Journal of Biomedical Informatics 44 (2011) 830–838

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Degree centrality for semantic abstraction summarization of therapeutic studies

Han Zhang [a,b,*], Marcelo Fiszman [b], Dongwook Shin [b], Christopher M. Miller [b], Graciela Rosemblat [b], Thomas C. Rindflesch [b]

[a] *Department of Medical Informatics, China Medical University, Shenyang, China*
[b] *National Library of Medicine, National Institutes of Health, Bethesda, MD, United States*

ARTICLE INFO

Article history:
Received 29 December 2010
Available online 8 May 2011

Keywords:
Automatic summarization
Natural language processing
Graph theory
Degree centrality
Semantic processing
Disease treatment

ABSTRACT

Automatic summarization has been proposed to help manage the results of biomedical information retrieval systems. Semantic MEDLINE, for example, summarizes semantic predications representing assertions in MEDLINE citations. Results are presented as a graph which maintains links to the original citations. Graphs summarizing more than 500 citations are hard to read and navigate, however. We exploit graph theory for focusing these large graphs. The method is based on degree centrality, which measures connectedness in a graph. Four categories of clinical concepts related to treatment of disease were identified and presented as a summary of input text. A baseline was created using term frequency of occurrence. The system was evaluated on summaries for treatment of five diseases compared to a reference standard produced manually by two physicians. The results showed that recall for system results was 72%, precision was 73%, and *F*-score was 0.72. The system *F*-score was considerably higher than that for the baseline (0.47).

Published by Elsevier Inc.

1. Introduction

With the continuing increase in the amount of online biomedical literature, it is difficult for researchers to utilize available resources effectively. Although information retrieval systems provide potentially useful documents to users, they do not help manage the often large amount of information returned in response to users' queries. For example, PubMed gives access to more than 19 million MEDLINE citations in the biomedical research literature. Queries can retrieve tens of thousands of documents (e.g. "breast cancer" returns 217,101 citations), and returned documents are not ranked by relevance but by date.

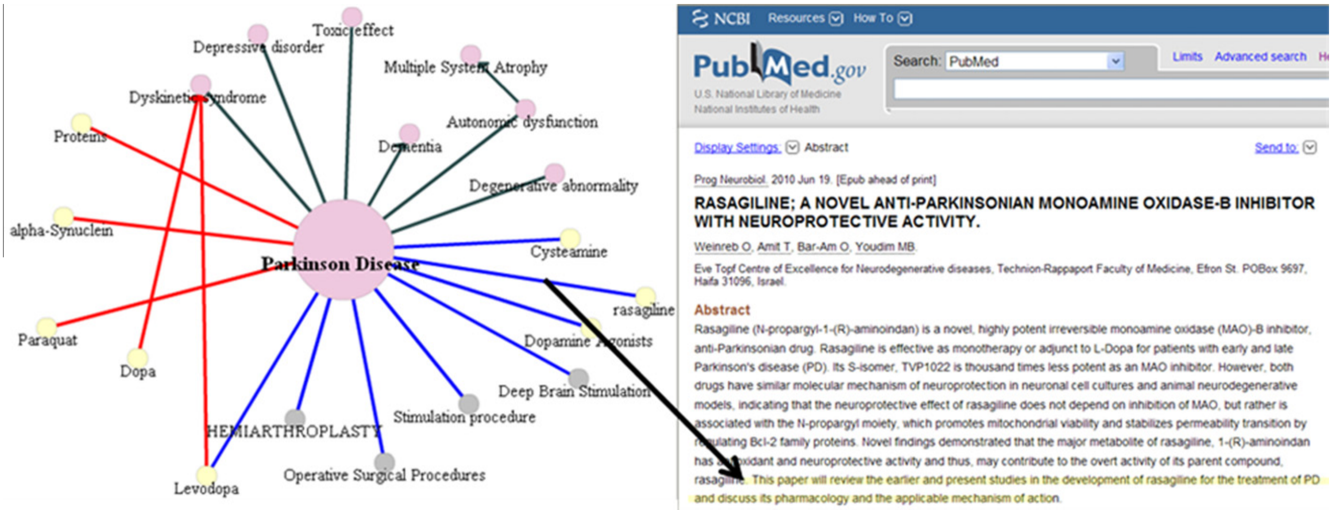
Recently, automatic summarization has been proposed as a way to help users extract needed information from large numbers of biomedical documents [1]. Automatic summarization [2,3] is "a reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source." There are two critical issues: selecting important content from the information source and presenting it to users effectively. Relying on concepts co-occurring in documents, several recent information extraction systems [4–6] visualize the biomedical literature retrieved using PubMed as a graph, with concepts represented as nodes and relations between them as edges. This way of displaying text information offers a new solution to present summary. By displaying the sentences that produced the links in the graph, a guide to investigating the underlying information of the summary is also available. This is especially effective in providing knowledge rich summaries for large numbers of biomedical documents.

To help medical researchers and practitioners keep current with the progress of biomedical research, Fiszman et al. devised a knowledge-rich abstraction summarization system [7] for MEDLINE citations based on semantic predications from SemRep [8]. Summarized predications are displayed in the Semantic MEDLINE application as a graph which maintains links to the original MEDLINE citations [9]. For example, Fig. 1 shows a graph representing the predications summarizing 500 citations on Parkinson's disease. Arguments are represented as nodes and predicates as color coded arcs: blue for TREATS, red for CAUSES, and green for COEXISTS_WITH. In Fig. 1, the arrow links the predication "rasagiline TREATS Parkinson Disease" to the sentence (highlighted) in a citation from which it was extracted.

The results of summarizing a small number of citations (500 or fewer) are promising, but the graph generated for large data sets is too dense. For example, Fig. 2 illustrates part of the summary of 2000 citations on Parkinson's disease. Although users need summaries for large sets of documents [10], this graph is too cluttered to be effective. In this paper, we propose a graph-theoretic method that renders the results of summarizing large numbers of MEDLINE citations more accessible and useful. Relying on a disease treatment schema, we sift the relevant information, and then condense it by keeping only predications with highly connected concepts.

* Corresponding author. Address: Department of Medical Informatics, China Medical University, 92, Bei 2 Road, Shenyang, China. Fax: +86 24 23211577.
E-mail address: zhanghan@mail.cmu.edu.cn (H. Zhang).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2011.05.001



Prog Neurobiol. 2010 Jun 19. [Epub ahead of print]

RASAGILINE; A NOVEL ANTI-PARKINSONIAN MONOAMINE OXIDASE-B INHIBITOR WITH NEUROPROTECTIVE ACTIVITY.

