A Relation Mining and Visualization Framework for Automated Text Summarization

Muhammad Abulaish[1,*], Jahiruddin[1], and Lipika Dey[2]

[1] Department of Computer Science, Jamia Millia Islamia, New Delhi, India
abulaish@ieee.org, jahir.jmi@gmail.com
[2] Innovation Labs, Tata Consultancy Services, New Delhi, India
lipika.dey@tcs.com

Abstract. In this paper, we present a relation mining and visualization framework to identify important semi-structured information components using semantic and linguistic analysis of text documents. The novelty of the paper lies in identifying key snippet from text to validate the interaction between a pair of entities. The extracted information components are exploited to generate semantic network which provides distinct user perspectives and allows navigation over documents with similar information components. The efficacy of the proposed framework is established through experiments carried out on biomedical text documents extracted through PubMed search engine.

Keywords: Relation mining, Text mining, Text visualization, Text summarization, Natural language processing.

1 Introduction

The rapidly growing repository of text information on any topic necessitates the design and implementation of strategies that enables fast and efficient access to relevant content. While search engines provide an efficient way of accessing relevant information, the sheer volume of the information repository on the Web makes assimilation of this information a potential bottleneck in the way its consumption. Thus, in the age of increasing information availability, it is essential to provide users with a convenient way to understand information easily. To achieve this goal, many techniques, including document classification and document summarization have been developed [7]. However, most of these methods only provide ways for users to easily access the information; they do not help users directly capture the key concepts and their relationships with the information. Understanding key concepts and their relationships is critical for capturing the conceptual structure of document corpora and avoiding information overload for users. Besides, development of intelligent techniques to collate the information extracted from various sources into a semantically related structure can aid the user for visualization of the content at multiple levels of complexity. Such a visualizer provides a semantically integrated view of the underlying text repository

* To whom correspondence should be addressed.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 249–254, 2009.

in the form of a consolidated view of the concepts that are present in the collection, and their inter-relationships as derived from the collection along with their sources. The semantic net thus built can be presented to the user at arbitrary levels of depth as desired.

In this paper, we propose a relation mining and visualization framework which uses linguistic and semantic analysis of text to identify key information components from text documents. The information components are centered on domain entities and their relationships, which are extracted using natural language processing techniques and co-occurrence-based analysis. The novelty of the paper lies in identifying key snippet from text to validate the interaction between a pair of entities. For example, consider the following sentence: "*Diallyl Sulfite (DAS) prevents cancer in animals (PMID: 8636192)*". In this sentence, "*prevents*" is identified as relational verb relating the biological entities "*Diallyl Sulfite (DAS)*" and "*cancer*" while "*animals*" is identified as key snippet to validate the interaction *prevents*(*Diallyl Sulfite (DAS), cancer)*. Our system also extracts the adverbs associated with relational verbs, which plays a very important role especially to identify the negation in sentences. We have also presented a scheme for semantic net generation which highlights the role of a single entity in various contexts and thus useful for a researcher as well as a layman. The efficacy of the proposed framework is established through experiment over documents from biomedical domain in which information overload also exists due to exponential growth of scientific publications and other forms of text-based data as a result of growing research activities. The remaining paper is structured as follows. Section 2 presents a brief introduction to related work. Section 3 presents the proposed framework and details about functioning of different modules. Section 4 presents the evaluation result of the proposed framework. Finally, section 5 concludes the paper with future directions.

2 Related Work on Relation Mining and Visualization

Context of entities in a document can be inferred from analysis of the inter-entity relations present in the document. Sekimizu *et al.* [5] proposed a linguistic-based approach to mine relationship among gene and gene products. They have collected the most frequently occurring verbs in a collection of abstracts and developed partial and shallow parsing techniques to find the verb's subject and object. Thomas *et al.* [6] modified a pre-existing parser based on cascaded finite state machines to fill templates with information on protein interactions for three verbs – *interact with, associate with, bind to*. The PASTA system [1] is a more comprehensive system that extracts relations between proteins, species and residues. Ono *et al.* [2] reports a method for extraction of *protein-protein interactions* based on a combination of syntactic patterns. Rinaldi *et al.* [4] have proposed an approach towards automatic extraction of a predefined set of seven relations in the domain of molecular biology, based on a complete syntactic analysis of an existing corpus.

