Published in final edited form as:
J Biomed Inform. 2014 December ; 0: 457–467. doi:10.1016/j.jbi.2014.06.009.

Text Summarization in the Biomedical Domain: A Systematic Review of Recent Research

Rashmi Mishra, BDS, MPH[1], **Jiantao Bian, M.S**[1,2], **Marcelo Fiszman, MD, PhD**[3], **Charlene R. Weir, B.S., Ph.D., R.N**[1,4], **Siddhartha Jonnalagadda, PhD**[5], **Javed Mostafa, PhD**[6], and **Guilherme Del Fiol, MD, PhD**[1]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

[2]Clinical Modeling Team, Intermountain Healthcare, Salt Lake City, Utah, USA

[3]Lister Hill Center, National Library of Medicine, Bethesda, MD, USA

[4]VA Medical Center, Salt Lake City, UT

[5]Department of Preventive Medicine-Health and Biomedical Informatics, Northwestern University, Chicago, IL

[6]School of Information and Library Science (SILS), University of North Carolina, Chapel Hill, NC

Abstract

Objective—The amount of information for clinicians and clinical researchers is growing exponentially. Text summarization reduces information as an attempt to enable users to find and understand relevant source texts more quickly and effortlessly. In recent years, substantial research has been conducted to develop and evaluate various summarization techniques in the biomedical domain. The goal of this study was to systematically review recent published research on summarization of textual documents in the biomedical domain.

Materials and methods—MEDLINE (2000 to October 2013), IEEE Digital Library, and the ACM Digital library were searched. Investigators independently screened and abstracted studies that examined text summarization techniques in the biomedical domain. Information is derived from selected articles on five dimensions: *input*, *purpose*, *output*, *method* and *evaluation*.

Results—Of 10,786 studies retrieved, 34 (0.3%) met the inclusion criteria. Natural Language processing (17; 50%) and a Hybrid technique comprising of statistical, Natural language processing and machine learning (15; 44%) were the most common summarization approaches. Most studies (28; 82%) conducted an intrinsic evaluation.

Discussion—This is the first systematic review of text summarization in the biomedical domain. The study identified research gaps and provides recommendations for guiding future research on biomedical text summarization.

Corresponding Author: Guilherme Del Fiol, MD, PhD, Assistant Professor, Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT, USA-84108, Phone: +1(801) 213-4129/Fax: +1(801) 581-4297, guilherme.delfiol@utah.edu.

conclusion—Recent research has focused on a Hybrid technique comprising statistical, language processing and machine learning techniques. Further research is needed on the application and evaluation of text summarization in real research or patient care settings.

Keywords

text summarization; intrinsic evaluation; language Processing; machine learning; biomedical domain

1. Introduction

The amount of information available for clinicians and clinical researchers is growing exponentially, both in the biomedical literature and patients' health records.(1, 2) To provide optimal patient care, clinicians need to efficiently and effectively retrieve, interpret, and integrate relevant information from multiple source.(2) Likewise, researchers need to navigate a vast amount of information from the biomedical literature for tasks such as generating new hypotheses and understanding the state-of-the-art in a given area. Electronic resources such as online literature databases and electronic health record (EHR) systems have been designed to help clinicians and researchers with their information management needs. However, the more resources grow, the harder it becomes for users to access information efficiently. Advances in information retrieval technology have shown some value in helping clinicians manage information overload.(3) Yet, information seekers often need to screen several documents and scan several pages of narrative content to find information that is relevant to their information needs.(2)

Automatic text summarization is a promising method for helping clinicians and researchers seeking information to efficiently obtain the “gist” in a given topic by producing a textual or graphical summary from one or multiple documents. A summary is “a reductive transformation of source text to summary text through content reduction selection and/or generalization on what is important in the source.”(4) The goal of text summarization is to present a subset of the source text, which expresses the most important points with minimal redundancy. The reduction of data accomplished by text summarization aims to allow users to identify and process relevant information more quickly and accurately. Thus, text summarization may become an important tool to assist clinicians and researchers with their information and knowledge management tasks.

Important advances have been achieved recently in text summarization. As a result, several applications that leverage text summarization techniques have become available to the general public.(5) There has been a growing interest in researching text summarization techniques in the biomedical domain. An informal literature survey conducted by Afantenos et al. identified ten biomedical text summarization studies published between 1999 and 2003.(6) Since then, there have been significant advances in the summarization tools and techniques employed in the biomedical domain. However, no systematic review on this topic has been conducted to date. A systematic review will promote improved understanding of the literature on this topic, identify gaps, and provide directions for future research. In the present study, we conducted a systematic review on text summarization methods applied to the biomedical literature and EHR systems. The systematic review is aimed at: 1)

identifying the different techniques, areas of application, and evaluation methods over the last decade; 2) identifying research trends; 3) identifying research gaps; and 4) proposing recommendations to guide future research.

2. Methods

We based the methodology of our study on the *Standards for Systematic Reviews* set by the Institute of Medicine.(7) The study protocol was iteratively designed and refined with input from the study co-authors. The following subsections describe each of the steps that were performed to identify, screen, and abstract data form the included studies.

2.1 Data Sources and Searches

The search strategies were developed with the help of the expert review committee and a medical librarian. The strategies were further tested and refined against a list of relevant citations from previous reviews on the topic. Three databases were searched: PubMed, IEEE, and ACM digital library. Searches were limited to the period between Jan 1st 2000 and October 16th 2013. The overall search strategy was to retrieve articles that included terms related to *text summarization*, such as *"medical text summarization", "clinical text summarization", and "biomedical summarization"*. The search time period was limited to avoid overlap with the review by Afantenos et al.(6) The search strategies applied are provided in the online supplement. In addition to searching literature databases, we inspected the citations of included articles with a special focus on previous relevant reviews. Finally, we requested input from the study co-authors for potentially relevant references that could have been missed by the literature search.

2.2 Study Selection

We included original research studies that developed and evaluated text summarization methods in the medical domain, including summarization of the biomedical literature and electronic health record documents.

We excluded studies that met any of the following criteria: 1) Summarization of content outside the biomedical domain; 2) summarization of the basic science literature, such as molecular biology; 3) not original research, such as editorials and opinion papers; 4) emphasis placed on text summarization tools, but without an evaluation component; 5) related techniques (e.g., text mining) that can be used to support text summarization, but that did not produce a summary; 6) not written in English; 7) image and multimedia summarization without a text summarization component; and 8) articles included in the survey by S. Afantenos et al.(6)

2.2.1 Abstract screening—The title and abstract of each article retrieved were reviewed independently by two of the study authors (JB, RM). Articles were labeled as "not relevant" or "potentially relevant." For calibration and refinement of the inclusion and exclusion criteria, 50 citations were randomly selected and independently reviewed. Disagreements were resolved by consensus with a third author (GDF). In a second round, another set of 50 articles was reviewed in a similar way. In a third round, 815 abstracts were independently

reviewed achieving a strong level of agreement (kappa=0.82). In a final round the remaining citations (7,871) were evenly assigned between the two reviewers and screened.

2.2.2 Article selection—Two authors (JB, RM) independently reviewed the full-text of a subset of 112 citations labeled as potentially relevant in the abstract screening phase. Disagreements between the two reviewers were reconciled with the help of a third reviewer (GDF). Since inter-rater agreement in this phase was high (kappa=0.78), the remaining full-text articles (120) were evenly assigned between the two reviewers and screened.