Weinreb O, Amit T, Bar-Am O, Youdim MB.

Eve Topf Centre of Excellence for Neurodegenerative diseases, Technion-Rappaport Faculty of Medicine, Efron St. POBox 9697, Haifa 31096, Israel.

Abstract

Rasagiline (N-propargyl-1-(R)-aminoindan) is a novel, highly potent irreversible monoamine oxidase (MAO)-B inhibitor, anti-Parkinsonian drug. Rasagiline is effective as monotherapy or adjunct to L-Dopa for patients with early and late Parkinson's disease (PD). Its S-isomer, TVP1022 is thousand times less potent as an MAO inhibitor. However, both drugs have similar molecular mechanism of neuroprotection in neuronal cell cultures and animal neurodegenerative models, indicating that the neuroprotective effect of rasagiline does not depend on inhibition of MAO, but rather is associated with the N-propargyl moiety, which promotes mitochondrial viability and stabilizes permeability transition by regulating Bcl-2 family proteins. Novel findings demonstrated that the major metabolite of rasagiline, 1-(R)-aminoindan has antioxidant and neuroprotective activity and thus, may contribute to the overt activity of its parent compound, rasagiline. This paper will review the earlier and present studies in the development of rasagiline for the treatment of PD and discuss its pharmacology and the applicable mechanism of action.

Fig. 1. Summary of 500 MEDLINE citations on Parkinson's disease.

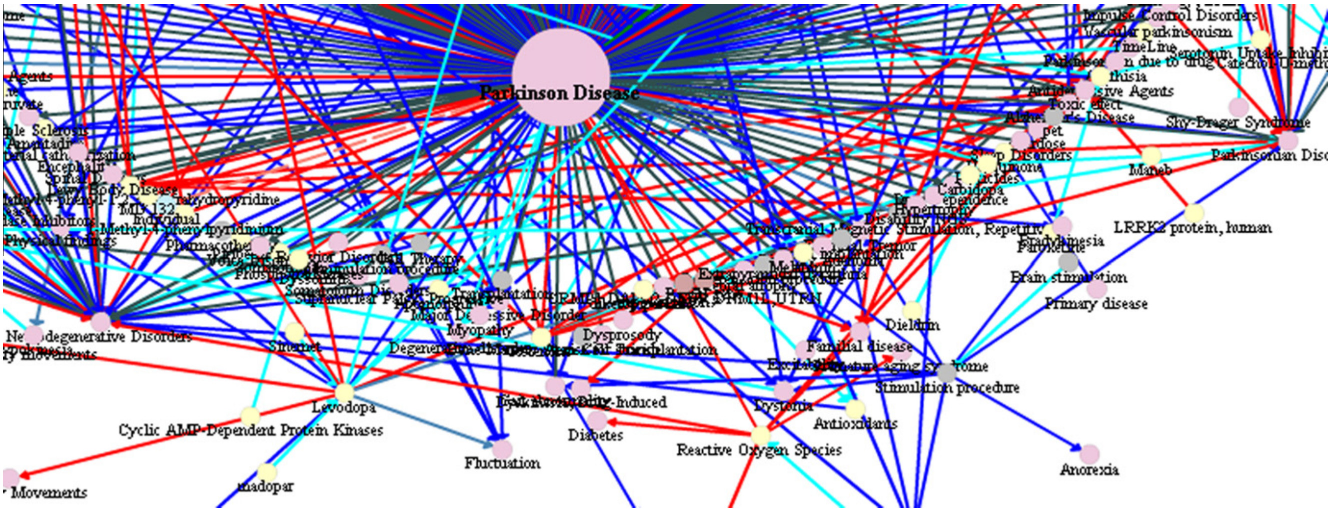


Fig. 2. Summary of 2000 MEDLINE citations on Parkinson's disease.

The core notion is that such concepts in a graphical representation of a summary convey information crucial to the summary. The proposed system is innovative in that it exploits graph theory to extend a semantic abstraction method for summarizing multiple biomedical texts. The principal aim of this work is to demonstrate the effectiveness of degree centrality in selecting information crucial for summarization. The underlying principle is that connectedness of arguments can be used to identify salient information for researchers and clinicians.

2. Background

2.1. Unified medical language system

The system described in this paper depends on the Unified Medical Language System® (UMLS)® [11,12] knowledge sources. The Metathesaurus is at the core of the UMLS and contains more than 1.3 million concepts compiled from more than 100 controlled vocabularies (in the 2006 version of the UMLS used for this project). The Metathesaurus combines terms in the constituent vocabularies into a set of synonyms, which constitutes a concept. Each concept is assigned at least one semantic type (such as 'Sign or Symptom', 'Disease or Syndrome', or 'Pharmacologic Substance'), which categorizes it in the biomedical domain. Semantic groups [13] organize semantic types into fifteen coarser aggregates such as Anatomy, Activities and Behaviors, Chemicals and Drugs, Disorders, and Living Beings.

UMLS semantic types are drawn from the Semantic Network, which also contains semantic predications with semantic types as arguments. The predications are semantic relations relevant to the biomedical domain, such as 'Pharmacologic Substance' TREATS 'Disease or Syndrome' and 'Disease or Syndrome' HAS_LOCATION 'Body Part, Organ, or Organ Component'.

2.2. SemRep

Following Fiszman et al., our summarizer relies on SemRep [8], a program that automatically extracts semantic predications from MEDLINE citations (titles and abstracts) using the UMLS. Based on an underspecified syntactic analysis that relies on the SPECIALIST Lexicon [14] and the MedPost tagger [15], the application maps noun phrases to concepts in the Metathesaurus using MetaMap [16] and finds relations between them, guided by the Semantic Network. MetaMap matches noun phrases to concepts in the

Metathesaurus and retrieves a semantic type for each concept found. When a phrase maps to more than one concept, a statistical word-sense disambiguation system [17] selects the best concept based on the semantic type appropriate for the context. SemRep predications are in the form subject–predicate–object, in which the subject and object are Metathesaurus concepts and the predicate is a relation from the Semantic Network. For example, SemRep identifies the predication (2) from (1):

(1) Advances in taxane therapy for breast cancer
(2) Taxanes TREATS Breast Carcinoma

2.3. Automatic summarization

An important aspect of automatic summarization is to recognize core content from source documents. Previous work has used several indicators of important words and sentences in biomedical documents, usually based on term or concept frequency [18–20]. Some systems exploit domain knowledge (such as the UMLS) to facilitate the representation of documents. For example, Reeve and his colleagues [21] select frequent concepts from relevant semantic types identified by experts in the oncology clinical trial domain, and extract sentences containing frequent concepts to serve as a summary. Combining information retrieval and summarization, Demner-Fushman [22] constructs a question answering system for clinical medicine. Semantic types are used to identify drug concepts from MEDLINE citations, and UMLS semantic relationships, such as hypernymy, are used to cluster drugs that share an ancestor.

