use data from Microsoft's search engine to gather triples (u, q, p) representing queries (q) issued by user (u) who then clicked through to page (p). In this work, a summary is constructed by selecting sentences from the web page. They experiment with two methods, one of which uses an adaptation of Luhn's method [111], computing the significance factor of words by a weighted mixture of TF*IDF computed by frequency on the web page and TF*IDF computed by frequency in the query set. Luhn's algorithm is then used to compute the significance factor of each sentence. They also adapt latent semantic analysis in a similar way by changing the computation for each term in the LSA matrix, weighting its frequency by the number of times it appears in the query set. To account for web pages which

are not covered by clickthrough data, they develop a thematic lexicon using the Open Directory Project, associating query terms with each node in the DMOZ hierarchy, where each node corresponds to one or more web pages. Thus, the query terms at the node can be used for web pages with no clickthrough data. Their comprehensive evaluation shows that query words dramatically improve performance and that the latent semantic analysis is better than the adaptation of Luhn's method. This is a strong piece of work with interesting results when clickthrough data is available. Note that for researchers outside of search engine facilities, such data is not readily available.

Research on web site summarization is ongoing. Choi et al. [30] use web site summarization to enhance sponsored ad presentation, augmenting the ad with all or part of the landing page, using unsupervised methods. It is clear that web site summarization is a very young area of research, with lots of room for both improvement and new ways in which summarization can be exploited. Furthermore, the methods for making use of web context have been only partially explored.

5.4.2 Summarization of Online Interactive Sites

A few people have looked at summarization of online discussion forums and of blogs. Zhou and Hovy [233] use a method similar to methods for speech and email in their work on summarization of online discussion forums. They note that discussion forums involve asynchronous communication between participants on multiple interleaved topics. Their work focuses on the identification of adjacency pairs (similar to [66]) to identify who talks to whom and uses topic segmentation to identify the different topics under discussion, finally producing a summary that is structured into sub-summaries for each topic. For blogs, they exploit the observation that in political blogs, posters base their discussion on news articles and different points in the discussion are linked directly to the article. Assuming that readers are interested in facts, they create a summary from the sentences that directly link to the associated news article. An alternate approach [83] creates a summary by extracting sentences from the blog post using word frequency to measure sentence importance, but this work critically differs from others in that

it measures word frequency in the comments on posts, thus aiming at extracting sentences that elicited discussion.

5.5 Summarization of Speech

Like most research in genre-specific summarization, summarization of speech also requires taking into account the specific features of the medium. That can include features from the speech signal or features that capture the dialog nature of exchanges. Techniques for automatic summarization of spoken input have been developed for many applications, most notably for summarization of broadcast news [31, 79, 124], dyadic conversations [71, 72], meetings [63, 142, 214] and lectures/presentations [68, 77, 93, 168, 230].

Some of the techniques developed for text summarization have been successfully applied to speech summarization without much modification, in the setting when a written transcript of the audio is available to work with. The most widely applied techniques for transcript summarization [31, 77, 142, 229] have been Maximal Marginal Relevance [25] and Latent Semantic Analysis [69]. Variations of frequency features (total frequency, TF*IDF) and positional features, as well as length features, similar to the ones developed for text summarization are employed in much of the existing work on speech summarization.

At the same time, there are some aspects unique to speech summarization which have led to the need for research in directions unexplored in work on text summarization. Probably the most striking finding that will need further verification and analysis in future work are results indicating that very good summarization performance can be achieved on the basis of *acoustic features alone*, with no recourse to transcripts or textual features [124, 158, 231].

We overview some of the unique characteristics of speech as input for summarization in the sections below.

5.5.1 Dealing with Errors From Automatic Speech Recognition

In most realistic situations, manual transcripts of the audio input to be summarized are not available. In this case, automatic speech

recognition (ASR) is used to obtain a textual transcript. Depending on the type of speech, ASR can be errorful, with word error rates ranging from 10% for more formal read speech to as high as 40% for spontaneous conversations or lectures [68]. Speech summarization systems intended to produce textual output need to develop techniques for finding portions of the input that are not only informative but also have lower recognition error rates.

To give an idea of how the use of automatic transcripts changes the clarity and content of summaries, we reproduce Murray et al.'s [142] example of a snippet of a human summary of a meeting, an LSA summary of manual and automatic transcripts:

Human The experiment consisted of leading a subject to believe she were talking to a computer, then having the "computer" break down and be replaced with a human.

LSA; manual transcript I should say the system was supposed to break down and then these were the remaining three tasks that she was going to solve with a human. One time to pretending to be a human which is actually not pretending.

LSA; automatic transcript Reverse should so the system were supposed to break down and then this would be remaining three tasks that she was going to solve with a human.

To address the problems with ASR issues during content selection, Zechner and Waibel [229] employ a combination of two features in selecting passages, using speaker turns as selection units: the fraction of words in the utterance whose confidence score from the automatic speech recognition system is greater than 0.95 and the importance of the passage computed using MMR [25]. They test the technique on four television shows consisting of conversations of multiple speakers. The trade-off between recognition confidence and segment importance led to an average word error rate reduction in the summary of about 10% and a relative improvement of summary accuracy of about 15% compared to the standard MMR approach developed for text summarization.

In later work, techniques for content selection accounting for recognition accuracy have been refined. For example Kikuchi et al. [93] use TF*IDF and trigram probability of words to measure importance and a recognition confidence score equal to the logarithm of the ratio of

the word hypothesis probability and that of all other hypotheses to measure confidence in the quality of the automatic transcript.

Alternatively, effort could be focused to improve ASR performance before summarization is performed [68]. Changing the output modality could also remove the need to explicitly deal with ASR errors. For example, errors would not matter much if the final summary is rendered by playing the audio snippets [204] determined to be important by a robust algorithm not sensitive to transcription errors [144].

5.5.2 Dealing with Disfluencies: Filled Pauses and Repetitions

Spontaneous speech is characterized by the presence of filled pauses ("um", "uh", "well") and repetitions ("i was going i was going to call you tonight"). The percentage of disfluent words in informal conversations can be high: 15–25% of the total words spoken [227], so detecting and removing disfluencies has the potential of improving the readability of summaries as well as to reduce noise in feature computation. Consider Zechner's example of original conversation utterance and its cleaned up version:

```
A: well I um I think we should
   discuss this you know with her
A': I think we should discuss this with her
```

Detection of disfluencies is necessary for many other applications besides summarization and has been addressed by numerous researchers as a stand alone task [91, 138, 196]. Such tools for disfluency detection can be used as initial preprocessing of the input, cleaning up the transcripts as the first task for the summarization system [229].

Another possible approach is that of Hori et al. [79] who include trigram probability as one of the terms used to calculate how desirable for selection an utterance is. The trigram probability of utterances containing ASR errors or disfluencies would be low, so this measure is effective in reducing out-of-context words in the summary.

In recent work, Zhu and Penn [236] put forward a suggestion for different treatment of disfluencies. They argued that instead of removing

disfluencies from the input, systems should use their presence as a feature in machine learning models for extractive summarization. They maintain that disfluencies could actually be a signal for important sentences because disfluencies are more likely to be inserted in the beginning of discourse segments or before important concepts are mentioned. In their experiments, using disfluency features — number of repetitions and filled pauses — leads to small improvements of less than 1% of ROUGE scores. A more detailed feature analysis is necessary in order to find out if disfluencies are positively or negatively correlated with the summary-worthy class, that is if they are indeed predictive of summary content rather than of content that should be avoided.

5.5.3 Units for Content Extraction

In text, the typical unit on which extractive systems operate are sentences. Sentence boundaries are usually easy to identify, especially for languages like English where punctuation is a relatively unambiguous delimiter of sentence boundaries. In speech, the decision about what unit to use for content selection is not that obvious. A speaker might pause in the middle of a sentence, or two consecutive sentences might not be delimited by a pause, so finding the right granularity for extractive summarization requires more sophisticated processing.

Several approaches have been developed for automatic segmentation of speech (transcripts) into sentences (for example see [108, 188]). The system of Liu et al. [108] simultaneously detects sentence boundaries and disfluencies, allowing for the transformation of original conversational speech that is hard to read into more usual textual form:

Original but uh i i i i think that you know i mean we always uh i mean ive ive had a a lot of good experiences with uh with many many people especially where theyve had uh extended family and i and an- i i kind of see that that you know perhaps you know we may need to like get close to the family environment.

Processed But ive had a lot of good experiences with many people especially where theyve had extended family. I kind of see that perhaps we may need to get close to the family environment.