2.3 Data Extraction

A data abstraction spreadsheet was developed based on the text summarization categories described by Mani which are summarized below.(8) Two authors (RM, JB) independently reviewed the included articles (34) to extract the data into the data abstraction spreadsheet. Next, the data were compared and disagreements were reconciled through consensus with the assistance of a third reviewer (GDF).

The data abstraction tool was adapted from a classification of text summarization methods described by Mani et al.(9) This classification consists of five dimensions: *input*, *purpose*, *output*, *method* and *evaluation*. The five classification categories are further described below.

2.3.1 Input—This dimension has been termed as “unit input parameter” or the “span parameter” by Sparck- Jones and Mani respectively. (4, 8) We categorized the Input dimension according to four attributes: 1) *single* versus *multiple* document summarizations; 2) *monolingual* (input and output on the same language) vs. *multilingual* summarization (input or output in multiple languages; 3) *abstract* versus *full- text*; 4) *biomedical research literature* versus *EHR documents*.

2.3.2 Purpose—*Purpose* denotes the stated main goal of the generated summary. This dimension was categorized according to two attributes: 1) *Generic* versus *user-oriented* summaries; and 2) *Broad spectrum* versus *Clinical decision support*.

Generic summaries take a predefined document or set of documents and produce a summary for these documents. *User-oriented* summaries are produced to address a user's specific information need. Typically, a user-oriented summary starts with a query submitted by a user and produces a summary that attempts to answer that query. *Broad spectrum* summaries could be used to support activities such as research and patient care, while *clinical* summaries aim specifically at helping clinicians' patient care decisions.

2.3.3 Output—The output of a summarization system may include information presented in a number of ways. We classified summarization output as *extract* versus *abstract* and *indicative* versus *informative* summaries. An *extractive* summary contains verbatim fragments from input document(s) while an *abstractive summary* produces new content inferred from the input documents. *Indicative* summaries provide users with an idea of the content available in the input source. Users still need to retrieve the input content for

understanding. *Informative* summaries contain complete enough content, so that users do not need to access the original input for understanding.

2.3.4 Method—There are a variety of text summarization approaches. In the present study, we classified the methods into four broad categories: *statistical, natural language processing, machine learning, and hybrid technique*. Statistical techniques are typically based on the Edmundsonian paradigm (10) where sentences are ranked based on a formula, which assigns a score to each sentence based on various factors such as cue phrases, keywords, and sentence location in the document. Unlike machine learning, methods that fall in the statistical category encompass manual design of the mathematical formulas used to calculate sentence scores. For example, Sarkar et al. combined several domain specific features such as term frequency, title and position and used a mathematical formula to produce extractive summaries in the medical domain.(11)

Natural Language processing techniques includes computational methods applied to understand human languages in a similar manner as it is processed in spoken and written medium.(12) This includes everything from simple applications like word counting to robust parsing. For the purpose of our study, we included studies that applied text processing, including steps such as extraction of lexical knowledge, lexical and structural disambiguation (e.g., part of speech tagging, word sense disambiguation), grammatical inference, and robust parsing. One example in this category is the work by Reeve et al. where summaries are produced by linking semantically related concepts in multiple sentences.(13) In our study, text mining methods purely based on machine learning techniques, such as supervised learning, were classified as machine learning methods as opposed to natural language processing. Machine learning methods produce summaries based on automated learning of logic from text corpora. For example, Chaung et al. used a supervised learning algorithm to train the summarizer to extract important sentence segments based on feature vectors.(14)

Hybrid methods employ two or more of the methods described above. Many studies in our review applied a hybrid technique for text summarization. For example, Plaza et al. used a combination of natural language processing and machine learning methods to generate extractive summaries. The algorithm aims to identify salient sentences in biomedical texts. They identified concepts and relations which were derived from the Unified Medical language Systems (UMLS) to construct a sematic graph and then applied a clustering algorithm to identify different themes and topics within the text to extract salient sentences for summarization.(15)

2.3.5 Evaluation—The evaluation of summaries has been broadly classified into two categories: *Intrinsic* and *extrinsic* methods.(16) *Intrinsic* evaluation methods assess the quality of the summarization output according to certain criteria, such as readability, comprehensiveness, accuracy, and relevancy. Output summaries are often rated by users or compared with a gold standard, typically hand-crafted by humans. *Extrinsic* methods assess the impact of a summarization system on specific information-seeking task performance based on measures such as success rate, time-to-completion, and decision-making accuracy.

3 Results

3.1 Description of studies

Of 10,786 unique citations retrieved, 232 were selected for full-text screening and 34 articles met the study criteria.(Figure 1) Agreement on abstract screening in the first, second, and third rounds was 74% (kappa=0.54), 88% (kappa=0.74), and 92% (kappa=0.82) respectively. Agreement on the full- text screening was 84% (kappa=0.78).

A list of the included studies along with their characteristics and description is available as part of the online supplement. Table 1 provides frequency of studies according to the data abstraction dimensions. Nineteen studies (56%) processed multiple documents. None of the studies consisted of multilingual summarization systems. Most studies used full-text articles (19; 56%) and the biomedical literature (31; 91%) as input for summarization. Sixteen studies (47%) produced a user-oriented summary and nineteen (56%) produced summaries for clinical decision support. The majority of the studies produced extractive summaries (23; 76%) and informative summaries (25; 74%). Natural Language processing (17; 50%) and combined methods (15; 44%) were the most common summarization approaches. One study was focused on usability evaluation of summarization systems. (17) Twenty-eight studies (82%) conducted an intrinsic evaluation. Tables 2 and 3 provide a list of the included studies along with their characteristics and description.

4. Discussion

According to our findings, this is the first systematic review of text summarization in the biomedical domain. The data abstraction was guided by a widely used framework for categorizing text summarization methods, which allowed comparison with a previous literature survey and examination of the current state-of-the-art in this field.(8, 9) Finally, the study identified research gaps and provides recommendations for guiding future research on biomedical text summarization.

4.1 State-of-the-art

Our review found several trends in biomedical text summarization research. First, research has shifted from a strong focus on single document summarization to both single and multi-document summarization. Multiple document summarizations are especially important in more recent times due to the exponential growth in the published scientific literature and the increasing popularity of the evidence-based medicine movement. In addition, relevant information is often distributed among multiple documents, such as clinical studies published in the primary literature and clinical notes in a patient's EHR. For integrating information with similar meaning and contrasting conflicting information specific methods are needed in multi-document summarization. For example, Johnson et al. designed a method which clusters similar sentences from multiple documents and consolidates these sentences into a single summary for those sentences.(27)

Second, while the extraction paradigm is still a dominant approach, there may be a growing attention to abstractive techniques. There were no studies based on abstractive methods prior to 2000 in the earlier review by Afantenos et al. In our systematic review, 24% of the studies

focused on abstractive techniques. This could be due to a number of reasons, such as availability of more sophisticated and semantic Natural language processing tools. For example, Fiszman et al. designed a method for generating graphical summarization of Medline citations based on semantic interpretation of biomedical text.(24) Abstractive techniques have the potential to be useful to clinicians and researchers, especially when summarizing multiple documents.

Third, a growing interest in knowledge rich methods compared to knowledge poor approaches was observed. The fact that a large number of publicly available knowledge resources, such as PubMed Central, the UMLS,(49) and natural language processing tools, such as MetaMap,(50) SemRep,(51) and cTAKES;(52) now exist and can be accessed conveniently may have contributed to the interest.

