Journal of Biomedical Informatics 52 (2014) 319–328

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Comparing different knowledge sources for the automatic summarization of biomedical literature

Laura Plaza

NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), C/ Juan del Rosal, 16, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 11 July 2014
Available online 24 July 2014

Keywords:
Biomedical knowledge sources
Unified Medical Language System
Semantic graph
Automatic summarization

ABSTRACT

Objective:: Automatic summarization of biomedical literature usually relies on domain knowledge from external sources to build rich semantic representations of the documents to be summarized. In this paper, we investigate the impact of the knowledge source used on the quality of the summaries that are generated.

Materials and methods:: We present a method for representing a set of documents relevant to a given biological entity or topic as a semantic graph of domain concepts and relations. Different graphs are created by using different combinations of ontologies and vocabularies within the UMLS (including GO, SNOMED-CT, HUGO and all available vocabularies in the UMLS) to retrieve domain concepts, and different types of relationships (co-occurrence and semantic relations from the UMLS Metathesaurus and Semantic Network) are used to link the concepts in the graph. The different graphs are next used as input to a summarization system that produces summaries composed of the most relevant sentences from the original documents.

Results and conclusions:: Our experiments demonstrate that the choice of the knowledge source used to model the text has a significant impact on the quality of the automatic summaries. In particular, we find that, when summarizing gene-related literature, using GO, SNOMED-CT and HUGO to extract domain concepts results in significantly better summaries than using all available vocabularies in the UMLS. This finding suggests that successful biomedical summarization requires the selection of the appropriate knowledge source, whose coverage, specificity and relations must be in accordance to the type of the documents to summarize.

1. Introduction

The amount of biomedical literature that is available on the Internet has experienced an unprecedent revolution during the last decade. Up-to-date MEDLINE contains over 19 million references to journal articles, and the US National Library of Medicine expects to index over one million articles annually within a few years [1]. In this context, researchers in biomedical-related disciplines find it extremely difficult to locate and read all the relevant literature that is published.

Aware of this situation, the text summarization community is actively working toward the development of domain-specific methods that help manage this information overload. Given a set of articles related to a topic (e.g., gene, disease, treatment, etc.), the aim is to produce a brief summary that condenses the relevant information. These summaries are not expected to fully replace the original documents but to increase the likelihood of researchers to identify the information most pertinent to their studies.

When applied to biomedical text, summarization methods usually rely on domain knowledge from external sources, such as GO [2], SNOMED-CT [3] or MeSH [4], to model the documents to be summarized. Previous works have demonstrated the benefit of such rich semantic representations compared to traditional approaches based on terms [5–7]. Texts are usually represented as sets of biomedical concepts linked by semantic relationships that are extracted from a given knowledge base. The selection of the knowledge source and the relations to be used, however, seems to be quite arbitrary or intuitive in spite of being a highly relevant decision in the summarization process.

In this work, the aim is to analyze the impact of the knowledge sources that are used to represent the text on the quality of the automatic summaries. Our hypothesis is that successful biomedical summarization (as other data mining and information extraction tasks) requires the use of the appropriate knowledge source, whose coverage, specificity and properties must be in accordance to the

E-mail address: lplaza@lsi.uned.es

http://dx.doi.org/10.1016/j.jbi.2014.07.014
1532-0464/

nature of the documents to summarize. To this end, we propose a method for representing a set of documents relevant to a given biological entity as a semantic graph of biomedical concepts and relations. Different knowledge sources and relations are used to build various graphs that are next used as input to a summarization system. The automatic summaries produced by each graph are then evaluated and compared to assess the impact of the knowledge source on the quality of the summaries. We find that the selection of the knowledge sources and the relationships to be used to build the graph has a significant impact on the summarization performance (≈15% of improvement in ROUGE-2 metric).

The article is organized as follows. The next section presents some related work in biomedical summarization along with the summarization system used in our experiments. We then describe the method for representing the documents as semantic graphs, as well as the evaluation methodology. We next report the results of the experiments and discuss these results. The final section provides concluding remarks.

2. Background

In this section, we first present some previous work in automatic summarization of biomedical texts. Next, we describe the sematic graph-based summarizer used to perform the experimentation.

2.1. Summarization of biomedical text

Text summarization is the process of automatically extracting the most relevant information from a document (or set of documents) [8]. Summaries may be *extractive* or *abstractive*. In extractive summarization, the most important sentences from the input document are taken as the summary. In contrast, abstractive summaries are built by paraphrasing the information in the original documents. Focusing on extractive summarization, this has been mostly addressed using traditional Information Retrieval (IR) heuristics, such as the frequency of the terms in the document [9,10], the position of the sentences [11–13], the presence of certain key words or indicative expressions [10], or the similarity of the different sentences with the title and the abstract of the document [14]. Using these simple criteria, sentences are scored, ranked and extracted for the summary.

More advanced works use graph representations and clustering techniques to produce the summaries [15,16]. These systems model the text as a graph, where the nodes represent lexical or semantic units (e.g., terms, sentences or concepts) and the links represent different types of relations between them (e.g., co-occurrence, lexical similarity, etc.). A ranking algorithm is then applied to sort the nodes according to their relevance, and the top-ranked units are extracted for building the summary. A representative example of graph-based method for summarization is LexRank [15]. Given a set of documents about a same topic (a *multi-document*), it builds a graph in which each node corresponds to a sentence represented by its TF-IDF vector and the edges are labeled with the cosine similarity between the sentences. Only the edges connecting sentences with a similarity above a predefined threshold are drawn in the graph. Under the hypothesis that the sentences that are similar to many others in the cluster are more central (or *salient*), those sentences represented by the most highly connected nodes are selected for the summary.

When these techniques are applied to summarize biomedical literature, texts are usually modeled using domain-specific knowledge sources, such as GO [2], SNOMED-CT [3] or MeSH [4], instead of using term-based representations, to better capture the meaning of the text. In this line, for instance, Demner-Fushman and Lin [17] present a hybrid approach to clinical question answering that combines summarization and information retrieval techniques. Given a set of MEDLINE citations, the system first identifies the drugs under study. Abstracts are then clustered using semantic types within the UMLS. For each abstract, an extractive summary is produced that includes information about (i) the main intervention described in the abstract, (ii) the title of the abstract, and (iii) the top-scoring outcome sentence (i.e., the sentence describing the "outcome" that asserts the clinical finding of the study). To extract the main intervention, the UMLS concepts falling under certain semantic types are considered as candidates, and each candidate is scored based on different features such as its position in the abstract or its frequency of occurrence. To extract the main outcome sentence, supervised machine learning is employed.