The use of systems for automatic sentence segmentation is necessary when human transcripts for the input are not available. In many cases, additional tuning of the sentence segmentation system for the specific purposes of summarization might be necessary [109, 139].

Some researchers have argued that segment granularities different from sentences would be even more appropriate for speech summarization. *Adjacency pairs* are motivated by pragmatic dependencies between utterances, while *intonational phrases* are more faithful representations of the properties of speech than sentence boundaries.

Maskey et al. [125] compared three ways of segmenting speech for extractive summarization: sentences, pause delimited fragments and intonational pharses. The automatic system of Liu et al. [108] was used to find sentences; pause delimited segments were defined as snippets of speech between pauses of 250 ms. Intonational phrases were also automatically detected by a machine learning predictor. For training of the predictor, one transcript of ABC "World News Tonight" broadcast was manually annotated. When used for extractive summarization, intonational phrases gave best results, significantly higher than either sentence or pause-delimited fragment extraction.

Adjacency pairs, and Question–Answer pairs in particular, consist of utterances that are pragmatically dependent on each other. Such pairs would usually involve utterances by two speakers and are informative only when taken together: a summary should not contain a question without its respective answer or an answer without the question it was given in response to. Zechner and Lavie [228] were the first to point out these issues. They proposed a heuristic method for identifying Question–Answer pairs and showed that including these as a single unit in the summary significantly improved the readability of the output as measured by subjective assessment of evaluators.

In later work, a machine learning model using structural, durational and lexical features for the identification of adjacency pairs in meetings was shown to achieve impressive accuracy of 90% [66]. The automatically identified adjacency pairs certainly augment the capabilities for summarization of conversational speech because they can be used to identify agreement and disagreement in conversations [66], as well as to better rank the importance of utterances in the conversation [63].

5.6 Discussion

Many of the summarization approaches discussed in this section diverge markedly from the generic sentence extraction approach. Instead, authors take advantage of the particular characteristics of the genre, media, or domain. In journal article summarization, researchers exploited consistency in structure of articles (e.g., conclusions, methods) and the use of citations. In email summarization, researchers exploited the asynchronous nature of dialog that typically appears, often using meta-data to help. In speech summarization, researchers exploit information from the speech signal itself as well as indicators about dialog structure. In web page summarization, researchers exploited the context of links and search results to aid in determining what is important to include in a summary. In many of these domains, additional resources are available to help researchers advance beyond pure sentence extraction. For example, in the medical domain, the combination of the semantic information available in the UMLS along with the specific structure of text (e.g., results sentences) enabled the use of interpretation followed by generation.

Just as the characteristics of the genre, media or domain enabled exploration of new approaches, other characteristics create difficulties for research in this area. Many of the genres feature noisy text, with partial or ungrammatical sentences, and disfluencies. Such hurdles sometimes spurred non-extractive approaches (e.g., gisting or compression) and at other times, required research into approaches that could process the data (e.g., the identification of who talks to whom).

As opposed to generic news summarization, whether single document or multi-document, summarization in these areas is only beginning. These are areas where the summarization field is poised for break-throughs and where new researchers can find interesting, unsolved problems.

6

Intrinsic Evaluation

Task-based evaluations are time-consuming, expensive and require a considerable amount of careful planning. As a result, they are not that suitable for system comparisons and evaluation during development. Intrinsic evaluations are normally employed in such cases, either by soliciting human judgments on the goodness and utility of a given summary, or by a comparison of the summary with a human-authored gold standard. When comparisons with a gold standard are involved, it is desirable that these be done automatically to further reduce the need for human involvement.

Here we give an overview of the main intrinsic approaches used in summarization evaluation and the motivation for developing certain evaluation methods and for abandoning others.

6.1 Precision and Recall

Most summarization systems select representative sentences from the input to form an *extractive* summary; the selected sentences are strung together to form a summary without any modification of their original wording. In such settings, the common information retrieval metrics of

precision and recall can be used to evaluate a new summary. A person is asked to select sentences that seem to best convey the meaning of the text to be summarized and then the sentences selected automatically by a system are evaluated against the human selections. Recall is the fraction of sentences chosen by the person that were also correctly identified by the system

$$\text{Recall} = \frac{|\text{system–human choice overlap}|}{|\text{sentences chosen by human}|} \tag{6.1}$$

and precision is the fraction of system sentences that were correct

$$\text{Precision} = \frac{|\text{system–human choice overlap}|}{|\text{sentences chosen by system}|} \tag{6.2}$$

The appeal of precision and recall as evaluation measure is that after a human defines the gold standard sentence selection, it can be repeatedly used to evaluate automatically produced summaries by a simple comparison of sentence identifiers. Yet, there are also several problems with these measures, as we discuss next.

6.1.1 Human Variation

Different people tend to choose different sentences when asked to construct an extractive summary of a text. Research as early as Rath et al. [178] reported that extracts selected by six different human judges for 10 articles from Scientific American had only 8% overlap on average. It has been shown [48] that the same summary can obtain a recall score with between 25% and 50% difference depending on which of two available human extracts are used for evaluation. Thus, a system can choose a good sentence, but still be penalized in precision/recall evaluation. In light of this observation, it also seems that in summarization evaluation it will be more beneficial to concentrate on recall rather than precision. Precision is overly strict — some of the sentences chosen by the system might be good, even if they have not been chosen by the gold standard creator. Recall, on the other hand, measures the overlap with already observed sentence choices.

6.1.2 Granularity

Another problem with the precision/recall (P/R) measures is the fact that sentences are not the best granularity for measuring content. Sentences differ in word length and convey different amounts of information. Selecting a longer and more informative sentence can be more desirable than selecting a short sentence. Imagine, for example, a human extract consisting of the sentences "(1) We need urgent help. (2) Fires have spread in the nearby forest, and threaten several villages in this remote area." Now imagine two systems, each choosing only one sentence appearing in the human extract, one choosing sentence (1) and the other choosing sentence (2). Both summaries will have the same P/R score, but can hardly be perceived as equally informative.

6.1.3 Semantic Equivalence

Yet another problem with using sentences as the unit of evaluation is that two distinct sentences can express the same meaning. This situation is very common in news, and is particularly pertinent in multi-document summarization of news, in which the input to the system consists of many articles on the same topic. Again, a human would select only one of the equivalent sentences but a system will be penalized for choosing an alternate sentence that expresses the same meaning.

Many of the alternative evaluation measures were designed to address the issues that were raised regarding P/R. For example, it has been suggested to use multiple models rather than a single person's judgment [85]. Smaller, more-semantically oriented units of analysis have been proposed, and more emphasis has been given on recall.

6.2 Relative Utility

Relative utility [172] has been proposed as a way to address the human variation and semantic equivalence problems in P/R evaluation. In this method, multiple judges score *each sentence in the input* on a scale from 0 to 10 as to its suitability for inclusion in a summary; highly ranked sentences are very suitable for a summary, and low ranked sentences should not be included in a summary. The judges also explicitly mark

which sentences are mutually substitutable because of semantic equivalence. Thus, each possible selection of sentences by a system can be assigned a score showing how good a choice of sentences it represents.

The approach seems intuitive and quite appealing, but requires a good deal of manual effort in sentence tagging. Moreover, at times it fails to discriminate between human and automatic summaries, a distinction which a good evaluation measure should be able to do. Particularly when applied to the evaluation of SWITCHBOARD summaries [235], automatic summarizers achieved a score higher than that of the humans, indicating that this approach for evaluation is not a good choice for evaluation of summarization of conversational speech. The approach has been used in only a handful of studies of text summarization and is in general not a preferred metric to report in a paper.

6.3 DUC Manual Evaluation

The Document Understanding Conference/Text Analysis Conference (DUC/TAC)[1] has been carrying out large-scale evaluations of summarization systems on a common dataset since 2001. On average, over 20 teams participate in this NIST-run evaluation each year and a lot of effort has been invested by the conference organizers to improve evaluation methods. DUC content evaluations performed in the period between 2001 and 2004 were based on a single human model. However, in order to mitigate the bias coming from using gold standards from only one person, different annotators were creating the models for different subsets of the test data, never having the same person creating gold standards for all test sets [75].

In order to address the need for more fine-grained analysis than the sentence level, DUC adopted elementary discourse units (EDUs), roughly correspond to clauses, as the basis for evaluation. The model summary was automatically split into EDUs, and machine summaries were evaluated by the degree to which they cover each EDU in the model. The overall score, called *coverage*, was the average score across

[1] `http://duc.nist.gov`, `http://www.nist.gov/tac/`.

all EDUs in the model. The measure was recall-oriented, in essence measuring what fraction of the model EDUs were covered.