Visualization is a key element for effective consumption of information. Semantic nets provide a consolidated view of domain concepts and can aid in this process.

In the information visualization literature, a number of exploratory visualization tools are described in [3]. Zheng et al. [7] have proposed an ontology-based visualization framework, GOClonto, for conceptualization of biomedical document collections.

Although, some visualization methods extract key concepts from document corpora, most of them do not explicitly exploit the semantic relationships between these concepts. The proposed method differs from all these approaches predominantly in its use of pure linguistic techniques rather than use of any pre-existing collection of entities and relations. Moreover, the knowledge visualizer module is integrated with the underlying corpus for comprehending the conceptual structure of biomedical document collections and avoiding information overload for users. On selecting a particular entity or relation in the graph the relevant documents are displayed with highlighting the snippet in which the target knowledge is embedded.

3 Proposed Relation Mining and Visualization Framework

Fig. 1 highlights the functional details of the proposed framework, which comprises of the following four main modules – *Document Processor*, *Concept Extractor*, *Information Component Extractor*, and *Knowledge Visualizer*. The design and working principles of these modules are presented in the following sub-sections.

3.1 Document Processor and Concept Extractor

The *document processor* cleans text documents by filtering unwanted chunks for Parts-Of-Speech (POS) analysis which assigns POS tags to every word in a sentence. The POS tags help in identifying the syntactic category of individual words. For POS analysis we have used the Stanford parser to convert every sentence into a phrase structure tree which is further analyzed by the *concept extractor module* to extract noun phrases. In order to avoid long noun phrases, which generally contain conjunctional words, we have considered only those noun words that appear at leaf- nodes in phrase structure tree. In case, more than one noun phrase appears at leaf node as siblings, the string concatenation function is applied to club them into a single noun phrase. Thereafter, based on a standard list of 571 stopwords, the words that occur frequently but have no meaning are removed. A phrase is removed from the list if it contains a stopword either at beginning or at end position. After extracting all noun phrases from all text documents in the corpus their weights are calculated as a function of term frequency (tf) and inverse document frequency (idf) by using the equation $\omega(p) = tf \times \log(N/n)$, where $\omega(p)$ represents the weight of the noun phrase *P*, *tf* denotes the total number of times the phrase *P* occurs in the corpus, $\log(N/n)$ is the idf of *P*, *N* is the total number of documents in the corpus, and *n* is the number of documents that contain *P*. A partial list of noun phrases in descending order of their weights extracted from a corpus of 500 PubMed abstracts on "Breast Cancer" is: *breast cancer*, *breast carcinoma*, *prostate cancer*, *zoledronic acid*, *tumor*, *estrogen*, *prognosis*, *tomoxifen*, *protein*.