Reeve et al. [18] adapt the lexical chaining [19] approach to use UMLS concepts rather than terms and apply it to single-document summarization. They identify UMLS concepts in the source and chain them so that each chain contains the concepts belonging to the same semantic type. The chains are next scored according to the frequency of their concepts, and the strongest chains are identified. Finally, the sentences are scored based on the number of concepts that they contain from strong chains.

Ling et al. [20] focus on a narrower domain, genomic, and present a system that ranks sentences according to three features: the relevance of six gene aspects, such as the DNA sequence, the relevance of the documents where the sentences are taken from, and the position of the sentences in the document. They use FlyBase's [21] annotations to enrich the text for summarization.

Yang et al. [22,23] describe an extractive approach to the summarization of mouse gene information that first clusters a set of genes by MeSH, GO and free text features into functionally related groups, and then ranks and extracts sentences for each gene group to produce the summaries. Ranking is done by weighting different features such as the presence of cue phrases, domain specific keywords and the length of the sentences.

Plaza et al. [24] propose a graph-based approach that generates single-document summaries of biomedical articles. Each document is represented as a semantic graph of UMLS Metathesaurus concepts and relations from both the UMLS Metathesaurus and the Semantic Network. A clustering algorithm based on degree centrality is used to identify topics within the text. The extraction of sentences for the summary is based on how much each sentence covers the different topics that are identified.

Shang et al. [25] combine IR techniques with information extraction methods to generate text summaries of sets of documents describing a certain topic. To do this, they use SemRep to extract relations among UMLS Metathesaurus concepts and a relation-level retrieval method to select the relations more relevant to a given query concept. They extract the most relevant sentences for each topic based on the previous ranking of relations and the location of the sentences in different sections of the document.

Fiszman et al. [26] propose an algorithm that makes use of semantic predications provided by SemRep [27] to interpret biomedical text and on the use of lexical and semantic information from the UMLS to produce a summary from biomedical scientific articles. This same method is adapted in a later work to summarize drug information in MEDLINE citations [28]. Unlike most works in the area that still follow an extractive paradigm (such as [18,20,24,25]), Fiszman et al. adopt an abstractive approach to produce a graph that summarizes the content of the documents. This graph shows the relevant UMLS concepts that describe the document and the semantic relations among them, providing a graphical summary that condensates the most important aspects of the content of the documents.

In Zhang et al. [29], a graph of predications from SemRep is used to represent multiple PubMed citations, with arguments as nodes

and predicates as arcs, so that, as in [24], nodes are UMLS Metathesaurus concepts and links are semantic relations among them (in this case, however, only relations from the Semantic Network are used). Degree centrality and co-occurrence of predications are used to select salient predications for the summary. Cliques (i.e., subsets of vertices within the graph such that every two vertices in the subset are connected by an edge) are next identified and clustered to find the different themes or points of view contained in the summary. Again, the summaries that are generated are not textual, but graphical.

Nonetheless, it is worth mentioning that not all summarization efforts in the biomedical domain rely on domain knowledge. Ruch et al. [30] address the summarization task as a classification problem. They train different term-based Bayessian classifiers to categorize sentences in MEDLINE abstracts into four argumentative moves: PURPOSE, METHODS, RESULTS and CONCLUSION, and to select the most representative sentence from them. Lu et al. [31] present a method for automatically generating GeneRIFs that scores sentences for extraction using simple features based on Edmundson work's [10], such as the position of the sentence in the document, the presence of "cue words" and the absence of "stigma words". Finallly, Jin et al. [32] propose a text summarization system that takes as input MEDLINE documents related to a given target gene and outputs a small set of genic information rich sentences. Sentences are ranked by the sum of two individual scores: (a) an authority score from a lexical PageRank algorithm and (b) a similarity score between the sentence and GO terms with which the gene is annotated. Redundant sentences are removed and top-ranked sentences are extracted for the summary.

Even though, as already mentioned, most approaches to biomedical text summarization make use of domain-specific knowledge to represent the semantics of the documents to be summarized, the selection of the source used to acquire such knowledge is not supported by any empirical study that demonstrates the adequacy of the selected source(s) compared to other alternatives. This is precisely the novel perspective given to this article, whose main objective is to evaluate the impact of the knowledge sources that are used to represent the documents on the quality of the final summaries. To this end, we make use of the summarizer presented in [24], which is described in the next section, and test different graph representations built by mapping the text into concepts from different knowledge sources and linking the concepts using different types of relations.

2.2. A semantic graph-based summarizer

The summarization system used for our experiments is based on the work presented in [24]. The original system has been adapted to (i) work with different knowledge sources and relations from the UMLS and (ii) use a different clustering method for topic detection. The method consists of three steps, which we briefly explain below. Fig. 1 illustrates the different steps.

- **Step I: Concept identification and document representation**. The summarizer takes as input a set of documents about a same entity or topic and merges them into a single *multi-document*. It then runs the MetaMap [33,34] program over the document to obtain the Metathesaurus concepts that are found within the text. It next builds a graph-based representation of the document, where the nodes are UMLS concepts and the links are different types of relationships between them. To do this, it first extends the UMLS concepts with their hierarchies of hypernyms (*is_a* relations) and merges the hierarchies of all concepts to build the *document graph*. This graph is next extended with further relations (e.g., co-occurrence relations or semantic relations from the UMLS Semantic Network). Finally, each edge is assigned a weight in [0,1] as shown in Eq. 1. The weight of an edge e representing an *is_a* relation between two vertices, v_i and v_j (where v_i is a parent of v_j), is calculated as the ratio of the depth of v_i to the depth of v_j from the root of their hierarchy. The weight of an edge representing any other relation is always 1.

$$weight(e, v_i, v_j) = \beta \quad \text{where} \begin{cases} \beta = \frac{depth(v_i)}{depth(v_j)} & \text{if } e \text{ represents an } is\_a \text{ relation} \\ \beta = 1 & \text{otherwise} \end{cases} \tag{1}$$

As an example that illustrates the document representation step, Fig. 3 shows the graph that represents a multi-document about the *EGFR* gene composed of the two MEDLINE abstracts shown in Fig. 2. This graph has been built using GO and SNOMED-CT as knowledge sources from the UMLS and three different types of relations: co-occurrence, Metathesaurus relations, and Semantic Network relations. These sources and relations are explained in detail in the Materials and Methods section.

