2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops

Text Mining through Entity-Relationship Based Information Extraction

Lipika Dey[1], Muhammad Abulaish[2], Jahiruddin[3] and Gaurav Sharma[4]
[1] *Innovation Labs, Tata Consultancy Services, New Delhi, India*
lipika.dey@tcs.com
[2] *Department of Mathematics,* [3]*Department of Computer Science,*
Jamia Millia Islamia (A Central University), New Delhi – 25, India
abulaish@ieee.org
[4] *Department of Mathematics, Indian Institute of Technology, New Delhi – 16, India*

Abstract

In this paper we describe an information extraction and text mining system which identifies key information components from text documents. The information components are centered on domain entities and their relationships. The components mined from a repository are chained using an n-gram-based algorithm. The information chains provide a comprehensive view of the collection and can be also used for inferential reasoning.

Keywords: Information extraction, Visualization, Text mining, Knowledge discovery

1. Introduction

Text mining, also known as text data mining [6] or knowledge discovery from textual databases [3] refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents from a fixed domain. Text mining involves the process of extracting semi-structured information components from text, usually from multiple documents, and then reason with these semi-structured components to derive patterns within them. The aim of text mining is to provide the following facilities : (i) Distill the meaning of a text in a concise form, (ii) View accurate summaries before plunging into full documents, (iii) Navigate efficiently through large textbases, and (iv) Perform natural language information retrieval. Automated text summarization and visualization are immensely popular with magazine editors, political and business analysts, and students who wish to see summaries before plunging into the full documents. The potential of text data mining had been visualized by Swanson who had shown how chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases [9].

In this paper we present methodologies to identify important semi-structured information components using semantic and linguistic analysis of text documents. We have also presented methodologies for integrating information from multiple sources that can help in easy comprehension and smooth navigation through the pile of information. We propose the use of entity-relationships as semi-structured information components. We have presented experimental results showing information mined from PUBMED documents related to Alzheimer's disease, Cancer and AIDS. In section 2 we review some of the related works. Section 3 elaborates on the text mining approach to extract relations and their arguments from text documents. In section 4 we present the mining process to discover information chains. Finally section 5 concludes the paper with future directions.

2. Related work

Fukuda *et al.* [5] had proposed a rule-based method called PROtein Proper-noun phrase Extracting Rules (PROPER) to extract material names from sentences using surface clue on character strings in medical and biological documents. Machine-learning based techniques like Hidden Markov Model, Naïve Bayes and Support Vector Machines (SVMs) have been successfully applied to identify and classify gene/protein names in text documents. Biological relationship extraction has been addressed in [2], [7], [8]. Friedman *et al.* [4] have developed a natural-language processing system, GENIES, for the extraction of molecular pathways from journal articles. [1] and [9] have presented semi-automated mechanisms for inferencing from information components, was partially automated. However, generic, efficient algorithms for relevant text information extraction, mining and inferencing all at one go are still rare. Hearst [6] had suggested that good algorithms will

0-7695-3028-1/07 $25.00 © 2007 IEEE
DOI 10.1109/WI-IATW.2007.75

have to take into account various kinds of semantic and linguistic constraints.

3. Relation Extraction through Text Mining

The proposed approach to extract semi-structured information components collected from a focused corpus explores *the roles of biological entities* in the corpus and integrates these into cohesive structures. We propose a multi-perspective collation mechanism to explore various aspects of the domain. *Roles* of entities are characterized by relations expressed in a sentence in which these entities occur. These relations are identified through semantic and linguistic analysis [2]. Natural Language Processing (NLP) tools are used to identify entities and relations in a document. Entities are identified as Noun Phrases. An information component is a relation triplet defined as <$\mathcal{S}$, $\mathcal{R}$, $\mathcal{O}$> where $\mathcal{R}$ is a relational verb representing an event/action/process and $\mathcal{S}$ and O are the associated subject and object respectively. Subject and object are entities represented by noun phrases. Relation extraction from text is a two step process. Text documents are parsed for Parts of Speech (POS) analysis. POS analysis tags are used for extracting information components. These steps are explained in brief in the following paragraphs.