Multidocument summarization provides a concise description of large numbers of documents. Extractive methods select salient sentences from the source and concatenate them into a summary, while abstractive techniques operate on a structured representation of the meaning of the source and produce novel sentences or terms for the summary [23]. Most genres for both methods are news articles covering, for example, accidents, natural disasters, and terrorist attacks. Only a few studies have focused on the biomedical literature [18–22,24–28].

Fiszman et al. [7,29] identify core content in input documents through abstraction processing that relies on a user-specified topic and a compressing technique composed of four steps: Relevance, Connectivity, Novelty, and Saliency. The Relevance step uses a schema that defines core semantic predications for several subdomains of biomedical research. There is a separate schema for treatment of disease [7], substance interactions [30], pharmacogenomics [31], and genetic etiology of disease [32]. The treatment schema, for example, would allow the predication "Rifampin TREATS Tuberculosis" to be retained for a summary on topic tuberculosis. The Connectivity step would include predications related to those identified with relevance, such as "Rifampin INTERACTS_WITH linezolid" for this topic. The novelty step eliminates predications with general, uninformative arguments, that is, those that occur close to the root node in the UMLS hierarchy. For example, the predication "Pharmaceutical Preparations TREATS Tuberculosis" would be eliminated because the concept "Pharmaceutical Preparations" is near the root. Finally, during Saliency, predications are removed which have a frequency of occurrence less than the average [33].

Fiszman et al. then represent the summarized list of predications as a directed graph in which nodes are arguments (UMLS Metathesaurus concepts) and arcs are predicates (UMLS Semantic Network relations). For each unique predication in the list, the node representing the subject concept is connected by an arc pointing to the node representing the object concept. Predications that share an argument (either subject or object) are represented in the graph as a node with multiple arcs. For example, unique predications "Penicillin–TREATS–Bell's Palsy" and "Acyclovir–TREATS–Bell's Palsy" may represent any number of instances of each predication in the summarized list. These two predications are represented in the graph as a node for "Bell's Palsy" with two arcs labeled "TREATS" connected to it, one directed from a node for "Penicillin" and the other from a node for "Acyclovir." Although the predications are represented as a graph, there is no graph metric used in Semantic MEDLINE to help the summarization process. In this paper, we exploit the graph theoretic notion of degree centrality to identify crucial concepts in a summary represented as a connected graph of semantic predications.

2.4. Graph theory for automatic summarization

Recently, graph theory has been combined with natural language processing and statistics for automatic summarization [34–39] and question answering task [40]. Systems usually represent aspects of the text being summarized as a graph. For example, in LexRank [35], sentences are represented as nodes and similarities between them as links. Centrality is computed taking into consideration similarity between sentences, and nodes with higher centrality are deemed as being more important for the summary.

There are several types of centrality, all based on the connectedness of a node to other nodes in the graph, and all are used to identify nodes important to the graph. Degree centrality is based on the degree of a node, that is, the number of arcs directly connected to it. Zhang et al. [41] compared several ways of computing centrality (degree centrality, shortest-path-based centrality and eigenvector centrality) and report that degree centrality is most effective in identifying nodes that humans judge to be important in a graph representing summarized information about vocabularies used in the Semantic Web. Erkan and Radev [35] also compared different centrality methods (degree centrality, LexRank and centroid) and found that those based on degree outperform other approaches. Although our graphs are not identical to those on which these comparisons were based, essential similarities led us to choose degree centrality as an effective metric for determining the importance of nodes for automatic summarization.

In a directed graph, the degree of a node can be computed based on a distinction between incoming and outgoing arcs, as is often done in social network analysis (e.g. Nooy [42] analyzing friendship patterns). Although our graphs are directed, in that arcs asymmetrically encode a connection between a subject and an object argument of a predication, we ignore arc direction when computing degree centrality. The consequences of this are minimal since the arcs in our graphs represent relationships that are inherently unidirectional, such as TREATS, LOCATION_OF, and CAUSES. For example, in the predication "levodopa TREATS Parkinson disease", the direction of the arc is uniquely determined by the meaning of TREATS (from the drug to the disease), and it is impossible for TREATS to join a drug and a disease in the opposite direction. Due to this fact, although our graph is displayed as a directed graph, we treat it as undirected when computing centrality.

For a graph $G: = (V, E)$ with n nodes, the formula for degree centrality $C_D(v)$ for node v is: $C_D(v) = deg\,(v)/(n-1)$, in which $deg\,(v)$ is the degree of node v (the number of lines connected to it). For example, the degree centrality of node "parkinson disease" in Fig. 3 is 0.83 (5/6).

3. Methods

3.1. Processing overview

Our method for modifying the summarization system of Fiszman et al. [7] based on degree centrality takes as input SemRep predications extracted from MEDLINE citations on some disease

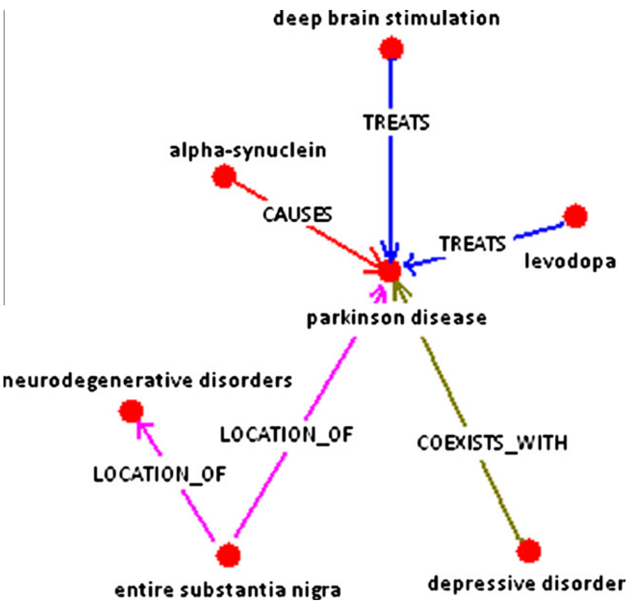


Fig. 3. Node degree.