In an attempt to encourage research in abstractive summarization, DUC also started using human-generated abstracts as models, rather than human selection of sentences. The above-described evaluation method supported this transition, at the expense of requiring more human involvement.

The availability of the output of many systems over many test inputs has allowed researchers to study the factors that influence summarization performance. It has been reported that in ANOVA analysis of coverage scores with system, input and the human creators of models as factors, the model creator turned out to be the most significant factor [128]. This once again raised concerns about the advisability of using a single human model for evaluation. The input document to be summarized was also a significant factor [148], suggesting that some inputs are easier to summarize than others.[2] Summary length was also a significant factor, with summary coverage tending to increase as the length of the summary increases.

Two lines of research on evaluation emerged in an effort to address some of the issues raised by the DUC evaluation protocol: developing cheap automatic methods for comparing gold standards with automatic summaries and better analysis of human variation in content selection, relying on multiple models to avoid dependence on the gold standard.

In later years (after 2004), the initial DUC evaluation protocol was no longer used. Instead, the Pyramid manual evaluation was used as a measure for content selection [153]. In addition, a simpler metric of content quality was introduced for the query-focused tasks in DUC and TAC: responsiveness. For this evaluation, human annotators were asked to rate on a scale from 1 to 5 how good they thought the summary was in terms of providing information relevant to the user query.

[2] This finding shows that paired tests should be used to compare the performance of two systems on the same test set. These tests eliminate the variation that is due to the input difficulty and better assesses the significance of difference between the systems.

6.4 Automatic Evaluation and ROUGE

Automatic evaluation measures have been known even before the widely used BLEU technique for machine translation evaluation [166] and the ROUGE technique derived from it [105] (see for example [176]). The problem has been that different automatic evaluation approaches give different results, so it was not clear what the scores mean and which automatic measure is to be preferred. In using BLEU for machine translation evaluation, however, researchers developed methods to validate automatic approaches. They took manual evaluations generally accepted in the research community, and looked for automatic measures which correlated well with the human scores *over a large set of test points*, especially when multiple human models were used.

Inspired by the success of the BLEU n-gram overlap based measure, similar n-gram matching was tried for summarization. Using DUC coverage scores to validate the method, the ROUGE[3] system for automatic evaluation of summarization was developed. ROUGE is also based on the computation of n-gram overlap between a summary and a set of models. ROUGE has been preferred for summarization because it is recall-oriented, unlike BLEU, which emphasizes precision [107]. ROUGE has numerous parameters, including word stemming, stop word removal and n-gram size. Different settings work best for different summarization tasks as can be seen from the detailed tables in [105]. This means that different parameters need to be tested for new tasks, such as speech summarization of spontaneous conversations or recordings of meetings.

ROUGE is the most commonly used metric of content selection quality used in research papers because it is cheap and fast. Many researchers feel more comfortable to supplement ROUGE figures with a manual evaluation of content such as the Pyramid method, at least on some small subset of the test data.

6.5 Pyramid Method

The Pyramid method [152, 153] is concerned with analysis of the content selection variation in human summaries [201, 211], as well

[3] Recall-oriented understudy for gisting evaluation.

as how evaluation results can be made less dependent on the model used for evaluation. Multiple human abstracts are analyzed manually to derive a gold standard for evaluation. The analysis is semantically driven: information with the same meaning, even when expressed using different wording in different summaries, is marked as expressing the same summary content unit (SCU). Each SCU is assigned a weight equal to the number of human summarizers who expressed the SCU in their summaries. The distribution of SCU weights is Zipfian, with few SCUs being included by many summarizers and a heavy tail of low-weight SCUs.[4] SCU analysis shows that summaries that differ in content can be equally good and assigns a score that is stable with respect to the models when 4 or 5 human summaries are used. The actual Pyramid score is equal to the ratio between the weight of content expressed in a summary and the weight of an ideally informative summary with the same number of SCUs.

A drawback of this approach is that it is very labor intensive, despite the fact that a special annotation tool (DUCView[5]) has been developed to facilitate the process. Also, the method was developed for evaluation of abstractive summaries, and requires analysis that is unnecessary for extractive summaries, as we see in later sections.

6.6 Linguistic Quality Evaluation

All the evaluation methods discussed so far have been focused on evaluating the information content of a summary, or its overall informativeness. But summary readability is also an important factor in summary evaluation, albeit often neglected by summarization researchers. In DUC, a set of questions was developed to evaluate readability aspects of summaries. Are they ungrammatical? Do they contain redundant information? Are the references to different entities clear? Does the summary build up sentence by sentence? While much progress has been seen in improving system content selection, most automatic summaries score rather poorly on readability aspects such as coherence and

[4] Hence the name of the method. If SCUs are ordered in tiers from low to high weight, we get a pyramid shape.

[5] `http://www1.cs.columbia.edu/~ani/DUCView.html`.

referential clarity [150]. Improving automatic summary readability is an open problem in summarization and developing suitable metrics that will allow tracking of progress in this direction is important. Recent interest in sentence ordering and referential cohesion have led to a proposal for automatic evaluation of cohesion [101]. Hopefully, more future effort will be focused on linguistic quality.

In TAC 2008, the "responsiveness" measure was re-defined, and included both the extent to which a summary succeeded in providing information relevant to the user defined topic *and* the overall linguistic quality of the summary.

Human assessments of linguistic quality on a scale, usually 1–5, are probably the fastest and cheapest to obtain. They do not require the collection of gold standard summaries, nor any annotation or manual analysis from the assessors in order to come up with a summary quality score. Because of these properties, this evaluation approach is rather attractive especially when many systems have to be compared on many inputs. At the same time, unlike for many other manual metrics, the properties of this metric, such as inter-annotator agreement and reproducibility, have not been well studied.

6.7 Intrinsic Evaluation for Speech Summarization

While many speech summarization researchers have used precision/recall of utterances [72, 124] or automatic measures such as ROUGE to evaluate their results, there have been two proposals for evaluation methods specifically designed for this new genre: summary accuracy and summarization accuracy.

6.7.1 Summary Accuracy

Summary accuracy was defined by Zachner and Waibel [229]: for each word in an utterance, they define a weight, which they call a relevance score, equal to the average number of times the word occurred in a phrase selected for inclusion in the summary by a human annotator.[6]

[6] Such a definition addresses the *granularity* problem discussed in Section 6.1 for precisions/recall, because using word-by-word comparison accounts for the possibly different informativeness of utterances.

So, if five annotators are asked to construct a summary, and three of them select the same span of text, all the words in this span will be assigned a relevance score equal to 3/5, even if some words appear in the other two text spans as well. Summary accuracy is then defined as equal to the sum of relevance scores of the words in a system-selected utterance that were correctly recognized by the ASR system, divided by the *maximum achievable relevance score* with the same number of words somewhere in the text. This definition of word relevance (weight) and overall summary score is very similar to the idea on which the pyramid evaluation method for news is based. In fact, while attempting to apply the pyramid method for evaluation of meeting transcripts, Galley [63] observed that, for meeting summarization, human summaries formed by sentence extraction convey the same information only when the two annotators extracted exactly the same sentence.[7] He then computed pyramid scores based on words rather than content units, with the restriction that a given word is assigned a non-zero relevance score only when it is part of *the same* utterance that the humans selected. This scoring worked out quite well for the meeting domain, and is almost equivalent to summary accuracy. Such reinvention of scoring metrics is indicative of the need for closer interaction between researchers tackling different types of summarization genres.

6.7.2 Summarization Accuracy

Summarization accuracy has been defined in the context of the evaluation of speech summarization systems performing sentence compaction [80]. The sentence compaction task is to eliminate "unnecessary" words from a sentence, similar to compression. Again, multiple annotators are asked to produce possible compactions; many possible compactions for a given sentence can be produced. In order to extrapolate more likely but unseen compactions from those produced by the annotators, all human productions for a sentence are merged into a single network and different traversals of the network can produce new compaction variants that were not produced by any of the humans, but that are

[7] This fact suggests that the *semantic equivalence problem* of precision/recall might not be an issue for meeting summarization evaluation.

considered possible. The thus enriched network is then used to evaluate the summarization accuracy of the automatic compaction. This evaluation procedure also allows for weighting of words that are included in the summary by many humans.

The summarization accuracy measure has been found to work well for high compression ratios where most of the original text is preserved, but results in problems for summaries at small ratios such as 10% [62]. In such cases, the authors propose the use of a score based on individual comparisons between the automatic summary and each of the manual summaries, choosing the best score among all the individual comparisons. This idea is very interesting and has not been explored in news summarization: it suggests that rather than using multiple human summaries for weighting, one can find the human summary that is most similar to the produced machine summary.

