Semantic Integration of Information Through Relation Mining - Application to Bio-medical Text Processing

Lipika Dey,[1] Muhammad Abulaish,[2] Rohit Goyel[3], and Jahiruddin[4]

[1] Innovation Labs, Tata Consultancy Services, New Delhi, India
lipika.dey@tcs.com
[2] Department of Mathematics, Jamia Millia Islamia, New Delhi, India
abulaish@ieee.org
[3] Department of Mathematics, Indian Institute of Technology, New Delhi, India
[4] Department of Computer Science, Jamia Millia Islamia, New Delhi, India

Abstract. Semantic frameworks can be used to improve the accuracy and expressiveness of natural language processing for the purpose of extracting meaning from text documents. Such a framework represents knowledge using semantic networks and can be generated using information mined from text documents. The key issue however is to identify relevant concepts and their inter-relationships. In this paper, we have presented a scheme for semantic integration of information extracted from text documents. The extraction principle is based on linguistic and semantic analysis of text. Entities and relations are extracted using Natural Language Processing techniques. A method for collating information extracted from multiple sources to generate the semantic net is also presented. The efficacy of the proposed semantic framework is established through experiments carried out for visualizing information embedded in biomedical texts extracted from PubMed database.

Keywords: Relation extraction, Semantic net, Knowledge visualization, NLP.

1 Introduction

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. One approach to overcome this difficulty could be to use intelligent techniques to collate the information extracted from various sources into a semantically related structure which the user can aid the visualization of the content at multiple levels of complexity. Such a visualiser provides a semantically integrated view of the underlying text repository 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. It may be noted that

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 365–372, 2007.

the proposed semantic net is different from a domain ontology [4] which is a more rigid structure aimed at presenting a shared conceptualization of a domain, usually representing knowledge that is ratified by domain experts. The semantic net based visualization system proposed in this paper is more of an aid towards assimilating knowledge from a large collection. However, the semantic net thus built can be used to build ontology, as proposed in [1], when used with restraint over a focused and authentic text corpus, armed with appropriate feasibility analysis mechanisms.

In this paper, we have presented a scheme for semantic integration of information extracted from text documents. The information components extracted from text are either concepts or inter-concept relations. The extraction principle is based on semantic analysis of text from which entities and relations are extracted using Natural Language Processing (NLP) techniques. Though there has been a lot of work on entity extraction from text documents, relation mining has received far less attention. We have shown that relation mining can yield significant information components from text whose information content is much more than entities. We have also proposed a method for collating information extracted from multiple sources and present them in an integrated fashion. The system functions are designed to work in a domain-independent way though here we predominantly present results from the biological domain. This has been chosen since the growth of articles in this area over the last decade has necessitated development of dedicated search engines for locating relevant documents to enable scientists and researchers assimilate information about ongoing research. The proposed system has been shown to be capable of collating and presenting information from multiple scientific abstracts to present a global view of the collection. It is possible to slice and dice or aggregate to get more detailed or more consolidated view as desired.

The remaining paper is structured as follows. Section 2 elaborates on the text mining approach to extract relations and their arguments from text documents. Section 3 presents the semantic net generation algorithm. Section 4 presents the experimental results. Section 5 is a review of the related works and finally section 6 concludes the paper with future directions.

2 Relation Extraction Through Text Mining

Over the last decade, a lot of work has been done on locating and recognizing biological entities in documents. The proposed approach to building a semantic net of information explores *the roles of biological entities* in a collection of document and integrates the information components thus extracted into a cohesive structure. *Roles* of entities are characterized by relations expressed in a sentence in which these entities occur. These relations can be identified through semantic and linguistic analysis [2]. The relation mining framework presented in this paper uses NLP tools to identify entities and relations in a document. It is domain-agnostic in nature. Entities within a document can be identified as Noun Phrases. For bio-medical documents one can additionally use biological

entity extractors. Relation extraction from text is a two step process which is explained in the following sub-sections.

2.1 Document Pre-processing and Parsing

The purpose of this step is to expedite the parsing process and facilitate the extraction of information components from text documents. While working with biological abstracts extracted through PubMed, we have eliminated author's names, their affiliations etc. The text documents are parsed whereby each word is assigned a Parts-Of-Speech (POS) tag and each sentences is converted into a dependency tree. We have used a statistical parser (Stanford Parser[1]) that has been developed by the standard natural language processing group, Stanford University. Two sample sentences, their tagged forms and corresponding dependency trees created by the Stanford parser are shown in Table 1. The dependency tree extracts linguistic relationships like *subject, object, possession, conjunction* etc. among words in a sentence.

2.2 Relation Extraction

The proposed approach to relation extraction traverses the dependency tree and analyzes the linguistic dependencies in order to trace biologically significant relations. A biological relation is usually manifested in a document as a relational verb. All relational verbs however do not represent biological relations and only those which are located in the proximity of biological entities are considered. In order to identify valid biological relations we apply a pattern-mining based technique. A biological relation along with the associated biological entities is termed as relation triplets (RT). A biological relation is characterized by verb and may occur in a sentence in its root form or as a variant of it. Different classes of variants of a relational verb are recognized by our system. *Morphological variants* of a root verb consist of self-modifications. For example, the root verb *"activate"*, has three inflectional verb forms: *"activates"*, *"activated"* and *"activating"*. In the context of biological relations, we also observe that the occurrence of a verb in conjunction with a preposition very often changes the nature of the verb. For example, the relation *"activates in"* denotes a significant class of biological reactions. Thus, we also consider a second category of biological relations, which are combinations of *root verbs* or their *morphological variants*, and *prepositions* that follow these. Typical examples of biological relations identified in this category include *"activated in"*, *"binds to"*, *"stimulated with"* etc.