- **Step II: Topic recognition**. This step consists of clustering the UMLS concepts in the document graph using the edge-betweenness clustering [35] algorithm to identify topics within the graph. This method has been widely used to discover communities in social and biological networks [36]. It identifies those edges that are most "between" topics, and progressively removes these edges from the original graph. In this way, the different topics are isolated. The betweenness centrality of a vertex i is defined as the number of shortest paths between pairs of other nodes that run through i. The edges connecting topics will have high edge betweenness. By removing these edges, the groups of nodes that describe each topic are separated from one another.
For each node in the graph, we also compute a *salience*. The salience of a vertex, v_i, is defined as the number of edges that are connected to it. This is shown in Eq. 2 where $connect(e, v_i, v_j)$ denotes that the edge e_j connects nodes v_i and v_j. Salience ranks the nodes according to their structural importance in the graph.

$$salience(v_i) = \sum_{\forall e_j \exists v_k \wedge connect(e_j, v_i, v_k)} 1 \tag{2}$$

Consider the two abstracts shown in Fig. 2. Consider also the graph corresponding to these two abstracts that is shown in Fig. 3. Table 1 shows the clusters generated after applying the edge-betweenness algorithm. Note that clusters with only one concept are not shown.

- **Step III: Sentence selection**. The aim of the last step is to select the sentences from the multi-document that best describe the content of each topic. These sentences are intended to help users to interpret and understand the meaning of the different topics. We first represent each sentence in the original documents as a graph. This is done by using the UMLS concepts identified within the sentence, extracting the complete hierarchy of hypernyms for each concept and merging the hierarchies to construct a single sentence graph. We next compute the similarity between each sentence graph and cluster, as the sum of the salience of the matching nodes between the sentence graph and the cluster. In this way, for each cluster, we obtain a ranking of sentences that reflects each sentence topic's coverage. A number n_i of sentences are selected from each cluster as the best topic's descriptors, where n_i is proportional to the cluster size. Finally, the selected sentences are concatenated to form the summary.
Fig. 4 shows the top 5 ranked sentences for the two abstracts shown in Fig. 2. The sentences have been ordered in the summary so that those belonging to the first abstract are presented in the first place, while those belonging to the second abstract

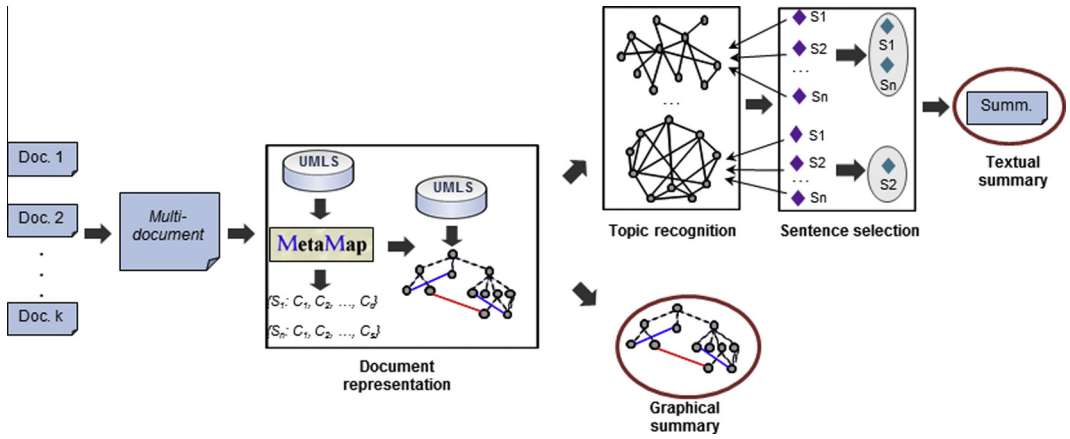

Fig. 1. Summarizer architecture.

PMID: *1322798*
Title: *The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling*

A cDNA clone encoding a novel, widely expressed protein (called growth factor receptor-bound protein 2 or GRB2) containing one src homology 2 (SH2) domain and two SH3 domains was isolated. Immunoblotting experiments indicate that GRB2 associates with tyrosine-phosphorylated epidermal growth factor receptors (EGFRs) and platelet-derived growth factor receptors (PDGFRs) via its SH2 domain. Interestingly, GRB2 exhibits striking structural and functional homology to the C. elegans protein sem-5. It has been shown that sem-5 and two other genes called let-23 (EGFR like) and let-60 (ras like) lie along the same signal transduction pathway controlling C. elegans vulval induction. To examine whether GRB2 is also a component of ras signaling in mammalian cells, microinjection studies were performed. While injection of GRB2 or H-ras proteins alone into quiescent rat fibroblasts did not have mitogenic effect, microinjection of GRB2 together with H-ras protein stimulated DNA synthesis. These results suggest that GRB2/sem-5 plays a crucial role in a highly conserved mechanism for growth factor control of ras signaling.

PMID: *1383230*
Title: *The EGF receptor is an actin-binding protein*

In a number of recent studies it has been shown that in vivo part of the EGF receptor (EGFR) population is associated to the actin filament system. In this paper we demonstrate that the purified EGFR can be cosedimented with purified filamentous actin (F-actin) indicating a direct association between EGFR and actin. A truncated EGFR, previously shown not to be associated to the cytoskeleton, was used as a control and this receptor did not cosediment with actin filaments. Determination of the actin-binding domain of the EGFR was done by measuring competition of either a polyclonal antibody or synthetic peptides on EGFR cosedimentation with F-actin. A synthetic peptide was made homologous to amino acid residues 984-996 (HL-33) of the EGFR which shows high homology with the actin-binding domain of Acanthamoeba profilin. A polyclonal antibody raised against HL-33 was found to prevent cosedimentation of EGFR with F-actin. This peptide HL-33 was shown to bind directly to actin in contrast with a synthetic peptide homologous to residues 1001-1013 (HL-34). During cosedimentation, HL-33 competed for actin binding of the EGFR and HL-34 did not, indicating that the EGFR contains one actin-binding site. These results demonstrate that the EGFR is an actin-binding protein which binds to actin via a domain containing amino acids residues 984-996.

Fig. 2. Example of two MEDLINE abstracts on the EGFR gene.

are presented next. The order of the sentences within each abstract is also kept in the summary.

3. Materials and methods

Our objective is to evaluate the impact of the knowledge source used in the summarizer on the quality of the summaries that are generated. To this end, given a set of documents about a biological entity, we build different graphs that represent the documents. This is done by using different knowledge sources to identify domain concepts and different types of relations to link the concepts in the graph.

In this section, we first present the knowledge sources and relations that are used to generate the graphs, and then explain how such graphs are build and provided as input to the summarizer presented in the previous section to produce the summaries. Next, we describe how the summaries are evaluated. In order to work in a narrow domain and evaluate the effect of the specificity and coverage of the knowledge base, we focus on documents reporting genetic studies.

