Knowl Inf Syst
DOI 10.1007/s10115-017-1042-4

SURVEY PAPER

Recent advances in document summarization

Jin-ge Yao[1,2] · **Xiaojun Wan**[1,2] · **Jianguo Xiao**[1,2]

Received: 27 October 2016 / Accepted: 17 March 2017

Abstract The task of automatic document summarization aims at generating short summaries for originally long documents. A good summary should cover the most important information of the original document or a cluster of documents, while being coherent, non-redundant and grammatically readable. Numerous approaches for automatic summarization have been developed to date. In this paper we give a self-contained, broad overview of recent progress made for document summarization within the last 5 years. Specifically, we emphasize on significant contributions made in recent years that represent the state-of-the-art of document summarization, including progress on modern sentence extraction approaches that improve concept coverage, information diversity and content coherence, as well as attempts from summarization frameworks that integrate sentence compression, and more abstractive systems that are able to produce completely new sentences. In addition, we review progress made for document summarization in domains, genres and applications that are different from traditional settings. We also point out some of the latest trends and highlight a few possible future directions.

Keywords Document summarization · Natural language generation · Natural language processing · Text mining

✉ Xiaojun Wan
wanxiaojun@pku.edu.cn

Jin-ge Yao
yaojinge@pku.edu.cn

Jianguo Xiao
xiaojianguo@pku.edu.cn

[1] Institute of Computer Science and Technology, Peking University, Beijing 100871, China

[2] The MOE Key Laboratory of Computational Linguistics, Peking University, Beijing, China

Published online: 28 March 2017

1 Introduction

With the rapid growth of information in the new era, people can obtain and share information almost instantly from a wide array of sources. The Web contains billions of documents and is growing at an exponential pace. As a result, we are facing an inevitable and challenging problem of information overload. Tools that provide timely access to, and digest of, various sources are necessary in order to alleviate the problem. Search engines have enabled users to retrieve information from digital collections by providing a ranked list of documents or web pages, given a user-specified query. However, even the most sophisticated search engines empowered by advanced information retrieval techniques lack the ability to synthesize information from multiple sources and present users with a concise yet informative response. 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.

Summarization systems are designed to take a single document or a cluster of documents as input and produce a concise and fluent summary conveying the most important information. Recent years have seen the development of numerous summarization tasks, approaches and applications. Such systems, imperfect as they are, have already been shown to help users and to enhance other automatic applications and interfaces. In the year 2013, Yahoo acquired the trendy and decidedly stylish news summarization app called Summly,[1] with an audacious bid of $30 million [176].

There are several distinctions typically made in summarization and here we introduce terminology that is often mentioned in the literature [139].

Extractive summarization produces summaries by concatenating several sentences taken exactly as they appear in the original documents being summarized. By contrast, *abstractive summarization* uses different words to describe the contents of the original documents rather than directly copying original sentences.

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, *multi-document summarization* emerged and applied to clusters of news articles on the same event, aiming at producing a one-paragraph short summary.

Much of the work to date has been in the context of *generic summarization*, making few assumptions about the audience or the goal for generating the summary. 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.

1.1 Outline and scope

The field of document summarization has moved forward in various aspects during recent years. Many papers have been published that focus on different aspects of document summarization systems. Given that there already exist a number of earlier survey papers [39,137,139] that provide comprehensive view of the field of document summarization, in this paper we are trying to give an overview of the most *important recent* progress that has been made within last 5 years. Significant progress has been made recently from traditional extractive summarization to more abstractive summarization, along with many more new interesting task settings and applications, but none of them has been covered or properly introduced in

[1] www.summly.com.

any previous survey paper. There exist very few similar attempts (such as Gambhir and Gupta, [54]) that unfortunately fail to cover the most significant study trends or are in shortage of clear organization of content, which partly motivates us for writing this survey for recent studies. We aim at a self-contained[2] description of the latest research progress for document summarization made roughly from 2011, as a solid complement of previous comprehensive reviews [39,137,139] written earlier than that.

As background information, we first briefly introduce classic approaches and paradigms, pointing out some key factors for the task (Sect. 1.2). Then we carefully review recent progress made on various important aspects. Section 2.2 describes the massive efforts made in the scope of extractive summarization. Due to the obvious limitations of sentence extraction, researchers have made many attempts to shift toward abstractive summarization, of which sentence compression plays an important role. As an intermediate step, compressive summarization that integrates sentence compression and extraction has aroused much attention, providing better coverage while almost retaining readability of original sentences. In Sect. 2.3.1 we describe recent work on compressive summarization. After that, we introduce more abstractive approaches that involve more operations other than compression in Sect. 2.3.2. Part of recent research also focuses on specific genres or applications beyond summarizing generic news documents. We give a brief overview of related progress in Sect. 2.4. We highlight some frontier research trends and discuss our perspectives on possible future directions in Sect. 3 and then conclude the paper.

The boundaries of research scope of different research papers are vague in terms of different taxonomies. As a result, we avoid rigorous categorization of approaches, but organize our descriptions according to the main significant streams of research progress (readers who prefer a tabular illustration may refer to Table 1 in the next section). Also, just as in previous survey papers in this area, we do not give quantitative comparisons for most methods, since: (1) most approaches may not be directly comparable as they evaluate on different subsets of standard benchmark datasets (especially for single-document summarization) while reporting results in subtly different ways as well, and (2) the commonly used automatic evaluation metrics are rather limited, with manual evaluation still being indispensable in standard shared task evaluation and in the most solid research studies.

1.2 Earlier Research and Classic Approaches

Since this paper mainly focuses on more recent advances and methodologies in document summarization, we only give a very brief overview of classic approaches to make this paper self-contained for reading, without providing a complete coverage for them. For more detailed descriptions of classic work, one may refer to earlier comprehensive survey papers [137,139].

Earlier research in the last decade is dominated by extractive summarization approaches, with a few of them also including other sentence-level operations such as sentence compression or reordering as a post-processing step after sentence extraction. The most typical frameworks can be roughly described with three key components:

- *Sentence scoring*: Each sentence is assigned a score which indicates its importance. Summarization aims at preserving the most important information via extracting the most important sentences.

[2] However, readers are still assumed to have some basic knowledge in natural language processing and text mining in general.

- *Sentence selection*: The summarizer has to select the best combination of important sentences to form a summary with paragraph length. Many global factors such as content coherence and redundancy in description must be considered in this part.
- *Sentence reformulation*: Sometimes sentences extracted from the original documents should be modified or paraphrased, in order to produce more clear, more coherent and more concise summaries.

The distinction of these components is sometimes vague, as some of them are implicitly considered or integrated in other modules.

In this subsection we briefly describe earlier approaches for these components and then slightly touch the common ways to evaluate summarization.

1.2.1 Sentence Scoring

Sentence scoring scheme is crucial for the summarization system to decide which sentences are more important and tend to be selected as summary sentences.

Earlier unsupervised approaches mostly rely on *frequency* and *centrality*.

Specifically, the assumption behind frequency-driven approaches is that the most important information will appear more frequently in the documents than less important detailed descriptions. For example, the earlier probabilistic system SUMBASIC [182] was driven by word probability estimation, assigning each sentence a weight equal to the average probability of the content words in the sentence. More powerful usages include log-likelihood ratio test for identifying topic signature words that are highly descriptive of the input documents [114]. In earlier coverage-based models, the concepts or word bigrams that are considered important are those with high document frequency [62].

Meanwhile, sentences which are more similar to other sentences are considered to be central, assumed to be carrying the most central ideas of the original documents. This assumption forms the basis of graph-based summarization frameworks, typically adapted from link analysis algorithms in network analysis. Both TextRank [133] and LexRank [47] run the PageRank algorithm in a weighted graph of words or sentences, with edge weights defined using literal or more semantic-driven similarities. In centroid-based summarization [157], a pseudo-sentence of the document called centroid is constructed, consisting of words with tf-idf[3] scores above a predefined threshold. The score of each sentence is defined by summing the scores based on different features including cosine similarity of the sentence with the centroid.

Probabilistic topic models based on co-occurrences have also been exploited in summarization. For example, the HIERSUM model [68] is presented based on hierarchical Latent Dirichlet Allocation (hLDA) to represent content specificity as a hierarchy of topic vocabulary distributions. A later work [21] also utilizes a hLDA-style model to devise a sentence-level probabilistic topic model and a hybrid learning algorithm for extracting salient features of sentences.

All these approaches have in common that they focus on selecting the most repeated information from a document. However, in noisy documents with significant amounts of redundant, unimportant texts, extracting the most central or most frequent parts may not be a good strategy.

[3] The *tf-idf* weighting scheme is a well-known concept in information retrieval that uses the *term frequency* (tf) in the document for each term and a complementary weight for each term which penalizes terms found in many documents in the collection by using the *inverse document frequency* (idf), i.e., the inverse of the number of documents that contain the term, as weights.

To date, various machine learning methods have been developed for extractive summarization by learning to extract sentences. Given sentences with labeled importance scores, it is straightforward to train regression models for importance prediction [53,72,209] or learning to rank models to train a model that is capable to assign high rank for the most important sentences [132,169,190]. To model possible inter-sentence dependency rather than predicting the important score for each sentence individually, document summarization can also be treated as a sequence labeling problem, with latent labels indicating whether to extract the sentence into the summary or not. As a result, hidden Markov models [37], conditional random fields [170] and structural SVMs [109] have all been applied in such settings. All these systems extract indicative features including sentence position, named entities, similarity or distance to query and content word frequency.

Supervised approaches rely on labeled training data. A typical way to construct labeled data for training is to set ROUGE (cf. Sect. 1.2.4), the most commonly used automatic evaluation metric, or its variants or approximations as prediction target for sentence scoring. This treatment is intuitive and has become more theoretically justified in a very recent study [153].

For query-focused summarization, the query information is typically considered via computations of similarity or overlap between each sentence and the query. These values can be either used in similarity-based approaches or act as features for importance prediction [209]. Supervised approaches have achieved more significant improvements for sentence scoring in query-focused settings as well due to better capturing the dependence with query terms [190].

1.2.2 Sentence Selection

Having predicted sentence importance scores, the most straightforward follow-up step is to directly select sentences that ranked at the top. However, for document summarization, especially multi-document scenarios, redundancy removal is a key issue. A good summary should never contain repeated descriptions for the same piece of information, even though the relevant sentences have all been treated as important ones.

One of the most popular approaches for sentence selection is *maximum marginal relevance* (MMR) [19]. It defines an objective function gain of adding text unit (e.g., sentence) k to set $S (k \notin S)$ as:

$$\lambda Sim_1(s_k, q) - (1 - \lambda) \max_{i \in S} Sim_2(s_i, s_k) \tag{1}$$

where $Sim_1(s_k, q)$ measures the similarity between unit s_k and a query q, while $Sim_2(s_i, s_k)$ measures the similarity between unit s_i and unit s_k , and $\lambda \in [0, 1]$ is a trade-off coefficient.

For probabilistic approaches [68,182], sentences are typically selected with the goal to minimize the Kullback–Leibler (KL) divergence between the probability distribution of words estimated from the summary and that from the input. Solving for the summary with the smallest KL divergence is computationally intractable, so greedy selection is often used.

Meanwhile, sentence scoring and selection can be modeled (sometimes implicitly) in the same framework, formulated as global optimization [62,131] rather than greedily adding sentences to form a summary. The most widely used practice is to formulate the problem as integer linear programming (ILP). The objective is usually to maximize coverage with constraints introduced to ensure the consistency between the selection of sentences and sub-sentential units, along with a knapsack constraint to limit the total length of the output summary. For example, in concept-based ILP for summarization [62], the goal is to maximize the sum of the weights of the *concepts* (usually implemented as bigrams) that appear in the summary.

The association between the concepts and sentences serves as the constraints. This ILP framework is formally represented as below:

$$\max \quad \textstyle\sum_i w_i c_i \tag{2}$$

$$\text{s.t.} \quad s_j o_{ij} \leq c_i, \tag{3}$$

$$\textstyle\sum_j s_j o_{ij} \geq c_i, \tag{4}$$

$$\textstyle\sum_j l_j s_j \leq L, \tag{5}$$

$$c_i \in \{0, 1\}, \forall i, \tag{6}$$

$$s_j \in \{0, 1\}, \forall j, \tag{7}$$

where c_i and s_j are binary variables that indicate the presence of a concept and a sentence, respectively, w_i is the weight for concept i, and o_{ij} means the occurrence of concept i in sentence j. The inequality constraints ensure consistencies that selecting a sentence leads to the selection of all the concepts it contain, and selecting a concept only happens when it is present in at least one of the selected sentences.

1.2.3 Sentence Reformulation and Ordering

Most of the earlier systems extract sentences and just leave them as they are. Systems targeting more practical usages also include additional operations as an additional step following sentence selection.

Sentences extracted from original documents usually contain unnecessary or redundant information, which makes them less suitable to be directly used as summary sentences. A popular solution is to pipeline sentence extraction and rule-based compression. More sophisticated operations may also be used to enhance compactness and informativeness, such as paraphrasing and sentence fusion [9]. Due to the immatureness of current natural language generation techniques, some of these operations may hurt readability of the final summary. As a result, very few progress in terms of sentence rewriting has been made in fully abstractive summarization in earlier work.

Meanwhile, the order in which information is presented to the reader critically influences the quality of a summary. In a single document, summary information can be presented by preserving the order in the original document [157]. However, extracted sentences do not always retain their precedence orders in manually written summaries. Reordering is a more significant issue for multi-document summarization as summary sentences are from multiple unaligned sources. Classic reordering approaches include inferring order from weighted sentence graph [36] or perform a chronological ordering algorithm [8] that sorts sentences based on timestamp and position.