Fourth, a combination of statistical, language processing and machine learning approaches is increasingly popular in text summarization. For example, Reeve et al. mapped terms to UMLS concepts and used UMLS semantic types to discover strong thematic chains.(37) Cao et al. used machine learning techniques along with language processing for developing AskHermes, an online question-answering system.(18) Another area that has received increased attention is graph-based summarization methods. These methods represent the text as a graph, where the nodes correspond to words or sentences and the edges represent various types of syntactic and semantic relations among them. Different clustering methods are then applied to identify salient nodes within the graph and to extract the sentences for the summary.(4) For example, Bio Squash developed by Zhongmin Shi et al. is a question-oriented multi-document summarizer for biomedical texts.(42) It constructs a graph that contains concepts of three types: ontological concepts, named entities, and noun phrases. Yoo et al. described an approach to multi-document summarization that uses MeSH descriptors and a graph-based method for clustering articles into topical groups and producing a multi-document summary of each group.(45)

Finally, most of the studies in the review conducted intrinsic evaluations. These evaluations often consisted of comparing summarization output with a reference standard developed by experts, typically in terms of measures such as precision and recall. However, this type of evaluation is expensive and time consuming. In addition, generating the reference standard is highly dependent on the experts who produce them and may lack consistency in quality. To address this limitation, research is being pursued to produce reference standards automatically.(30) Many researchers use the summary or the abstract of the paper as the reference standard. Plaza et al. compared their summarization output with abstracts included in the articles.(15) Sarkar et al. compared the performance of their proposed summarization system and employed as a reference standard the output of a broad summarization system called MEAD.(40) Another common set of metrics and software package used for evaluating automatic summaries is ROUGE (Recall-Oriented Understudy for Gisting Evaluation), developed by the University of Southern California.(53) A small number of studies conducted extrinsic evaluations. For example, Elhadad et al. included both intrinsic and extrinsic components in their study.(22)

4.2 Gaps and implications for research

Despite advances in biomedical text summarization research, this systematic review identified some important gaps that need to be filled in order to enable future progress. Several text summarization techniques depend heavily on the quality of annotated corpora and reference standards available for training and testing. However, our review found only one study which reported on a generalizable biomedical summarization corpus with the potential of being used by other researchers.(30) Thus, more research is needed to enable summarization corpora and reference standards to support the development of summarization tools in various applications.(30) Further research is also needed to enable publicly available summarization corpora and reference standards to support the development of summarization tools.

Another gap is the extensive reliance on English documents as the input for summarization. One of the major causes could be due to limitations of lexical and semantic tools in any other language. The summary presentation was also usually in the form of text. Very few studies focused on producing a visual output.

Another gap was the scarcity of studies that conducted extrinsic evaluations. This may be an indication that most of the research is still focused on the components used for summarization as and not on testing the impact of more mature summarization systems. As a possible consequence of the nascent status of the biomedical text summarization research, none of the studies identified in our systematic review have been assessed in patient care settings or in actual research applications. To advance the field, more attention is needed on the cognitive implications of text summarization. This could be accomplished through methods such as usability studies, simulations, and studies that aim at integrating text summarization tools into routine workflows. A further advance would be studies that focus on deployment of text summarization systems in real research and patient care settings, and evaluate the impact of such systems on the users' decision-making performance and on patient outcomes.

4.3 Limitations

This systematic review has several limitations. First, research trends were inferred by comparing our findings with the review conducted in by Afantenos et al. The latter study was not a systematic review and may have missed important past research. Second, although we did not find any study that deployed a text summarization system in operational settings, it is still possible that some of the systems described in the included studies have been deployed in real settings after the study was published. Likewise, there may be commercial summarization systems that are available in work settings, but that have not been formally studied and published in peer-reviewed forums. Third, a meta-analysis comparing the performance of different approaches was not possible due to the heterogeneity of the evaluation methods. The lack of widely used standard evaluation methods is possibly indicative of the low level of maturity of the field compared to other similar areas such as information retrieval.(5) Fourth, our data abstraction was guided by the dimensions included in Mani's framework. As a consequence, we might have missed other important dimensions and trends that did not receive sufficient attention in the data abstraction process. Last, by

excluding articles not written in English we may have missed systems that summarize text in other languages.

5. Conclusions

We systematically reviewed the literature on text summarization methods in the biomedical domain. Our study found a predominance of methods that produce extractive summaries; use multiple documents as the source for summarization; employ a combination of statistical, language processing, and machine learning techniques; utilize knowledge rich approaches that leverage a range of publicly available tools and knowledge resources; and that are evaluated through intrinsic techniques. We also found a growing interest in abstractive summaries and graph-based methods. To advance knowledge in this field, further research is needed in the cognitive aspects of text summarization, including visualization techniques, and evaluations of the impact of text summarization systems in work settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge Alice Weber for providing insights on the search strategy of this systematic review. This project was supported by grant number 1R01LM011416-01 from the National Library of Medicine.

References

1. Smith R. Strategies for coping with information overload. BMJ. 2010; 341:c7126. [PubMed: 21159764]
2. Davidoff F, Miglus J. Delivering clinical evidence where it's needed: building an information system worthy of the profession. JAMA. 2011; 305(18):1906–7. [PubMed: 21558524]
3. Pluye P, Grad RM, Dunikowski LG, Stephenson R. Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies. International journal of medical informatics. 2005; 74(9):745–68. [PubMed: 15996515]
4. Spärck Jones K. Automatic summarising: The state of the art. Information Processing & Management. 2007; 43(6):1449–81.
5. Mani I, Klein G, House D, Hirschman L, Firmin T, Sundheim B. SUMMAC: a text summarization evaluation. Natural Language Engineering. 2002; 8(01):43–68.
6. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artificial intelligence in medicine. 2005; 33(2):157–77. [PubMed: 15811783]
7. Eden, J.; Levit, L.; Berg, A.; Morton, S. Finding what works in health care: standards for systematic reviews. National Academies Press; 2011.
8. Mani, I. Automatic summarization. John Benjamins Publishing; 2001.
9. Mani, I.; Maybury, MT. Advances in automatic text summarization. the MIT Press; 1999.
10. Edmundson HP. New methods in automatic extracting. Journal of the ACM (JACM). 1969; 16(2): 264–85.
11. Sarkar K. Using Domain Knowledge for Text Summarization in Medical Domain. International Journal of Recent Trends in Engineering. 2009; 1(1):200–5.
12. Jurafsky, D.; Martin, JH. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall; 2000. Speech & Language Processing.