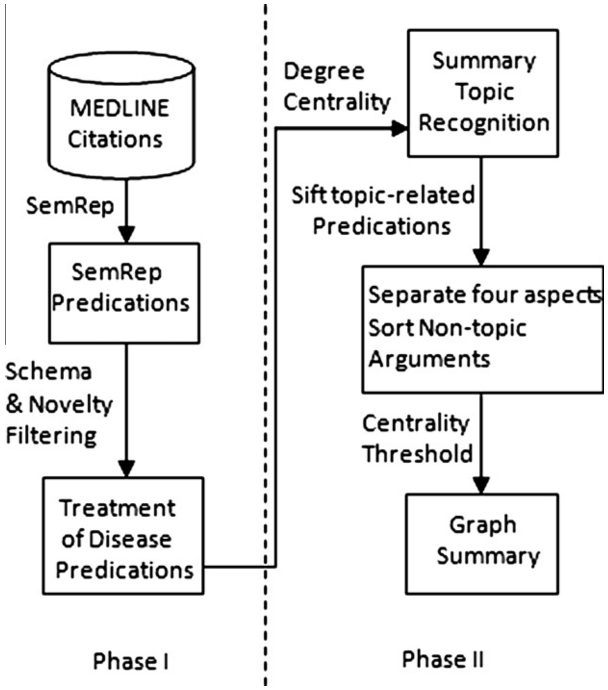

MEDLINE Citations
SemRep
SemRep Predications
Schema & Novelty Filtering
Treatment of Disease Predications
Phase I
Degree Centrality
Summary Topic Recognition
Sift topic-related Predications
Separate four aspects Sort Non-topic Arguments
Centrality Threshold
Graph Summary
Phase II

Fig. 4. General overview of the summarizing procedure.

and then proceeds in two phases. The first exploits a schema to keep only predications in the four major aspects of treatment: Comorbidities, Location, Drugs, and Procedures. Novelty processing then eliminates predications with generic arguments. In the second phase, degree centrality is first calculated and is then used to determine the topic of the summary and to eliminate predications with arguments of low connectivity. This filtering is done based on a formula for setting a degree centrality threshold, which is applied separately in each of the four treatment aspects. The overall process is illustrated in Fig. 4.

3.2. Schema with four aspects

In order to implement a schema on treatment of disease, we rely on formally defined metapredications for each of four major aspects of therapy. For each metapredication, a predicate is first stipulated and then arguments are defined generally as domains based on UMLS Semantic Network semantic groups [13]. Four argument domains are used: *Disorders* includes all semantic types in the semantic group Disorders (such as 'Disease or Syndrome' and 'Neoplastic Process'); *Location* has semantic types such as 'Body Part, Organ, or Organ Component'; *Drugs* has 'Antibiotic', 'Pharmacologic Substance', and 'Organic Chemical', among others; and *Procedures* has 'Therapeutic or Preventive Procedure'. The metapredications for each aspect are shown in Table 1.

In filtering SemRep predications through the schema, SemRep predicates match predicates in the metapredication, and SemRep arguments having UMLS semantic types as defined for the argument domain match that domain. For example, the predication "Depressive disorder COEXISTS_WITH Parkinson Disease" matches the metapredication for comorbidities because "Depressive disorder" has semantic type 'Mental or Behavioral Dysfunction' and "Parkinson Disease" has 'Disease or Syndrome', both of which are in the Disorders argument domain.

3.3. Data set

Citations for four diseases were used as a development set: Parkinson's disease (10,497 citations), breast cancer (48,900), hepatitis B (7252), and rheumatoid arthritis (14,141), while five were selected for testing: Alzheimer's disease (16,413), migraine (4245), peptic ulcer (3693), heart failure (16,358), and melanoma (13,951). For each disease, MEDLINE citations were retrieved with a PubMed query using the disease name as the major MeSH topic, limited to English, along with publication dates between 2000 and 2009. Of the 54,660 citations retrieved for testing, 516 also appeared in the development data. These were eliminated before evaluation, so that there was no overlap between development and testing data.

3.4. Processing

Processing is illustrated with Parkinson's disease. SemRep extracted 69,142 predication tokens from the 10,497 citations retrieved for that disorder. Predication tokens were represented by 19,885 unique predication types, which were subjected to further processing. These had predicates on all aspects of the disease, including genetic etiology (e.g. PREDISPOSES and ASSOCIATED_WITH) and related substance interactions (e.g. INHIBITS, STIMULATES, and INTERACTS_WITH). In order to focus on treatment, predications were filtered through the metapredications of the schema, keeping only those with predicates COEXISTS_WITH, LOCATION_OF, TREATS, and PREVENTS. After this step, 6757 predications remained. Novelty eliminated a further 4702 predications with uninformative arguments, leaving 2055 for the second summarization phase.

Degree centrality was computed for all nodes (1088) in the graph of the remaining predications, and nodes were sorted in

Table 1
Metapredications for the four aspects of treatment.

Aspect	Metapredication
Comorbidities	{Disorders} CO-EXISTS_WITH {Disorders}
Location	{Anatomy} LOCATION_OF {Disorders}
Drugs	{Drugs} TREATS or PREVENTS {Disorders}
Procedures	{Procedures} TREATS or PREVENTS {Disorders}

descending order of degree centrality. The node with the highest degree centrality was designated as the topic of the summary. In this graph, that node is "Parkinson Disease" with degree centrality of 0.158, which is many times greater than that of the next highest, "Dementia" (0.025).

The list of predications was further condensed by restricting one of the arguments to the topic concept. For example, "Levodopa TREATS Parkinson Disease" satisfies this criterion, but "Antipsychotic Agents TREATS Nonorganic psychosis" does not. Predications not having "Parkinson Disease" as an argument were eliminated, leaving 675 predications. These were then separated into the four aspects of the summary by matching to the metapredications; Comorbidities had 304 predications, Location 93, Drugs 167, and Procedures 111. For each aspect, predications were sorted in descending order of degree centrality of their non-topic arguments.