1.2.4 Evaluation

A good summary must be easy to read and give a good overview of the content of the source text. Manual evaluation for document summarization is time-consuming and difficult; hence, a series of proposals have been made to partially or fully automate the evaluation. Currently the ROUGE (Recall-Oriented Understudy for Gisty Evaluation) metrics [115] are the de facto standard for automatic evaluation of summarization. The ROUGE metrics are based on the comparison of n-grams between the summary to be evaluated and one or several human-written reference summaries. There are several variants of ROUGE, including ROUGE-*n* (n-grams), ROUGE-L (the longest common sequence) and ROUGE-SU (skip-bigrams and uni-grams). For example, the most commonly used ROUGE-N is an n-gram-based metric with the recall-oriented score, the precision-oriented score and the F-measure score for ROUGE-N computed, respectively, as follows:

$$\text{ROUGE-N}_{\text{recall}} = \frac{\sum\limits_{s \in \text{ref\_sum}} \sum\limits_{\text{Ngram} \in S} Count_{\text{match}}(Ngrams)}{\sum\limits_{s \in \text{ref\_sum}} \sum\limits_{\text{Ngram} \in S} Count(Ngrams)} \tag{8}$$

$$\text{ROUGE-N}_{\text{precision}} = \frac{\sum\limits_{s \in \text{ref\_sum}} \sum\limits_{\text{Ngram} \in S} Count_{\text{match}}(Ngrams)}{\sum\limits_{s \in \text{cand\_sum}} \sum\limits_{\text{Ngram} \in S} Count(Ngrams)} \tag{9}$$

$$\text{ROUGE-N}_{\text{F-score}} = \frac{2 \times \text{ROUGE-N}_{\text{recall}} \times \text{ROUGE-N}_{\text{precision}}}{\text{ROUGE-N}_{\text{precision}} + \text{ROUGE-N}_{\text{recall}}} \tag{10}$$

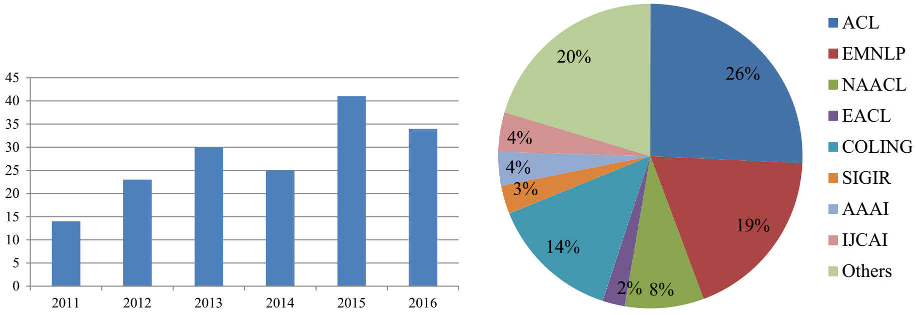
Other commonly used evaluation metrics also exist. Hovy et al. [75] propose a method where they represent each sentence as a set of semantic units called basic elements (BE) and calculate the coverage of BEs in the system outputs with regard to the reference summary. Nenkova and Passonneau [138] develop the pyramid evaluation approach by using Summarization Content Units (SCUs) to calculate weighted scores. An SCU has a higher weight if it is mentioned more frequently by human summaries. Consequently, a summary covering SCUs with higher weights will have a higher pyramid score. Intrinsic evaluation on other important aspects of summaries still very much relies on human judgment. For DUC or TAC conferences, human judges are asked to rate on various aspects of the system summaries, e.g., grammaticality, non-redundancy, clarity or coherence. Currently none of these aspects can be properly modeled by automatic approaches; therefore, manual evaluation is still indispensable in principle.

2 Recent Advances

2.1 Overview of Recent Progress

Document summarization tasks require systems to consider multiple factors when producing a summary, e.g., coverage of information, coherence, non-redundancy and conciseness. Progress has been made in recent years for document summarization from various aspects. In this section, we carefully survey the most significant streams of recent contributions made from relevant research, with more focus on methodologies that yield strong performance on standard benchmark evaluations. When organizing the descriptions in this section, we do not explicitly separate methods proposed for single-document summarization and multi-document summarization, although they may emphasize different aspects slightly differently.

In later DUC/TAC evaluation tasks, query-focused document summarization and guided summarization are starting to receive more attention. They differ from generic summarization in that a pre-specified query sentence is provided to describe the specific information need and thereby guide the summarization process. Until now, query-focused systems are mostly proposed with merely surface-level treatment for queries: using term overlap or literal similarity between document sentences and query sentences to integrate unsupervised systems or serve as features for supervised summarization. Many papers for query-focused summarization have no special treatment other than these, and there are actually more contri-



(a) **(b)**

Fig. 1 Bibliographic statistics for the recent research papers covered in this article. **a** Number of papers in each year. **b** Main published venues

butions made for generic summarization in these papers. Therefore we do not describe them separately in an individual section.

Here we make some bibliographic statistics for this survey, as illustrated in Fig. 1. We may observe that there exists a continuous trend in the community toward summarization tasks, and the number of surveyed papers appears to be relatively stable with an overall increase of interest. The main stream conferences on natural language processing including ACL, EMNLP, NAACL, EACL, COLING and others are the main publication venues for recent progress in document summarization and related topics. Some also appear in AI/IR venues as they are topically relevant. We also cover a few machine learning papers that model some aspects that are important for summarization tasks, as the authors of those papers also show the effectiveness of their approaches in summarization scenarios. There are also some relevant journal publications mentioned in this survey.

To provide a more explicit view of the organization structure of our descriptions and help readers to get a big picture of recent advances, we summarize the most important recent relevant papers we have covered in Table 1. The way we organize this survey is according to the most significant research streams made in document summarization in recent years and therefore need not be strictly categorized according to certain taxonomy of the different types of summarization tasks themselves.

2.2 Advances in extractive summarization

Much progress has been made within traditional frameworks of extractive summarization. We organize the descriptions in this section according to the most significant lines of research made in recent years.

As we mentioned earlier, we will not provide quantitative comparisons for different methods. Interested readers may refer to a recent quantitative analysis [73] on the performance of different extractive systems on the DUC 2004 multi-document summarization task, which conduct consistent, thorough analysis on the system outputs from a few representative papers with the state-of-the-art performance, and the conclusions are relatively reliable.

2.2.1 Improving concept-based ILP

Classic concept-based ILP systems optimizing for bigram concept coverage are based on concept weights derived from document frequency counts, with the assumption that frequently

Table 1 Main stream recent studies, corresponding sections and main bibliographic references

Recent research directions (section)		(Partial) main references
Traditional extractive summarization (Sect. 2.2)	Concept-based ILP (Sect. 2.2.1)	Supervised weights [53,97], learning with external resources [100], non-bigram concepts [167]
	Diversity prompting submodular functions (Sect. 2.2.2)	Submodular maximization [116,117], parameterized [118,172], volume maximization [61,93,94,207]
	Coherence modeling (Sect. 2.2.3)	Topic models [22,104], discourse [71,208], G-FLOW [32], semi-CRF [144]
	Other aspects (Sect. 2.2.4)	System ensemble [74,152,186], indirect supervision [87,162], neural network rankers [16,124,206]
Beyond sentence extraction (Sect. 2.3)	Compressive (Sect. 2.3.1)	Supervised compression [96,98,190], joint learning [2,12,45,134,156], discourse trimming [45,88,108,145]
	Abstractive (Sect. 2.3.2)	Caseframes [29], grammar-based [195], recombining units [13,30,121], extraction templates [59,154,166]
	End-to-end (Sect. 2.3.3)	Sentence simplification [31,135,165], hierarchical attention [28,135]
Beyond traditional summarization (Sect. 2.4)	New task settings (Sect. 2.4.1)	Comparative [79], update [42,99], evolutionary [198], multilingual [119,183]
	New domains/genres (Sect. 2.4.2)	Microblogs [24,168], meetings [188], opinions [189,191], scientific papers [35,155], etc.
	New applications (Sect. 2.4.3)	Generating slides [78], news [212], poetry [200], etc.

appeared bigrams will mostly contain important concepts. We have already pointed out the limitations of the frequency assumption in the introductory section. Introducing supervised learning may better predict which pieces of information are more important and should be preserved in the summary.

One possible way to inject supervision is to learn weights for sentences [53] by: training a regression model to predict sentence-level importance scores while assigning same weights for each bigram, and let the ILP model select important sentences while covering more frequent bigrams. The ROUGE scores can be used as prediction target. One may also directly predict importance scores for bigram concepts: for example, using discriminative training to learn a regression model to minimize the distance between the ground truth bigram frequency statistics in the reference summary and the estimated frequency [97].

Bigram-based ILP summarization methods may be further improved from different aspects [100]: Rather than using a predefined threshold to filter concepts as in previous practice [63], using syntactic information to select more important bigrams has been proved to be more effective, based on the intuition that in most cases nouns, verbs and adjectives are more indicative for document analysis. In addition to the internal features such as document frequency or bigram positions, features derived from external resources may also be helpful. The authors of [100] propose to extract features by leveraging multiple external resources such as pretrained word embeddings from large external corpus or relatively more informative resources such as Wikipedia, Dbpedia, WordNet and SentiWordNet. The bigram weights are then trained discriminatively in a joint learning model that predicts the bigram weights and selects the summary sentences in the ILP framework at the same time. It has also been shown that relevant public posts can provide useful information and can be effectively leveraged to improve news article summarization by helping to determine bigrams weights or even directly used as candidate sentences [101].

Another study finds that pruning low-weight concepts may not only lead to lower ROUGE scores but also multiple optimal solutions for ILP with very different real summary quality [14]. The authors introduce a small term into the objective function of ILP based on frequency of non-stopwords in the document set and prompt a single solution with improved performance.

On the other hand, concepts other than bigrams have also been studied. It has been shown that using syntactic and semantic concepts (e.g., frame semantics) instead of bigram concepts may not improve document summarization in classic settings of summarizing news clusters, but may become extremely useful in other genres such as lawsuits and Wikipedia texts [167].

2.2.2 Diversity prompting via submodular maximization

Another angle to improve information coverage is to promote diversity when selecting important individuals. By balancing itemwise important scores and overall selectional diversity, more items with high importance will be packed in the summary with more diverse coverage and thereby less redundant descriptions. This idea in the context of document summarization is typically implemented in the mathematical framework of submodular function maximization.

Submodular functions are set functions that satisfy the property of *diminishing returns*: Given a finite ground set V, for $\forall A \subseteq B \subseteq U \backslash u$, a set function $f : 2^U \rightarrow \mathbb{R}$ is said to be *submodular* iff[4]

$$f(A \cup \{u\}) - f(A) \geq f(B \cup \{u\}) - f(B). \tag{11}$$

The concept of submodularity fits content selection in summarization tasks well: There will be less gain by introducing an information unit into the current partial solution once we have already selected certain number of information units. Especially when scoring a summary at the sub-sentence level, submodularity naturally arises. For instance, concept-based summarization usually maximizes the weighted credit of concepts covered by the summary.

The problem of maximizing submodular functions is usually approximately solved via simple greedy algorithms, often packed with theoretical guarantees for worst-case approximation. For instance, a famous result is that the problem of maximizing a monotone submodular function under a cardinality constraint (restricting total number of selected elements) can be solved using a greedy algorithm to get an approximate solution which is at

[4] There is an equivalent definition which provides less intuition in the context of document summarization: f is submodular iff for $\forall A, B \subseteq V$ we have $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$.

least ($1 - 1/e \approx 0.63$) of the optimal value [136]. There are also studies for various performance guarantees for having knapsack constraints, monotone[5] or non-monotone objective functions, etc. (See [117], for more discussions.)

Lin and Bilmes [116] first treat the document summarization problem as maximizing a submodular function under a budget constraint. They show both theoretically and empirically that a modified greedy algorithm can efficiently solve the budgeted submodular maximization problem near-optimally, at least as good as $1/2(1 - 1/e) f(S^*)$ for the optimal solution $f(S^*)$.[6] Inspired by MMR (Eq. 1), the authors used an objective consisting of a graph cut function combined with penalty for redundancy:

$$f_{\mathrm{MMR}}(S) = \sum_{i \in V \setminus S} \sum_{j \in S} w_{i,j} - \lambda \sum_{i,j \in S : i \neq j} w_{i,j}, \lambda \geq 0. \tag{12}$$

Intuitively, many objective functions for document summarization are submodular. For example, the MMR sentence selection function (1) clearly satisfies the diminishing returns property. Lin and Bilmes [117] studied a class of submodular functions targeting for document summarization tasks. These functions each combine two terms: one which encourages the summary to be representative of the corpus and the other which positively rewards diversity. They model the summary score as

$$\mathcal{F}(S) = \mathcal{L}(S) + \lambda \mathcal{R}(S), \tag{13}$$

where $\mathcal{L}(S)$ measures coverage and $\mathcal{R}(S)$ rewards diversity in S. The authors propose the following objective that does not rely on concepts:

$$\mathcal{L}(S) = \sum_{i \in V} \min\{\mathcal{C}_i(S), \alpha \mathcal{C}_i(V)\}, \tag{14}$$

where $\mathcal{C}_i : 2^V \to \mathbb{R}$ is a monotone submodular function (designed as $\mathcal{C}_i(S) = \sum_{j \in S} w_{i,j}$ in the paper with $w_{i,j}$ for pairwise similarity) and $\alpha \in [0, 1]$ is a threshold coefficient. Instead of penalizing redundancy by subtracting from the objective, the authors propose to reward diversity by adding the following to the objective:

$$\mathcal{R}(S) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap S} r_j}, \tag{15}$$

where $P_i, i = 1, \ldots, K$ is a partition of the ground set V (i.e., $\bigcup_i P_i = V$ and $P_i \cap P_j = \emptyset \; \forall i, j$) into separate clusters and $r_i \geq 0$ indicates the reward of adding i into the empty set. The function $\mathcal{R}(S)$ rewards diversity in that there is usually more benefit to selecting a sentence from a cluster not yet having one of its elements already chosen. As soon as an element is selected from a cluster, other elements from the same cluster start having diminishing gain due to the square root function, hence the submodularity.