6.7.3 What About Using ROUGE for Speech Summarization?

The use of a generally agreed upon and automatic metric such as ROUGE is hugely appealing. It allows for cheap evaluation and ease in comparing results from different research efforts. For these reasons, researchers have investigated the degree to which ROUGE scores correlate with human judgments of informativeness of such summaries. In Murray et al. [143], subjective human judgments were collected for summaries of meetings (six test meetings), and compared with several of the popular ROUGE variants. ROUGE scores were not found to correlate with the human judgments on this data. More disturbingly, when Galley [63] compared automatic and human summaries for the same test meetings, ROUGE scores were not able to distinguish between the two types. Both results suggest that the use of ROUGE is not advisable for this type of data especially with so few test points.[8] In a larger study on 30 spoken presentations [62], ROUGE, as well as summarization

[8] The results are also consistent with those reported in [235], indicating that ROUGE scores were not correlated with summary accuracy when evaluating summaries of telephone conversations.

accuracy and F-score, measures were found to highly correlate with human judgments on a five point scale. Such findings suggest that ROUGE should be used only when a large number of test points is available and that its applicability should be tested for new types of data.

7

Conclusions

The vast majority of summarization systems today continue to rely on sentence extraction. While approaches do resemble those that originated in the 1960s, we have seen numerous advances. Many of these advances have been pushed forward because of the establishment of an annual evaluation for summarization, DUC, which later became TAC. Evaluation in DUC has established that, of the various unsupervised methods for determining sentence salience, the use of the log-likelihood ratio test for topic signatures is the best measure of word importance when used in the context of multi-document summarization of news. The availability of previous years' evaluation data for DUC has also made possible the exploration of supervised methods for summarization. Machine learning approaches allowed for automatic experimentation with the importance of different features.

Multi-document summarization is another landmark in the progression of summarization research. It was introduced as a problem in the 1990s when the web became a presence and people began to be overwhelmed by the vast amount of information available on any one topic. Research interest in multi-document summarization further increased when it was established as a task in DUC in 2001. New methods for

sentence extraction also have arisen. They include graph-based methods that identify salience in part by links between sentences, the words they contain or between the concepts within sentences. Researchers have also questioned the traditional approach to sentence extraction in which sentences for the summary are selected one at a time in a greedy fashion and have shown how, alternatively, a system can globally optimize the selection of summary sentences.

Independently of DUC, other major advances within the extractive paradigm were also introduced. One was the introduction of more sophisticated natural language analysis for the extraction of features. The use of discourse to determine the importance of any individual sentence was a signature theme of a number of approaches in single document summarization. By measuring how connected a sentence is to the remainder of the article, whether through coreferential or lexical chains, a system could determine its importance. More sophisticated uses of discourse investigate the detection of rhetorical relations which can serve as guide for extracting salient information. While promising, identification of rhetorical relations is not yet robust enough for summarization to use reliably. The use of semantics and discourse knowledge are clearly areas where more research is needed. This area holds the potential for further improvement of summarization systems.

Following a rise in research on generic single document and multi-document summarization, a concern arose that summarization was being carried out in a vacuum, without taking specific user characteristics and needs into account. Researchers claimed that people usually create summaries in the context of a task, with some information about the needs of the summary reader. The field reacted and a task was created for use in subsequent evaluations: the task of question answering. This move gave rise to query-focused summarization. Query-focused summarization has been carried out naturally in both information retrieval and web-based question answering contexts. It allows for nice synergy between the fields to address the issues. It could, however, use a further push through changes in how the query-focused evaluations are carried out. In the evaluations, participants are usually provided with a set of documents that are exactly on topic to use as input to their summarization system. This allows them to test summarization

independently of errors in information retrieval. In this context, however, it appears that generic summarization approaches perform almost as well — and sometimes, as well — as approaches developed specifically for the query-focused task. A change in task scenario that required summarizers to deal with the realities of irrelevant results from information retrieval could help to push the field forward again in new ways because it would demand researchers to develop new approaches instead of re-using generic approaches.

While sentence extraction still dominates the field, there have been major advances in research on the generation of sentences for a summary. One approach has been to remove information from an extracted sentence that is not needed in the summary. The need for this kind of compression has spawned an entire new research field on compression, sometimes divorced from summarization itself. There is extensive data for research on compression, particularly in the area of headline generation, since news articles are almost always published with headlines, and thus, statistical approaches to compression are the norm. The problem of sentence fusion, or combining information from several document sentences to create a summary sentence, is further behind compression, but has also seen a surge in approaches in just the last few years. Research usually investigates how a system can take repeated phrases from different sentences and combine them in a new sentence, thus forming a sentence that conveys the "intersection" of different sentences. Common approaches involve pair-wise comparison of aligned parse trees, with differences in approaches centering on how the phrases identified in this way are finally combined. In very recent work, researchers have also looked at how to generate sentences that convey the "union" of information found in different sentences and have shown how this approach would be useful for question answering. This is a field that is still wide open for follow up research.

Domain and genre-specific approaches provide some note-worthy lessons. Unlike news summarization, where most of the work has been done, summarization research for specific domains and specific genres often uses more sophisticated approaches and often has access to semantic resources. While these approaches may use sentence extraction, the approaches for finding salient sentences are quite different.

In the medical domain, there are examples where actual generation of sentences from semantics is performed. This is possible because of expectations about the type of information required in a summary, how it is typically expressed and where it typically appears in journal articles. Research on summarization of the web typically relies on massive resources of web page/summary pairs and thus, statistical, supervised approaches are the norm. It also makes use of the structure on the web, exploiting links and information around URLs, to determine summary content. Speech summarization exploits information from the acoustic signal in deciding what information is salient. Both speech and email summarization use information about dialog structure to determine what is important.

The initiation of large-scale, cross-site evaluation, led by NIST, changed the summarization field in major ways. Data for both training and testing was systematically gathered, resulting in sets of input documents paired with one, and, more often, several human summaries. This infusion of data into the community fueled learning-based approaches, allowed comparison of different methods, and probably most importantly, enabled research on evaluation itself. Several new methods for summary evaluation arose, including manual (e.g., Pyramid) and automatic methods (e.g., ROUGE). While there are still many open questions, the emergence of large-scale evaluation overall pushed the field forward at a much faster speed than would otherwise have been possible.

Some problematic issues remain. Given the large number of metrics that are used at the DUC and TAC evaluations, system developers are free to choose and report the metric in follow-on publications that makes their system look best. Furthermore, authors often make their comparisons with the systems that reported results at the evaluation workshop, omitting the fact that DUC/TAC participants developed their systems in a very short time-frame and had no access to the test data, failing to mention that other systems have been developed since the specific evaluation that improve on the original performance. So, it can still be difficult to determine whether a new approach is actually state of the art. The evaluation metrics also clearly need further work. ROUGE, for example, rarely detects significant differences

between systems and, in some cases, shows that a system performs better than or equal to humans. Manual methods show that humans outperform systems (particularly in the single document case), but can require too much effort to use. These problems highlight the need for additional evaluation research.

Where should the field go from here? One obvious additional area would be more intensive use of natural language analysis and generation. Very few systems do much in the way of semantic analysis. Additional semantic analysis would be helpful in determining when a phrase is relevant to a query in query-focused summarization or in determining saliency. Semantic analysis could also help in determining when two phrases convey the same meaning. Semantics in conjunction with generation could be used for better determining how to combine phrases without creating errors in reference.

There is beginning to be research on update summarization. An update summary follows a stream of news over time and the summary attempts to include information on what is new compared to what was reported in previous days. This is a problem that information retrieval researchers tried some 10 years back [1] and concluded was not a feasible task. Recently, it has received attention again as part of the TAC evaluation framework. It proves to be a difficult task because, in contrast to typical multi-document summarization, a system must be able to determine not only what is important but also what is different. Almost anything could be different and yet, not everything should be included in the summary. Only the important differences should be selected. How this is best done remains a topic of current research.

The growth in new types of online information also changes the type of summarization that is of interest. Summarization has recently been combined with work on sentiment analysis [21, 26, 102]. In this context, it is desirable to have summaries which list the pros and cons of a service or product. For example, for a restaurant, one might want to hear that the service was good, but the food bad. Given the many different reviews that one can find on the web, the problem is to identify common opinions. Some of the approaches that have been tried so far include determining semantic properties of an object, determining the intensity of an opinion, or determining whether an opinion is important.