The next step was to determine a degree centrality cutoff for each aspect to filter out the non-topic arguments with lower connectedness. This was based on informal, provisional annotation of the development data by comparing the sorted lists of non-topic arguments to concepts found in published review articles relevant to each disorder. Using this annotation, we found that the distribution of degree centrality varies among the four aspects, and concepts in the Comorbidities aspect usually have the highest degree centrality. The next highest is in Location, while Drugs and Procedures have relatively low degree centrality. The final determination of the cutoff formula, which is the mean of the sum of the degree centrality values plus the standard deviation, takes into consideration the distribution of degree centrality in each aspect. The cumulative degree centrality of concepts with values above the cutoff averaged 47.5% over all the aspects in development data. Subsequent cutoff filtering was applied to each aspect individually and non-topic arguments falling below the cutoff point were eliminated.

For example the nodes of the non-topic arguments in the 111 predications from the Procedures aspect had degree centrality values ranging from 0.09% to 4.22%. The mean of the sum of these is 0.30%, with standard deviation 0.53%. The cutoff is thus 0.83%. Table 2 illustrates part of the ranked list with cutoff for this aspect.

After eliminating predications below the cutoff in each aspect, 45 total predication types remained (17 in the Comorbidities aspect, 8 in Location, 14 in Drugs, and 6 in Procedures). Taken together, these predications constitute the therapeutic summary for this disease topic. The connected graph of the predications is shown in Fig. 5. In this graph, line length is not significant. The color of the nodes represents the semantic type of the concepts (for example, white: Pathologic Function; pink: Cell or Molecular Dysfunction, etc.).

3.5. Evaluation

3.5.1. Overview

In most evaluation studies of multidocument summarization [43–45], reference standard summaries are produced by experts, and measures of intra- and inter-rater agreement are provided. The systems are contrasted quantitatively with the reference standards and performance measures are computed. Other evaluation studies are user-centered, which seek to assess a summary on how well a user can exploit it to perform a given information retrieval task [46,47]. Recently, Amigo et al. [48] proposed an "information synthesis" task, defined as "given a specific information need, the multidocument summary should extract, organize, and synthesize an answer that satisfies that need." Based on this proposal, the annual Document Understanding Conference (DUC) [49] was reengineered to address a more focused, topic-oriented approach to evaluating automatic summarization systems. The topic invokes questions and human assembly of answers so they can be compared against the results of the summarizers.

Table 2
Part of the predications for the Procedures aspect.

Predications	Degree centrality (%)
Deep Brain Stimulation TREATS Parkinson Disease	4.23
Stimulation procedure TREATS Parkinson Disease	3.49
Pallidotomy TREATS Parkinson Disease	1.29
Injection procedure TREATS Parkinson Disease	1.01
Transplantation TREATS Parkinson Disease	0.92
Transcranial Magnetic Stimulation, Repetitive TREATS Parkinson Disease	0.83
Cutoff = 0.83	
Replacement therapy TREATS Parkinson Disease	0.64
Thalamotomy TREATS Parkinson Disease	0.64
Observation TREATS Parkinson Disease	0.55
Implantation procedure TREATS Parkinson Disease	0.55
Pet TREATS Parkinson Disease	0.46
Monitoring TREATS Parkinson Disease	0.46
Neurosurgical Procedures TREATS Parkinson Disease	0.46
Transcranial magnetic stimulation TREATS Parkinson Disease	0.46
Supplementation TREATS Parkinson Disease	0.46
Detection TREATS Parkinson Disease	0.46

Following Amigo, our evaluation is topic-oriented based on questions physicians might have about the diseases in the testing set (Alzheimer's disease, migraine, peptic ulcer, heart failure, and melanoma). A reference standard was constructed, and system output and baseline results were evaluated against it.

3.5.2. Reference standard

Two of the authors (MF and CMM), physicians not involved in system design, constructed a reference standard for the five testing diseases. Four questions (topics) correlated with the aspects of the summarization schema were articulated for each disease, according to the following patterns:

- What are the comorbidities (related disorders) of disease X?
- What are the anatomic locations of disease X?
- What are the drugs used to treat or prevent disease X?
- What are the therapeutic procedures used to treat disease X?

For each question, for each disease, the physicians identified answer terms (independently) after consulting two electronic books widely used in internal medicine (Harrison's Principles of Internal Medicine [50] and Current Medical Diagnosis and Treatment [51]). Inter-rater agreement was measured and disagreements were resolved by consensus [44].

Table 3 shows inter-rater agreement for each of the five diseases. Overall agreement was good; most of the problems arose in considering phenomena such as "behavioral changes" and "cognitive changes" as pathologic functions or as comorbidities to the tested diseases. After discussion, it was decided that they were to be assigned as comorbidities.

An illustration of part of the reference standard is given in Table 4.

Table 5 shows the total number of terms in the final reference standard. The average number of terms assigned to each disease is 32.

3.5.3. Baseline

The baseline was constructed using MetaMap to extract Metathesaurus concepts from the citations being summarized. For each question, a list of concepts was generated as follows: Topic concepts were eliminated, as were uninformative concepts,