Slight modifications of the above functions can be easily made to adapt for query-focused summarization, taking similarity or overlap with query terms into account. Despite simple structures, they already achieve competitive performance on DUC datasets [117].

Further improvements have been made via designing parameterized submodular functions that can utilize explicit supervision from data to learn model parameters. An additional benefit is that structured learning under a structured SVM framework makes it easy to introduce

[5] A set function f is called monotone, if $f(A) \leq f(B)$ whenever $A \subseteq B$.

[6] The original paper [116] incorrectly proved a better $(1 - 1/\sqrt{e})$ bound, as pointed out in a later work from a different research group [134].

the ROUGE metrics into the training process, using (ROUGE-1) as the loss function for loss-augmented inference. For instance, one may design a mixture of "submodular shells" (classes of submodular functions with varying parameters) [118] whose mixture weights are learned directly from data. Another conceptually simpler way is to use linear models to parameterize the basic units in submodular functions [172]. In document summarization, the building blocks for submodular objective functions mostly involve two kinds of units: pairwise similarity scores $\sigma(i, j)$ and unit-level coverage scores $\omega(v)$. We can parameterize $\sigma(i, j)$ and $\omega(v)$ using linear models, allowing that each depends on the full set of input sentences x:

$$\sigma_x(i, j) = \mathbf{w}^\top \phi_x^p(i, j), \text{or} \quad \omega_x(v) = \mathbf{w}^\top \phi_x^c(v), \tag{16}$$

where $\mathbf{w}$ is the weight vector to be learned and ϕ denotes feature vectors. We (the authors of this survey) test their open implementation[7] on DUC 2004 and find that the system can achieve the state-of-the-art performance when compared with the currently published strongest results, in terms of both ROUGE and manual ratings of quality.

Other than explicit parameterizations, policy learning has also been studied in contextual submodular prediction [164]. By learning a contextual prediction policy based on a single no-regret learner, the system can produce a near-optimal list of predictions. This has been verified on document summarization task as sequentially predicting a list of sentences to construct the summary.

There are also studies that try to extend the concept of submodularity, with very similar framework but slightly different design and analysis. For example, Dasgupta et al. [40] formulate the objective function as a sum of a submodular function and a non-submodular function called dispersion, with the latter using inter-sentence dissimilarities in different ways in order to ensure non-redundancy of the summary.

There is another special type of submodular functions, derived from a probabilistic model called determinantal point processes [94], which jointly model the quality (importance) of each item and overall diversity in a set of items. Determinantal point processes (DPPs) are distributions over subsets that jointly prefer quality of each item and diversity of the whole subset. Formally, a DPP is a probability measure defined on all possible subsets of a group of items $\mathcal{Y} = \{1, 2, \ldots, N\}$. For every $Y \subseteq \mathcal{Y}$ we have:

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L + I)}$$

where L is a positive semidefinite matrix typically called an *L-ensemble*. $L_Y \equiv [L_{ij}]_{i,j \in Y}$ denotes the restriction of L to the entries indexed by elements of Y, and $\det(L_\emptyset) = 1$. The term $\det(L + I)$ is the normalization constant which obviously has a succinct closed form and is therefore easy to compute. We can define the entries of L as follows:

$$L_{ij} = q_i \phi_i^\top \phi_j q_j = q_i \cdot \text{sim}(i, j) \cdot q_j \tag{17}$$

where we can think of $q_i \in \mathbb{R}^+$ as the *quality* of an item i and $\phi_i \in \mathbb{R}^n$ with $\|\phi_i\|_2 = 1$ denotes a normalized feature vector such that $\text{sim}(i, j) \in [-1, 1]$ measures *similarity* between item i and item j. This simple definition gives rise to a distribution that places most of its mass on sets that are both high quality and diverse. This is intuitive in a geometric sense since determinants are closely related to volumes: In particular, $\det(L_Y)$ is proportional to the volume spanned by the vectors $q_i\phi_i$ for $i \in Y$. Thus, item sets with both high quality and diverse items will have the highest probability (Fig. 2).

[7] Available at http://www.cs.cornell.edu/~rs/sfour/.

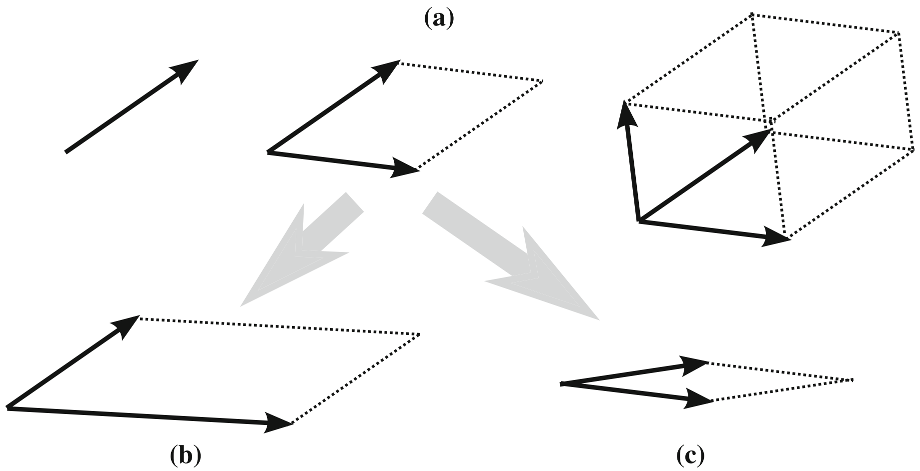


Fig. 2 Geometric intuitions of DPPs: **a** The probability of a set Y depends on the volume spanned by vectors $q_i\phi_i$ for $i \in Y$. **b** As length increases, so does volume. **c** As similarity increases, volume decreases

Specifically, DPPs combine a per-sentence quality model that prefers relevant or important sentences with a global diversity model encouraging non-overlapping content. This setup has several advantages. First, by treating these opposing objectives probabilistically, there is a unified framework for trading off between them. Second, the sentence quality model can depend on arbitrary features, and its parameters can be efficiently learned from reference summaries via maximum likelihood training [93]. Finally, because a DPP is a probabilistic model, at test time it is possible to sample multiple summaries and apply minimum Bayes risk decoding, thus improving ROUGE scores [94].

A closely related approach is maximizing the semantic volume [207]. The authors use singular value decomposition on bigram vectors to get vectorial representations for sentences and then maximize the volume spanned by the vectors via a Gram–Schmidt process. This volume maximization procedure has been shown to be more effective than MMR selection, for the purpose of redundancy removal and diversity prompting.

2.2.3 Methods for improving summary coherence

Coherence is an important property when producing a summary. Understanding the descriptive structure of the original documents is crucial for prompting coherence in the generated summaries. Early approaches construct lexical chains [7], which represent sentence relatedness through word and synonym that overlap across sentences. The hypothesis is that each chain represents a topic, and topics that are pursued for greater lengths are likely to be more salient.

Unsupervised probabilistic approaches, usually variants of Bayesian topic models, can be adapted to model the hidden abstract concepts across documents as well as their correlations, to generate topically coherent and non-redundant summaries. These approaches are suitable for query-focused summarization, integrating query relatedness in the generative models [22, 104].

The G- FLOW system [32] estimates coherence by using an approximate discourse graph, where each node is a sentence from the input documents and each edge represents a discourse

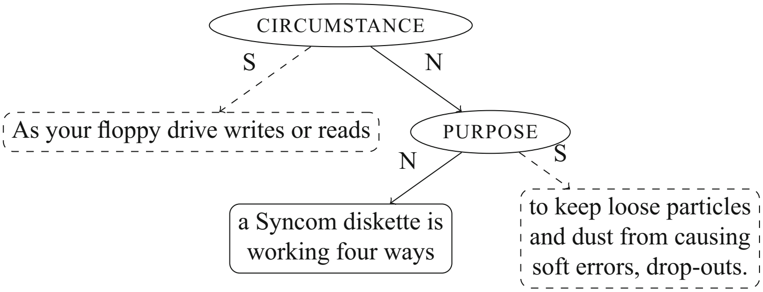

CIRCUMSTANCE
S
N
As your floppy drive writes or reads
PURPOSE
N
S
a Syncom diskette is working four ways
to keep loose particles and dust from causing soft errors, drop-outs.

Fig. 3 (An example from Mann and Thompson, [130]) An RST discourse tree with EDUs as leaf nodes

relationship between two sentences. Relationship between all described entities in the sentences can be used to calculate edge-level scores or used more globally to score a full candidate summary containing multiple sentences [194]. Coherence scores can also be parameterized, for example, using structured linear models like CRFs or semi-Markov HMMs [144]. Then the summarization problem can be formulated as combinatorial optimization with the objective function consisting of both parameterized coverage scores and parameterized coherence scores, both jointly learned from data.

Many single-document summarization systems utilize analysis of the discourse structure of the input document to produce more coherent single-document summaries. *Rhetorical structure theory* (RST) [130] is a commonly mentioned concept, which requires the overall structure of a text to be represented by a tree. RST trees have the smallest units of text analysis, called *elementary discourse units* (EDU), as leaf nodes. EDUs are essentially sub-sentential clauses derived from a segmentation of sentences, including dependencies such as clausal subjects and complements. The more central units to each RST relation are *nuclei* (N), while the more peripheral are *satellites* (S). Figure 3 describes an example discourse tree with EDUs.

After discourse parsing and getting the RST dependency tree, the single-document summarization problem can be formulated as a tree knapsack problem [71], in which sentence selection must follow the rule that once any sentence has been selected, so must its head sentence be. The discourse parser itself can also be trained using summarization data [193,208].

2.2.4 Other advances in extractive approaches

We describe some relatively more discrete advances for extractive summarization in this section.

Many studies also try to improve document summarization from other aspects that have not been explicitly considered in traditional approaches, for example, to extract more certain sentences [185] or to utilize timeline information to enhance summarization [142].

Some studies focus on integrating the power of different summarization systems, trying to promote weighted consensus [186], or directly perform supervised aggregation [152] or reranking outputs from different base systems [74].

There are also a few recent studies that focus on improving graph-based summarization. Li and Li [111] integrate topic models into graph ranking, utilizing relations between topics and sentences. Parveen and Strube [149] use a bipartite graph connecting sentences and topics to represent a document and apply the HITS algorithm to calculate importance. Graph-based topical coherence can be naturally introduced in graph-based frameworks. By building sentence-entity bipartite graphs, coherence scores can be derived from node degrees (possibly weighted) and integrated in a ILP objective function [150]. Meanwhile, using rich syntactic/semantic information to derive frequent sub-patterns for similarity calculations may also improve the performance of graph ranking models [201].

Machine learning under indirect supervision, such as using reinforcement learning [162] or learning to search [87], has also been adapted to summarization tasks and shows great potential by defining proper reward functions. Such approaches can directly utilize relevant metrics (such as ROUGE) during training for defining proper reward signals, while the non-differentiability of relevant metrics makes it difficult for direct numerical optimization in other frameworks. Also, such methodologies can naturally fit many scenarios where data are in large scale and come in streams.

Representation learning based on neural networks with multiple layers has made significant progress in many subfields of artificial intelligence, especially in computer vision and speech recognition. In recent years, it also starts to show some potential in natural language processing. There starts to emerge a bunch of work that tries to model summarization tasks in neural network architectures, with less or no dependence on handcrafting features. Until now, neural network approaches for document-level summarization are mostly playing partial roles, acting as one component such as sentence scoring in essentially a traditional extractive framework. Deep Boltzmann machines have been adapted for document summarization to learn hierarchical concept representations and to predict concept importance and select sentences accordingly [124]. A few studies explored directly measuring similarity based on distributed representations, using the sum of trained word embeddings to represent sentences or documents [91,92]. Convolutional architectures have been designed for sentence modeling and selection [16,206], used as sentence scoring modules for extractive summarization. A later work [18] also uses convolutional sentence embeddings to model sentence-level attentive behaviors, using a layered neural network to learn query relevance ranking and sentence saliency ranking simultaneously. Sentence ranking framework can also be built upon recursive neural networks, formulating scoring as hierarchical bottom-up regression [15]. Recently, it has been shown effective to use even the simplest form of neural network, i.e., generic multilayer perceptron, to directly predict the relative importance of a sentence given a set of selected sentences, considering importance and redundancy simultaneously [160].

Meanwhile, a few unsupervised approaches have also been proposed. However, unsupervised approaches have mostly been overperformed by supervised approaches, even though the size of available training data is currently still relatively small. Zhang et al. [215] utilize the density peaks clustering algorithm [163] for scoring representativeness and diversity, yielding relatively strong ROUGE results as an unsupervised framework. Empirically, the OCCAMS system [41] gives currently the best performance in unsupervised methods on standard DUC datasets. It first derives the term weights via latent semantic analysis and then selects sentences that cover the maximum combined weights. Another recently explored idea is data reconstruction [70], based on an assumption that a good summary may consist of those sentences that can best reconstruct the original document. The mathematical formulation is straightforward, while being rather easy to extend as shown in a bunch of follow-up papers or ideas [110,122,129,204]. However, efforts from this stream of study fail to achieve convincing performance as shown by experimental evaluation on standard multi-document DUC datasets.[8] The reported results are inferior to OCCAM and far less comparable to the state-of-the-art supervised approaches, and one of them [204] actually was later found to

[8] Starting from [70], all these papers weirdly evaluate their systems merely on query-focused datasets although they are designed for generic cases.

perform even worse than reported due to incorrect length control in the experiments. In fact, apart from lacking task-specific supervision, there exists a conceptual gap between the reconstruction assumption and practice. Data reconstruction approaches encourage summaries to cover information as much as possible, while in practice good summaries should only cover a small portion of original information. We cannot expect even a human to recover most of the information described by a full document only after reading a short paragraph of summary.