3.1. Knowledge sources and relations

We investigate different combinations of ontologies and vocabularies from the UMLS. In particular, we consider both gene-specific databases (such as the Gene Ontology [2] and the HUGO Gene Nomenclature [37]), and other more general biomedical nomenclatures (such as the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [3]). We also investigate the effect of including concepts from all vocabularies in the UMLS Metathesaurus.[1]

Concerning relations, we investigate three different types of relations and their combinations:

[1] A complete list of the source vocabularies present in the current version of the UMLS Metathesaurus can be found in http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html.

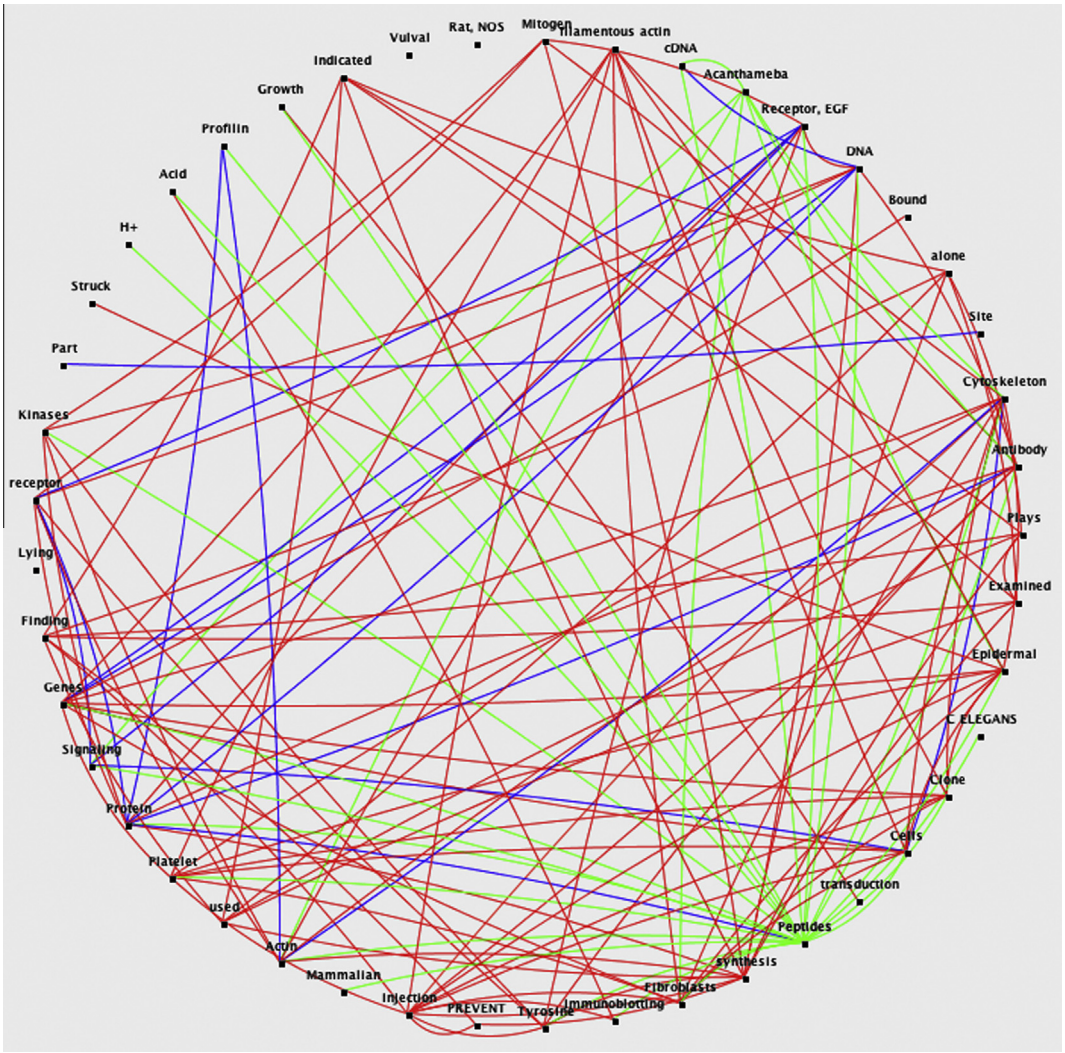

Fig. 3. Semantic graph for the two MEDLINE abstracts shown in Fig. 2. The graph has been built using Gene Ontology and SNOMED-CT as knowledge sources and the following relations: co-occurrence (in blue color), Metathesaurus relations (in green color) and Semantic Network relations (in red color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Clusters generated by the edge-betweenness algorithm for the two MEDLINE abstracts shown in Fig. 2.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Site [Spatial Concept]	Immunoblotting [Laboratory Procedure]	Growth [Organism Function]	Platelet [Cell]	Used [Finding]	DNA [Biologically Active Substance]
Part [Spatial Concept]	Acid [Chemical]	Transduction [Organism Function]	Genes [Gene or Genome]	Indicated [Finding]	Protein [Amino Acid]
			Epidermal [Tissue]	Plays [Finding]	Receptor [Amino Acid]
			Clone [Cell]	Examined [Finding]	Antibody [Amino Acid]
			Cells [Cell]	Alone [Finding]	Peptides [Amino Acid]
			Fibroblasts [Cell]	Finding [Sign or Symptom]	
			Cytoskeleton [Cell Component]	Synthesis [Biologic Function]	
			Injection [Therapeutic Procedure]		
			filamentous actin [Cell Component]		

- **Co-occurrence relations** (*CO*): These relations are derived from the co-occurrence distribution of UMLS concepts in different corpora of documents. The co-occurrence values used in this work are those available in the UMLS Metathesaurus, and represent the number of times concepts have co-occurred as key topics within the same articles. We will consider that two UMLS concepts are linked by a *CO* relation if they present a co-occurrence value above five.

1. *While injection of GRB2 or H-ras proteins alone into quiescent rat fibroblasts did not have mitogenic effect, microinjection of GRB2 together with H-ras protein stimulated DNA synthesis.* **[Abstract 1]**
2. *Immunoblotting experiments indicate that GRB2 associates with tyrosine-phosphorylated epidermal growth factor receptors (EGFRs) and platelet-derived growth factor receptors (PDGFRs) via its SH2 domain.* **[Abstract 1]**
3. *To examine whether GRB2 is also a component of ras signaling in mammalian cells, microinjection studies were performed.* **[Abstract 1]**
4. *A truncated EGFR, previously shown not to be associated to the cytoskeleton, was used as a control and this receptor did not cosediment with actin filaments.* **[Abstract 2]**
5. *Determination of the actin-binding domain of the EGFR was done by measuring competition of either a polyclonal antibody or synthetic peptides on EGFR cosedimentation with F-actin.* **[Abstract 2]**

Fig. 4. Automatic summary for the abstracts presented in Fig. 2.