Each text document is parsed using the Stanford Parser which is a probabilistic parser. Based on the parser output each sentence is converted into a dependency tree. The dependency tree encodes linguistic relationships like subject-object, possession, conjunction etc. among words in a sentence. The relation triplet extraction process has been presently implemented as a rule-based system. One sample rule is presented below to highlight the functioning of the system.

Table 1. Sample sentences and corresponding dependency trees generated by the Stanford parser

Sentence No. 1. [PMID: 17446028]
Alzheimer's disease (AD) is the commonest form of degenerative dementia and is characterised by progressive cognitive decline.

Dependency Tree:
poss(disease-3, Alzheimer-1), nsubj(is-7, disease-3), dep(disease-3, AD-5), det(form-10, the-8), amod(form-10, commonest-9), dobj(is-7, form-10), amod(dementia-13, degenerative-12), of(form-10, dementia-13), dep(characterised-16, is-15), and(is-7, characterised-16), amod(decline-20, progressive-18), amod(decline-20, cognitive-19), by(characterised-16, decline-20)

Table 2. A partial list of relation triplets extracted by using rules 1 and 2 from a collection of bio-medical text documents

IC	Domain	Subject	Relational Verb	Preposition	Object
1	Alzheimer’s Disease	Alzheimer 's disease	is		the commonest form of degenerative dementia
2	Alzheimer’s Disease	Formation of beta-amyloid plaques	is		a crucial feature of Alzheimer 's disease
3	Alzheimer’s Disease	BACE1	is		the protease responsible for the production of amyloid-beta peptides that accumulate in the brain of Alzheimer 's disease (AD) patients
4	Alzheimer’s Disease	Down-regulation of HIF-1alpha	reduced		the level of BACE1
5	Alzheimer’s Disease	Hypoxia	facilitates		Alzheimer 's disease pathogenesis by up-regulating BACE1 gene expression
6	AIDS and Cancer	HIV infected people and AIDS patients	develop		Cancer
7	AIDS	Streptococcus pneumoniae and Legionella pneumophila	pneumonia	in	HIV infected patients
8	Cancer	both estrogen(E (2)) and hypoxia	Involved	in	tumor development and progression
9	Cancer	siRNA targeting MGr1-Ag	showed		a markedly decreased VCR-induced HIF-1alpha expression and transcriptional activity

Sample Rule: If there exist two dependencies involving two different entities $\mathcal{E}_i$ and $\mathcal{E}_j$ associated with single verb $\mathcal{V}$ satisfying the condition [Subj ($\mathcal{V}$, $\mathcal{E}_i$) $\wedge$ Obj ($\mathcal{V}$, $\mathcal{E}_j$)], then $\mathcal{V}$ is identified as a relational verb between the two entities $\mathcal{E}_i$ and $\mathcal{E}_j$. It is characterized as an instance of binary relation represented by $\mathcal{E}_i \rightarrow \mathcal{V} \leftarrow \mathcal{E}_j$. During triplet extraction, $\mathcal{E}_i$ is treated as a head noun and the longest noun phrase containing this is considered as the subject of the information component. Similarly, the longest noun phrase containing head noun $\mathcal{E}_j$ forms the object of the