As social networking grows, summaries may be helpful in navigating a network, determining who talks to who, or summarizing activities. Blogs or chat are a new form of media that, like email, share characteristics of both text and speech. They typically involve interaction between participants, often separated in time. There can be multiple responses to the same posting. They often involve informal language and in the case of chat, many abbreviations. New words appear and given their appeal to young people, the language can be quite different from the kind of language that summarizers typically handle, such as newswire. Summarization of this kind of media is just beginning to appear and we expect to see more of this in the near future.

References

[1] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 314–321, 2003.

[2] E. Amitay and C. Paris, "Automatically summarizing web sites - Is there a way around it?," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 173–179, 2000.

[3] R. H. Baayen, *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

[4] B. Baldwin and T. Morton, "Dynamic coreference-based summarization," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1–6, 1998.

[5] M. Banko, V. O. Mittal, and M. J. Witbrock, "Headline generation based on statistical translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 318–325, 2000.

[6] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," *Computational Linguistics*, vol. 34, no. 1, pp. 1–34, 2008.

[7] R. Barzilay and M. Elhadad, "Text summarizations with lexical chains," in *Advances in Automatic Text Summarization*, (I. Mani and M. Maybury, eds.), pp. 111–121, MIT Press, 1999.

[8] R. Barzilay and N. Elhadad, "Sentence alignment for monolingual comparable corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 25–32, 2003.

[9] R. Barzilay, N. Elhadad, and K. McKeown, "Inferring strategies for sentence ordering in multidocument news summarization," *Journal of Artificial Intelligence Research*, vol. 17, pp. 35–55, 2002.

[10] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 113–120, 2004.

[11] R. Barzilay and K. R. McKeown, "Sentence fusion for multidocument news summarization," *Computational Linguistics*, vol. 31, no. 3, pp. 297–328, 2005.

[12] M. Becher, B. Endres-Niggemeyer, and G. Fichtner, "Scenario forms for web information seeking and summarizing in bone marrow transplantation," in *Proceedings of the International Conference on Computational Linguistic*, pp. 1–8, 2002.

[13] A. Berger and V. Mittal, "OCELOT: A system for summarizing web pages.," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 144–151, 2000.

[14] G. Berland, M. Eilliot, L. Morales, J. Algazy, R. Kravitz, M. Broder, D. Kanouse, J. M. noz, J.-A. Puyol, M. Lara, K. Watkins, H. Yang, and E. McGlynn, "Health information on the Internet: Accessibility, quality and readability in English and Spanish," *American Medical Association*, vol. 285, no. 20, pp. 2612–2621, 2001.

[15] F. Biadsy, J. Hirschberg, and E. Filatova, "An Unsupervised Approach to Biography Production Using Wikipedia," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 807–815, 2008.

[16] S. Blair-Goldensohn, K. R. McKeown, and A. H. Schlaikjer, "DefScriber: A hybrid system for definitional QA," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 462–462, 2003.

[17] B. K. Boguraev and M. S. Neff, "Discourse segmentation in aid of document summarization," in *Proceedings of the Hawaii International Conference on System Sciences-Volume 3*, p. 3004, 2000.

[18] B. Boguraev and C. Kennedy, "Salience-based content characterization of text documents," in *Advances in Automatic Text Summarization*, pp. 2–9, The MIT Press, 1997.

[19] D. Bollegala, N. Okazaki, and M. Ishizuka, "A machine learning approach to sentence ordering for multidocument summarization and its evaluation," in *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 624–635, 2005.

[20] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," in *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pp. 385–392, 2006.

[21] S. Branavan, H. Chen, J. Eisenstein, and R. Barzilay, "Learning document-level semantic properties from free-text annotations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 263–271, 2008.

[22] R. Brandow, K. Mitze, and L. F. Rau, "Automatic condensation of electronic publications by sentence selection," *Information Processing and Management*, vol. 31, no. 5, pp. 675–685, 1995.

[23] J. Burstein and D. Marcu, "Toward using text summarization for essay-based feedback," in *Proceedings of Le Conference Annuelle sur Le Traitement Automatique des Langues Naturelles*, pp. 51–59, 2000.

[24] O. Buyukkokten, O. Kaljuvee, H. Garcia-Molina, A. Paepcke, and T. Winograd, "Efficient web browsing on handheld devices Using page and form summarization," in *ACM Transactions on Information Systems*, vol. 20, no. 1, pp. 82–115, January 2002.

[25] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based rerunning for reordering documents and producing summaries," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[26] G. Carenini and J. C. K. Cheung, "Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality," in *Proceedings of the International Natural Language Generation Conference*, pp. 33–41, 2008.

[27] G. Carenini, R. T. Ng, and X. Zhou, "Summarizing email conversations with clue words," in *Proceedings of the International Conference on World Wide Web*, pp. 91–100, 2007.

[28] Y. Chali, S. Hasan, and S. Joty, "Do automatic annotation techniques have any impact on supervised complex question answering?," in *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP*, pp. 329–332, 2009.

[29] Y. Chali and S. Joty, "Improving the performance of the random walk model for answering complex questions," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Short Papers*, pp. 9–12, 2008.

[30] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang, "Using landing pages for sponsored search ad selection," in *Proceedings of the International Conference on World Wide Web*, 2010.

[31] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarization to style-specific summarization for broadcast news," in *Proceedings of the European Conference on IR Research*, 2004.

[32] J. Clarke and M. Lapata, "Models for sentence compression: A comparison across domains, training requirements and evaluation measures," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 377–384, 2006.

[33] J. Clarke and M. Lapata, "Modelling compression with discourse constraints," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1–11, 2007.

[34] A. Cohen, "Unsupervised gene/protein entity normalization using automatically extracted dictionaries," in *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pp. 17–24, Association for Computational Linguistics, 2005.

[35] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," *Journal of Artificial Intelligence Research*, vol. 10, pp. 243–270, 1998.

[36] T. Cohn and M. Lapata, "Sentence compression beyond word deletion," in *Proceedings of the International Conference on Computational Linguistic*, pp. 137–144, 2008.

[37] J. Conroy, J. Schlessinger, D. O'Leary, and J. Goldstein, "Back to basics: CLASSY 2006," in *Proceedings of the Document Understanding Conference*, 2006.

[38] J. Conroy, J. Schlesinger, J. Goldstein, and D. O'Leary, "Left-brain/right-brain multi-document summarization," in *Proceedings of the Document Understanding Conference*, 2004.

[39] J. M. Conroy and D. P. O'Leary, "Text summarization via hidden Markov models," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 406–407, 2001.

[40] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary, "Topic-focused multi-document summarization using an approximate oracle score," in *Proceedings of the International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics*, pp. 152–159, 2006.

[41] T. Copeck and S. Szpakowicz, "Leveraging pyramids," in *Proceedings of the Document Understanding Conference*, 2005.

[42] S. Corston-Oliver, E. Ringger, M. Gamon, and R. Campbell, "Task-focused summarization of email," in *Proceedings of the ACL Text Summarization Branches Out Workshop*, pp. 43–50, 2004.

[43] H. Daume III and D. Marcu, "Generic Sentence Fusion is an Ill-Defined Summarization Task," in *Proceedings of the ACL Text Summarization Branches Out Workshop*, pp. 96–103, 2004.

[44] H. Daumé III and D. Marcu, "A phrase-based HMM approach to document/abstract alignment," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 119–126, 2004.

[45] H. Daumé III and D. Marcu, "Bayesian query-focused summarization," in *Proceedings of the International Conference on Computational Linguistics and the annual meeting of the Association for Computational Linguistics*, pp. 305–312, 2006.

[46] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, pp. 391–407, 1990.

[47] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi, "Enhanced web document summarization using hyperlinks," in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, pp. 208–215, 2003.

[48] R. L. Donaway, K. W. Drummey, and L. A. Mather, "A comparison of rankings produced by summarization evaluation measures," in *Proceedings of the NAACL-ANLP Workshop on Automatic summarization*, pp. 69–78, 2000.

[49] B. Dorr, D. Zajic, and R. Schwartz, "Hedge Trimmer: A parse-and-trim approach to headline generation," in *Proceedings of the HLT-NAACL Workshop on Text Summarization*, pp. 1–8, 2003.

[50] P. A. Duboue, K. R. McKeown, and V. Hatzivassiloglou, "ProGenIE: Biographical descriptions for intelligence analysis," in *Proceedings of NSF/NIJ Symposium on Intelligence and Security Informatics*, vol. 2665, pp. 343–345, Springer-Verlag, 2003.