- **Metathesaurus relations** (*MT*): These relations link Metathesaurus concepts and are categorized as two main types: *intra-source*, which links concepts that come from the same source vocabulary, and *inter-source*, which links concepts from different source vocabularies. Some types of relationships in the Metathesaurus are *synonymy*, *child of*, *narrower*, *broader* and *qualifier of*, to name a few.
- **Semantic Network relations** (*SN*): They link Semantic Types in the UMLS. The primary link is the "is_a" relation, which defines the hierarchy of types within the network. The remaining non-hierarchical relations may be grouped into five major categories: *physically related to*, *spatially related to*, *temporally related to*, *functionally related to*, and *conceptually related to*. We will consider that two UMLS concepts are linked by a SN relation if their semantic types are related in the UMLS Semantic Network.

3.2. Document representation

Our aim is to model a set of documents relevant to a given biological entity (a gene) as a semantic graph. In this graph, nodes represent the UMLS concepts found within the text, while the arcs represent relations between concepts.

To map the text to UMLS concepts, the MetaMap [34] program is used. MetaMap is invoked using the *Restrict to Sources (-R)* option, which allows us to specify the knowledge sources to be used. The hierarchy of hypernyms for each concept is retrieved from the *MRHIER* table of the UMLS Metathesaurus. A node in the graph is created for each retrieved concept and a link is added between each pair of *parent–child* concepts.

Next, we extend the graph with further relations. To this end, for each pair of concepts identified in the previous step, we extract from the UMLS database all relationships that exist between them. In particular, Metathesaurus relations are retrieved from the *MRREL* table, Semantic Network relations are retrieved from the *SRSTR* table, and co-occurrence relations are retrieved from the *MRCOC* table. For each relation, an edge linking the concepts is added to the graph.

In this way, for instance, the concepts *"cells"* and *"signaling"* are linked by a co-occurrence relation in the UMLS, *"protein"* and *"peptides"* are linked by a *has narrower* relation in the Metathesaurus, and "cytoskeleton" and "cells" are linked by a *part of* relation in the Semantic Network.

Table 2 lists all combinations of UMLS sources and relations that are used to build the graphs. For each source or combination of sources listed in the right column of this table, all combinations of relations listed in the left column are used, which means that 25 different graphs are built.

Fig. 5 shows different graphs for both (a) a single document and (b) a group of 5 related documents (in particular, MEDLINE abstracts), using different combinations of knowledge sources: (i) Gene Ontology (GO), (ii) GO + SNOMED-CT and (iii) all sources available in the UMLS. These graphs have been built using all the types of relationships described above (i.e., *CO*, *MT* and *SN*). The purpose of this figure is to give readers an idea of the size of the graphs and how they evolve with the number of documents and knowledge sources taken into account for building them. As it may be seen, using just GO as knowledge source for representing a single document produces a very small and sparse graph. As the number of knowledge sources increases, the graph becomes more connected. The same occurs when the number of abstracts used to generate the graph increases.

Table 2
List of combinations of UMLS sources and relations used to build the document graphs.

Sources	Relations
Gene Ontology	CO
SNOMED-CT	MT
Gene Ontology + SNOMED-CT	SN
Gene Ontology + SNOMED-CT + HUGO	MT + SN
ALL UMLS SOURCES	CO + MT + SN

3.3. Summary generation

Given a set of documents to summarize, the 25 graphs are given as input to the second step of the graph-based summarization system presented in the previous section and described in detail in [24]. For each graph, a summary is generated. Note that we keep all parameters of the summarizer unchanged, so that we only vary the graphs that represent the multi-document.

3.4. Evaluation methodology

To evaluate the impact of the knowledge sources on the final summaries, we test and compare the automatic summaries that are produced by using the different graph configurations. The next subsections describe the evaluation collection and metrics, as well as the baselines used for comparison.

3.4.1. Evaluation collection

For the evaluation, we selected a set of 25 genes from the human genome and retrieved from the MEDLINE database 10 citations for each gene. To this end, we used the gene-articles associations provided by GO to manually retrieve Pubmed identifiers of 10 citations related to each of the genes. From each citation, we extracted the title and abstract sections. The 10 abstracts describing each gene were represented as 25 different graphs using different knowledge sources and relations, and for each graph, we generate a summary of 10 sentences. We do not impose the restriction that each sentence has to be selected from a different citation,

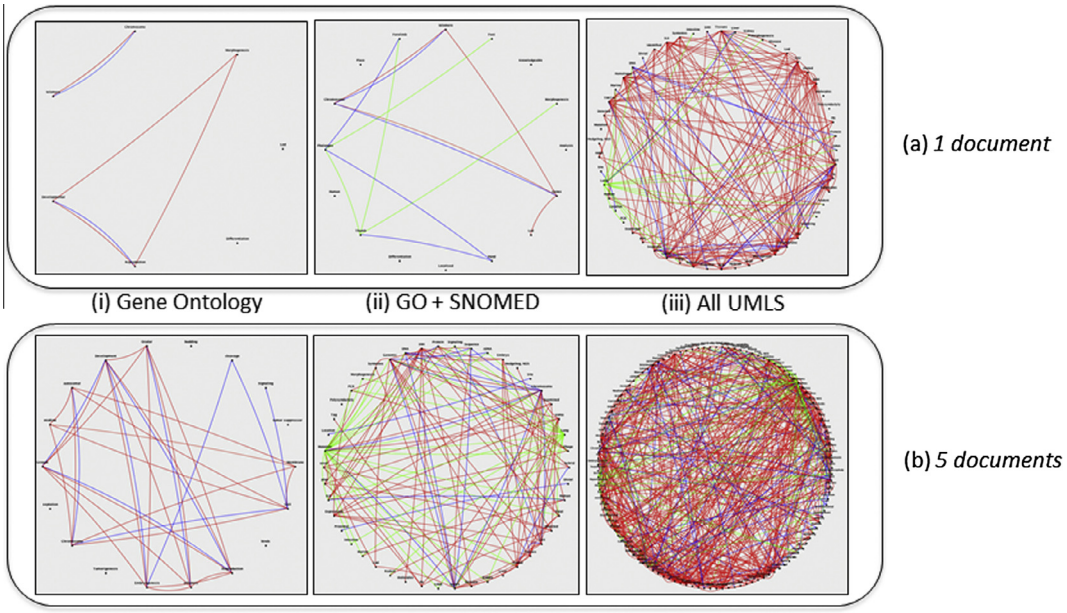

Fig. 5. Examples of semantic graphs for (a) a single MEDLINE abstract and (b) a group of 5 related MEDLINE abstracts, using different combinations of knowledge sources: (i) Gene Ontology (GO), (ii) GO + SNOMED-CT and (iii) all sources available in the UMLS. The following relationships between concepts are used: co-occurrence (in blue color), Metathesaurus relations (in green color) and Semantic Network relations (in red color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

because we want to evaluate the ability of the summarizer to detect and weight the different topics that are covered in the different citations.