2.3 Beyond sentence extraction

Although much progress has been made in extractive summarization, one of the problems that extractive approaches suffer from is that they unavoidably include secondary or redundant information. More importantly, it is still far from the way humans write summaries. For single-document summarization in particular, the well-known *Lead* baseline, i.e., extracting the first sentences of the document, has already been close to the 99% percentile of the ROUGE score distribution over all possible extractive summaries for newswire and scientific domains [23], showing that it is difficult to significantly improve over the *Lead* system on standard benchmarks (e.g., see standard DUC/TAC evaluations). Similar percentile ranks have also been observed for the TextRank system [133]. These results may not suggest that additional improvements cannot be made in these domains, but that making further improvements based on only sentence extraction will be considerably difficult.

Abstractive summarization is generally considered to be much more difficult, involving sophisticated techniques for meaning representation, content organization, surface realization, etc. There has been a surge of interest in recent years on compressive document summarization that tries to compress original sentences to form a summary, as an intermediate, viable step toward abstractive summarization.

2.3.1 Compressive summarization

Compressive summarization includes sentences which are compressed from original sentences, by extracting partial sentences from the original documents, without further modifications other than word deletion. Compressive summaries often contain more information than sentence extraction, since they can remove less important sentence components and make room for more salient information that is otherwise dropped due to the total length constraint. To form grammatical compressions, sentence compression is typically implemented as trimming syntax trees produced by a constituent parser or a dependency parser, while following certain linguistically motivated rules. Figure 4 shows an example via trimming a constituent parse tree.

Two general strategies have been used for compressive summarization. One is pipelining, where sentence extraction is followed or proceeded by sentence compression [113,190,211]. Another line of work uses joint compression and summarization. Such methods have been shown to achieve promising performance but typically computationally much more expensive.

Chali and Hasan [25] study the effectiveness of sentence compression under an ILP framework for query-focused summarization. A comprehensive set of query-related and importance-oriented measures are used as well as various sentence similarity measures to define the relevance constraints and redundancy constraints. They show that jointly performing compression and extraction via optimizing a combined objective function outperforms pipeline approaches.

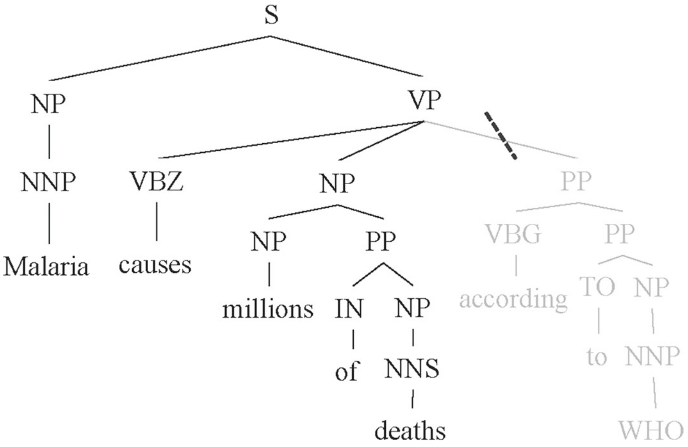


Fig. 4 An example of constituent tree trimming for sentence compression. The nodes to be dropped are grayed out. The sentence is compressed as *Malaria causes millions of deaths*

In earlier work, sentence compression is usually done in an unsupervised fashion based on frequency-driven scores and tree-trimming rules or being supervised from external sentence compression datasets. Such general-purpose sentence compression is somewhat independent or inconsistent to the goal of summarization. Improvements can naturally be achieved with supervision or guidance from summarization data when training compression models.

One treatment is to use summarization data to provide training targets for compression models. Li et al. [96] train conditional random fields for sentence compression, using data annotated with word importance derived from manually written summaries. They show that including sentences with such guided compression in ILP models improves over including sentences with generic compression. For sentence compression based on trimming constituent trees, the reference label for every node in the tree can also be obtained automatically from the bottom to the top of the tree [98]. In a pipeline framework where sentences are first compressed via trimming expanded constituent trees using the learned model, the system achieves similar ROUGE scores but better linguistic quality on TAC data.

Another way is to combine multiple scoring models with the guidance of summarization data. Wang et al. [190] investigated the role of supervised sentence compression techniques for query-focused multi-document summarization. A compression scoring function is constructed to incorporate multiple task-specific scorers, including scores from their proposed tree-based compression, query relevance, significance and redundancy, with combination weights tuned on held-out data. Their system shows statistically significant improvements over pure extraction-based approaches, achieving the current state-of-the-art results on query-focused DUC datasets (DUC 2006 and DUC 2007), in terms of both ROUGE scores and pyramid scores, along with reasonably good manual evaluation scores.

Currently the most popular way for supervised compressive summarization is to perform multi-task learning or jointly learn an extraction model and a compression model in the same framework.

Berg-Kirkpatrick et al. [12] first propose an approach to score the candidate summaries according to a combined linear model of extractive sentence selection and compression. They train the model using a margin-based objective whose loss captures the final summary quality. Since the search space is way larger than pure sentence selection for ILP solvers, they perform some sentence filtering in the first step to reduce the number of candidates as more practical approximation.

As the scale of problem grows significantly larger in joint extraction and compression settings, various alternatives to the ineffective ILP solvers have been studied. A recently proposed framework enables independent decoding for compression while dealing with knapsack constraint separately, based on alternating direction dual decomposition (AD^3) [2]. The authors propose multi-task learning to train compressive summarizers, using auxiliary data for extractive summarization and sentence compression. Their framework yields high ROUGE scores and consumes running time as short as extractive systems. Another approximate inference strategy is to cast the original ILP into a more tractable formulation, such as graph cuts [156]. The authors modify the objective function with supermodular binary quadratic functions to eliminate subtree deletion constraints and relax the length constraint using Lagrangian relaxation. The relaxed objective function is bounded by the supermodular binary quadratic programming problem which can be approximately solved using graph max-flow/min-cut. Morita et al. [134] try to produce compressive summarization by extracting a set of dependency subtrees in the document cluster, under the budgeted submodularity framework, with dependency constraints to guarantee readability. They propose an efficient greedy algorithm for approximate inference with performance guarantee, calling a dynamic programming procedure for subtree extraction.

Compressive summarization in single-document case can also integrate discourse-level compression, which may lead to more coherent compressed sentences. A natural way is to consider both the syntactic dependency tree for words and discourse dependency tree between sentences (rhetorical structures) as a nested tree structure, then formulate this nested tree-trimming problem as combinatorial optimization [88] and generate compressive summaries using ILP solvers or more carefully designed dynamic programming procedure [145].

A very recent system [45] tries to combine discourse-level compression based on RST tree and syntactic compression based on constituent trees. To improve cross-sentence coherence, the system incorporates a model of anaphora resolution and gives the ability to rewrite pronominal mentions and then integrates pronoun coreference constraints in the ILP formulation. Specifically, the model incorporates (1) constraints from coreference ensuring that critical pronoun references are clear in the compressed summary and (2) constraints from syntactic and discourse parsers ensuring that sentence realizations are well formed. The ILP objective function contains weighted scores for both unit extraction and anaphoric references. Weights are directly trained using manual abstractive summaries via structured SVM with ROUGE-based loss function. On the New York Times dataset and the RST Treebank which contain reasonably sufficient scale of document-summary pairs for supervised training, the system significantly outperforms the baseline that extracts leading sentences.

Actually there exist other justifications for utilizing discourse parsing and discourse units for compressive summarization. By studying the compatibility of EDUs with human-labeled summarization units from pyramid evaluations and by assessing their utility in reconstructing manually written document previews, a recent study [108] demonstrates that segmenting EDUs (*elementary discourse units*, cf. Sect. 2.2.3) is effective in preserving human-labeled summarization concepts, while using EDUs as units of content selection instead of sentences leads to stronger summarization performance, especially under tight budgets.

In all, compressive systems are currently producing competitive results with syntactic and discourse constraints directly guiding the results toward being concise and coherent, achieving a good trade-off between content compactness and readability.

2.3.2 Toward full abstraction

Fully abstractive summarization attempts to understand the input and generate the summary from scratch, usually including sentences or phrases that may not appear in the original document. It actually involves multiple subproblems, each of its own can be made a relatively independent research topic, including: simplification, paraphrasing, merging or fusion. Cheung and Penn [29] conduct a series of studies comparing human-written model summaries to system summaries at the semantic level of *caseframes*, which are shallow approximations of the semantic role structure of a proposition-bearing unit like a verb and are derived from the dependency parse of a sentence. They find that human summaries are: (1) more abstractive, using more aggregation, (2) contain less caseframes and (3) cannot be reconstructed solely from original documents but are able to if in-domain documents are added.

Due to the inherent difficulty and complexity of full abstraction, current research in abstractive document summarization mostly restricts in one or a few of the subproblems. It is also less active compared with compressive summarization, since merely considering compressions has already boosted system performance, as discussed in the last section.[9]

Woodsend and Lapata [195] propose a model that allows paraphrases induced from a quasi-synchronous tree substitution grammar (QTSG) to be selected in the final ILP model covering content selection, surface realization, paraphrasing and stylistic conventions. For document summarization that involves paraphrasing and fusing multiple sentences simultaneously, other than grammar-based rewriting, one simpler more typical approach is to merge information contained in sub-sentence-level units. For instance, one can cluster sentences, build word graphs and generate (shortest) paths from each cluster to produce candidates for making up a summary [6,51]. More sophisticated treatments can also be built on syntactic or semantic analysis. One may build sentences via merging consistent noun phrases and verb phrases [13] or linearizing graph-based semantic units derived from semantic formalisms such as abstract meaning representation (AMR) [121].

There also exist psychologically motivated studies [48] trying to implement cognitive human comprehension models based on *propositions*, which are elements extracted from an original sentence, each containing one functor and several arguments. Propositions form a tree where a proposition is attached to another proposition with which they share at least one argument. Summaries are then generated from selected important propositions. Currently the systems have mostly been evaluated on over-specific datasets and rely heavily on various components including parsing, coreference resolution, distributional semantics, lexical chains [49] and natural language generation from semantic graphs [50].

In order to better guide alignment and merging processes, supervised learning-based methods have been investigated [46,178]. A later work [30] expands the sentence fusion process with external resources beyond the input sentences by combining the subtrees of many sentences, allowing for relevant information from sentences that are not similar to the original input sentences to be added during fusion.

Abstractive summarization has also been studied in information extraction (IE) perspective, for example, IE-informed metrics have also been shown to be useful to rerank the output of high performing baseline summarization systems [83]. In the context of guided summarization where predefined categories and readers' intent have been predefined, preliminary full abstraction can be achieved by extracting templates using predefined rules for different types of events [59,166].

[9] Nevertheless, in some specific domains and genres such as meeting summarization or opinion summarization, the system has to produce abstractive summaries. We will briefly give some relevant introduction in next section.

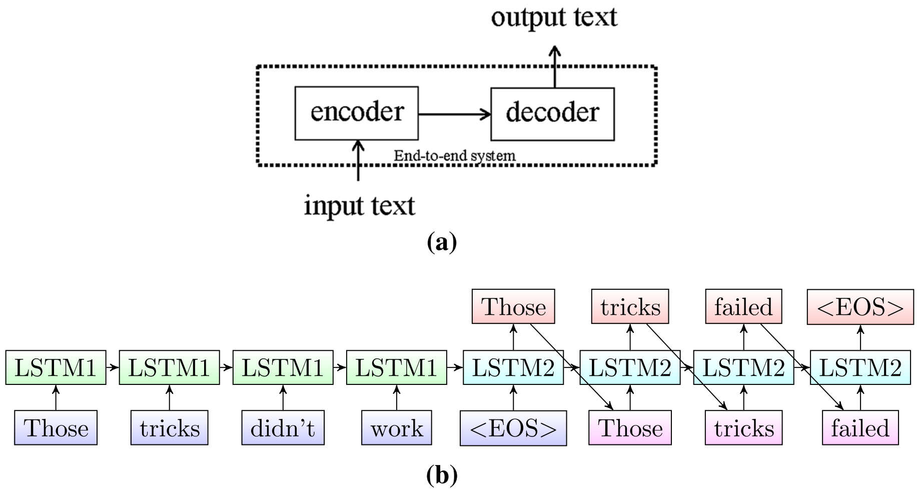


Fig. 5 Encoder–decoder architecture. **a** General architecture. **b** An RNN-based instance for sequence-to-sequence transduction (see [4,175] for technical details)

A large part of existing work in abstractive summary generation is actually limited to more specific domains, where fixed templates or rules are manually crafted for generating the sentences. For example, abstract-based approaches have been studied for product reviews where graph-based algorithms can be designed to merge reviews with similar textual content [55]. Sentence realization templates can be designed to ensure grammaticality [60]. Meanwhile, instead of generating a summary consisting of multiple sentences, some research focuses on only generating a headline for each news article sentence [1,154]. The authors first cluster or learn the event templates from a large number of news articles and then fill the entities into appropriate templates to form the headline. Headline generation has also become a test bed for modern neural abstractive generation, described in the next section.

2.3.3 Toward end-to-end abstractive summarization

Recently end-to-end training with encoder–decoder neural networks [175] have achieved huge success in data sufficient sequence transduction tasks such as machine translation, which brings potential applications for summarization tasks, especially for abstractive settings. Figure 5a gives a high-level description of encoder–decoder architecture. Input texts are encoded in a encoder network and then pass to decoder network to produce the desired output. Such architecture will be made more specified (typically implemented using basic building blocks such as recurrent neural networks with gated units and attention weighting [4]) to adapt to different sequence-to-sequence tasks, including machine translation and text rewriting. The inputs are typically just raw texts, making the whole system free from heavy manual feature engineering.[10] Figure 5b depicts an instance based on two LSTM recurrent networks as encoder (the sequence with green states LSTM1) and decoder (the sequence with cyan states LSTM2), used for rewriting the input text (blue squares) into a more concise output (red squares; each output token is also reused as input for the next decoder state to generate the next token).

[10] That said, designing architectures that actually work is commonly reckoned to be equally labor-intensive.