13. Reeve L, Han H, Brooks AD. BioChain: Lexical chaining methods for biomedical text summarization. Proceedings of the 2006 ACM symposium on Applied computing; 2006. 2006:180–4.
14. Chuang, WT.; Yang, J. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval; 2000. ACM; 2000. Extracting sentence segments for text summarization: a machine learning approach; p. 152-9.
15. Plaza L, Díaz A, Gervás P. A semantic graph-based approach to biomedical summarisation. Artificial intelligence in medicine. 2011; 53(1):1–14. [PubMed: 21752612]
16. Jones, KS.; Galliers, JR. Evaluating natural language processing systems: an analysis and review. Springer; 1995.
17. Kushniruk AW, Kan MY, McKeown K, et al. Usability evaluation of an experimental text summarization system and three search engines: implications for the reengineering of health care interfaces. Proc AMIA Symp. 2002:420–4. [PubMed: 12463858]
18. Cao Y, Liu F, Simpson P, et al. AskHERMES: An online question answering system for complex clinical questions. Journal of biomedical informatics. 2011 Apr; 44(2):277–88. [PubMed: 21256977]
19. Chen P, Verma RA. Query-Based Medical Information Summarization System Using Ontology Knowledge. 19th IEEE Symposium on Computer-Based Medical Systems. 2006:37–42. 2006.
20. Cruz YR, Llavori RB, García RG. A Framework for Obtaining Structurally Complex Condensed Representations of Document Sets in the Biomedical Domain. Procesamiento de Lenguaje Natural. 2012; 49:21–8.
21. Demner-Fushman, D.; Lin, J. 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics; 2006. Association for Computational Linguistics; 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering; p. 841-8.
22. Elhadad N, McKeown K, Kaufman D, Jordan D. Facilitating physicians' access to information via tailored text summarization. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium. 2005:226–30. [PubMed: 16779035]
23. Elhadad N, Kan MY, Klavans JL, McKeown KR. Customization in a unified framework for summarizing medical literature. Artificial intelligence in medicine. 2005; 33(2):179–98. [PubMed: 15811784]
24. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL workshop on computational lexical semantics. 2004:76–83. 2004.
25. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium. 2006:254–8. [PubMed: 17238342]
26. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. Journal of biomedical informatics. 2009; 42(5):801–13. [PubMed: 19022398]
27. Johnson DB, Zou Q, Dionisio JD, Liu VZ, Chu WW. Modeling medical content for automated summarization. Ann N Y Acad Sci. 2002; 980:247–58. [PubMed: 12594094]
28. Lin J, Wilbur WJ. Syntactic sentence compression in the biomedical domain: facilitating access to related articles. Information Retrieval. 2007; 10(4-5):393–414.
29. McKeown KR, Elhadad N, Hatzivassiloglou V. Leveraging a common representation for personalized search and summarization in a medical digital library. 3rd ACM/IEEE-CS joint conference on Digital libraries. 2003:159–70. 2003.
30. Molla D, Santiago-Martinez ME. Development of a corpus for evidence based medicine summarisation. Australasian Language Technology Association Workshop. 2011:86–94. 2011.
31. Morales, LP.; Esteban, AD.; Gervás, P. 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing; 2008. Association for Computational Linguistics; 2008. Concept-graph based biomedical automatic summarization using ontologies; p. 53-6.

32. Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium. 2006:599–603. [PubMed: 17238411]
33. Plaza, L.; Stevenson, M.; Diaz, A. Improving summarization of biomedical documents using word sense disambiguation. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing; Uppsala, Sweden. Association for Computational Linguistics; 2010. p. 55-63.
34. Plaza L, Jimeno-Yepes AJ, Diaz A, Aronson AR. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. BMC bioinformatics. 2011; 12:355. [PubMed: 21871110]
35. Plaza L, Stevenson M, Díaz A. Resolving ambiguity in biomedical text to improve summarization. Information Processing and Management. 2012; 48(4):755–66.
36. Plaza L, Carrillo-de-Albornoz J. Evaluating the use of different positional strategies for sentence selection in biomedical literature summarization. BMC bioinformatics. 2013; 14:71. [PubMed: 23445074]
37. Reeve, L.; Han, H.; Brooks, AD. Proceedings of the 2006 ACM symposium on Applied computing; 2006. ACM; 2006. BioChain: lexical chaining methods for biomedical text summarization; p. 180-4.
38. Reeve, LH.; Han, H.; Nagori, SV.; Yang, JC.; Schwimmer, TA.; Brooks, AD. Concept frequency distribution in biomedical text summarization. 15th ACM international conference on Information and knowledge management; 2006. p. 604-11.2006
39. Reeve LH, Han H, Brooks AD. The use of domain-specific concepts in biomedical text summarization. Information Processing & Management. 2007; 43(6):1765–76.
40. Sarkar K, Nasipuri M, Ghose S. Using Machine Learning for Medical Document Summarization. International Journal of Database Theory and Application. 2011; 4:31–49.
41. Sarker A, Molla D, Paris C. Extractive summarisation of medical documents using domain knowledge and corpus statistics. The Australasian medical journal. 2012; 5(9):478–81. [PubMed: 23115581]
42. Shi, Z.; Melli, G.; Wang, Y., et al. Advances in Artificial Intelligence. Springer; 2007. Question answering summarization of multiple biomedical documents; p. 284-95.
43. Summerscales, RL.; Argamon, S.; Bai, S.; Huperff, J.; Schwartzff, A. Automatic Summarization of Results from Clinical Trials. 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2011. p. 372-7.2011
44. Workman TE, Fiszman M, Hurdle JF. Text summarization as a decision support aid. BMC medical informatics and decision making. 2012; 12:41. [PubMed: 22621674]
45. Yoo I, Hu X, Song IY. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC bioinformatics. 2007; 8(Suppl 9):S4. [PubMed: 18047705]
46. Shang Y, Li Y, Lin H, Yang Z. Enhancing biomedical text summarization using semantic relation extraction. PloS one. 2011; 6(8):e23862. [PubMed: 21887336]
47. Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. Journal of biomedical informatics. 2011 Oct; 44(5):830–8. [PubMed: 21575741]
48. Zhang H, Fiszman M, Shin D, Wilkowski B, Rindflesch TC. Clustering cliques for graph-based summarization of the biomedical research literature. BMC bioinformatics. 2013; 14:182. [PubMed: 23742159]
49. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. Journal of the American Medical Informatics Association : JAMIA. 1998; 5(1):1–11. [PubMed: 9452981]
50. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA. 2010; 17(3):229–36. [PubMed: 20442139]
51. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics. 2003; 36(6):462–77. [PubMed: 14759819]

52. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010; 17(5):507–13. [PubMed: 20819853]

53. Lin CY. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004:74–81. 2004.

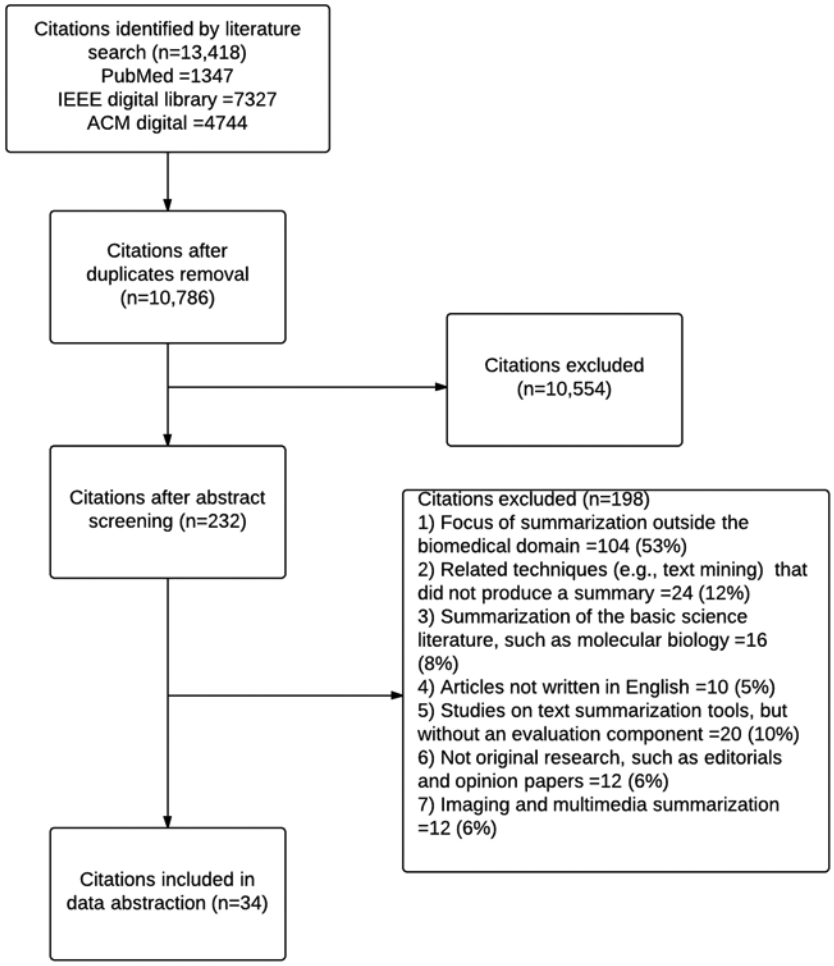

Figure 1.