We make use of the titles in the MEDLINE citations as the reference/model summary; i.e., for each gene, we get a reference summary composed of the titles of the 10 related citations. The title given to a document by its author is intended to represent the most significant information in the document, and thus it is expected to summarize the main content of the document [14].

3.4.2. Evaluation metrics

We use ROUGE [38] as the metric for evaluating the summaries. ROUGE is an evaluation method for summarization which uses the proportion of n-grams between a *peer* and one or more *reference* summaries to estimate the content that is shared between them.

We used the following ROUGE metrics: ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4), since they have been shown to be robust and highly correlated with human judgments, and have been widely used in the DUC[2] and TAC[3] communities. R-2 computes the number of bigrams that are shared by the peer and the reference summaries. R-SU4 measures the overlap of skip-bigrams between the peer and the reference summaries, allowing a skip distance of four.

However, it must be noted that ROUGE metrics present two important limitations: (1) they depend on the length of the peer summaries (i.e., the longer is the peer with respect to the model, the higher are expected to be the ROUGE scores), and (2) since they use lexical matching instead of semantic matching, peer summaries that are worded different but have the same semantic information may be assigned different ROUGE scores. Thus, these metrics should only be used in a comparative fashion on the same dataset and should not be interpreted as absolute measures.

A Wilcoxon Signed Ranks test with a 95% confidence interval is used to test the significance of the results for both R-2 and R-SU4 metrics.

[2] http://duc.nist.gov.
[3] http://www.nist.gov/tac/.

3.4.3. Baselines

We compare our results with two baselines: *LEAD*, which extracts the first sentence from each citation and concatenates them to build the summary; and *FREQ*, which implements the concept frequency-based summarizer described in [39] (without using the sentence location feature) to generate the automatic summaries. To briefly illustrate how the *FREQ* summarizer works, consider the following text fragment from [39]:

> *Interactions among LRF-1, JunB, c-Jun, and c-Fos define a regulatory program in the G1 phase of liver regeneration. In regenerating liver, a physiologically normal model of cell growth, LRF-1, JunB, c-Jun, and c-Fos among Jun/Fos/LRF-1 family members are induced posthepatectomy. In liver cells, high levels of c-Fos/c-Jun, c-Fos/JunB, LRF-1/c-Jun, and LRF-1/JunB complexes are present for several hours after the G0/G1 transition, and the relative level of LRF-1/JunB complexes increases during G1.*

If $\{C_1, C_2, \ldots, C_n\}$ is the set of n Metathesaurus concepts that appear in the text d, and $f_i(d)$ is the number of times that C_i appears in it, then the text may be represented by the vector $D = \{f_1(d), f_2(d), \ldots, f_n(d)\}$. Similarly, each individual sentence may be represented by a concept frequency vector, S_j. A concept frequency score, $CF(S_j)$, may be calculated for each sentence S_j as the sum of the frequency of all the concepts in the sentence multiplied by the frequency of those concepts in the whole text. In this way, the text above is represented by the following vectors:

$$D = \{LRF = 5, c-Jun = 3, c-fos = 3, Liver = 3, Cell = 2, Complexes = 2, Program = 1, Regeneration = 1, growth = 1, LRF-1 = 1, Transition = 1\}$$

$$S_1 = \{LRF = 1, c-Jun = 1, c-fos = 1, Program = 1, Liver = 1, Regeneration = 1\}$$

$$S_2 = \{LRF = 2, c-Jun = 1, c-fos = 1, Liver = 1, Cell = 1, growth = 1\}$$

$S_3 = \{LRF = 2, c-Jun = 1, c-fos = 1, Liver = 1, Cell = 1, LRF - 1 = 1, Transition = 1, Complexes = 2\}$

The concept frequency scores of the sentences are, respectively, $CF(S_1) = 16, CF(S_2) = 22$, and $CF(S_3) = 27$. The N top scoring sentences are then selected for the summary.

4. Results and discussion

The results of our experiments are summarized in Tables 3–7. We can distinguish two parts in these tables. The first part (columns `#nodes`, `#edges`, `#clusters`) shows the different graph configurations that arise from varying the knowledge sources and the relationships used to build the graphs, as well as the number of clusters (ignoring those having a single element) that are generated by the clustering algorithm. The second part (columns `R-2`, `R-SU4`) shows the average ROUGE scores for the summaries generated using the different graphs.

As may be seen from these tables, the choice of the knowledge sources and relations used to build the graph strongly influences its structural properties, as well as the results of the clustering algorithm, and thus, the quality of the automatic summaries. In the next subsections, we discuss these results in detail. We first focus on how the selection of the knowledge source and relations affects the properties of the graph and the output of the clustering method, and next we focus on the results of the summarization method.

4.1. Graph properties and clustering

Tables 3–7 show how the topology of the graphs vary with the knowledge sources and relations. Concerning the knowledge sources that are used to build the graphs, it may be seen that using the Gene Ontology (GO) alone produces very small graphs (≈30 nodes per graph on average), that do not seem to be enough to capture the semantic of the documents for the purpose of summarization (see Table 3). This finding was expected since, as we can observe in the abstracts presented in Fig. 2, texts reporting gene studies usually contain concepts related to clinical procedures, findings, diseases, etc., that are not annotated in GO but are frequently relevant enough as to be included in the summary. When SNOMED-CT is used instead of GO, the vocabulary in the abstracts is better covered and the size of the graphs significantly increases (≈116 nodes per graph on average). However, aggregating both sources or adding new ones does not increase the size of the graphs significantly: when GO and SNOMED-CT are used together (see Table 5), only ≈ 10 nodes are added to the graph; when the HUGO concepts are included (see Table 6), only ≈8 new nodes are added; using all the knowledge sources that are available in the UMLS (see Table 7) produces graphs of ≈162 nodes on average, that is, only 38 nodes more than just using SNOMED-CT.

Table 3
ROUGE scores for the summaries generated using Gene Ontology as knowledge source and different types of relationships. Significance is calculated with respect to the *CO* relation baseline. The best ROUGE scores achieved are highlighted in bold.