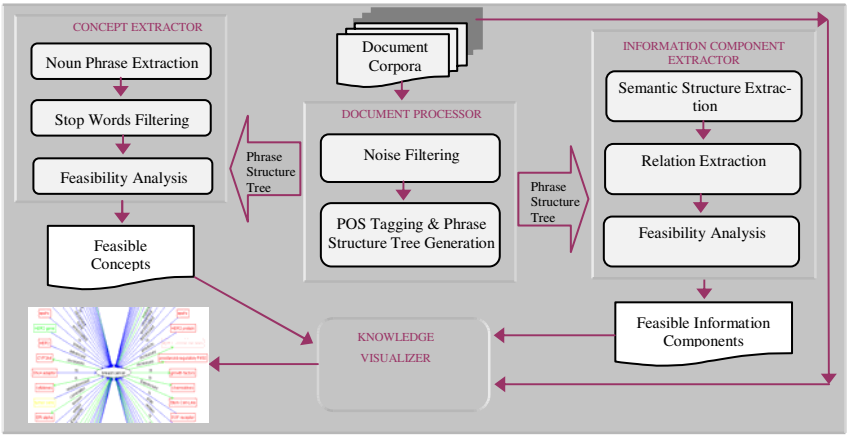

Fig. 1. Relation mining and visualization framework

3.2 Information Components Extractor

An information component can be defined formally as a 7-tuple of the form $<\mathcal{E}_i, \mathcal{A}, \mathcal{V}, \mathcal{P}_v, \mathcal{E}_j, \mathcal{P}_c, \mathcal{E}_k>$ where, $\mathcal{E}_i$, $\mathcal{E}_j$, and $\mathcal{E}_k$ are biological entities identified as noun phrases; $\mathcal{E}_i$ and $\mathcal{E}_j$ forms the *subject* and *object* respectively for $\mathcal{V}$; $\mathcal{A}$ is adverb; $\mathcal{V}$ is relational verb, $\mathcal{P}_v$ is verbal-preposition associated with $\mathcal{V}$; $\mathcal{P}_c$ is conjunctional-preposition linking $\mathcal{E}_j$ and $\mathcal{E}_k$. The information component extraction module traverses the phrase structure tree and analyzes phrases and their linguistic dependencies in order to trace relational verbs and other constituents. Since entities are marked as noun phrases in the phrase structure tree, this module exploits phrase boundary and proximitivity to identify relevant information components. Initially all tuples of the form $<\mathcal{E}_i, \mathcal{A}, \mathcal{V}, \mathcal{P}_v, \mathcal{E}_j, \mathcal{P}_c, \mathcal{E}_k>$ are retrieved from the documents. Thereafter, feasibility analysis is applied to filter out non-relevant verbs and thereby the corresponding information component.

Rather than focusing only on root verbs, morphological variants of a relational verb and associated prepositions are also recognized by our system. Information component extraction process is implemented as a rule-based system. Four sample information components extracted from PubMed documents are presented in Table 1.

Table 1. Four sample information components extracted from PubMed documents

Left Entity	Adv.	Rel. Verb	Verb Prep	Right Entity	Conj Prep	Key Snippet
AO enzymes	---	associated	with	breast cancer and aging	---	---
the levels of MDC1 and BRIT1	---	correlated	with	centrosome amplification, defective mitosis and cancer metastasis	in	human breast cancer
oral glutamine (GLN)	---	inhibited	---	tumor growth	through	Stimulation of host
BMP-4	not	stimulate	---	cell proliferation	by	itself

3.3 Biological Knowledge Visualizer

The major idea of visualization using semantic net is to highlight the role of a concept in a text corpus by eliciting its relationship to other concepts. The nodes in a semantic net represent entities/concepts and links indicate relationships. The whole graph is centered around a selected entity by the user. For a relation triplet $<\mathcal{E}_i, \mathcal{R}, \mathcal{E}_j>$, which is a subset of information component (IC), the entities $\mathcal{E}_i$ and $\mathcal{E}_j$ are used to define classes and $\mathcal{R}$ is used to define relationships between them. The other constituents of the IC can be used to handle biomedical queries which design is one of our future works. While displaying $\mathcal{E}_i$ and $\mathcal{E}_j$, the biological entities are marked with ABNER – a biological entity recognizer. To define a relationship map, the user selects an entity, say ξ, around which the graph is to be created. The selected entity ξ is used to extract all those ICs which contain ξ either as a part of $\mathcal{E}_i$, or $\mathcal{E}_j$ or both. A snapshot of the semantic net generated around the biological entity "breast cancer" is shown in Fig. 2. The semantic net is also designed to facilitate users to navigate through the pile of documents in an efficient way. While double-clicking on a node all the ICs containing the entity stored in that node are selected. Thereafter, the documents containing these ICs are displayed in which the relevant parts of the sentences are highlighted. Similarly, on double-clicking an edge, all the ICs centered around the relation appearing as edge label are selected. Thereafter, the documents containing these ICs are displayed with properly highlighting the relevant snippet of text. The ICs are also displayed separately in the bottom pan of the same window.