Currently, this line of research, sometimes under the term *sentence summarization* (started from Rush et al. [165] and somewhat misleadingly called *text summarization* in some follow-up research work), is in fact essentially *sentence simplification* working on short text inputs such as microblogs, tweets or single sentences. Therefore the applications are mainly still in microblog summarization, sentence simplification and headline generation.

Relevant advances typically contribute more on extremely focused aspects to improve sequence-to-sequence learning or more specifically attention-based RNN encoder–decoder structures [31,135] in general. For example, since many words in a simplified sentence are retained from the original input sentence, it has been proved to be useful to incorporate the *copying* mechanism [66,67,135] that allows a word to be generated by directly copying an input word rather than producing from the hidden state. Meanwhile, directly optimizing ROUGE via reinforcement learning has been shown to be more effective than optimizing likelihood for the decoder generation [3,159]. For sentence simplification tasks usually there exists a predefined length constraint. As it is difficult to pose hard constraints on decoder generation, one recent work [89] studies various solutions, including direct truncation on generated sequence, discarding out-of-range generations in the decoding beam and directly embedding length information in the LSTM units.

Unfortunately, it is still long way to go to adapt such architectures to document summarization. Encoding for generic documents, which typically contains multiple paragraphs or a collection of related documents, currently still lacks satisfactory solutions. This hampers the generalizability and usability of sequence-to-sequence approaches. Currently there are a few attempts for generic document summarization under end-to-end neural architectures. To challenge the problem of currently longer inputs, hierarchical encoding and multiple levels of attention have been designed [28,135]. However, recent proposals of architectural designs have yet to achieve competitive performance for fully abstractive summary generation.

On the other hand, another unfortunately less noticed drawback in this stream of study of neural sentence simplification is that most papers equate performance and quality with the ROUGE metrics and simply just omit manual evaluations on meaning preservation and linguistic quality, even though there exists no proof that the quality of simplification correlates well with single-reference ROUGE on sentence-level output. As a result, one has to take related progress with a grain of salt. A recent study [179] introduces a manually created *multi-reference* dataset for abstractive compression of sentences or short paragraphs. Empirical evaluation on the dataset shows the importance of multiple reference as well as suitable units in order to make automatic metrics more reliable, while neural models currently are still inferior to classic deletion-based ILP frameworks [34] in terms of human ratings.

Nevertheless, sequence-to-sequence frameworks have been shown to be effective for some specific genres with short output, for example generating abstracts for opinion and arguments [189]. At the encoder part, importance sampling is performed to limit the input to consist only a few possible sentences, with the importance weights estimated from an external regression model.

2.3.4 Comments

Recent years have witnessed some progress beyond sentence extraction, with a number of studies shifting focus toward compressive summarization and more abstractive summarization to directly generate sentences. Compared with sentence extraction, compressive summarization can produce more concise summaries, but not as flexible as more abstractive approaches. Meanwhile, research in non-extractive approaches is still at the beginning. Current fully abstractive approaches cannot always ensure grammatical abstracts, which is also one major limitation of language generation in general.

2.4 Progress beyond traditional news summarization

The most typical settings of traditional generic summarization studies are based on standard benchmarks that are collected from news data. However, there exist various types of different tasks settings, domains and genres that worth some efforts of research. Meanwhile, traditional summarization techniques have been adapted and applied for many related but different applications. In this section we will describe progress beyond the most standard settings of summarization tasks.

2.4.1 New settings for document summarization

In recent years there have been massive studies that explore beyond traditional generic document summarization, addressing different use cases for document summarization with specific settings.

For example, *comparative summarization* requires the system to provide short summaries from multiple comparative aspects. For extractive approaches, sentences with high representativeness and comparativeness should have more tendency to be selected [79]. Wang et al. [187] propose a discriminative sentence selection approach based on a multivariate normal generative model to extract sentences best describing the unique characteristics of each document group, aiming at summarizing the differences. Ren and de Rijke [161] explicitly consider contrast, relevance and diversity for summarizing contrastive themes. They employ hierarchical nonparametric Bayesian model to infer hierarchical relations among topics and enhance the diversity of themes by using structured determinantal point processes [61]. They pair contrastive themes and employ an iterative optimization algorithm to select sentences. Recently differential topic models have also been explored to measure sentence discriminative capability for comparative summarization [69].

Update summarization addresses another goal to generate a short and concise summary for the latest updating topic-related documents, assuming that the user has already read some earlier documents on the same topic. In this setting, both salience and novelty should be considered. Graph-based methods can be adapted to capture the relation between the information in earlier documents and the latest documents and derive salience scores for sentence ranking [184]. There are also studies using structured topic models as an unsupervised probabilistic approach to model novelty in a document collection and applying it to the generation of update summaries [42]. In particular, hierarchical Dirichlet processes have been shown to be flexible for integrating temporal information and inferring the relationships between sentences and multiple aspects [103,105]. A recent study addresses the task by modifying the classic concept-based ILP framework for traditional summarization, using supervised concept weights and discriminative reranking to produce more competitive results [99].

An extension of update summarization with multiple steps is called *evolutionary timeline summarization* (ETS). Given the massive collection of time-stamped Web documents related to a general news query, ETS aims to return the evolution trajectory along the timeline, consisting of individual but correlated summaries of each date. This setting emphasizes multiple factors including relevance, coverage, coherence and (cross-date) diversity. The task can be formulated as a constrained optimization problem to select and substitute sentences that satisfy multiple requirements [199] or can be addressed in a graph ranking framework unifying inter-date and intra-date dependencies between sentences [198]. Related ideas can be used to track large-scale events across time, in frameworks such as pipelining sentence salience prediction and clustering-based multi-document summarization [86].

Currently most summarization research settings are monolingual. A few exceptions try to explore the multilingual summarization setting in which the system should be able to process several languages in source documents with a summary in the same language as input [119]. Litvak and Last [119] describe cross-lingual methods for training an extractive single-document text summarizer called MUltilingual Sentence Extractor (MUSE), using a genetic algorithm to find the best linear combination of a rich set of language-independent sentence scoring metrics.[11] Another related but different setting is cross-language summarization, where source and target languages are different. Currently proposed solutions include bilingual graph coranking [183] and other approaches inspired by statistical machine translation [205,213].

2.4.2 Summarization in specific domains and genres

Since different tasks are defined under different domains or text genres, researchers may develop approaches that differ substantially from the generic summarization approaches designed for news documents. In some specific genres, the input documents are usually short and therefore not considered as "document summarization." We will slightly touch these settings as well for integrity. Typically there exist new challenges, compared with generic document summarization. For example, microblog data may come in massively large scale, consisting multiple items that repeatedly and redundantly describe the same event. Texts are far less formal and contain huge noise. Information might be time-variant, while user information needs are diverse.

Microblog timeline summarization and twitter stream summarization serve as an example. Microblog data are individually short, but often highly redundant as a collection, and are often aligned on timeline. Extractive approaches are predominant on tweet summarization. They are first used for streams following simple and structured events such as sports and games [24,143,177]. In particular, Chakrabarti and Punera [24] utilize temporal structural properties by designing modified hidden Markov models to automatically learn differences in language models of sub-events. Date selection is also proved to be important in timeline summarization [180]. More abstractive studies start from the Phrase Reinforcement algorithm [168] which extracts frequently used sequences of words and is first applied in summarizing topic streams. Subsequent research emphasizes improving word graphs by using dependency parses [85], sequential summarization over evolving topics [56] or having online stream data as input [146]. Due to the specific properties of microblogs, personalization and social context can also be introduced in the model to enhance performance for twitter summarization [77,102,123,203] or leverage both social factors and content quality [44,216]. There also exists research that studies summarizing the repost structures of popular tweets [106], leveraging both the content of repost messages and different reposting relations between commenters and followers. A related task is indicative tweet generation, which aims at generating indicative tweets that contain a link to an external web page. There has been some work within extractive frameworks [125]. However, it has been shown recently that word extraction is rather limited for this task [171]: The less formal the source article is, the less extractive the tweets seem to be.

[11] The authors of [119] use ROUGE-1 recall as the fitness function for measuring summarization quality. The discreteness of objective function (ROUGE) hampers the use of linear programming solutions. In principle, other more advanced and more efficient global optimization techniques such as Bayesian optimization [173] may also be applicable.

Summarizing spoken data or transcripts poses the extreme challenge of noise and redundancy. Other than information coverage, special treatments are necessary to extend beyond utterance extraction. For meetings [148,188] and conversations [181], more compact and more abstractive generations are required. However, unlike generic summarization, they typically have relatively fixed patterns and procedures, making template extraction and information fusion slightly easier and more feasible. Typical frameworks consist of templates extraction from the training set and template filling.

Opinion summarization is the task of producing a summary that also preserves the sentiment of the text,[12] therefore posing a trade-off between summarization and opinion mining or sentiment analysis: The demand of extraction or compression may drop sentiment- bearing sentences, while the demand of sentiment detection may bring in redundant sentences. Submodular functions or modifications can be designed to address the conflicting requirements, balancing the coverage of both topics and polarities [81,191]. Product review summarization can also be implemented via ILP based on phrase selection, optimizing for both popularity and descriptiveness of phrases [210]. Additional information for reviews such as review helpfulness ratings has also been proved useful to guide review summarization [196]. Meanwhile, abstractive approach has been shown to be more appropriate for summarizing evaluative text [20,43]. In particular, graph-based method has been explored to produce ultra- concise opinion summaries [55]. To improve fluency for abstraction, Carenini et al. [20] try to generate well-formed grammatical abstracts that describe the distribution of opinion over the entity and its features, with a handcrafted feature taxonomy for each product as input. Di Fabbrizio et al. [43] propose a hybrid abstractive/extractive sentiment summarizer to select salient quotes from the input reviews and embed them into the abstractive summary to exemplify, justify or provide evidence for the aggregate positive or negative opinions. End-to-end encoder–decoder RNNs have also showed effectiveness in producing short, abstractive summaries for opinions [189]. For longer reviews it is feasible to perform discourse parsing and aggregate discourse units in a graph; then, review summarization will reduce to sequentially perform subgraph selection and template-based generation [60].

Summarizing scientific articles has been a popular research topic in recent years. Although author-written abstracts are usually available, they are considered to be less structured, vary significantly in terms of length, are often not self-contained and sometimes even have been written independently of the main document. Apart from features in generic summarization, many other information can be explored in scientific articles. For example, automatically annotated argumentative zones[13] can be used as features to guide extractive summarization for scientific articles [38]. More fine-grained aspects of the content and conceptual structure of the papers might be more useful than argumentative zones in certain cases by providing a greater level of detail in terms of categories denoting objectives, methods and outcomes [112]. Citations to a particular article can also be aggregated to construct its summary, e.g., performing centrality-based summarization after clustering citations [155]. Recent studies also try to combine both sources, i.e., utilizing the citation sources while reflecting the content and the discourse structure of the original paper [35]. More careful treatment for discovering salient keywords and information-rich citation sentences may further improve scientific summarization as well [197]. A related application is to perform scientific survey generation.

[12] For a more specific, comprehensive discussion on opinion summarization, readers may refer to existing survey papers (e.g., [90,120]).

[13] A scheme of information structure that classifies sentences in scientific text into categories (such as Aim, Background, Own, Contrast and Basis) based on their rhetorical status in scientific discourse.

Link analysis models such as HITS can be adapted to exploit the lexical network structure between sentences from citing and cited papers [82].

Besides the aforementioned studies, there also exists research on summarizing e-mails [127], community question answering [26], student responses [128], movie scripts [64], entity descriptions in knowledge graphs [27] and source codes descriptions [80]. Different scenarios pose variously different requirements and objectives on summarization systems.

2.4.3 New applications of document summarization techniques

There also exists research that explore new applications of classic document summarization techniques. For instance, traditional summarization framework including sentence scoring and selection has been applied in new scenarios such as automatically generating presentation slides for scientific papers [78] and automatically constructing sports news from live commentary scripts [212]. More crafted content selection and organization have even enlightened the possibility to automatically compose poetry [200]. There also exist studies for generating topically relevant event chronicles, mainly consisting of event detection module followed by learning-to-rank extractive summarization to select salient events and construct the final chronicle [57].

Summarization techniques have also been used to help interpreting predictions from neural networks, which are commonly treated as black boxes that make predictions without explicitly readable justifications. For example, it is useful to extract or generate short rationales to explain why a neural network model predicts certain sentiment classes for a paragraph of user-generated reviews. Sentences generated for such scenarios should be concise and coherent, while being sufficient for making the same predictions when only using these sentences alone without referring to the full passage of review [95].

There exists another kind of high-level document summarization that tries to produce a summary of huge topic hierarchies. Bairi et al. [5] recently study this task to summarize topics over a massive topic hierarchies (a huge directed acyclic graph) such that the summary set of topics produced represents the objects in the collection. The representation is characterized through various classes of monotone submodular functions with learned mixture weights capturing coverage, similarity, diversity, specificity, clarity, relevance and fidelity of the topics.

3 On future trends and directions

The fast development of related fields has brought some new possibilities for document summarization. However, there still exist many remaining challenges unsolved. In this section we will give a brief overview on some of the significant trends and possible important future directions in the research frontier.

3.1 Collecting data for summarization

Currently the standard datasets for document summarization tasks, especially for multi-document cases, are mostly in small scale, consisting of only tens of topics per task. This hampers the progress of machine learning-based approaches. The shortage of data appears more obviously in domains other than news, as well as in languages other than English. As a consequence, current research lacks focus on other domains and languages. Building high-quality datasets for summarization will be an important future direction that will largely boost the development of this field. There exists some preliminary progress in collecting large-scale data for producing extremely short summaries using microblogs related to specific news articles [17,76], but data collection for more generic summarization or other different genres is still a topic to be explored. As a temporary solution to the data shortage problem, it is also worthwhile to consider better utilizing external resources or additional background corpora to help summarizers in capturing important information [100,218].