[51] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1994.

[52] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM*, vol. 16, no. 2, pp. 264–285, 1969.

[53] N. Elhadad, M.-Y. Kan, J. Klavans, and K. McKeown, "Customization in a unified framework for summarizing medical literature," *Journal of Artificial Intelligence in Medicine*, vol. 33, pp. 179–198, 2005.

[54] N. Elhadad, K. McKeown, D. Kaufman, and D. Jordan, "Facilitating physicians' access to information via tailored text summarization," in *Proceedings of the AMIA Annual Symposium*, pp. 226–300, 2005.

[55] G. Erkan and D. Radev, "The University of Michigan at DUC 2004," in *Proceedings of the Document Understanding Conference*, 2004.

[56] G. Erkan and D. R. Radev, "LexRank: Graph-based centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, 2004.

[57] D. Feng and E. Hovy, "Handling biographical questions with implicature," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 596–603, 2005.

[58] E. Filatova and V. Hatzivassiloglou, "A formal model for information selection in multi-sentence text extraction," in *Proceedings of the International Conference on Computational Linguistic*, pp. 397–403, 2004.

[59] K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 177–185, 2008.

[60] M. Fuentes, E. Alfonseca, and H. Rodríguez, "Support Vector Machines for query-focused summarization trained and evaluated on Pyramid data," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Companion Volume: Proceedings of the Demo and Poster Sessions*, pp. 57–60, 2007.

[61] P. Fung and G. Ngai, "One story, one flow: Hidden Markov Story Models for multilingual multidocument summarization," *ACM Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1–16, 2006.

[62] S. Furui, M. Hirohata, Y. Shinnaka, and K. Iwano, "Sentence extraction-based automatic speech summarization and evaluation techniques," in *Proceedings of the Symposium on Large-scale Knowledge Resources*, pp. 33–38, 2005.

[63] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 364–372, 2006.

[64] M. Galley and K. McKeown, "Improving word sense disambiguation in lexical chaining," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1486–1488, 2003.

[65] M. Galley and K. McKeown, "Lexicalized Markov grammars for sentence compression," in *Human Language Technologies: The Conference of the*

North American Chapter of the Association for Computational Linguistics, pp. 180–187, 2007.

[66] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 669–676, 2004.

[67] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4769–4772, 2009.

[68] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2553–2556, 2007.

[69] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 19–25, 2001.

[70] S. Gupta, A. Nenkova, and D. Jurafsky, "Measuring importance and query relevance in topic-focused multi-document summarization," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions*, pp. 193–196, 2007.

[71] I. Gurevych and T. Nahnsen, "Adapting lexical chaining to summarize conversational dialogues," in *Proceedings of the Recent Advances in Natural Language Processing Conference*, pp. 287–300, 2005.

[72] I. Gurevych and M. Strube, "Semantic similarity applied to spoken dialogue Summarization," in *Proceedings of the International Conference on Computational Linguistic*, pp. 764–770, 2004.

[73] B. Hachey, G. Murray, and D. Reitter, "Dimensionality reduction aids term co-occurrence based multi-document summarization," in *SumQA '06: Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pp. 1–7, 2006.

[74] D. Hakkani-Tur and G. Tur, "Statistical sentence extraction for information distillation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. IV–1 –IV–4, 2007.

[75] D. Harman and P. Over, "The effects of human variation in DUC summarization evaluation," in *Proceedings of ACL Text Summarization Branches Out Workshop*, pp. 10–17, 2004.

[76] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M. yen Kan, and K. R. McKeown, "SIMFINDER: A flexible clustering tool for summarization," in *Proceedings of the NAACL Workshop on Automatic Summarization*, pp. 41–49, 2001.

[77] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence-extractive automatic speech summarization and evaluation techniques," *Speech Communication*, vol. 48, no. 9, pp. 1151–1161, 2006.

[78] C. Hori and S. Furui, "Speech summarization: An approach through word extraction and a method for evaluation," *IEICE Transactions on Information and Systems*, vol. 87, pp. 15–25, 2004.

[79] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–9, 2002.

[80] C. Hori, T. Hori, and S. Furui, "Evaluation methods for automatic speech summarization," in *Proceedings of the European Conference on Speech Communication and Technology*, pp. 2825–2828, 2003.

[81] E. Hovy and C.-Y. Lin, "Automated Text Summarization in SUMMARIST," in *Advances in Automatic Text Summarization*, pp. 82–94, 1999.

[82] G. Hripcsak, J. Cimino, and S. Sengupta, "WebCIS: Large scale deployment of a web-based clinical information system," in *Proceedings of the AMIA Annual Symposium*, pp. 804–808, 1999.

[83] M. Hu, A. Sun, and E.-P. Lim, "Comments-oriented blog summarization by sentence extraction," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 901–904, 2007.

[84] D. Ji and Y. Nie, "Sentence ordering based on cluster adjacency in multi-document summarization," in *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 745–750, 2008.

[85] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad, "Summarization evaluation methods: Experiments and analysis," in *AAAI Symposium on Intelligent Summarization*, pp. 51–59, 1998.

[86] H. Jing and K. McKeown, "The Decomposition of Human-Written Summary Sentences," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 129–136, 1999.

[87] H. Jing, "Sentence reduction for automatic text summarization," in *Proceedings of the Conference on Applied Natural Language Processing*, pp. 310–315, 2000.

[88] H. Jing, "Using Hidden Markov modeling to decompose human-written summaries," *Computational linguistics*, vol. 28, no. 4, pp. 527–543, 2002.

[89] H. Jing and K. R. McKeown, "Cut and paste based text summarization," in *Proceedings of the North American chapter of the Association for Computational Linguistics Conference*, pp. 178–185, 2000.

[90] H. Jing, *Cut-and-paste text summarization*. PhD thesis, Columbia University, 2001.

[91] M. Johnson and E. Charniak, "A TAG-based noisy-channel model of speech repairs," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 33–39, 2004.

[92] M.-Y. Kan, "Combining visual layout and lexical cohesion features for text segmentation," Tech. Rep. CUCS-002-01, Columbia University, 2001.

[93] T. Kikuchi, S. Furui, and C. Hori, "Automatic speech summarization based on sentence extraction and compaction," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 384–387, 2003.

[94] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artiitial Intelligence*, vol. 139, no. 1, pp. 91–107, 2002.

[95] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, vol. 2, no. 1, pp. 1–24, 2005.

[96] E. Krahmer, E. Marsi, and P. van Pelt, "Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 193–196, 2008.

[97] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 68–73, 1995.

[98] F. Lacatusu, A. Hickl, S. Harabagiu, and L. Nezda, "Lite_GISTexter at DUC2004," in *Proceedings of the Document Understanding Conference*, 2004.

[99] D. Lam, S. L. Rohall, C. Schmandt, and M. K. Stern, "Exploiting e-mail structure to improve summarization," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2002.

[100] A. M. Lam-Adesina and G. J. F. Jones, "Applying summarization techniques for term selection in relevance feedback," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1–9, 2001.

[101] M. Lapata and R. Barzilay, "Automatic evaluation of text coherence: models and representations," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1085–1090, 2005.

[102] K. Lerman, S. Blair-Goldensohn, and R. McDonald, "Sentiment summarization: evaluating and learning user preferences," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 514–522, 2009.

[103] J. Leskovec, N. Milic-frayling, and M. Grobelnik, "Impact of linguistic analysis on the semantic graph coverage and learning of document extracts," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 1069–1074, 2005.

[104] C. Lin and E. Hovy, "From single to multi-document summarization: A prototype system and its evaluation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 457–464, 2002.

[105] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of ACL Text Summarization Branches Out Workshop*, pp. 74–81, 2004.

[106] C.-Y. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the International Conference on Computational Linguistic*, pp. 495–501, 2000.

[107] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurance statistics," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003.

[108] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1526–1540, 2006.

[109] Y. Liu and S. Xie, "Impact of automatic sentence segmentation on meeting summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5009–5012, 2008.

[110] A. Louis, A. Joshi, and A. Nenkova, "Discourse indicators for content selection in summarization," in *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 147–156, 2010.

[111] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[112] M. Mana-López, M. D. Buenaga, and J. M. Gómez-Hidalgo, "Multidocument summarization: An added value to clustering in interactive retrieval," *ACM Transactions on Informations Systems*, vol. 22, no. 2, pp. 215–241, 2004.

[113] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "SUMMAC: A text summarization evaluation," *Natural Language Engineering*, vol. 8, no. 1, pp. 43–68, 2002.