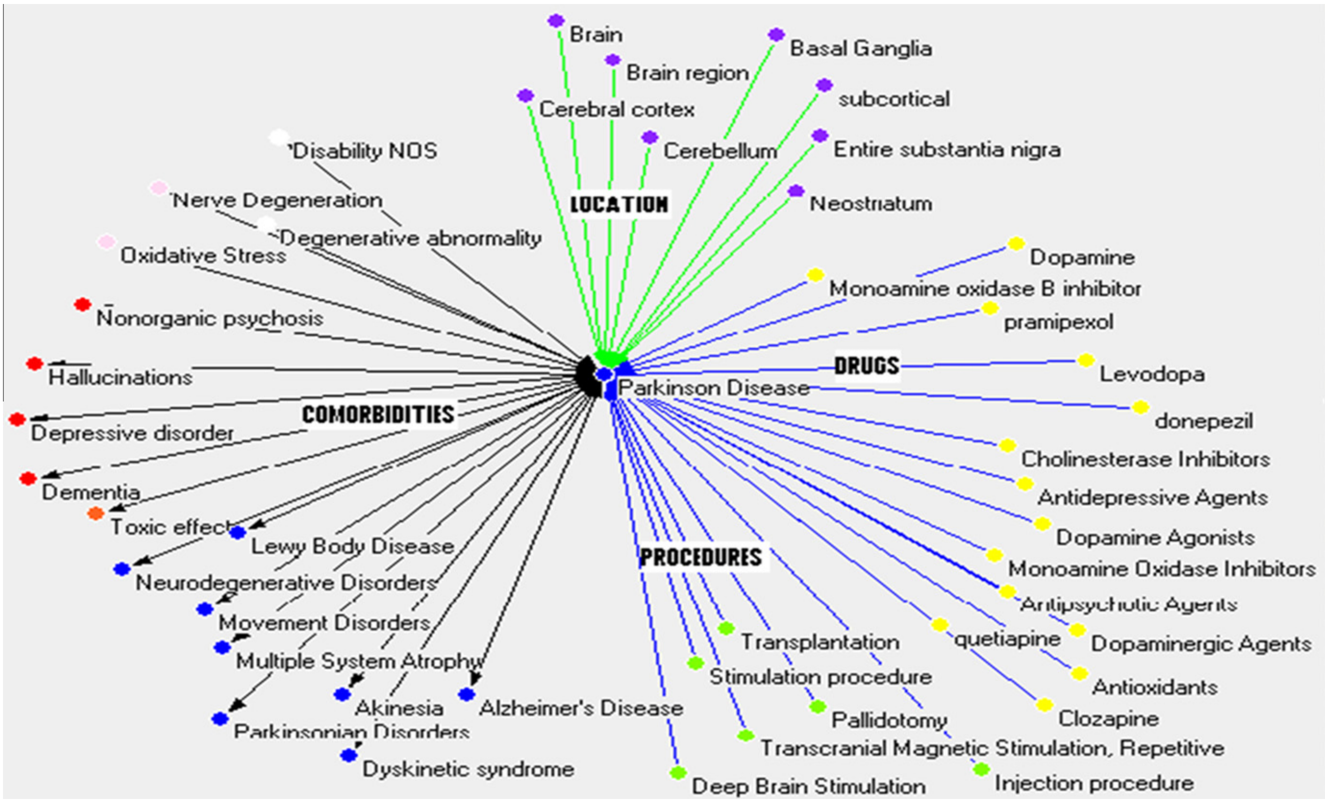


Fig. 5. Parkinson's disease: summary of 10,497 citations.

Table 3
Agreement between two physician judges for the five diseases.

Disease	Agreement (%)
Alzheimer's disease	89
Heart failure	88
Melanoma	86
Migraine	90
Peptic ulcer	79
Overall	87

using Novelty processing. Further, concepts whose semantic type did not match the metapredication representing the aspect corresponding to the question (see Section 3.2) were eliminated. Remaining concepts were ordered by frequency of occurrence, and those with a value at or above a cutoff were considered to be a treatment summary answering the relevant question. The formula for computing the cutoff for the baseline was similar to that used for system results: mean frequency of occurrence of the concepts in the baseline summary was added to the standard deviation.

3.5.4. Comparing results to the reference standard

Since the results of both the system and the baseline are UMLS concepts and the reference standard contains terms, results were matched manually to the reference standard by the second author (MF). Matching comorbidities, locations, and procedures was straightforward; however, drugs were matched hierarchically. Drug concepts that were class members returned by the summarization system and baseline were allowed to match to their respective classes in the reference standard and were counted as true positives for the whole class. For example, the concept "donepezil" in system output as a drug used to treat Alzheimer's disease was counted as a true positive for the class "cholinesterase inhibitors" in the reference standard. After matching, standard informatics performance metrics were computed for both system results and baseline: recall, precision, and *F*-score.

Table 4
Reference standard for "What are the anatomic locations of melanoma?".

Terms
Palms
Soles
Nail
Face
Back
Lower leg
Hands
Forearm
Scalp
Feet
Regional lymph nodes
Liver
Lung
Bone
Brain
Eye
Skin

4. Results

For each disease evaluated, Table 6 provides an overview of the distribution of citations retrieved with PubMed, predications initially extracted with SemRep, and nodes in the summarized graphs. The number of citations ranged from 4250 to 16,697, while the number of predications initially extracted from these citations varied from 17,183 to 136,677. After summarization, the total

Table 5
Number of terms in the reference standard by disease and aspect.

	Alzheimer's disease	Heart failure	Melanoma	Migraine	Peptic ulcer	Overall
Comorbidities	23	22	2	2	7	56
Locations	6	3	17	2	2	30
Drugs	10	17	10	15	5	57
Procedures	1	4	6	2	5	18
Overall	40	46	35	21	19	161

Table 6
Citations, predications, and nodes in system summary and baseline.

Disease	No. of citations	No. of predications	No. of nodes	
			System summary	Baseline
Alzheimer's disease	16,697	107,807	59	116
Heart Failure	16,403	136,677	54	174
Melanoma	14,118	101,603	62	229
Migraine	4250	17,183	19	62
Peptic ulcer	3708	33,412	15	113
Average	11,035	79,336	42	139

Table 7
Performance metrics on the five diseases for the summarization system (SS) and baseline (BL).

Disease	Recall		Precision		*F*-score	
	SS	BL	SS	BL	SS	BL
Alzheimer's disease	0.74	0.74	0.67	0.34	0.70	0.46
Heart failure	0.71	0.89	0.77	0.42	0.74	0.57
Melanoma	0.88	0.90	0.72	0.23	0.79	0.37
Migraine	0.50	0.75	0.72	0.39	0.59	0.52
Peptic ulcer	0.61	0.93	0.79	0.35	0.69	0.51
Overall	0.72	0.85	0.73	0.33	0.72	0.47

Table 8
Performance metrics for the four aspects for the summarization system (SS) and baseline (BL).

Aspect	Recall		Precision		*F*-score	
	SS	BL	SS	BL	SS	BL
Comorbidities	0.51	0.77	0.67	0.42	0.58	0.54
Locations	0.86	0.93	0.91	0.52	0.88	0.67
Drugs	0.75	0.87	0.91	0.27	0.82	0.41
Procedures	0.92	0.92	0.41	0.27	0.56	0.41

number of nodes for the five diseases in the summarized results ranged from 15 to 62. In the baseline, the total number of nodes varied from 62 to 229, with the average 139.

The results of comparing the summarization system and baseline to the reference standard for the five diseases are shown in Table 7. Although recall was better in the baseline, there was an overall improvement of .25 in the *F*-score for the summarization system due to better precision.