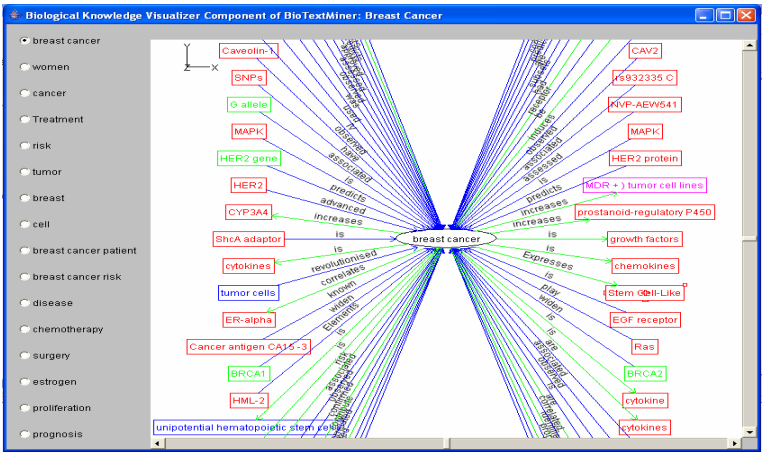

Fig. 2. Semantic net generated around *"breast cancer"*

4 Evaluation of the Relation Extraction Process

The performance of the whole system is analyzed by taking into account the performance of the relation extraction process. A relation is said to be "correctly identified" if its occurrence within a sentence along with its left and right entities is grammatically correct, and the system has been able to locate it in the right context. To judge the

performance of the system, it is not enough to judge the extracted relations only, but it is also required to analyze all the correct relations that were missed by the system. The system was evaluated for its *recall* and *precision* values for ten randomly selected relation triplets. The precision and recall value of the proposed system is found to be 88.77% and 84.96% respectively. On analysis, it was found that most of the incorrect identifications and misses occur when the semantic structure of a sentence is wrongly interpreted by the parser.

5 Conclusion and Future Work

In this paper, we have proposed the design of a relation mining and visualization framework which uses linguistic and semantic analysis of text to identify feasible information components from unstructured text documents. We have also proposed a method for collating information extracted from multiple sources and present them in an integrated fashion with the help of semantic net. Presently, we are refining the rule-set to improve the *precision* and *recall* values of the system. Moreover, a query processing module is being developed to handle biomedical queries on underlying repository of extracted information components.

References

1. Gaizauskas, R., Demetriou, G., Artymiuk, P.J., Willett, P.: Protein Structures and Information Extraction from Biological Texts: the PASTA System. Bioinformatics 19(1), 135–143 (2003)
2. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated Extraction of Information on Protein-Protein interactions from the Biological Literature. Bioinformatics 17(2), 155–161 (2001)
3. Plaisant, C., Fekete, J.-D., Grinsteinn, G.: Promoting Insight-based Evaluation of Visualizations: From Contest to Benchmark Repository. IEEE Transactions on Visualization and Computer Graphics 14(1), 120–134 (2008)
4. Rinaldi, F., Scheider, G., Andronis, C., Persidis, A., Konstani, O.: Mining Relations in the GENIA Corpus. In: Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, pp. 61–68 (2004)
5. Sekimizu, T., Park, H.S., Tsujii, J.: Identifying the Interaction between Genes and Genes Products based on Frequently Seen Verbs in Medline Abstract. Genome Informatics 9, 62–71 (1998)
6. Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M.: Automatic Extraction of Protein Interactions from Scientific Abstracts. In: Pacific Symposium on Biocomputing, pp. 538–549 (2000)
7. Zheng, H.-T., Borchert, C., Kim, H.-G.: Exploiting Gene Ontology to Conceptualize Biomedical Document Collections. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 375–389. Springer, Heidelberg (2008)