Table 1

Study frequency according to the data abstraction dimensions.

Parameters	Category	Frequency
Location	USA	16 (47 %)
	Multinational	9 (26%)
	Asia	2 (6%)
	Australia	2 (6%)
	Spain	3 (9 %)
	Canada	2 (6%)
Input	Single document (SD)	13 (38%)
	Multiple documents (MD)	19 (56%)
	Single and Multiple documents (SD and MD)	2 (6%)
	Monolingual (Mono)	34 (100%)
	Multilingual (Multi)	0%
	Full Text (FT)	19 (56%)
	Abstract (Ab)	12 (35%)
	Full Text and Abstract (FT and Ab)	3 (9%)
	Biomedical Research Literature (Lit)	31 (91%)
	Literature and EHR (Lit and EHR)	3 (9%)
Purpose	Clinical decision support (CDS)	19 (56%)
	Broad spectrum (BS)	15 (44%)
	User-oriented (U)	16 (47%)
	Generic (G)	18 (53%)
Output	Text (T)	28 (82%)
	Graph (Gr)	6 (18%)
	Informative (Inf)	25 (74%)
	Informative and indicative (Inf and Ind)	9 (26%)
	Extract (Ext)	26 (76%)
	Abstract (Abs)	8 (24%)
Method	Statistical (Stats)	1 (3%)
	Statistical and Natural Language Processing (Stats and NLP)	7 (20%)
	Natural Language Processing (NLP)	17 (50%)
	Machine Learning and Statistical (ML and Stats)	1 (3%)
	Natural Language Processing and Machine learning (NLP and ML)	5 (15%)
	Statistical, machine learning and natural language processing (Stats, ML and NLP)	2 (6%)
Evaluation	Intrinsic (I)	28 (82%)
	Extrinsic (E)	4 (12%)
	Intrinsic and Extrinsic (I and E)	2 (6%)

Table 2

Included studies and their characteristics.

Author, year	Location	Input	Purpose	Output	Evaluation
Cao, 2011 (18)	USA	MD, Mono, Ab and FT, Lit	CDS, U	T, Ind and Inf, Ext	E
Chen, 2006 (19)	USA	SD, Mono, FT, Lit	BS, U	T, Inf, Ext	I
Cruz, 2012 (20)	Multinational	MD, Mono, FT, Lit	CDS, U	T, Inf, Ext	I
Demner-Fushman, 2006 (21)	USA	MD, Mono, Ab, Lit	CDS, U	T, Ind and Inf, Ext	I
Elhadad, 2005 (22)	USA	MD, Mono, FT, Lit and EHR	CDS, U	T, Inf, Ext	E
Elhadad, 2005 (23)	Multinational	MD, Mono, FT, Lit and EHR	CDS, U	T, Ind and Inf, Abs	E
Fiszman, 2004 (24)	USA	SD and MD, Mono, Ab and FT, Lit	CDS, G	Gr, Inf and Ind, Abs	I
Fiszman, 2006 (25)	USA	MD, Mono, Abs, Lit	CDS, U	Gr, Inf and Ind, Abs	I
Fiszman, 2009 (26)	USA	MD, Mono, Ab, Lit	CDS, G	Gr, Inf, Abs	I and E
Johnson, 2002 (27)	USA	SD and MD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Kushniruck, 2002 (17)	Multinational	MD, Mono, FT, Lit	BS, U	T, Ind and Inf, Ext	E
Lin, 2007 (28)	USA	SD, Mono, Ab, Lit	BS, G	T, Inf, Ext	I and E
Mckeown, 2003 (29)	USA	MD, Mono, FT, Lit and EHR	CDS, U	T, Ind and Inf, Ext	I
Molla, 2011 (30)	Australia	MD, Mono, Ab and FT, Lit	CDS, G	T, Inf, Ext	I
Morales, 2008 (31)	Spain	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Niu, 2006 (32)	Canada	MD, Mono, Ab, Lit	CDS, U	T, Inf, Ext	I
Plaza, 2010 (33)	UK, Spain	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Plaza, 2011 (15)	Spain	MD, Mono, FT, Lit	BS, U	T, Inf, Ext	I
Plaza, 2011 (34)	Multinational	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Plaza, 2012 (35)	Multinational	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Plaza, 2013 (36)	Spain	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Reeve, 2006 (37)	USA	SD, Mono, FT, Lit	CDS, G	T, Inf, Ext	I
Reeve, 2006 (38)	USA	SD, Mono, FT, Lit	CDS, G	T, Inf, Ext	L
Reeve, 2007 (39)	USA	SD, Mono, FT, Lit	CDS, G	T, Inf, Ext	I
Sarkar, 2009 (11)	Asia	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Sarkar, 2011(40)	Asia	SD, Mono, FT, Lit	BS, G	T, Inf, Ext	I
Sarker, 2012 (41)	Australia	MD, Mono, Ab, Lit	CDS, G	T, Inf, Ext	I
Shi, 2007 (42)	Canada	MD, Mono, Ab, Lit	CDS, U	T, Inf, Ext	I

Author, year	Location	Input	Purpose	Output	Evaluation
Summerscales, 2011 (43)	USA	SD, Mono, FT, Lit	CDS, G	T, Inf, Ext	I
Workman, 2012 (44)	USA	MD, Mono, Ab, Lit	CDS, U	Gr, Ind and Inf, Abs	I
Yoo, 2007 (45)	USA	MD, Mono, Ab, Lit	BS, G	T, Inf, Ext	I
Yue, 2011 (46)	Multinational	MD, Mono, Ab, Lit	BS, U	T, Inf, Ext	I
Zhang, 2011 (47)	Multinational	MD, Mono, Abs, Lit	CDS, U	Gr, Inf, Abs	I
Zhang, 2013 (48)	Multinational	MD, Mono, Ab, Lit	BS, U	Gr, Ind and Inf, Abs	I

Abbreviations : SD- Single document, MD- Multiple document, Mono-Monolingual, ML-Multilingual, FT-Full text, Ab- Abstract, CDS- Clinical decision support, BS- Broad spectrum, G- Generic, U-User oriented, T- Text, Gr-Graph, Inf- Informative, Ind- Indicative, Ext- Extract, Abs-Abstract, I- Intrinsic, E- Extrinsic