[114] I. Mani, *Automatic Summarization*. John Benjamins, 2001.

[115] I. Mani and E. Bloedorn, "Summarizing similarities and differences among related documents," *Information Retrieval*, vol. 1, no. 1-2, pp. 35–67, April 1999.

[116] I. Mani, B. Gates, and E. Bloedorn, "Improving summaries by revising them," in *Proceedings of the annual meeting of the Association for Computational Linguistics*, pp. 558–565, 1999.

[117] W. Mann and S. Thompson, "Rhetorical Structure Theory: Towards a functional theory of text organization," *Text*, vol. 8, pp. 243–281, 1988.

[118] C. D. Manning and H. Schutze, *Foundations of Natural Language Processing*. MIT Press, 1999.

[119] D. Marcu, "From discourse structure to text summaries," in *Proceedings of ACL/EACL 1997 Summarization Workshop*, pp. 82–88, 1997.

[120] D. Marcu, "To build text summaries of high quality, nuclearity is not sufficient," in *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pp. 1–8, 1998.

[121] D. Marcu, "The automatic construction of large-scale corpora for summarization research," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 137–144, 1999.

[122] D. Marcu, *The Theory and Practice of Discourse and Summarization*. The MIT Press, 2000.

[123] E. Marsi and E. Krahmer, "Explorations in sentence fusion," in *Proceedings of the European Workshop on Natural Language Generation 2005*, pp. 109–117, 2005.

[124] S. Maskey and J. Hirschberg, "Summarizing speech without text using Hidden Markov Models," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 89–92, 2006.

[125] S. Maskey, A. Rosenberg, and J. Hirschberg, "Intonational phrases for speech summarization," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2430–2433, 2008.

[126] R. McDonald, "Discriminative sentence compression with soft syntactic evidence," in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 297–304, 2006.

[127] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proceedings of the European Conference on IR Research*, pp. 557–564, 2007.

[128] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, B. Schiffman, and S. Teufel, "Columbia multi-document summarization: Approach and evaluation," in *Proceedings of the Document Understanding Conference*, 2001.

[129] K. McKeown, S.-F. Chang, J. Cimino, S. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Johnson, D. Jordan, J. Klavans, A. Kushniruk, V. Patel, and S. Teufel, "PERSIVAL, a system for personalized search and summarization over multimedia healthcare information," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 331–340, 2001.

[130] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman, "Tracking and summarizing news on a daily basis with Columbia's Newsblaster," in *Proceedings of the International Conference on Human Language Technology Research*, 2002.

[131] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg, "Do summaries help? A task-based evaluation of multi-document summarization," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 210–217, 2005.

[132] K. Mckeown, L. Shrestha, and O. Rambow, "Using question-answer pairs in extractive summarization of email conversations," in *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 542–550, 2007.

[133] K. R. McKeown, J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, "Towards multidocument summarization by reformulation: progress and prospects," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 453–460, 1999.

[134] Q. Mei and C. Zhai, "Generating impact-based summaries for scientific literature," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 816–824, 2008.

[135] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, 2004.

[136] R. Mihalcea and P. Tarau, "An algorithm for language independent single and multiple document summarization," in *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 19–24, 2005.

[137] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography (special issue)*, vol. 3, no. 4, pp. 235–312, 1990.

[138] T. Miller and W. Schuler, "A syntactic time-series model for parsing fluent and disfluent speech," in *Proceedings of the International Conference on Computational Linguistic*, pp. 569–576, 2008.

[139] J. Mrozinski, E. W. D. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 981–984, 2006.

[140] H. Murakoshi, A. Shimazu, and K. Ochimizu, "Construction of deliberation structure in email conversation," in *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pp. 570–577, 2004.

[141] S. Muresan, E. Tzoukermann, and J. L. Klavans, "Combining linguistic and machine learning techniques for email summarization," in *Proceedings of the Workshop on Computational Natural Language Learning*, pp. 1–8, 2001.

[142] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proceedings of 9th European Conference on Speech Communication and Technology*, pp. 593–596, 2005.

[143] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proceedings of the ACL Workshop on Evaluation Measures for MT/Summarization*, 2005.

[144] G. Murray and S. Renals, "Term-weighting for summarization of multi-party spoken dialogues," in *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, vol. 4892, pp. 155–166, 2007.

[145] H. Nanba and M. Okumura, "Towards multi-paper summarization using reference information," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 926–931, 1999.

[146] H. Nanba and M. Okumura, "Producing more readable extracts by revising them," in *Proceedings of the International Conference on Computational Linguistic*, pp. 1071–1075, 2000.

[147] A. Nenkova and K. McKeown, "References to named entities: A corpus study," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 70–72, 2003.

[148] A. Nenkova, "Automatic text summarization of newswire: lessons learned from the document understanding conference," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 1436–1441, 2005.

[149] A. Nenkova, "Entity-driven rewrite for multi-document summarization," in *Proceedings of the International Joint Conference on Natural Language Processing*, 2008.

[150] A. Nenkova, *Understanding the process of multi-document summarization: content selection, rewrite and evaluation.* PhD thesis, Columbia University, January 2006.

[151] A. Nenkova and A. Bagga, "Facilitating email thread access by extractive summary generation," in *Proceedings of the Recent Advances in Natural Language Processing Conference*, 2003.

[152] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 145–152, 2004.

[153] A. Nenkova, R. Passonneau, and K. McKeown, "The Pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 2, 2007.

[154] A. Nenkova, A. Siddharthan, and K. McKeown, "Automatically learning cognitive status for multi-document summarization of newswire," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 241–248, 2005.

[155] A. Nenkova, L. Vanderwende, and K. McKeown, "A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 573–580, 2006.

[156] P. S. Newman and J. C. Blitzer, "Summarizing archived discussions: A beginning," in *Proceedings of the International Conference on Intelligent user Interfaces*, pp. 273–276, 2003.

[157] M. L. Nguyen, A. Shmazu, S. Horiguchi, T. B. Ho, and M. Fukushi, "Probabilistic sentence reduction using support vector machines," in *Proceedings of the International Conference on Computational Linguistic*, pp. 743–49, 2004.

[158] K. Ohtake, K. Yamamoto, Y. Toma, S. Sado, S. Masuyama, and S. Nakagawa, "Newscast Speech Summarization via Sentence Shortening Based on Prosodic Features," in *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 247–254, 2003.

[159] K. Ono, K. Sumita, and S. Miike, "Abstract generation based on rhetorical structure extraction," in *Proceedings of the International Conference on Computational Linguistic*, pp. 344–348, 1994.

[160] M. Osborne, "Using maximum entropy for sentence extraction," in *Proceedings of the ACL Workshop on Automatic Summarization*, pp. 1–8, 2002.

[161] J. Otterbacher, G. Erkan, and D. Radev, "Using random walks for question-focused sentence retrieval," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922, 2005.

[162] J. Otterbacher, D. Radev, and A. Luo, "Revisions that improve cohesion in multi-document summaries: A preliminary study," in *Proceedings of the Workshop on Automatic Summarization*, pp. 2–7, 2002.

[163] P. Over, H. Dang, and D. Harman, "DUC in context," *Information Processing and Managemant*, vol. 43, no. 6, pp. 1506–1520, 2007.

[164] C. D. Paice, "The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 172–191, 1981.

[165] C. D. Paice, "Constructing literature abstracts by computer: Techniques and prospects," *Information Processing and Management*, vol. 26, no. 1, pp. 171–186, 1990.

[166] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[167] F. Peng, R. Weischedel, A. Licuanan, and J. Xu, "Combining deep linguistics analysis and surface pattern learning: A hybrid approach to Chinese definitional question answering," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 307–314, 2005.

[168] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 470–478, 2008.

[169] V. Qazvinian and D. R. Radev, "Scientific paper summarization using citation summary networks," in *Proceedings of the International Conference on Computational Linguistic*, pp. 689–696, 2008.

[170] D. Radev and K. McKeown, "Generating natural language summaries from multiple on-line sources," *Computational Linguistics*, vol. 24, no. 3, pp. 469–500, 1998.

[171] D. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn, "NewsInEssence: Summarizing online news topics," *Communications of the ACM*, vol. 48, no. 10, pp. 95–98, 2005.

[172] D. Radev and D. Tam, "Single-document and multi-document summary evaluation via relative utility," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 508–511, 2003.

[173] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, and R. S. Raghavan, "NewsInEssence: A system for domain-independent, real-time news clustering and multi-document summarization," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 1–4, 2001.