Gene Ontology					
Relations	# nodes	# edges	# clusters	R-2	R-SU4
CO	30.4	0.36	1	0.239	0.205
SN	30.4	119.36	2.32	0.260	0.216
MT	30.4	49.84	5.4	0.264*	0.223*
CO + MT + SN	30.4	176.64	4	0.272*	0.226*
MT + SN	30.4	132.4	3.76	**0.279***	**0.230***

* $p < 0.05$.

Table 4
ROUGE scores for the summaries generated using SNOMED-CT as knowledge source and different types of relationships. Significance is calculated with respect to the *CO* relation baseline. The best ROUGE scores achieved are highlighted in bold.

SNOMED-CT					
Relations	# nodes	# edges	# clusters	R-2	R-SU4
CO	116.1	95.2	1	0.235	0.210
SN	116.1	638.1	4.23	0.268*	0.219
MT	116.1	160.4	17.62	0.265*	0.217
CO + MT + SN	116.1	926.2	11.48	0.270*	0.222
MT + SN	116.1	683.7	10.9	**0.274***	**0.228***

* $p < 0.05$

Table 5
ROUGE scores for the summaries generated using Gene Ontology and SNOMED-CT as knowledge sources and different types of relationships. Significance is calculated with respect to the *CO* relation baseline. The best ROUGE scores achieved are highlighted in bold.

SNOMED-CT + Gene Ontology					
Relations	# nodes	# edges	# clusters	R-2	R-SU4
CO	126.32	109.40	1	0.247	0.205
SN	126.32	796.56	4.64	0.276*	0.229*
MT	126.32	189.92	21.72	0.246	0.204
CO + MT + SN	126.32	1095.88	13.6	0.260	0.221
MT + SN	126.32	844	13.04	**0.282***	**0.234***

* $p < 0.05$

Table 6
ROUGE scores for the summaries generated using Gene Ontology, SNOMED-CT and HUGO as knowledge sources and different types of relationships. Significance is calculated with respect to the *CO* relation baseline. The best ROUGE scores achieved are highlighted in bold.

Gene Ontology + SNOMED-CT + HUGO					
Relations	# nodes	# edges	# clusters	R-2	R-SU4
CO	134.92	106.44	1	0.254	0.210
SN	134.92	1119.52	8.12	0.291*	0.242*
MT	134.92	198.4	22.8	0.268	0.227
CO + MT + SN	134.92	1424.36	11.24	0.286*	0.241*
MT + SN	134.92	1191.6	12.56	**0.315***	**0.260***

* $p < 0.05$

Table 7
ROUGE scores for the summaries generated using all vocabularies available in the UMLS and different types of relationships. Significance is calculated with respect to the *CO* relation baseline. The best ROUGE scores achieved are highlighted in bold.

All UMLS sources					
Relations	# nodes	# edges	# clusters	R-2	R-SU4
CO	162.56	104.4	1	0.245	0.209
SN	162.56	1465.6	5.08	0.277*	0.235*
MT	162.56	293.76	23.84	0.259	0.218
CO + MT + SN	162.56	1863.76	13.04	0.282*	0.236*
MT + SN	162.56	1531.64	13.56	**0.286***	**0.239***

* $p < 0.05$

Regarding the relationships that are used to link the concepts in the graphs, it may be seen in Tables 3–7 that using the *CO* relation alone produces very sparse graphs, being the average degree centrality of the nodes 0.63 (i.e., the average number of links incident upon a node). This means that most of the nodes in the graph are isolated and only a small number of nodes have a few links connecting them with other nodes in the graph. Thus, the graph does not present a community structure and, as a result, the clustering method does not work properly. In this situation, the clustering method is usually returning a number of clusters with a single

concept but only one cluster with multiple concepts. A similar result was found in our previous work [24], where the use of the UMLS hypernym relation alone showed to produce very disconnected graphs and, as a result, the summaries generated got poor ROUGE scores. In contrast, the use of the *SN* relation produces highly connected graphs (6.61 links per node on average). Somewhere in the middle is the *MT* relation, which produces graphs with an average degree centrality of 1.56.

4.2. Summarization performance

We next discuss the results that concern the generation of automatic summaries. To facilitate understanding of the results, Table 8 shows the ROUGE scores achieved by the best graph configuration for each combination of knowledge sources. This table also shows the results of the two baselines methods: *LEAD* and *FREQ*.

It may be seen from Table 8 that the best summaries are generated when the documents are represented as graphs of concepts from *GO*, *SNOMED-CT* and *HUGO*, and links between concepts are relations from both the UMLS Metathesaurus and the Semantic Network. Besides, the summaries produced by this configuration are significantly better ($p < 0.05$) for both ROUGE metrics than all other combinations. In contrast, the worst ROUGE scores are obtained when only SNOMED-CT concepts are represented in the graph. This seems to indicate that highly specialized concepts from the genomic domain that are not captured in SNOMED-CT are important to identify the relevant content from the documents. Similarly, very poor results are obtained when only *GO* concepts are included in the graph, which also suggests that, besides this highly specialized knowledge, more general medical knowledge is also needed, which is not covered in GO, which is a highly specialized and sparse resource.

Therefore, these results indicate that automatic summarization of gene-related literature benefits from both gene-specific vocabulary and general biomedical information. However, using all knowledge sources in the UMLS to build the graph introduces non-relevant concepts in the graph that negatively affect the summarization process.

Concerning the relations, it is observed that the best combination of relationships (from the summarization perspective) is that which attaches concepts from related semantic types in the UMLS Semantic Network and concepts which are related in the UMLS Metathesaurus (i.e., *SN* + *MT*). This is the best combination of relations regardless of the knowledge sources that are used to extract the domain concepts. This combination produces a highly connected graph (average degree centrality = 7,03). Note that even if the combination of the three relationships produces a more connected graph, the use of the co-occurrence relation seems to link concepts with low semantic similarity, which decreases the summarization performance.

Table 8
ROUGE scores for the summaries generated using different combinations of knowledge sources and the best set of relationships (*MT* + *SN*), as well as those generated by the two baseline systems. Significance is calculated with respect to the *ALL UMLS* summarizer.

Summarizer	Relations	R-2	R-SU4
GO	MT + SN	0.279	0.230
SNOMED	MT + SN	0.274	0.228
SNOMED + GO	MT + SN	0.282	0.234
SNOMED + GO + HUGO	MT + SN	**0.315***	**0.260***
ALL UMLS	MT + SN	0.286	0.239
FREQ		0.241	0.211
LEAD		0.212	0.201

* $p < 0.05$

5. Conclusions and future work

In this work, we have analyzed the influence of the knowledge sources and relations used to model biomedical text on the quality of the automatic summaries. To this aim, we have presented a graph-based summarization algorithm and evaluated its performance on different text representations that consider different combinations of biomedical knowledge sources and relations from the UMLS.