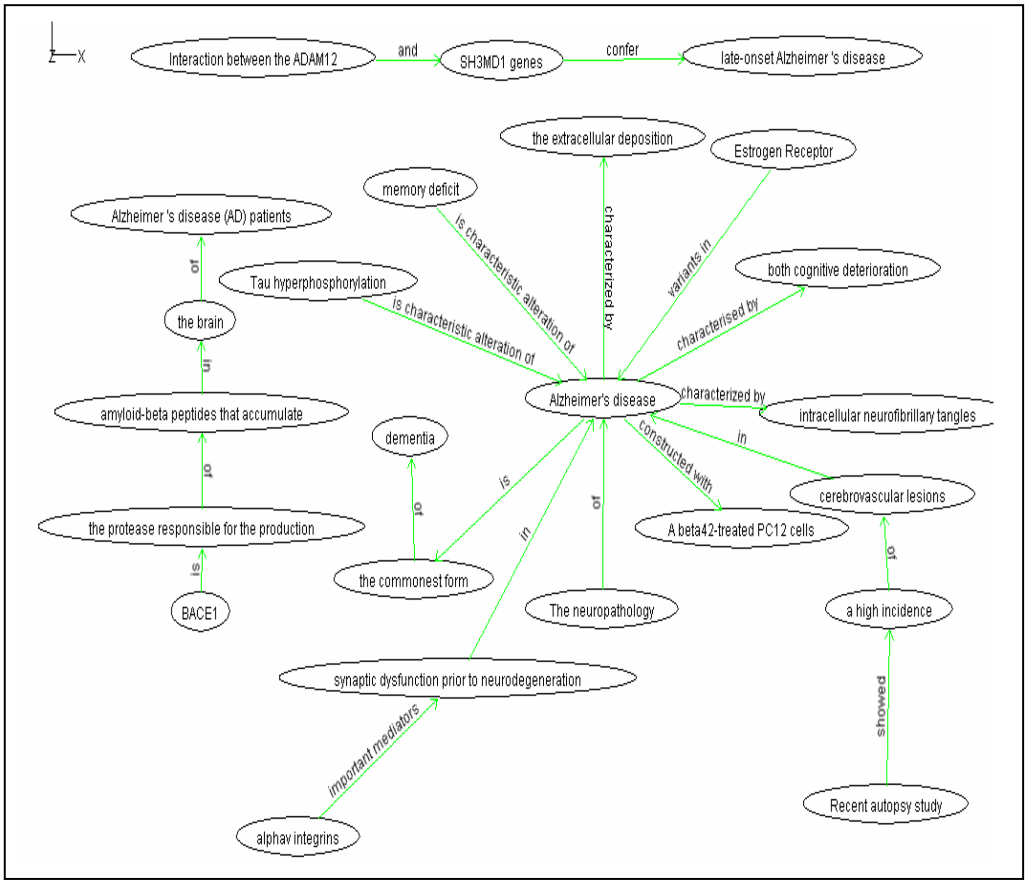

Figure 1. Semantic Net generated around the biological concept “Alzheimer’s disease”

information component. By applying this rule, the two relation triplets identified from the first sentence in table 1 is <Alzheimer's disease (AD) → is ← the commonest form of degenerative dementia> and <Alzheimer's disease (AD) → characterised by ← progressive cognitive decline>.

Table 2 shows a partial list of relation triplets <subject, relation, object>, extracted by using our rule set from a collection of abstracts which were returned by PubMed on Alzheimer’s Disease, Cancer and AIDS.

4. Text Summarization Using Information Components

The information components extracted from a corpus can be used to provide a summarized view of information both document-wise and corpus-wise. Figure 1 displays a set of relations, where either the subject or the object contained the term “Alzheimer’s Disease”. These relations were extracted from top hundred documents extracted from PUBMED with the query “Alzheimer’s Disease”. It is obvious that these relations capture various key information about “Alzheimer’s Disease” including its manifestation, cause, pathological condition etc. very succinctly. When information chains are computed from a single homogeneous corpus, the corpus can be navigated incrementally using the chains. For example, starting with relation number 3 of Table 1, one can further look at information components containing BACE1 in subject or object. One such relation found is “*Mitochondrial respiratory inhibition and oxidative stress elevate beta-secretase(BACE1) proteins*”, which provide valuable information about the reasons.

When information chains are computed over a heterogeneous collection, knowledge gained from one domain can be extended to another domain through similarity over components. Table 3 shows the role played by “hypoxia” that is lack of oxygen and “Hypoxia Inducing Factors” as summarized from a heterogeneous corpus on all diseases. Two types of chaining, forward and backward are used. While forward chaining identifies inferential reasoning path by establishing similarity between the object of earlier component and the subject of the next, backward chains represent similar information components.