The necessity to collect high-quality data also naturally appears when evaluating summarization systems, where certain scale of evaluation data are needed to reach statistically convincing conclusions. There even exist additional issues other than scale. As a concrete example, a recent study [11] shows that when evaluated on traditional query-focused summarization datasets, state-of-the-art algorithms designed for query-focused summarization do not significantly improve upon generic summarization methods which simply ignore query relevance. They introduce a new dataset with controlled levels of *topic concentration* and report strong performance improvements for algorithms that properly model query relevance as opposed to generic summarizers.

3.2 Improving evaluation for summarization

Even if ROUGE metrics are currently the de facto standard for automatic evaluation, they are not perfect in many aspects. For example, ROUGE scores will remain unchanged after arbitrarily disordering the sentences in a summary, since ROUGE metrics are designed mostly for detecting information coverage rather than coherence or other important quality factors. Studies have shown that for lower-order ROUGE scores they tend to detect significant differences among systems, even though human judges find that they are actually comparable [158]. Also, a recent study [65] reveals some inconsistency for using different ROUGE variants, using pairwise Williams significant test to show that previously recommended best variants of ROUGE (average ROUGE-2 recall without stop word removal) [73,147] might be suboptimal.

Some strategies for improvements on ROUGE as well as other automatic or semiautomatic metrics have also been proposed. For example, since ROUGE scores are unfairly biased toward surface lexical similarities, word embeddings can be used to compute the semantic similarity of the words used in summaries and better correlations with human judgements have been achieved [140]. Distributional semantics have also been used to perform automatic pyramid scoring [151]. Louis and Nenkova [126] propose to use four classes of easily computable features that are supposed to capture aspects of the input text without the need of gold-standard summaries, showing that their approach correlates favorably with pyramid scores.

3.3 Summarization via understanding the documents and queries

Currently identifying important information still mostly relies on occurrence frequency or surface-level features. There still exists huge quality gap between automatic summaries and human-written summaries. Good summaries should contain all semantically important information described in the original documents, rather than those most frequent word sequences. Unfortunately current systems mostly involve no semantic-level processing.

The issue becomes more obvious in guided summarization or query-focused summarization, as current methods mostly make use of shallow calculations of similarities or overlaps between document sentences and query terms without any effort to understand the information needs and response accordingly. Some attempts have been made to explicitly deal with query guidance for news with manually provided category information [141,214], but more general-purpose solutions are still eagerly needed. Meanwhile, discourse parsing is inevitable to explicitly capture the structure of the document, which will be crucial for generating more coherent and more organized summaries as well. Current relevant research mainly exists in single-document summarization as the task reduces to trimming a single discourse tree in some sense. Properly utilizing discourse relations between text units in other scenarios is currently a topic still to be explored.

3.4 End-to-end neural architectures for abstractive summarization

Representation learning based on neural network architectures has proved to be useful in some natural language processing tasks that involve text rewriting, such as machine translation. At the moment, some initial attempts on document summarization have been made for end-to-end training but have yet to achieve real performance gain. Naive RNN encoder–decoder structures currently fail to encode documents, which are way much longer and more structured compared with sentences as input. Better hierarchical encoding and attention with multiple levels on both words and sentences [107] are perhaps needed, along with possible external memory units [174] for storing distant but more significant information. Explicitly designing latent variable structures to capture discourse relations between sentences [84] may also help the document encoding process.

3.5 Summarization at scale

The motivation for automatic summarization is the explosion of information. Current research focuses more on generic summarization on standard benchmarks of news data, with a relatively small number of documents already provided as data source. However, real data sometimes come in stream and may have different formats including news texts and all kinds of user-generated contents [58,146,217]. Most proposed methods for generic summarization may not be trivially adaptable for large-scale streaming data with possible loss of either efficiency or effectiveness. More specific treatments are needed to handle the challenges of events detection, dynamics modeling, contextual dependency, information fusion and credibility assessment.

3.6 Summarization with user interactions

Another research direction is to develop summarization systems that involve user interactions. Different users may have different requirements for summarization systems; hence, certain level of personalization or user interaction is needed. From users' point of view, one may modify his/her queries based on the previous summaries generated from the system. This idea has been studied as a query-chain summarization task [10], where a series of relevant queries are considered, and an update summary is constructed for each query in the chain. Summaries can also be made hierarchical. A user may click on a sentence from a global summary and get to see a more detailed, focused summarization for the point of that sentence [33].

3.7 Multi-modal summarization

Most documents on the Web are not all in the format of texts. They also contain multimedia information such as images, audio or videos. Since the current wave of deep learning approaches has made more significant achievements in visual information processing and audio processing than in natural language processing, utilizing multimedia data sources may also be helpful for learning text representations and thereby be helpful for text summarization. A recent study proposes a joint embedding scheme for images and texts in news for generating multimedia story timelines [192]. We are expected to find more progress from this line of research in the future.

3.8 Summarization for non-factoid question answering

Search engines are currently surpassing the traditional keyword-based document retrieval, providing direct answers for certain kinds of simple factoid questions as queries. However, there are still many types of questions that cannot be answered using simple phrases or one single sentence. These non-factoid questions include definitions, reasons, procedures and opinions. Giving credible, comprehensive answers to these questions requires aggregating and summarizing information from one or many documents. Due to the complexity of these problems, there exists few breakthrough in recent years that surpasses traditional information retrieval systems [52,202]. Hopefully some progress can be made with the development of discourse analysis and natural language understanding.

4 Conclusion

In this paper we survey recent efforts and progress made for document summarization. While many research papers are still focusing on improving extractive summarization from various aspects, there is also a strong emerging favorite toward more abstractive summarization, with compressive summarization being particularly popular as an intermediate step. Also much progress has been made in summarizing under various settings or genres of documents, extending the field beyond traditional news documents and English texts.

Although many papers on document summarization have been published each year, there are still many important issues remaining unsolved and slightly neglected. There exists much space for improvement in almost every aspect, such as the scale of available data, the quality of evaluation, responsiveness to given query or implicit user needs, too much reliance on shallow features (e.g., term frequency) or patterns (e.g., manually written templates) in most current solutions. With the fast development of natural language understanding (semantic parsing), discourse analysis, growth of various kinds of data and data collection platforms, as well as neural network-based representation learning as new powerful modeling tools, new chances for overcoming previous difficulties emerge. We are optimistic that more progress will be witnessed in this field in the near future.

Acknowledgements The work was supported by National Hi-Tech Research and Development Program (863 Program) of China (2015AA015403), National Natural Science Foundation of China (61331011) and IBM Global Faculty Award Program. We would like to thank anonymous reviewers for feedbacks and Jiwei Tan for reporting typos in an earlier draft of this paper.

References

1. Alfonseca E, Pighin D, Garrido G (2013) Heady: news headline abstraction through event pattern clustering. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1243–1253

2. Almeida M, Martins A (2013) Fast and robust compressive summarization with dual decomposition and multi-task learning. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 196–206
3. Ayana, Shen S, Liu Z, Sun M (2016) Neural headline generation with minimum risk training. CoRR abs/1604.01904
4. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representations (ICLR)
5. Bairi R, Iyer R, Ramakrishnan G, Bilmes J (2015) Summarization of multi-document topic hierarchies using submodular mixtures. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 553–563
6. Banerjee S, Mitra P, Sugiyama K (2015) Multi-document abstractive summarization using ilp based multi-sentence compression. In: International joint conference on artificial intelligence
7. Barzilay R, Elhadad M (1999) Using lexical chains for text summarization. Advances in automatic text summarization, pp 111–121
8. Barzilay R, Elhadad N (2002) Inferring strategies for sentence ordering in multidocument news summarization. J Artif Intell Res 17:35–55
9. Barzilay R, McKeown K (2005) Sentence fusion for multidocument news summarization. Comput Linguist 31(3):297–328. doi:10.1162/089120105774321091
10. Baumel T, Cohen R, Elhadad M (2014) Query-chain focused summarization. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Baltimore, pp 913–922
11. Baumel T, Cohen R, Elhadad M (2016) Topic concentration in query focused summarization datasets. In: AAAI Conference on Artificial Intelligence
12. Berg-Kirkpatrick T, Gillick D, Klein D (2011) Jointly learning to extract and compress. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 481–490
13. Bing L, Li P, Liao Y, Lam W, Guo W, Passonneau R (2015) Abstractive multi-document summarization via phrase selection and merging. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 1587–1597
14. Boudin F, Mougard H, Favre B (2015) Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1914–1918
15. Cao Z, Wei F, Dong L, Li S, Zhou M (2015) Ranking with recursive neural networks and its application to multi-document summarization. In: AAAI conference on artificial intelligence
16. Cao Z, Wei F, Li S, Li W, Zhou M, Wang H (2015) Learning summary prior representation for extractive summarization. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers). Association for Computational Linguistics, Beijing, pp 829–833
17. Cao Z, Chen C, Li W, Li S, Wei F, Zhou M (2016) Tgsum: build tweet guided multi-document summarization dataset. In: AAAI conference on artificial intelligence
18. Cao Z, Li W, Li S, Wei F, Li Y (2016) Attsum: Joint learning of focusing and summarization with neural attention. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee. Osaka, pp 547–556
19. Carbonell JG, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, August 24–28, 1998, Melbourne, Australia, pp 335–336. doi:10.1145/290941.291025
20. Carenini G, Cheung JCK, Pauls A (2013) Multi-document summarization of evaluative text. Comput Intell 29(4):545–576. doi:10.1111/j.1467-8640.2012.00417.x
21. Celikyilmaz A, Hakkani-Tur D (2010) A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala, pp 815–824
22. Celikyilmaz A, Hakkani-Tur D (2011) Discovery of topically coherent sentences for extractive summarization. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 491–499
23. Ceylan H, Mihalcea R, Özertem U, Lloret E, Palomar M (2010) Quantifying the limits and success of extractive summarization systems across domains. In: Human language technologies: the 2010 annual

conference of the North American chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Los Angeles, pp 903–911

24. Chakrabarti D, Punera K (2011) Event summarization using tweets. In: International AAAI conference on web and social media
25. Chali Y, Hasan SA (2012) On the effectiveness of using sentence compression models for query-focused multi-document summarization. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 457–474
26. Chan W, Zhou X, Wang W, Chua TS (2012) Community answer summarization for multi-sentence question with group l1 regularization. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Jeju Island, pp 582–591
27. Cheng G, Xu D, Qu Y (2015) Summarizing entity descriptions for effective and efficient human-centered entity linking. In: Proceedings of the 24th international conference on World Wide Web, WWW 2015, Florence, Italy, May 18–22, 2015, pp 184–194. doi:10.1145/2736277.2741094
28. Cheng J, Lapata M (2016) Neural summarization by extracting sentences and words. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 484–494
29. Cheung JCK, Penn G (2013) Towards robust abstractive multi-document summarization: In: A caseframe analysis of centrality and domain. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1233–1242
30. Cheung JCK, Penn G (2014) Unsupervised sentence enhancement for automatic summarization. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 775–786
31. Chopra S, Auli M, Rush AM (2016) Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, San Diego, pp 93–98
32. Christensen J, Mausam Soderland S, Etzioni O (2013) Towards coherent multi-document summarization. In: Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Atlanta, pp 1163–1173
33. Christensen J, Soderland S, Bansal G, Mausam, (2014) Hierarchical summarization: Scaling up multi-document summarization. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Baltimore, pp 902–912
34. Clarke J, Lapata M (2008) Global inference for sentence compression: an integer linear programming approach. J Artif Intell Res 31:399–429. doi:10.1613/jair.2433
35. Cohan A, Goharian N (2015) Scientific article summarization using citation-context and article's discourse structure. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 390–400
36. Cohen WW, Schapire RE, Singer Y (1999) Learning to order things. J Artif Intell Res 10:243–270. doi:10.1613/jair.587
37. Conroy JM, O'Leary DP (2001) Text summarization via hidden markov models. In: SIGIR 2001: proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, September 9–13, 2001, New Orleans, Louisiana, USA, pp 406–407. doi:10.1145/383952.384042
38. Contractor D, Guo Y, Korhonen A (2012) Using argumentative zones for extractive summarization of scientific articles. In: Proceedings of COLING 2012, The COLING 2012 Organizing Committee. Mumbai, India, pp 663–678
39. Das D, Martins AF (2007) A survey on automatic text summarization. Lit Surv Lang Stat II Course CMU 4:192–195
40. Dasgupta A, Kumar R, Ravi S (2013) Summarization through submodularity and dispersion. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1014–1022
41. Davis ST, Conroy JM, Schlesinger JD (2012) Occams–an optimal combinatorial covering algorithm for multi-document summarization. In: 2012 IEEE 12th international conference on data mining workshops. IEEE, pp 454–463