Overall, the results presented in this paper suggest that the selection of the knowledge sources and the relationships to be used to model the text has a significant impact on the summarization results. In particular, we found that using GO, SNOMED-CT and HUGO to extract domain concepts allows for significantly better summaries than using all available vocabularies in the UMLS. Therefore, the knowledge base to be used should be an important parameter to take into account when developing summarization systems in the biomedical domain.

Concerning future work, we will apply the main lessons learned from this work to our research in other bioNLP tasks. In the short term we plan to evaluate the effect of knowledge source selection in the automatic classification and indexing of MEDLINE citations [40]. We want to test, for instance, whether indexing of MeSH terms related to genes and proteins improves when the documents are represented using specialized resources, such as GO and HUGO, while indexing of more general terms (such as humans, mice, pregnancy, to name a few) are better captured by using more general resources for knowledge representation (e.g., SNOMED-CT or the NCI Thesaurus [41]).

We will also apply the results of these experiments to improve our summarization system, by selecting the *GO* + *SNOMED-CT* + *HUGO* combination of knowledge sources for building the document graph, and testing the performance of the summarizer when the graph-based method is combined with other summarization techniques that have proved to be of great use for summarization of biomedical literature, such as sentence position [39], the similarity with the title [24] and the frequency of the concepts in the document [39].

References

[1] MEDLINE. <http://www.nlm.nih.gov/databases/databases_medline.html>.
[2] Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[3] Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). <http://www.ihtsdo.org/snomed-ct/>.
[4] Medical Subject Headings. <http://www.nlm.nih.gov/mesh/>.
[5] Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med 2005;33:157–77.
[6] Shi Z, Melli G, Wang Y, Liu Y, Gu B, et al. Question answering summarization of multiple biomedical documents. In: Proceedings of the Canadian conference on artificial intelligence; 2007. p. 284–95.
[7] Plaza L, Díaz A, Gervás P. Concept-graph based biomedical automatic summarization using ontologies. In: Proceedings of the 3rd textgraphs workshop on graph-based algorithms for natural language processing; 2008. p. 53–6.
[8] Sparck-Jones K. Automatic summarising: factors and directions. The MIT Press; 1999.
[9] Luhn HP. The automatic creation of literature abstracts. IBM J Res Dev 1958;2:159–65.
[10] Edmundson HP. New methods in automatic extracting. J Assoc Comput Mach 1969;2:264–85.
[11] Brandow R, Mitze K, Rau L. Automatic condensation of electronic publications by sentence selection. Inform Process Manage 1995;5:675–85.
[12] Lin C, Hovy E. Identifying topics by position. In: Proceedings of the fifth conference on applied natural language processing; 1997. p. 283–90.
[13] Ko Y, Park J, Seo J. Improving text categorization using the importance of sentences. Inform Process Manage 2004;40:65–79.
[14] Bawakid A, Oussalah M. A semantic summarization system: university of birmingham at TAC 2008. In: Proceedings of the first text analysis conference; 2008.
[15] Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. J Artif Intell Res 2004;22:457–79.

[16] Plaza L, Díaz A, Gervás P. Automatic summarization of news using wordnet concept graphs. IADIS Int J Comput Sci Inform Syst 2010;5:45–57.
[17] 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.
[18] Reeve L, Han H, Brooks A. The use of domain-specific concepts in biomedical text summarization. Inform Process Manage 2007;43:1765–76.
[19] Barzilay R, Elhadad M. Using lexical chains for text summarization. In: Proceedings of the ACL workshop on intelligent scalable text summarization; 1997. p. 10–7.
[20] Ling X, Jiang J, He X, Mei Q, Zhai C, et al. Generating gene summaries from biomedical literature: a study of semi-structured summarization. Inform Process Manage 2007;43:1777–91.
[21] Drysdale R, Crosby M. Flybase: genes and gene models. Nucl Acids Res 2005;33:390–5. the FlyBase Consortium.
[22] Yang J, Cohen AM, Hersh W. Automatic summarization of mouse gene information by clustering and sentence extraction from medline abstracts. In: Proceedings of the AMIA annual symposium; 2007. p. 831–5.
[23] Yang J, Cohen AM, Hersh W. Evaluation of a gene information summarization system by users during the analysis process of microarray datasets. BMC Bioinform 2009;10(Suppl 2):S5.
[24] Plaza L, Díaz A, Gervás P. A semantic graph-based approach to biomedical summarisation. Artif Intell Med 2011;53:1–15.
[25] Shang Y, Li Y, Lin H, Yang Z. Enhancing biomedical text summarization using semantic relation extraction. PLoS One 2011;6.
[26] 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.
[27] Rindflesch T, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36:462–77.
[28] Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in medline citations. In: Proceedings of the AMIA annual symposium; 2006. p. 254–8.
[29] Zhang H, Fiszman M, Shin D, Wilkowsk B, Rindflesch TC. Clustering cliques for graph-based summarization of the biomedical research literature. BMC Bioinform 2013;14:182.
[30] Ruch P, Boyer C, Chichester C, Tbahriti I, Geissbuhler A, et al. Using argumentation to extract key sentences from biomedical abstracts. Int J Med Inform 2007;76:195–200.
[31] Lu Z, Cohen B, Hunter L. Finding GeneRIFs via Gene Ontology annotations. In: Pacific symposium on biocomputing; 2006. p. 52–63.
[32] Jin F, Huang Z Mand Lu, Zhu X. Towards automatic generation of gene summary. In: Proceedings of the BioNLP workshop; 2009. p. 97–105.
[33] Aronson AR, Lang FM. An overview of metamap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–36.
[34] MetaMap. <http://metamap.nlm.nih.gov/>.
[35] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci 2002;99:7821–6.
[36] Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways. Bioinformatics 2003;19:532–8.
[37] The HUGO Gene Nomenclature Committee (HGNC). <http://www.genenames.org/>.
[38] Lin CY. Rouge: a package for automatic evaluation of summaries. In: Proceedings of the ACL 2004 workshop: text summarization branches out; 2004. p. 74–81.
[39] Plaza L, Carrillo-de Albornoz J. Evaluating the use of different positional strategies for sentence selection in biomedical literature summarization. BMC Bioinform 2013;14:71.
[40] Jimeno-Yepes AJ, Plaza L, Mork JG, Aronson AR, Daz A. Mesh indexing based on automatically generated summaries. BMC Bioinform 2013;14:208.
[41] Nci thesaurus. <http://ncit.nci.nih.gov/>.