The proposed similarity computation algorithm locates possibly similar entities even when there is variation in naming of a single entity. For example, the algorithm ensures that the term “amyloid beta plaques” have a good match with “beta amyloid plaques” or even “abeta”. The algorithm uses a novel weighted n-gram method and works as follows:

- Step 1: For all 2-grams and 3-grams extracted from the entity tokens, its frequency in each token is stored.
- Step 2: All full token matches are rewarded while token misses are penalized

Table 3. Summarizing study of role of Hypoxia and Hypoxia Inducing factor (HIF) across diseases

17121991	<**Hypoxia,** facilitates, **Alzheimer 's disease** pathogenesis by up-regulating BACE1 gene expression> <**Hypoxia**, is, a direct consequence of hypoperfusion> <**Hypoxia**, facilitate, AD pathogenesis> <**Hypoxia** treatment markedly, increased, Abeta deposition and neuritic plaque formation>
17303576	<Acute **hypoxia**, increases, the expression and the enzymatic activity of BACE1 by up-regulating the level> <Results demonstrate an important role for **hypoxia**\/HIF-1alpha, in, modulating the amyloidogenic processing of APP and provide a molecular mechanism for increased incidence of AD following cerebral ischemic and stroke injuries>
CANCER1	<Both estrogen(E (2)) and **hypoxia**, involved, tumor development and progression>
CANCER2	<VCR, induce, a significant expression of HIF-1alpha> <VCR-resistant SGC7901\/VCR cells, had, higher expression of **HIF**-1alpha> <VCR, enhance, DNA binding activity and transcriptional activity of **HIF**-1alpha by 5.42 - and 9.42-fold> Further study, upregulated, HIF-1alpha protein expression and transcriptional activity in gastric cancer cell
CANCER3	<siRNA targeting MGr1-Ag, showed, a markedly decreased VCR-induced HIF-1alpha expression and transcriptional activity>
	<SiRNA, be, the major signaling molecules in MGr1-Ag \/ 37LRP-induced **HIF-1**alpha expression>

5. Conclusion

In this paper we have proposed a text mining mechanism which extracts entity-relationships from documents. Summarizing a document through relations help in easy visualization of contents of a repository and experimental results for PUBMED documents have been presented. Chaining through relations can also help in the discovery of potentially interesting information from the vast text repository. This is an attractive idea since it can direct future research and provide interesting insights into a domain.

References

[1] Beeferman, D., Lexical discovery with an enriched semantic network, in: Proceedings of the ACL/COLING Workshop on Applications of WordNet in Natural Language Processing Systems, 1998, pp. 358-364.

[2] Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I., Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI’05), pp. 659-664.

[3] Feldman, R. & Dagan, I.: 1995, Knowledge Discovery in Textual Databases, in: *Proceedings of KDD’95*, pp. 112-117.

[4] Friedman, C., Kra, P., Yu, H., Krauthammer, M., A. Rzhetsky, GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles, Bioinformatics, vol. 17, Suppl. 1, 2001, pp. s74-s82.

[5] Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T., Toward Information Extraction: Identifying Protein Names from Biological papers, in: Proceedings of the Pacific Symposium on Biocomputing, Hawaii, 1998, pp. 707-718.

[6] Hearst, M. A., Untangling Text Data Mining, in: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, 1999, pp. 3-10.

[7] Jenssen, T. -K., Laegreid, A., Komorowski, J., Hovig, E., A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression, Nat. Genet 2001, pp. 28: 21-28.

[8] Sekimizu, T., Park, H. S., Tsujii, J., Identifying the Interactions Between Genes and Genes Products based on Frequently Seen Verbs in MEDLINE Abstract, Genome Informatics 9, 1998, pp. 62–71.

[9] Swanson, D. R., and Smalheiser, N. R., An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence* 91, 1997, pp. 183-203.