42. Delort JY, Alfonseca E (2012) Dualsum: a topic-model based approach for update summarization. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Avignon, pp 214–223
43. Di Fabbrizio G, Stent A, Gaizauskas R (2014) A hybrid approach to multi-document summarization of opinions in reviews. In: Proceedings of the 8th international natural language generation conference (INLG). Association for Computational Linguistics, Philadelphia, pp 54–63
44. Duan Y, Chen Z, Wei F, Zhou M, Shum HY (2012) Twitter topic summarization by ranking tweets using social influence and content quality. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 763–780
45. Durrett G, Berg-Kirkpatrick T, Klein D (2016) Learning-based single-document summarization with compression and anaphoricity constraints. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 1998–2008
46. Elsner M, Santhanam D (2011) Learning to fuse disparate sentences. In: Proceedings of the workshop on monolingual text-to-text generation. Association for Computational Linguistics, Portland, pp 54–63
47. Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479
48. Fang Y, Teufel S (2014) A summariser based on human memory limitations and lexical competition. In: Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, pp 732–741
49. Fang Y, Teufel S (2016) Improving argument overlap for proposition-based summarisation. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Berlin, pp 479–485
50. Fang Y, Zhu H, Muszyńska E, Kuhnle A, Teufel S (2016) A proposition-based abstractive summariser. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee. Osaka, pp 567–578
51. Filippova K (2010) Multi-sentence compression: Finding shortest paths in word graphs. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, pp 322–330
52. Fried D, Jansen P, Hahn-Powell G, Surdeanu M, Clark P (2015) Higher-order lexical semantic models for non-factoid answer reranking. Trans Assoc Comput Linguist 3:197–210
53. Galanis D, Lampouras G, Androutsopoulos I (2012) Extractive multi-document summarization with integer linear programming and support vector regression. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 911–926
54. Gambhir M, Gupta V (2016) Recent automatic text summarization techniques: a survey. Artif Intell Rev 47:1–66
55. Ganesan K, Zhai C, Han J (2010) Opinosis: a graph based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, pp 340–348
56. Gao D, Li W, Zhang R (2013) Sequential summarization: A new application for timely updated twitter trending topics. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Sofia, pp 567–571
57. Ge T, Pei W, Ji H, Li S, Chang B, Sui Z (2015) Bring you to the past: Automatic generation of topically relevant event chronicles. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 575–585
58. Ge T, Cui L, Chang B, Li S, Zhou M, Sui Z (2016) News stream summarization using burst information networks. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 784–794
59. Genest PE, Lapalme G (2012) Fully abstractive approach to guided summarization. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Jeju Island, pp 354–358
60. Gerani S, Mehdad Y, Carenini G, Ng RT, Nejat B (2014) Abstractive summarization of product reviews using discourse structure. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 1602–1613
61. Gillenwater J, Kulesza A, Taskar B (2012) Discovering diverse and salient threads in document collections. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, Jeju Island, pp 710–720

62. Gillick D, Favre B, Hakkani-Tur D (2008) The ICSI summarization system at TAC 2008. In: Proceedings of the text understanding conference
63. Gillick D, Favre B, Hakkani-Tur D, Bohnet B, Liu Y, Xie S (2009) The ICSI/UTD summarization system at TAC 2009. In: Proceedings of the second text analysis conference. National Institute of Standards and Technology, Gaithersburg
64. Gorinski PJ, Lapata M (2015) Movie script summarization as graph-based scene extraction. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 1066–1076
65. Graham Y (2015) Re-evaluating automatic summarization with bleu and 192 shades of rouge. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 128–137
66. Gu J, Lu Z, Li H, Li VO (2016) Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 1631–1640
67. Gulcehre C, Ahn S, Nallapati R, Zhou B, Bengio Y (2016) Pointing the unknown words. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 140–149
68. Haghighi A, Vanderwende L (2009) Exploring content models for multi-document summarization. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Boulder, pp 362–370
69. He L, Li W, Zhuge H (2016) Exploring differential topic models for comparative summarization of scientific papers. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 1028–1038
70. He Z, Chen C, Bu J, Wang C, Zhang L, Cai D, He X (2012) Document summarization based on data reconstruction. In: AAAI conference on artificial intelligence
71. Hirao T, Yoshida Y, Nishino M, Yasuda N, Nagata M (2013) Single-document summarization as a tree knapsack problem. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, pp 1515–1520
72. Hong K, Nenkova A (2014) Improving the estimation of word importance for news multi-document summarization. In: Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Gothenburg, pp 712–721
73. Hong K, Conroy J, Favre B, Kulesza A, Lin H, Nenkova A (2014) A repository of state of the art and competitive baseline summaries for generic news summarization. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, pp 1608–1616, aCL Anthology Identifier: L14-1070
74. Hong K, Marcus M, Nenkova A (2015) System combination for multi-document summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 107–117
75. Hovy E, Lin CY, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. In: Proceedings of the Fifth conference on language resources and evaluation (LREC 2006), Citeseer, pp 604–611
76. Hu B, Chen Q, Zhu F (2015) Lcsts: A large scale chinese short text summarization dataset. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1967–1972
77. Hu P, Ji D, Teng C, Guo Y (2012) Context-enhanced personalized social summarization. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee, Mumbai, pp 1223–1238
78. Hu Y, Wan X (2015) Ppsgen: Learning-based presentation slides generation for academic papers. IEEE Trans Knowl Data Eng 27(4):1085–1097. doi:10.1109/TKDE.2014.2359652
79. Huang X, Wan X, Xiao J (2011) Comparative news summarization using linear programming. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 648–653
80. Iyer S, Konstas I, Cheung A, Zettlemoyer L (2016) Summarizing source code using a neural attention model. In: Proceedings of the 54th annual meeting of the Association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 2073–2083
81. Jayanth J, Sundararaj J, Bhattacharyya P (2015) Monotone submodularity in opinion summaries. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 169–178

82. Jha R, Finegan-Dollak C, King B, Coke R, Radev D (2015) Content models for survey generation: a factoid-based evaluation. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 441–450
83. Ji H, Favre B, Lin WP, Gillick D, Hakkani-Tur D, Grishman R (2013) Open-domain multi-document summarization via information extraction: challenges and prospects. In: Poibeau T, Saggion H, Piskorski J, Yangarber R (eds) Multi-source, multilingual information extraction and summarization. Springer, Berlin, pp 177–201
84. Ji Y, Haffari G, Eisenstein J (2016) A latent variable recurrent neural network for discourse-driven language models. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, San Diego, pp 332–342
85. Judd J, Kalita J (2013) Better twitter summaries? In: Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Atlanta, pp 445–449
86. Kedzie C, McKeown K, Diaz F (2015) Predicting salient updates for disaster summarization. In: Proceedings of the 53rd annual meeting of the Association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 1608–1617
87. Kedzie C, Diaz F, McKeown K (2016) Real-time web scale event summarization using sequential decision making. In: International joint conference on artificial intelligence, pp 3754–3760
88. Kikuchi Y, Hirao T, Takamura H, Okumura M, Nagata M (2014) Single document summarization based on nested tree structure. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Baltimore, pp 315–320
89. Kikuchi Y, Neubig G, Sasano R, Takamura H, Okumura M (2016) Controlling output length in neural encoder-decoders. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 1328–1338
90. Kim HD, Ganesan K, Sondhi P, Zhai CX (2011) Comprehensive review of opinion summarization. UIUC Technical Report, USA
91. Kobayashi H, Noguchi M, Yatsuka T (2015) Summarization based on embedding distributions. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1984–1989
92. Kågebäck M, Mogren O, Tahmasebi N, Dubhashi D (2014) Extractive summarization using continuous vector space models. In: Proceedings of the 2nd workshop on continuous vector space models and their compositionality (CVSC). Association for Computational Linguistics, Gothenburg, pp 31–39
93. Kulesza A, Taskar B (2011) Learning determinantal point processes. In: Proceedings of the 27th conference on uncertainty in artificial intelligence
94. Kulesza A, Taskar B (2012) Determinantal point processes for machine learning. Found Trends Mach Learn 5(2–3):123–286
95. Lei T, Barzilay R, Jaakkola T (2016) Rationalizing neural predictions. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 107–117
96. Li C, Liu F, Weng F, Liu Y (2013) Document summarization via guided sentence compression. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, pp 490–500
97. Li C, Qian X, Liu Y (2013) Using supervised bigram-based ilp for extractive summarization. In: Proceedings of the 51st Annual Meeting of the Association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1004–1013
98. Li C, Liu Y, Liu F, Zhao L, Weng F (2014) Improving multi-documents summarization by sentence compression based on expanded constituent parse trees. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 691–701
99. Li C, Liu Y, Zhao L (2015) Improving update summarization via supervised ilp and sentence reranking. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 1317–1322
100. Li C, Liu Y, Zhao L (2015) Using external resources and joint learning for bigram weighting in ilp-based multi-document summarization. In: Proceedings of the 2015 conference of the North American chapter of the Association for computational linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 778–787

101. Li C, Wei Z, Liu Y, Jin Y, Huang F (2016) Using relevant public posts to enhance news article summarization. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee. Osaka, pp 557–566
102. Li J, Cardie C (2014) Timeline generation: tracking individuals on twitter. In: 23rd international world wide web conference, WWW '14, Seoul, Republic of Korea, April 7–11, 2014, pp 643–652. doi:10.1145/2566486.2567969
103. Li J, Li S (2013) Evolutionary hierarchical dirichlet process for timeline summarization. In: Proceedings of the 51st annual meeting of the Association for Computational linguistics (volume 2: short papers). Association for Computational Linguistics, Sofia, pp 556–560
104. Li J, Li S (2013) A novel feature-based bayesian model for query focused multi-document summarization. Trans Assoc Comput Linguist 1:89–98
105. Li J, Li S, Wang X, Tian Y, Chang B (2012) Update summarization using a multi-level hierarchical dirichlet process model. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 1603–1618
106. Li J, Gao W, Wei Z, Peng B, Wong KF (2015) Using content-level structures for summarizing microblog repost trees. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 2168–2178
107. Li J, Luong T, Jurafsky D (2015) A hierarchical neural autoencoder for paragraphs and documents. In: Proceedings of the 53rd annual meeting of the Association for Computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 1106–1115
108. Li JJ, Thadani K, Stent A (2016) The role of discourse units in near-extractive summarization. In: Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, Los Angeles, pp 137–147
109. Li L, Zhou K, Xue G, Zha H, Yu Y (2009) Enhancing diversity, coverage and balance for summarization through structure learning. In: Proceedings of the 18th international conference on world wide web, WWW 2009, Madrid, Spain, April 20–24, 2009, pp 71–80. doi:10.1145/1526709.1526720
110. Li P, Bing L, Lam W, Li H, Liao Y (2015) Reader-aware multi-document summarization via sparse coding. In: International joint conference on artificial intelligence
111. Li Y, Li S (2014) Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. Dublin City University and Association for Computational Linguistics, Dublin, pp 1197–1207
112. Liakata M, Dobnik S, Saha S, Batchelor C, Rebholz-Schuhmann D (2013) A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, pp 747–757
113. Lin CY (2003) Improving summarization performance by sentence compression—a pilot study. In: Proceedings of the sixth international workshop on information retrieval with Asian languages. Association for Computational Linguistics, Sapporo, pp 1–8
114. Lin CY, Hovy E (2000) The automated acquisition of topic signatures for text summarization. In: Proceedings of the 18th conference on computational linguistics—volume 1. Association for Computational Linguistics, pp 495–501
115. Lin CY, Hovy E (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on human language technology—volume 1. Association for Computational Linguistics, pp 71–78
116. Lin H, Bilmes J (2010) Multi-document summarization via budgeted maximization of submodular functions. In: Human language technologies: the 2010 annual conference of the North American chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Los Angeles, pp 912–920
117. Lin H, Bilmes J (2011) A class of submodular functions for document summarization. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 510–520
118. Lin H, Bilmes JA (2012) Learning mixtures of submodular shells with application to document summarization. In: Proceedings of the 28th conference on uncertainty in artificial intelligence
119. Litvak M, Last M (2013) Cross-lingual training of summarization systems using annotated corpora in a foreign language. Inf Retr 16(5):629–656. doi:10.1007/s10791-012-9210-3
120. Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167

121. Liu F, Flanigan J, Thomson S, Sadeh N, Smith NA (2015) Toward abstractive summarization using semantic representations. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 1077–1086
122. Liu H, Yu H, Deng ZH (2015) Multi-document summarization based on two-level sparse representation model. In: AAAI conference on artificial intelligence
123. Liu X, Li Y, Wei F, Zhou M (2012) Graph-based multi-tweet summarization using social signals. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee, Mumbai, pp 1699–1714
124. Liu Y, hua Zhong S, Li W (2012) Query-oriented multi-document summarization via unsupervised deep learning. In: AAAI conference on artificial intelligence, pp 1699–1705
125. Lloret E, Palomar M (2013) Towards automatic tweet generation: a comparative study from the text summarization perspective in the journalism genre. Expert Syst Appl 40(16):6624–6630. doi:10.1016/j.eswa.2013.06.021
126. Louis A, Nenkova A (2013) Automatically assessing machine summary content without a gold standard. Comput Linguist 39(2):267–300
127. Loza V, Lahiri S, Mihalcea R, Lai PH (2014) Building a dataset for summarization and keyword extraction from emails. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, pp 2441–2446, aCL Anthology Identifier: L14-1028
128. Luo W, Litman D (2015) Summarizing student responses to reflection prompts. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1955–1960
129. Ma S, Deng ZH, Yang Y (2016) An unsupervised multi-document summarization framework based on neural document model. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 1514–1523
130. Mann WC, Thompson SA (1988) Rhetorical structure theory: toward a functional theory of text organization. Text Interdiscip J Study Discourse 8(3):243–281
131. McDonald RT (2007) A study of global inference algorithms in multi-document summarization. In: Advances in information retrieval, 29th European conference on IR research, ECIR 2007, Rome, Italy, April 2–5, 2007, proceedings, pp 557–564
132. Metzler D, Kanungo T (2008) Machine learned sentence selection strategies for query-biased summarization. In: SIGIR learning to rank workshop, pp 40–47
133. Mihalcea R, Tarau P (2004) Textrank: bringing order into texts. In: Lin D, Wu D (eds) Proceedings of EMNLP 2004. Association for Computational Linguistics, Barcelona, pp 404–411
134. Morita H, Sasano R, Takamura H, Okumura M (2013) Subtree extractive summarization via submodular maximization. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1023–1032
135. Nallapati R, Zhou B, glar Gulcehre C, Xiang B, (2016) Abstractive text summarization using sequence-to-sequence rnns and beyond. In: Proceedings of the 20th SIGNLL conference on computational natural language learning. Association for Computational Linguistics, Berlin, pp 280–290
136. Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functionsi. Math Program 14(1):265–294
137. Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Aggarwal CC, Zhai CX (eds) Mining text data. Springer, Berlin, pp 43–76
138. Nenkova A, Passonneau R (2004) Evaluating content selection in summarization: the pyramid method. In: Susan Dumais DM, Roukos S (eds) HLT-NAACL 2004: main proceedings. Association for Computational Linguistics, Boston, pp 145–152
139. Nenkova A, McKeown K et al (2011) Automatic summarization. Found Trends Inf Retr 5(2–3):103–233
140. Ng JP, Abrecht V (2015) Better summarization evaluation with word embeddings for rouge. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1925–1930
141. Ng JP, Bysani P, Lin Z, Kan MY, Tan CL (2012) Exploiting category-specific information for multi-document summarization. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 2093–2108
142. Ng JP, Chen Y, Kan MY, Li Z (2014) Exploiting timelines to enhance multi-document summarization. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Baltimore, pp 923–933
143. Nichols J, Mahmud J, Drews C (2012) Summarizing sporting events using twitter. In: Proceedings of the 2012 ACM international conference on intelligent user interfaces. ACM, pp 189–198