Table 3

A summary of study methods and their evaluation.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
Cao, 2011 (18)	Question answering system with 5 components:(1) *question analysis* includes a support vector machine (SVM) classifier based on lexical features that map the user question to a question type (e.g., diagnosis, procedure); and a conditional random fields model that identifies keywords and UMLS concepts to retrieve documents and extract answers; (2) *related questions extraction* retrieves a list of similar questions; (3) *information retrieval* retrieves documents that are relevant to the question; (4) *information extraction* extracts passages comparing the similarity between the question and the retrieved sentences; (5) *summarization and answer presentation* clusters the extracted passages according to content-rich keywords in the question.	Comparison with Google and UpToDate answering 60 questions from a clinical questions database	Three physicians rated the system according to ease of use, quality of answer, time spent, and overall performance (according to 1 to 5 scale)	Similar rating in terms of ease of use. Lowest rating on quality of answer. Best rating in terms of time spent. No statistical comparisons were made.
Chen, 2006 (19)	User query-based text summarization system. Documents are retrieved from databases such as Medline and indexed with UMLS concepts. As a user inputs a query, the system represents a retrieved document in a concept network. The summary consists of a set of sentences extracted from the original document based on their similarity to the user's query.	Sentences included in the abstract and conclusions of an article selected for the evaluation	Precision and recall	Method with expanded keywords using the UMLS obtained a precision of 60% and recall of 80%, compared with a baseline of 30% precision and 40% recall with no key word expansion.
Cruz, 2012 (20)	Framework for structurally complex condensed representations of document sets, which can be used as a basis for summarization. The framework	Top 20 facts obtained from a Medline search were manually rated as relevant or not relevant	Precision at top k ranked elements	Precision at different levels varied from 40% to 100%.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	extracts a ranked list of facts (entity-relation-entity triples) that are relevant to a focus concept, applying a dependency parser with pattern-based heuristics. Top ranked facts are organized into a bipartite graph-based representation, with one set of node for entities and another for relations.			
Demner-Fushman,2006 (21)	Question-answering system, in which answers obtained from Medline citations are presented in a hierarchical and interactive fashion. The system has 3 components: (1) *answer extraction* identifies drugs of interest in a set of retrieved citations using MetaMap; (2) s*emantic clustering* uses a hierarchical clustering algorithm to organize retrieved articles according to the interventions under investigation; (3) extractive summarization generates a summary with the main intervention, article title, and the top-scoring outcome (i.e., study findings) sentence according to a supervised machine learning algorithm.	Gold standard: test collection of abstracts for 30 questions from Clinical Evidence (a resource of manually digested clinical evidence)	Manual rating of system output and ROUGE (abstracts cited in Clinical Evidence) for top 3 abstracts.	System increased twice as many answers with beneficial drugs as a PubMed search. 86% of the answers retrieved with the system were rated as “good” versus 60% from PubMed.
Elhadad, 2005 (23)	PErsonalized Retrieval and Summarization of Images, Video and Language (PERSIVAL) system. The system produces a user model, which tailors summarization according to 3 dimensions: 1) level of expertise; 2) characteristics of a given patient; and 3) user's access task (e.g., browsing, searching, obtain a briefing). Characteristics of the patient are automatically extracted and represented as UMLS	TAS: Study results manually extracted from a set of 40 articles. Centrifuser: evaluation described below in Kushniruck, 2002 (17).	TAS: Precision and recall of results	TAS: Precision of 90%, recall of 65%.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	concepts from the patient's electronic health record using natural language processing. A personalized summary for medical experts is generated by the Technical Article Summarizer (TAS) system based on the user's query and the patient's characteristics. TAS consists of a pipeline that includes: *classification of articles* according to a main clinical task (e.g., diagnosis, treatment); *extraction of study results* leveraging the typical structure of biomedical articles and using shallow syntactic parsing to match sentences to pre-defined patterns; *matching* of study results to the patient's characteristics; *merging and ordering* of results into a semantic graph (nodes are UMLS concepts and vertices are relations); and *surface generation* to produce a textual summary from the semantic graph. A separate component (Centrifuser) is used to produce summaries for lay users.			
Elhadad, 2005 (22)	Technical Article Summarizer (TAS) system described above in Elhadad 2005 (23). The focus of the study was an extrinsic evaluation of the system.	Task-based evaluation: 12 clinicians extracted all findings relevant to a patient from a set of 5 relevant articles. Performance was compared using 3 options: search results, a generic summary, and a summary personalized to the patient's characteristics.	F_2 measure and user satisfaction.	Performance with personalized summaries was significantly better than with generic summaries (F-measure = 28 versus 14; p=0.07). Subjects preferred personalized over generic summaries (p=0.001)
Fiszman, 2004 (24)	Semantic abstraction summarization system for Medline citations. The summarization relies on semantic predications extracted by SemRep, an NLP parser based on underspecified linguistic analysis and domain knowledge represented in the UMLS. The abstract	Manual linguistic evaluation of the quality of generated abstracts for four diseases.	Precision and compression (difference between initial and final number of semantic predications after transformation principles are applied).	66% average precision across four diseases.Total number of predications reduced from 11245 in the baseline to 306 after transformation.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	generation is based on four transformation principles: relevance, connectivity, novelty, and saliency. The same process used for single document summarization is applied for multiple document summarization. The summarization output is presented in graphical form.			
Fiszman, 2006 (25)	Semantic abstraction summarization system: described above in Fiszman, 2004 (24), extended to summarize drug information from Medline citations.	Manual linguistic evaluation of the quality of generated abstracts for 10 drugs.	Precision	Precision of 58% before and 78% after saliency transformation.
Fiszman, 2009 (26)	Semantic abstraction summarization system: described above in Fiszman, 2004 (24). The study focused on a large scale topic-based evaluation of summarization of drug interventions.	Gold standard: questions and synthesized answers on the treatment of 53 diseases obtained from Clinical Evidence (a resource of manually digested clinical evidence) and the Physicians' Desk Reference (PDR).	Mean average precision and clinical usefulness	Mean average precision of 50% versus 33% in the baseline ($p <$ 0.01).Clinical usefulness score of 0.64 versus 0.25 in the baseline ($p < 0.05$).
Johnson, 2002 (27)	Method based on words and n-word combinations found in document sentences. Sentences are ranked according to the overall frequency of its word combinations in the document. An alternate method included a clustering step, in which similar documents are clustered in the same group. Clusters are then analyzed for key features, which are used to rank sentences.	Collection of radiology reports.	Precision and recall of three queries representing specific medical findings	The 3-word method had better precision than 2-word and 1-word combinations. For the Salton method, the average precision rates are 8% for 1-word, 24% for 2-word, and 40% for 3-word. The result shows that for both methods the precision is better for multiword combinations than for 1-word models.
Kushniruck, 2002 (17)	Usability study comparing Centrifuser with three search engines (Google, Yahoo, and About.com). Centrifuser is described above in Elhadad, 2005 (23).	Centrifuser compared with Google, Yahoo and About.com	Users' think-a-loud expressions coded with a usability coding scheme	None of the systems contained features that were positively rated by all subjects. Centrifuser received the highest number of positive comments on content usefulness and understanding, and the highest number of negative comments on understanding user interface labels.
Lin, 2007 (28)	Sentence compression algorithm consisting of a series of	Corpus of 200 manually compressed Medline article titles and manual rating of compressed	BLEU metrics; fluency and content ratings of compressed article titles, and consistency of human	Algorithm compresses article titles by 30% on average without