Results for each of the four aspects (Comorbidities, Location, Drugs, and Procedures) for the system and baseline are presented in Table 8. As with overall results, the baseline had slightly better recall for each aspect, but the summarization system had much better precision in all cases. The improvement in the *F*-score was 0.04 for Comorbidities, 0.21 for Location, 0.41 for Drugs, and 0.15 for Procedures. Note that the least improvement was for Comorbidities, while the greatest was for Drugs.

5. Discussion

Evaluation results suggest that degree centrality computed on a connected graph of semantic predications provides an effective mechanism for selecting clinically useful information from MEDLINE citations, especially for large data sets. As indicated by the *F*-score, overall system performance is better than the baseline, which exploits concept frequency of occurrence. Although system precision is significantly higher than that of the baseline (73% versus 33%), the baseline produced somewhat better recall (85% compared to 72%).

5.1. Error analysis

System recall is dependent on the cutoff value determined by the formula applied to predications extracted from citations on the four diseases in the development set: Parkinson's disease (10,497 citations), breast cancer (48,900), hepatitis B (7252), and rheumatoid arthritis (14,141). This formula did not produce optimal recall results, particularly for comorbidities. For example, important comorbidities of Alzheimer's disease, such as stroke and diabetes, were below the cutoff.

5.1.1. False negatives

In specific instances examined, lowering the cutoff would significantly improve recall without diminishing precision. For example, Table 9 lists the drugs for peptic ulcer above the cutoff, with recall and precision of 0.57 and 0.67 respectively. If the cutoff is lowered to include three times as many concepts (18, rather than 6), recall increases to 0.86, while precision remains the same. Additional development data is required to calibrate the cutoff formula to maintain a high *F*-score generally. Such a case only occurs in

Table 9
Drugs for peptic ulcer.

System output	True positive
Proton pump inhibitors	Y
Omeprazole	Y
Anti-inflammatory agents, non-steroidal	N
Ranitidine	Y
Rabeprazole	Y
Antioxidants	N
Cut off = 0.176%	
Famotidine	Y
Histamine H2 antagonists	Y
Esomeprazole	Y
Celecoxib	N
Antibiotics	Y
Rebamipide	Y
Misoprostol	Y
Sucralfate	Y
Indomethacin	N
Nizatidine	Y
Ethanol	N
Analgesics	N

drugs for peptic ulcer and migraine. As for the other three diseases in the testing set, lowering the cutoff does not improve recall, yet lowers precision.

5.1.2. False positives

Analysis revealed that false positives fall into three categories. In descending order of frequency they are: (1) concepts that do not appear in the reference standard (37%), (2) concepts that are too general (33%), and (3) infelicitous mappings to the UMLS Metathesaurus (30%).

The first error type included concepts which have only recently appeared in the research literature. For example, the concept "Inflammation" was found by the summarization system as related to Alzheimer's disease. This relationship is clearly discussed in MEDLINE (PMID 19738171), despite not appearing in either of the medical textbooks that were used.

Errors of the second type are often due to missing concepts in the Metathesaurus. For example, SemRep extracted the predication (2) from text (1) based on mapping the noun phrase *androgen supplementation* to the concept "Supplementation." The more specific concept "Androgen Supplementation" does not occur in the Metathesaurus

(1) Androgen supplementation may be beneficial in Alzheimer's disease.
(2) Supplementation TREATS Alzheimer Disease

Errors of the third type are caused by word sense ambiguity in the Metathesaurus. For example, text phrases *division* and *power* are mapped to concepts "Division (procedure)" and "Power (procedure)" (both with semantic type 'Therapeutic or Preventive Procedure') due to incorrect word sense disambiguation when SemRep extracts predications from titles and abstracts.

5.2. Limitations

Although our results showed that most of the important concepts for disease treatment research could be recognized and were useful for physicians, there are limitations to the system. We have so far considered concepts in the final output as isolated terms, although some of them in hierarchical relationship to each other. For example, both "Cholinesterase inhibitors" and children, "donepezil," "rivastigmine," "galantamine," and "tacrine" were recognized for the question, "What are the drugs used to treat or prevent Alzheimer's disease?" However, the system does not overtly indicate that they are related. Although SemRep provides hierarchical relations between concepts with predications such as "Donepezil ISA Cholinesterase Inhibitors," they are currently eliminated during summarization processing. In addition, after the topic concept was defined, selection of related predication was based on exact match, which means that predications having an argument that is child of the topic concept were not included. This may result in some information loss.

Since the process of manually creating a reference standard is onerous, we limited the evaluation to results for five diseases. A related limitation is that we focused on treatment, and did not test whether degree centrality is an effective mechanism to summarize research on diagnosis or etiology, for example. Nonetheless therapy is at the core of medicine, and the presentation of treatment modalities and other pertinent information regarding a specific disease would likely be of use as an adjunct to clinical guidelines. Another potential use for this automatic technique might be as an aid to experts involved in the establishment of guidelines, to both first-line, as well as infrequently used or experimental therapies.

5.3. Future work

Although the results of this study showed that degree centrality can select important concepts for summarization, the method depends on a manually created schema. We are exploring a graph-based approach that exploits degree centrality in addition to other graph-theoretic constructs, such as cliques (clustered with statistical methods), as well as frequency of occurrence. This has the potential to automatically partition the whole semantic predication graph into meaningful subgraphs (subsets) for any given topic without the need for predefined schemas.

6. Conclusion

We exploited a graph theoretic method for extracting the most important information from large graphs in Semantic MEDLINE, an application which uses automatic summarization to help manage citations returned by PubMed. Semantic MEDLINE presents summarized results to the user as a graph which maintains links to the original citations. However, graphs summarizing more than 500 citations are hard to read and navigate. Our method isolates the most important information from large graphs based on degree centrality, which measures node connectedness in a graph and correlates well with information likely to be important for a summary.

The system was tested on summaries containing four aspects of treatment of five diseases. Physicians manually produced lists of clinically important concepts for those four aspects of each disease, which served as a reference standard. A baseline was created by identifying frequently occurring concepts in relevant MEDLINE citations. The system and the baseline were compared to the reference standard, and results showed that the overall performance of the system was significantly better than the baseline.