144. Nishikawa H, Arita K, Tanaka K, Hirao T, Makino T, Matsuo Y (2014) Learning to generate coherent summary with discriminative hidden semi-markov model. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. Dublin City University and Association for Computational Linguistics, Dublin, pp 1648–1659
145. Nishino M, Yasuda N, Hirao T, Si Minato, Nagata M (2015) A dynamic programming algorithm for tree trimming-based text summarization. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 462–471
146. Olariu A (2014) Efficient online summarization of microblogging streams. In: Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics, volume 2: short papers. Association for Computational Linguistics, Gothenburg, pp 236–240
147. Owczarzak K, Conroy JM, Dang HT, Nenkova A (2012) An assessment of the accuracy of automatic evaluation in summarization. In: Proceedings of workshop on evaluation metrics and system comparison for automatic summarization. Association for Computational Linguistics, Montréal, pp 1–9
148. Oya T, Mehdad Y, Carenini G, Ng R (2014) A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In: Proceedings of the 8th international natural language generation conference (INLG). Association for Computational Linguistics, Philadelphia, pp 45–53
149. Parveen D, Strube M (2015) Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In: International joint conference on artificial intelligence
150. Parveen D, Ramsl HM, Strube M (2015) Topical coherence for graph-based extractive summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1949–1954
151. Passonneau RJ, Chen E, Guo W, Perin D (2013) Automated pyramid scoring of summaries using distributional semantics. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Sofia, pp 143–147
152. Pei Y, Yin W, Fan Q, Huang L (2012) A supervised aggregation framework for multi-document summarization. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee. Mumbai, pp 2225–2242
153. Peyrard M, Eckle-Kohler J (2016) Optimizing an approximation of rouge - a problem-reduction approach to extractive multi-document summarization. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 1825–1836
154. Pighin D, Cornolti M, Alfonseca E, Filippova K (2014) Modelling events through memory-based, open-ie patterns for abstractive summarization. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Baltimore, pp 892–901
155. Qazvinian V, Radev DR, Mohammad S, Dorr BJ, Zajic DM, Whidby M, Moon T (2013) Generating extractive summaries of scientific paradigms. J Artif Intell Res 46:165–201. doi:10.1613/jair.3732
156. Qian X, Liu Y (2013) Fast joint compression and summarization via graph cuts. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, pp 1492–1502
157. Radev DR, Jing H, Sty M, Tam D (2004) Centroid-based summarization of multiple documents. Inf Process Manag 40(6):919–938. doi:10.1016/j.ipm.2003.10.006
158. Rankel PA, Conroy JM, Dang HT, Nenkova A (2013) A decade of automatic content evaluation of news summaries: reassessing the state of the art. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 2: short papers). Association for Computational Linguistics, Sofia, pp 131–136
159. Ranzato M, Chopra S, Auli M, Zaremba W (2016) Sequence level training with recurrent neural networks. In: International conference on learning representations (ICLR)
160. Ren P, Wei F, CHEN Z, MA J, Zhou M (2016) A redundancy-aware sentence regression framework for extractive summarization. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 33–43
161. Ren Z, de Rijke M (2015) Summarizing contrastive themes via hierarchical non-parametric processes. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Chile, August 9–3, 2015, pp 93–102. doi:10.1145/2766462.2767713
162. Rioux C, Hasan SA, Chali Y (2014) Fear the reaper: A system for automatic multi-document summarization with reinforcement learning. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 681–690

163. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. Science 344(6191):1492–1496
164. Ross S, Zhou J, Yue Y, Dey D, Bagnell D (2013) Learning policies for contextual submodular prediction. In: Proceedings of the 30th international conference on machine learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013, pp 1364–1372
165. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 379–389
166. Saggion H (2013) Unsupervised learning summarization templates from concise summaries. In: Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Atlanta, pp 270–279
167. Schluter N, Søgaard A (2015) Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers). Association for Computational Linguistics, Beijing, pp 840–844
168. Sharifi B, Hutton MA, Kalita J (2010) Summarizing microblogs automatically. In: Human language technologies: the 2010 annual conference of the North American chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Los Angeles, pp 685–688
169. Shen C, Li T (2011) Learning to rank for query-focused multi-document summarization. In: 2011 IEEE 11th international conference on data mining (ICDM). IEEE, pp 626–634
170. Shen D, Sun JT, Li H, Yang Q, Chen Z (2007) Document summarization using conditional random fields. In: International joint conference on artificial intelligence, vol 7, pp 2862–2867
171. Sidhaye P, Cheung JCK (2015) Indicative tweet generation: an extractive summarization problem? In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 138–147
172. Sipos R, Shivaswamy P, Joachims T (2012) Large-margin learning of submodular summarization models. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Avignon, pp 224–233
173. Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems. Curran Associates, Inc., Lake Tahoe, Nevada, pp 2951–2959
174. Sukhbaatar S, Szlam A, Weston J, Fergus R (2015) End-to-end memory networks. Adv Neural Inf Process Syst 28:2440–2448
175. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 27:3104–3112
176. Swisher K (2013) Yahoo paid $30 million in cash for 18 months of young summly entrepreneur's time. http://allthingsd.com/20130325/yahoo-paid-30-million-in-cash-for-18-months-of-young-summly-entrepreneurs-time/. Accessed 30 Dec 2016
177. Takamura H, Yokono H, Okumura M (2011) Summarizing a document stream. In: Advances in information retrieval—33rd European conference on IR research, ECIR 2011, Dublin, Ireland, April 18–21, 2011. Proceedings, pp 177–188
178. Thadani K, McKeown K (2013) Supervised sentence fusion with single-stage inference. In: Proceedings of the sixth international joint conference on natural language processing. Asian Federation of Natural Language Processing, Nagoya, pp 1410–1418
179. Toutanova K, Brockett C, Tran KM, Amershi S (2016) A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 340–350
180. Tran G, Herder E, Markert K (2015) Joint graphical models for date selection in timeline summarization. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Beijing, pp 1598–1607
181. Trione J, Favre B, Béchet F (2016) Beyond utterance extraction: summary recombination for speech summarization. Interspeech 2016:680–684
182. Vanderwende L, Suzuki H, Brockett C, Nenkova A (2007) Beyond sumbasic: task-focused summarization with sentence simplification and lexical expansion. Inf Process Manag 43(6):1606–1618. doi:10.1016/j.ipm.2007.01.023
183. Wan X (2011) Using bilingual information for cross-language document summarization. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Portland, pp 1546–1555

184. Wan X (2012) Update summarization based on co-ranking with constraints. In: Proceedings of COLING 2012: posters. The COLING 2012 Organizing Committee, Mumbai, pp 1291–1300
185. Wan X, Zhang J (2014) CTSUM: extracting more certain summaries for news articles. In: The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, Gold Coast , QLD, Australia, July 06–11, 2014, pp 787–796. doi:10.1145/2600428.2609559
186. Wang D, Li T (2012) Weighted consensus multi-document summarization. Inf Process Manag 48(3):513–523
187. Wang D, Zhu S, Li T, Gong Y (2013) Comparative document summarization via discriminative sentence selection. TKDD 7(1):21–218. doi:10.1145/2435209.2435211
188. Wang L, Cardie C (2013) Domain-independent abstract generation for focused meeting summarization. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1395–1405
189. Wang L, Ling W (2016) Neural network-based abstract generation for opinions and arguments. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, San Diego, pp 47–57
190. Wang L, Raghavan H, Castelli V, Florian R, Cardie C (2013) A sentence compression based framework to query-focused multi-document summarization. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Sofia, pp 1384–1394
191. Wang L, Raghavan H, Cardie C, Castelli V (2014) Query-focused opinion summarization for user-generated content. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. Dublin City University and Association for Computational Linguistics, Dublin, pp 1660–1669
192. Wang WY, Mehdad Y, Radev DR, Stent A (2016) A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, San Diego, pp 58–68
193. Wang X, Yoshida Y, Hirao T, Sudoh K, Nagata M (2015) Summarization based on task-oriented discourse parsing. IEEE/ACM Trans Audio Speech Lang Process 23(8):1358–1367. doi:10.1109/TASLP.2015.2432573
194. Wang X, Nishino M, Hirao T, Sudoh K, Nagata M (2016) Exploring text links for coherent multi-document summarization. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 213–223
195. Woodsend K, Lapata M (2012) Multiple aspect summarization using integer linear programming. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, Jeju Island, pp 233–243
196. Xiong W, Litman D (2014) Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers. Dublin City University and Association for Computational Linguistics, Dublin, pp 1985–1995
197. Xu H, Martin E, Mahidadia A (2015) Extractive summarisation based on keyword profile and language model. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 123–132
198. Yan R, Kong L, Huang C, Wan X, Li X, Zhang Y (2011) Timeline generation through evolutionary trans-temporal summarization. In: Proceedings of the 2011 conference on empirical methods in natural language processing. Association for Computational Linguistics, Edinburgh, pp 433–443
199. Yan R, Wan X, Otterbacher J, Kong L, Li X, Zhang Y (2011) Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In: Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2011, Beijing, China, July 25–29, 2011, pp 745–754, doi:10.1145/2009916.2010016
200. Yan R, Jiang H, Lapata M, Lin SD, Lv X, Li X (2013) I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In: Proceedings of the twenty-third international joint conference on artificial intelligence. AAAI Press, pp 2197–2203
201. Yan S, Wan X (2014) Srrank: leveraging semantic roles for extractive multi-document summarization. IEEE/ACM Trans Audio Speech Lang Process 22(12):2048–2058

202. Yang L, Ai Q, Spina D, Chen RC, Pang L, Croft WB, Guo J, Scholer F (2016) Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In: European conference on information retrieval, Springer, Berlin pp 115–128
203. Yang Z, Cai K, Tang J, Zhang L, Su Z, Li J (2011) Social context summarization. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 255–264
204. Yao J, Wan X, Xiao J (2015) Compressive document summarization via sparse optimization. In: International joint conference on artificial intelligence
205. Yao J, Wan X, Xiao J (2015) Phrase-based compressive cross-language summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 118–127
206. Yin W, Pei Y (2015) Optimizing sentence modeling and selection for document summarization. In: International joint conference on artificial intelligence
207. Yogatama D, Liu F, Smith NA (2015) Extractive summarization by maximizing semantic volume. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1961–1966
208. Yoshida Y, Suzuki J, Hirao T, Nagata M (2014) Dependency-based discourse parser for single-document summarization. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 1834–1839
209. You O, Li W, Li S, Lu Q (2011) Applying regression models to query-focused multi-document summarization. Inf Process Manag 47(2):227–237. doi:10.1016/j.ipm.2010.03.005
210. Yu N, Huang M, Shi Y, zhu x, (2016) Product review summarization by exploiting phrase properties. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 1113–1124
211. Zajic DM, Dorr B, Lin J, Schwartz R (2006) Sentence compression as a component of a multi-document summarization system. In: Proceedings of the 2006 document understanding workshop, New York
212. Zhang J, Yao J, Wan X (2016a) Towards constructing sports news from live text commentary. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Berlin, pp 1361–1371
213. Zhang J, Zhou Y, Zong C (2016b) Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. IEEE/ACM Trans Audio Speech Lang Process 24(10):1842–1853
214. Zhang R, Li W, Gao D (2013) Towards content-level coherence with aspect-guided summarization. TSLP 10(1):2:1–2:22. doi:10.1145/2442076.2442078
215. Zhang Y, Xia Y, Liu Y, Wang W (2015) Clustering sentences with density peaks for multi-document summarization. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Association for Computational Linguistics, Denver, pp 1262–1267
216. Zhao WX, Guo Y, Yan R, He Y, Li X (2013) Timeline generation with social attention. In: The 36th international ACM SIGIR conference on research and development in information retrieval, SIGIR ’13, Dublin, Ireland, July 28–August 01, 2013, pp 1061–1064. doi:10.1145/2484028.2484103
217. Zopf M, Loza Mencía E, Fürnkranz J (2016) Sequential clustering and contextual importance measures for incremental update summarization. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, pp 1071–1082
218. Zopf M, Mencıa EL, Fürnkranz J (2016b) Beyond centrality and structural features: Learning information importance for text summarization. In: Proceedings of the 20th SIGNLL conference on computational natural language learning. Association for Computational Linguistics, Berlin, pp 84–94

Jin-ge Yao is currently a Ph.D. student at Institute of Computer Science and Technology of Peking University. He received the B.S. degree from Chongqing University in 2011. His research interests include natural language processing and text mining.

Xiaojun Wan is currently a professor at Institute of Computer Science and Technology of Peking University. He received his B.S., M.S. and Ph.D. degrees from Peking University in 2000, 2003 and 2006, respectively. His research interests include natural language processing and text mining. He has published more than 60 publications in major international conferences and journals, including ACL, SIGIR, AAAI, IJCAI, COLING, EMNLP, ICDM, CIKM, ACM TOIS, Computational Linguistics, JASIST, KAIS and Information Sciences. He has served as PC member or area chair of major conferences such as ACL, SIGIR, EMNLP, COLING, CIKM, IJCNLP and NLPCC.

Jianguo Xiao is a professor at Institute of Computer Science and Technology of Peking University. He received the M.S. degree in Department of Computer Science and Technology of Peking University in 1990. His research interests include natural language processing and Chinese computing.