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	syntactic compression rules applied over the output of the Stanford parser.	titles by 2 domain experts.	judgments between original and compressed titles (precision and recall)	compromising task performance.
Mckeown, 2003 (29)	PERSIVAL allows clinicians to search the biomedical literature within the context of a patient's record. User queries are sent to online literature resources along with specific characteristics of the patient of interest. The search results are ranked according to these characteristics. Articles in the search output are summarized with the Technical Article Summarizer (TAS), which is described above in Elhadad, 2005 (23).	Manual examination of search results and summarization output for one patient scenario. A more comprehensive evaluation is reported in Elhadad, 2005 (23).	Precision of articles retrieved for the summarization step.	Patient-specific re-ranking yielded 89% precision and 65% without re-ranking.
Molla, 2011 (30)	Development of a corpus to support text summarization research in medical texts. The source of the corpus was the Clinical Inquiries section of the Journal of Family Practice, which contains an evidence digest for a set of medical questions along with evidence rating and citations that support the evidence. The process included automated, manual annotation, and crowd-sourced annotation.	Description of the corpus and baseline performance based on random sentence selection and last three sentences of each clinical inquiry.	ROUGE metrics	The ROUGE-L score for random sentences was 0.188 and for last 3 sentences was 0.193.
Morales, 2008 (31)	Ontology and graph-based extractive method. Each document is represented as a graph, where nodes are UMLS concepts and edges are relations. A weight is calculated for each edge based on the specificity of the associated concepts. Next, sentences are grouped into clusters that reflect a common theme. Last, sentences are extracted based on a set of heuristics. Three heuristics were tested: 1) top sentences for each cluster; 2) top sentences from the	Exploratory comparison of summarization output for a document from BioMed Central with the document abstract using the three proposed heuristics.	Sentence scores for each heuristic	Heuristics 1 and 3 retrieve sentences that cover all the topics included in the article's original abstract.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	cluster with larger number of most concepts; and 3) top sentences with a highest score.			
Niu, 2006 (32)	Sentence extraction algorithm for a question-answering system. The method detects clinical outcomes in sentences and uses a support vector machine (SVM) classifier to determine the polarity of the outcome (e.g., positive, negative, or neutral). The classifier is based on a set of phrases that denote a change, such as “increase” or “enhanced”. The method also extracts most important sentences from the source based on an SVM classifier with the following features: sentence position, sentence length, presence of numeric values, and maximal marginal relevance (a measure of novelty).	2298 annotated sentences from 197 Medline abstracts cited in Clinical Evidence.	Accuracy for classification of the presence of outcomes. Precision and recall curve and ROUGE metrics for sentence-level evaluation.	Accuracy of 82.5% for the presence of outcomes in a sentence and 78.3% for the outcome polarity. Identification of outcomes improves F-measure by 5 points, but benefit from identification of polarity was small.
Plaza, 2010 (33)	Word sense disambiguation (WSD) algorithm added to the summarization system described above in Morales, 2008 (31). The algorithm was adapted from the Personalized PageRank algorithm, a graph-based algorithm in which nodes represent ambiguous words in a document with links to their possible meanings.	150 articles randomly selected from BioMed Central corpus. Comparison of summarization output with and without WSD.	ROUGE metrics	ROUGE scores were significantly better for the summarizer enhanced with WSD.
Plaza, 2011 (15)	Enhanced version of the graph-based approach described above in Morales, 2008 (31). The study also included an experiment to identify optimal parameters for three sentence extraction heuristics: 1) selects the top sentences from the cluster with the most concepts (focuses on the main topic of the article); 2) selects the	300 full-text articles randomly selected from the BioMed Central corpus. Article abstracts were used as reference summaries. Comparisons of the three heuristics with three summarizers: SUMMA, LexRank, and Microsoft Autosummarize.	ROUGE metrics	The three heuristics performed significantly better than the comparison summarizers. Heuristic 3 had the best performance overall.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	top sentences from each cluster according to the cluster size (covers all topics of the article); 3) computes a score for each sentence based on the votes for each cluster (focuses on the main topic of the article, but with some information from other secondary topics).			
Plaza, 2011 (34)	Comparison of three knowledge-based WSD approaches integrated with the summarizer described above in Morales, 2008 (31), Plaza 2010 (33), and Plaza, 2011 (15): 1) Journal Descriptor Index (JDI), an unsupervised technique; 2) Machine Readable Dictionary (MRD), a knowledge-based method; and 3) Automatic Extracted Corpus (AEC), a supervised learning technique.	150 articles randomly selected from BioMed Central corpus. Article abstracts were used as reference summaries.	ROUGE metrics	Overall, WSD improved ROUGE scores of the summarization output compared with a summarizer without WSD. MRD and AEC had equivalent performance. Both performed better than JDI.
Plaza, 2012 (35)	Comparison of three knowledge-based WSD approaches integrated with the summarizer described above in Morales, 2008 (31): 1) first concept returned by MetaMap (equivalent to no WSD); 2) Journal Descriptor Index (JDI); 3) Personalized PageRank (PPR); 4) all candidate concept mappings returned by MetaMap; and 5) all mappings, but with concepts weighted based on the output of a WDS algorithm.	150 articles randomly selected from BioMed Central corpus. Article abstracts were used as reference summaries.	ROUGE metrics	All WSD approaches led to summaries with higher ROUGE scores than the approach without WSD. JDI led to summaries with higher ROUGE scores than PPR. The differences between JDI, all mappings, and weighted mappings were not statistically significant.
Plaza, 2013 (36)	Comparison of three sentence position approaches for sentence extraction integrated with the summarizer described above in Morales, 2008 (31): (1) preference for sentences close to the beginning of the document; (2) preference for sentences close to the beginning and end of	100 articles randomly selected from the PMC Open Access Subset. Article abstracts were used as reference summaries.	ROUGE metrics	Algorithms (1) and (2) did not improve ROUGE metrics compared to not using positional information. Section-based weighting (3) improved the performance over the non-positional approaches by 17% points for the graph-based summarizer.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	the document; and (3)weighting the sentences accordingto the section (e.g., introduction, methods, results) in which they appear.			
Reeve, 2006 (37)	BioChain: Lexical chaining summarization system based on UMLS concepts rather than terms. UMLS concepts are placed into chains based on their UMLS semantic type. Then, chains are scored based on a method that includes features such as reiteration (repetition of concepts), density (proximity of concepts in the text), length (number of concepts), and the importance of their semantic type. Chains are then sorted according to their scores and most frequent concepts in the top chains are identified. Sentences that contain these most frequent concepts are extracted for the output.	24 articles from an article collection of oncology clinical trial.	Precision and recall of strong chains compared with the article human abstract.	Average precision and recall of 0.90 and 0.92 respectively.
Reeve, 2006 (38)	FreqDist: Summarization approach based on the frequency distribution of concepts in sentences. Sentences are selected iteratively in an effort to include new sentences that are not similar to the ones previously included in the summary. The study compared the performance when using five different similarity methods: cosine similarity, Dice's coefficient, Euclidean distance, vector subtraction, and an approach to vector comparison that only considers unit item frequency.	24 articles randomly selected from a collection of oncology clinical trials. Four model summaries were used per article: the article's abstract and 3 different summaries created by medical students.	ROUGE metrics	FreqDist with Dice's coefficient obtained the highest ROUGE-2 and ROUGE-SU4 scores compared with other general summarizers. FreqDist with vector subtraction and Euclidean distance similarity methods had the lowest scores.
Reeve, 2007 (39)	ChainFreqSum: hybrid method that includes BioChain (37)and FreqDist (38) (both described above). BioChain identifies sentences	Same as in (38).	ROUGE metrics	The ChainFreqSum and FreqDist approaches, both using Dice's similarity, had similar performance, and better than other approaches.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	that represent the main theme of an article. FreqDist takes the BioChain ouput and reduces redundant sentences.			
Sarkar, 2009 (11)	Combines domain specific features with commonly used features for sentence ranking and summary generation. Sentence ranking is based on medical cue phrases and terms, term frequency, similarity with article title, sentence position, presence of novel terms, and sentence length. For summary generation, top sentences are selected after applying a variant of the maximal marginal relevance (MMR) re-ranking method.	50 medical articles obtained from the Internet. Article abstracts were used as reference summaries. Summarization with and without domain-specific features was compared.	ROUGE metrics	Domain specific features increased the ROUGE-1 score by 0.12 to 0.14 points.
Sarkar, 2011 (40)	Supervised learning method based on features such as domain specific cue phrases, centroid overlap, sentence position in the text, and sentence length. The method consisted of training a bagging meta-classifier with a dataset composed of sentences from medial news articles and class labels derived from human-written summaries. Sentences are ranked according to the class and probability output of the classifier. Summary generation is performed as described above in Sarkar, 2011 (11).	75 medical news articles downloaded from Web sites, such as MedlinePlus. Two model summaries were created for each article. Proposed approach was compared with the general purpose MEAD summarizer.	Precision and recall, ROUGE metrics.	The supervised learning approach performed better than MEAD both in terms of precision and recall and ROUGE scores.
Sarker, 2012 (41)	Hybrid sentence extraction method based on 1) a machine learning method that classifies sentences into types (i.e., population, intervention, background, outcome, study); 2) relative sentence position; 3) sentence length; and 4) presence of semantic types relevant to the type of query (e.g., diagnosis, treatment).	Described above in Molla, 2011 (30). The proposed method was compared with domain independent and domain specific techniques.	ROUGE metrics	The ROUGE-L score for the proposed method was 0.1653 (statistically equivalent or better than other methods). The performance improved with the addition of query-specific information (0.025 points for treatment and 0.019 for diagnosis queries).