Acknowledgments

The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the US Department of Energy and the National Library of Medicine. This study was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

[1] Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform 2009;42(5):801–13.
[2] Sparck Jones K. Automatic summarizing: factors and directions. In: Mani I, Maybury MT, editors. Advances in automatic text summarization. London: The MIT Press; 1999. p. 1–12.
[3] Marcu D. The theory and practice of discourse parsing and summarization. London: The MIT Press; 2000.
[4] Jensen TK, Laegreid A, Komorowski J, Hovig. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001;28(1):21–8.
[5] Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. ALIBABA: PubMed as a graph. Bioinformatics 2006;22(19):2444–5.
[6] Feldman R, Regev Y, Hurvitz E, Finkelstein-Landau M. Mining the biomedical literature using semantic analysis and natural language processing techniques. Biosilico 2003;1(2):69–80.
[7] Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. In: Proceedings of the HLT-NAACL workshop on computational lexical semantics; 2004. p. 76–83.
[8] Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller SS, Friedman C, Hersh W, editors. Medical informatics: knowledge management and data mining in biomedicine. New York: Springer; 2005. p. 399–422.
[9] Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple AM, Rindflesch TC. Semantic MEDLINE: a web application to manage the results of PubMed searches. In: Proceedings of the third international symposium for semantic mining in biomedicine; 2008. p. 69–76.

[10] Stein GS, Strzalkowski T, Wise GB. Interactive, text-based summarization of multiple documents. Comput Intell 2000;16(4):606–13.
[11] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. J Am Med Inform Assoc 1998;5(1):1–11.
[12] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue): D267–270.
[13] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Proc Medinfo 2001;10(Pt 1):216–20.
[14] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Symposium on computer applications in medicine; 1994. p. 235–9.
[15] Smith LH, Rindflesch TC, Wilbur WJ. The importance of the lexicon in tagging biological text. Nat Lang Eng 2005;12(4):335–51.
[16] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
[17] Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. J Am Soc Inf Sci Technol 2006;57(1):96–113.
[18] 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.
[19] Reeve L, Han H, Nagori S, Yang JC, Schwimmer TA, Brooks AD. Concept frequency distribution in biomedical text summarization. In: Proceedings of the 15th ACM international conference on information and knowledge management; 2006. p. 604–11.
[20] Reeve L, Han H, Brooks AD. Biochain: lexical chaining methods for biomedical text summarization. The 21st annual ACM symposium on applied computing; 2006. p. 23–7.
[21] Reeve L, Han H. Biomedical text summarisation using concept chains. Int J Data Min Bioinform 2007;1(4):389–407.
[22] Demner-Fushman D, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics; 2006. p. 841–8.
[23] Hahn U, Mani I. The challenges of automatic summarization. IEEE Comput 2000;33(11):29–36.
[24] Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical document: a survey. Artif Intell Med 2005;33(2):157–77.
[25] Reichert D, Kaufman D, Bloxham B, Chase H, Elhadad N. Cognitive analysis of the summarization of longitudinal patient records. In: AMIA annu symp proc; 2010. p. 667–71.
[26] Vleck TV, Elhadad N. Corpus-based problem selection for EHR note summarization. In: AMIA annu symp proc; 2010. p. 817–21.
[27] Yu H. Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. In: AMIA annu symp proc; 2006. p. 834–8.
[28] Yang J, Cohen AM, Hersh W. Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. In: AMIA annu symp proc; 2007. p. 831–5.
[29] Fiszman M, Rindflesch TC, Kilicoglu H. Summarization of an online medical Encyclopedia. MedInfo 2004:506–10.
[30] Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in medline citations. In: AMIA annu symp proc; 2006. p. 254–8.
[31] Ahlers C, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from MEDLINE citations for pharmacogenomics. Pac Symp Biocomput 2007;12:209–20.
[32] Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc 2010;98(4):273–81.
[33] Hahn U, Reimer U. Knowledge-based text summarization: salience and generalization operators for knowledge base abstraction. In: Mani I, Maybury MT, editors. Advances in automatic text summarization. Cambridge: MIT Press; 1999. p. 215–32.
[34] Stein GC, Bagga A, Wise GB. Multi-document summarization: methodologies and evaluations. In: Proceedings of the 7th conference on automatic natural language processing; 2000. p. 337–46.
[35] Erkan G, Radev DR. LexRank: graph-based centrality as salience in text summarization. J Artif Intell Res 2004;22:457–79.
[36] Leskovec J, Grobelnik M, Milic-frayling N. Learning sub-structures of document semantic graphs for document summarization. Workshop on link analysis and group detection; 2004.
[37] Yoo I, Hu X, Song I. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinform 2007;8(Suppl. 9):S4.
[38] Chen Z, Ji H. Graph-based clustering for computational linguistics: a survey. In: Proceedings of the 2010 workshop on graph-based methods for natural language processing. ACL; 2010. p. 1–9.
[39] Plaza L, Stevenson M, Diaz A. Improving summarization of biomedical documents using word sense disambiguation. In: Proceedings of the 2010 workshop on biomedical natural language processing. ACL; 2010. p. 55–63.
[40] Chali Y, Hasan SA, Joty SR. Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. Inf Process Manage. doi:10.1016/j.ipm.2010.10.002.
[41] Zhang X, Cheng G, Qu Y. Ontology summarization based on RDF sentence graph. In: Proceedings of the 16th international conference on world wide web; 2007. p. 707–16.
[42] Nooy W, Mrvar A, Batagelj V. Exploratory social network analysis with Pajek. New York: Cambridge University Press; 2005.
[43] Jing H, Barzilay R, McKeown K, Elhadad M. Summarization evaluation methods: experiments and analysis. AAAI symposium on intelligent summarization; 1998. p. 60–8.
[44] Lin CY, Hovy E. Manual and automatic evaluation of summaries. In: Proceedings of the ACL workshop on automatic summarization; 2002. p. 45–51.
[45] Radev D, Teufel S, Saggion H, et al. Evaluation challenges in large-scale document summarization. In: Proceedings of the 41st annual meeting on association for computational linguistics; 2003. p. 375–82.
[46] Elhadad N, McKeown K, Kaufman D, Jordan D. Facilitating physicians' access to information via tailored text summarization. In: AMIA annu symp proc; 2005. p. 226–30.
[47] McKeown H, Passonneau RJ, Elson DK, et al. Do summaries help? A task-based evaluation of multidocument summarization. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval; 2005. p. 210–17.
[48] Amigo E, Gonzalo J, Peinado V. An empirical study of information synthesis tasks. In: Proceedings of the 42nd annual meeting of the association for computational linguistics; 2004. p. 207–14.
[49] http://duc.nist.gov/duc2005/tasks.html.
[50] http://www.accessmedicine.com/resourceTOC.aspx?resourceID=4.
[51] http://www.accessmedicine.com/resourceTOC.aspx?resourceID=1.