[174] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies," in *Proceedings of the NAACL-ANLP Workshop on Automatic summarization*, pp. 21–30, 2000.

[175] D. R. Radev and K. R. McKeown, "Building a generation knowledge source using internet-accessible newswire," in *Proceedings of the Conference on Applied Natural Language Processing*, pp. 221–228, 1997.

[176] D. R. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, H. Qi, A. Çelebi, D. Liu, and E. Drabek, "Evaluation challenges in large-scale document summarization: The MEAD project," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 375–382, 2003.

[177] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen, "Summarizing email threads," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2004.

[178] G. J. Rath, A. Resnick, and R. Savage, "The formation of abstracts by the selection of sentences: Part 1: Sentence selection by man and machines," *American Documentation*, vol. 2, no. 12, pp. 139–208, 1961.

[179] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tur, "Packing the meeting summarization knapsack," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 2434–2437, 2008.

[180] S. Riezler, T. H. King, R. Crouch, and A. Zaenen, "Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for

lexical-functional grammar," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 118–125, 2003.

[181] D. G. Roussinov and H. Chen, "Information navigation on the web by clustering and summarizing query results," *Information Processing and Management*, vol. 37, no. 6, pp. 789–816, 2001.

[182] T. Sakai and K. Sparck Jones, "Generic summaries for indexing in information retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 190–198, 2001.

[183] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Information Processing and Management*, vol. 33, no. 2, pp. 193–208, 1997.

[184] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, pp. 513–523, 1988.

[185] B. Schiffman, I. Mani, and K. Concepcion, "Producing biographical summaries: Combining linguistic knowledge with corpus statistics," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 458–465, 2001.

[186] B. Schiffman, A. Nenkova, and K. McKeown, "Experiments in multidocument summarization," in *Proceedings of the International Conference on Human Language Technology Research*, pp. 52–58, 2002.

[187] L. Shrestha and K. Mckeown, "Detection of question-answer pairs in email conversations," in *Proceedings of the International Conference on Computational Linguistic*, pp. 889–895, 2004.

[188] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[189] A. Siddharthan and K. McKeown, "Improving multilingual summarization: using redundancy in the input to correct MT errors," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 33–40, 2005.

[190] A. Siddharthan, A. Nenkova, and K. Mckeown, "Syntactic simplification for improving content selection in multi-document summarization," in *Proceedings of the International Conference on Computational Linguistic*, pp. 896–902, 2004.

[191] A. Siddharthan and S. Teufel, "Whose idea was this, and why does it matter? Attributing scientific work to citations," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 316–323, 2007.

[192] H. G. Silber and K. F. McCoy, "Efficiently computed lexical chains as an intermediate representation for automatic text summarization," *Computational Linguistics*, vol. 28, no. 4, pp. 487–496, 2002.

[193] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[194] K. Sparck Jones, "Automatic summarizing: factors and directions," in *Advances in Automatic Text Summarization*, pp. 1–12, MIT Press, 1998.

[195] J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Jeek, "Two uses of anaphora resolution in summarization," *Information Processing and Management*, vol. 43, no. 6, pp. 1663–1680, 2007.

[196] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 405–408, 1996.

[197] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen, "Web-page summarization using clickthrough data," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 194–201, 2005.

[198] W. tau Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1776–1782, 2007.

[199] S. Teufel, "Task-based evaluation of summary quality: describing relationships between scientific papers," in *Proceedings of the NAACL Workshop on Automatic Summarization*, pp. 12–21, 2001.

[200] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational Linguisics.*, vol. 28, no. 4, pp. 409–445, 2002.

[201] S. Teufel and H. van Halteren, "Evaluating information content by factoid analysis: Human annotation and stability," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[202] D. R. Timothy, T. Allison, S. Blair-goldensohn, J. Blitzer, A. elebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, A. Winkel, and Z. Zhang, "MEAD — a platform for multidocument multilingual text summarization," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2004.

[203] A. Tombros and M. Sanderson, "Advantages of query biased summaries in information retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[204] S. Tucker, N. Kyprianou, and S. Whittaker, "Time-compressing speech: ASR transcripts are an effective way to support gist extraction," in *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, pp. 226–235, 2008.

[205] S. Tucker and S. Whittaker, "Temporal compression of speech: An evaluation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 790–796, 2008.

[206] J. Turner and E. Charniak, "Supervised and unsupervised learning for sentence compression," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (Ann Arbor, Mi.), pp. 290–297, June 2005.

[207] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams, "Fast generation of result snippets in web search," in *Proceedings of the Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 127–134, 2007.

[208] E. Tzoukermann, S. Muresan, and J. L. Klavans, "GIST-IT: Summarizing email using linguistic knowledge and machine learning," in *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, pp. 1–8, 2001.

[209] J. Ulrich, G. Murray, and G. Carenini, "A publicly available annotated corpus for supervised email summarization," in *Proceedings of the AAAI EMAIL Workshop*, pp. 77–87, 2008.

[210] UMLS, "UMLS Knowledge Sources," National Library of Medicine, Bethesda, Maryland, 9th edition, 1998.

[211] H. van Halteren and S. Teufel, "Examining the consensus between human summaries: Initial experiments with factoid analysis," in *Proceedings of the HLT-NAACL Workshop on Automatic Summarization*, 2003.

[212] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing and Managment*, vol. 43, pp. 1606–1618, 2007.

[213] R. Varadarajan and V. Hristidis, "A system for query-specific document summarization," in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2006.

[214] A. Waibel, M. Bett, and M. Finke, "Meeting browser: Tracking and summarizing meetings," in *Proceedings of the DARPA Broadcast News Workshop*, pp. 281–286, 1998.

[215] S. Wan and K. McKeown, "Generating state-of-affairs summaries of ongoing email thread discussions," in *Proceedings of the International Conference on Computational Linguistic*, 2004.

[216] X. Wan and J. Yang, "Improved affinity graph based multi-document summarization," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 181–184, 2006.

[217] R. Weischedel, J. Xu, and A. Licuanan, "A hybrid approach to answering biographical questions," in *New Directions In Question Answering*, (M. Maybury, ed.), pp. 59–70, 2004.

[218] M. J. Witbrock and V. O. Mittal, "Ultra-summarization (poster abstract): A statistical approach to generating highly condensed non-extractive summaries," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 315–316, 1999.

[219] F. Wolf and E. Gibson, "Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 383–390, 2004.

[220] K.-F. Wong and W. Wu, Mingli sand Li, "Extractive summarization using Supervised and semi-supervised learning," in *Proceedings of the International Conference on Computational Linguistic*, pp. 985–992, 2008.

[221] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4985–4988, 2008.

[222] E. Yamangil and S. M. Shieber, "Bayesian synchronous tree-substitution grammar induction and its application to sentence compression," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 937–947, 2010.

[223] J. Yang, A. Cohen, and W. Hersh, "Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts," in *Proceedings of the AMIA Annual Symposium*, pp. 831–835, 2007.

[224] S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," *Information Processing and Management*, vol. 43, no. 6, pp. 1643–1662, 2007.

[225] W. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1776–1782, 2007.

[226] D. Zajic, B. J. Dorr, J. Lin, and R. Schwartz, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks," *Information Processing and Management*, vol. 43, no. 6, pp. 1549–1570, 2007.

[227] K. Zechner, "Summarization of spoken language - challenges, methods, and prospects," *Speech Technology Expert eZine*, January 2002.

[228] K. Zechner and A. Lavie, "Increasing the coherence of spoken dialogue summaries by cross-speaker information linking," in *Proceedings of the NAACL Workshop on Automatic Summarization*, pp. 22–31, 2001.

[229] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proceedings of the North American chapter of the Association for Computational Linguistics Conference*, pp. 186–193, 2000.

[230] J. Zhang, H. Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 195–200, 2007.

[231] J. Zhang and P. Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 213–216, 2007.

[232] L. Zhou and E. Hovy, "A web-trained extraction summarization system," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 205–211, 2003.

[233] L. Zhou and E. Hovy, "On the summarization of dynamically introduced information: Online discussions and blogs," in *Proceedings of AAAI Spring Symposium 2006*, 2006.

[234] L. Zhou, M. Ticrea, and E. Hovy, "Multi-document biography summarization," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 434–441, 2004.

[235] X. Zhu and G. Penn, "Evaluation of sentence selection for speech summarization," in *Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research*, pp. 39–45, 2005.

[236] X. Zhu and G. Penn, "Summarization of spontaneous conversations," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1531–1534, 2006.