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
Shang, 2011 (46)	The method consists of three steps: 1) Extracting semantic relations in each sentence using the SemRep system; 2) retrieving relations that are relevant to a given query and forming a core set of most frequent relations; 3) ranking and retrieving the most relevant sentences for each relation above, and removing redundant sentences.	A subset of Medline abstracts published in 2009 was used as the study corpus. Definitions of diseases in Wikipedia were used as reference summaries. The method was compare with the MEAD summarizer.	ROUGE metrics	The proposed method performed better than MEAD and a baseline in all ROUGE metrics. Relation expansion, noise reduction, and redundancy removal improved performance in 0.02 to 0.03 points in ROUGE-1, ROUGE-2 and ROUGE-L scores.
Shi, 2007 (42)	BioSquash: Question oriented extraction summarization system based on domain-specific and domain-independent knowledge. The method constructs a semantic graph based on UMLS and WordNet concepts and relations in multiple documents and the question. Concepts in the graph are assigned a significance score based on the frequency of each concept in questions, documents, and semantic relations. Sentences are ranked according to a score based on the significance of the relations contained in each sentence. Sentences are extracted in an iterative fashion, in which sentences that are similar to sentences already extracted are penalized to avoid redundancy. The final summarization step takes the top ranked sentences; re-orders them based on features such as sentence significance, overlap, and length; and groups similar sentences together.	Adapted from the Ad-hoc Retrieval Task of the Genomics track at the 2005 Text Retrieval Conference (TREC). The corpus includes a set of questions and a subset of 4.5 million Medline citations. For the BioSquash evaluation, a subset of 18 questions was selected along with a subset of Medline citations relevant to those questions. The article abstracts were compared with summaries.	ROUGE metrics	Automated, question-oriented summaries achieved higher ROUGE-2 and SU4 scores (0.0697 versus 0.069; 0.13 versus 0.1118 respectively) in relation to the question than the human written summaries.
Summerscales, 2011 (43)	Automatically calculates summary statistics (i.e., absolute risk reduction (ARR) and number needed to treat (NNT)) from data (number of bad	Corpus of 263 abstracts of RCTs published in the British Medical Journal. Performance detecting treatments and outcomes was compared with a named entity recognition	Precision, recall, and F-measure	The proposed method had higher F-measure than the baseline for detecting groups (0.76 vs. 0.74), outcomes (0.42 vs. 0.38), group sizes (0.8 vs. 0.66), and outcome numbers (0.71 vs. 0.52). The F-

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	outcomes and individuals in the intervention and control groups) available in the abstract text of randomized controlled trial (RCT) articles. First, the method identifies sentences that contain numbers. Then, it uses a conditional random field (CRF) classifier to identify candidates for treatment groups and outcome labels and to identify the quantities for outcomes and group sizes. Templates are filled out with information extracted in the previous steps. If sufficient information is available, summary statistics are calculated from information in the templates.	system (BANNER) as a baseline.		measure for the calculation of summary statistics was 0.53.
Workman, 2012 (44)	A statistical abstraction summarization method based on predications form the SemanticMedline database. In Fiszman 2004 (24) described above, predication saliency is determined based on predefined and manually crafted predication schemas for each point-of-view (e.g., treatment, diagnosis). Instead, the alternative method dynamically computes saliency using statistical metrics derived from the distribution of query-specific semantic predications in the Semantic Medline database.	Corpus: citations retrieved from 8 Medline searches focused on the treatment and prevention of 4 conditions. Reference standard: 225 interventions extracted from an online clinical knowledge resource (DynaMed) and verified independently by two physicians. Disagreements were resolved by a third physician.	Precision, recall, F-measure	For drug treatment, average precision and recall of dynamic versus conventional summarization were 0.85 and 0.38 versus 0.58 and 0.71 respectively. For prevention, comparison with a baseline yielded 0.66 and 0.33 for dynamic summarization versus 0.27 and 0.24 for baseline.
Yoo, 2007 (45)	CSUGAR: Clustering-based document summarization. Documents and sentences are represented as an ontology-enriched graph. Semantically similar documents are clustered. Text summarization is performed for each document cluster.	Document sets retrieved from Medline searches for various diseases.	Misclassification index (MI) and cluster purity	Compared with baseline algorithms, CSUGAR obtained an MI of 0.053 versus 0.096 to 0.429. For cluster purity, the performance was 0.947 versus 0.601 to 0.944.

Author, year	Study method	Corpus/Gold Standard	Performance Measured	Performance Achieved
	Sentences from each document in the cluster are extracted based on their centrality in the network.			
Zhang, 2011 (47)	A modification of the method above in Fiszman, 2004 (24). The method applies degree centrality to determine the main topic of a summary and to remove predications with low connectivity. Summaries are focused on 4 aspects of a disease: comorbidities, locations, drugs, and procedures.	Dataset: 54,144 Medline citations for 5 diseases. Reference standard: 4 questions and answers for each of the 5 diseases (e.g., what drugs treat condition X, what are the anatomic locations of condition X). Baseline: an algorithm based on MetaMap.	Precision, recall, F-measure	Compared with the baseline, the average recall for degree centrality was 0.72 versus 0.85 and precision was 0.73 versus 0.33. The best recall was for procedures (0.92). The best precision was for locations and drugs (0.91).
Zhang, 2013 (48)	The method uses a graph-based, clustering approach to automatically identify themes in multi-document summarization. The method is especially aimed at supporting summarization of a large set of Medline citations. The method represents a summary as a graph composed of semantic predications. The approach uses the degree centrality method described in Zhang, 2011 (47) to remove non-salient predications from the graph and then identifies cliques in the summarized graph. For theme identification, the method applies a hierarchical clustering algorithm to cliques. A semantic theme is assigned to each cluster.	Dataset: Medline citations focused on 11 topics. Reference standard for cluster theme labels: based on MeSH terms in the citations that were sources for predications in the summary clusters. Baseline: clusters determined by the silhouette coefficient.	Cluster cohesion and separation. Precision, recall, F-measure.	Compared with the baseline, the cohesion was 0.47 versus 0.51 (not statistically significant) and separation was 0.4 versus 0.24 (statistically significant). For theme labels, recall, precision, and F-measure were 0.64, 0.65, and 0.65 respectively.