RT extraction process is implemented as a rule-based system. Dependencies output by the Parser are analyzed to identify *subject, object, verb, preposition*, and various other relationships among elements in a sentence. Some sample rules are presented below to highlight the functioning of the system.

Rule 1: If there exist two dependencies involving two different entities E_i and E_j associated with single verb V satisfying the condition $[Subj(V, E_i) \wedge Obj(V, E_j)]$,

[1] http://nlp.stanford.edu/downloads/lex-parser.shtml

Table 1. Sample sentences, their tagged form and corresponding dependency tree 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. **Taged Sentence:** Alzheimer/NNP 's/POS disease/NN -LRB-/-LRB- AD/NNP -RRB-/-RRB- is/VBZ the/DT common-est/JJ form/NN of/IN degenerative/JJ dementia/NN and/CC is/VBZ characterised/VBN by/IN progres-sive/JJ cognitive/JJ decline/NN ./. **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)
Sentence No. 2. [PMID: 17445916] Alzheimer's disease is characterised by both cognitive deterioration and the development of a wide range of neuropsychiatric disturbances... **Taged Sentence:** Alzheimer/NNP 's/POS disease/NN is/VBZ characterised/VBN by/IN both/DT cognitive/JJ deteriora-tion/NN and/CC the/DT development/NN of/IN a/DT wide/JJ range/NN of/IN neuropsychiatric/JJ disturbances/NNS ... **Dependency Tree:** poss(disease-3, Alzheimer-1) nsubjpass(characterised-5, disease-3), aux(characterised-5, is-4), det(deterioration-9, both-7), amod(deterioration-9, cognitive-8), by(characterised-5, deterioration-9), det(development-12, the-11), and(deterioration-9, development-12), det(range-16, a-14), amod(range-16, wide-15), of(development-12, range-16), ...

then V is identified as a relational verb between the two entities E_i and E_j. It is characterized as an instance of binary relation represented by $E_i \rightarrow V \leftarrow E_j$. During RT extraction, E_i is treated as head noun and along with other related words in its proximity forms the subject of the sentence. Similarly, the head noun E_j along with related words in its proximity forms the object of the sentence. By applying this rule, the two relation triplets identified from the first sentence in table 1 is $\langle$*Alzheimer's disease (AD)* $\rightarrow$ *is* $\leftarrow$ *the commonest form of degenerative dementia*$\rangle$ and $\langle$*Alzheimer's disease (AD)* $\rightarrow$ *characterised by* $\leftarrow$ *progressive cognitive decline*$\rangle$.

Rule 2: If there exist two dependencies involving two different entities E_i and E_j associated with single verb V satisfying the condition $[Subj(V, E_i) \wedge P(V, E_j)]$, where P is a prepositional word, then the verb V along with the prepositional word P is identified as a relational verb between the entities E_i and

E_j. It can be characterized as an instance of a binary relation represented by $E_i \rightarrow V - P \leftarrow E_j$.

Rule 2 presents another kind of rule in which the object component is not explicitly marked by the parser in the dependency tree. Rule 2 helps in identifying the relation triplet ⟨ *Alzheimer's disease* → *characterised by* ← *both cognitive deterioration and the development of a wide range of neuropsychiatric disturbances* ⟩ from sentence 2 shown in table 1.

Table 2 shows a partial list of relation triplets ⟨*subject, relation, object*⟩, extracted by using rules 1 and 2 from a collection of abstracts which were returned by PubMed for the query term *"Alzhemer's Disease"*. Relations thus extracted are used to generate a semantic net as explained in the next section.

Table 2. A partial list of relation triplets extracted by using rules 1 and 2 from a collection of text documents describing Alzheimer's disease

Subject	**Relation**	**Object**
BACE1	is	the *protease* responsible for the production of *amyloid-beta peptides* that accumulate in the brain of *Alzheimer 's disease (AD) patients*
Alzheimer's disease	is	the commonest form of *degenerative dementia*
Alphav integrins	be	important mediators of *synaptic dysfunction* prior to *neurodegeneration* in *Alzheimer 's disease*
Interaction between the *ADAM12* and *SH3MD1* genes	confer	*Late-onset Alzheimer 's disease*
Alzheimer 's disease (AD)	characterised by	Both *cognitive deterioration* and the development a wide range of *neuropsychiatric disturbances*
Alzheimer's disease	characterised by	progressive *cognitive decline*

3 Semantic Net Generation

The major idea of generating a 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. While concept ontologies are specialized types of semantic net, which also highlight the taxonomical and partonimical relations among concepts, the proposed semantic net is designed only to represent the biological relations mined from the text corpus. For an extracted relation triplet $\langle E_i, R_a, E_j \rangle$, the entities E_i and E_j are used to define classes and R_a is used to define relationships. Biological entities can be *complex* in nature which includes relations among *atomic* entities. For example, in the current scenario a Noun Phrase *"Interaction between the ADAM12 and SH3MD1 genes"* represents a complex entity, which contains

atomic entities like *ADAM12* and *SH3MD1*. Linguistic analysis rules based on *preposition* and *conjunctive* analysis are recursively applied over complex entities to identify atomic entities and their relationships as well. A relation of the type ¡Ec1, Rc, Ec2¿ where Ec1 (or Ec2) is a complex entity ¡E1, Ri, E2¿ is represented as an n-ary relation whose arguments include the atomic entities and relations extracted from the complex entities. A formal algorithm for semantic net generation is given below:

Algorithm: Semantic Net Generation
Input: *The set of relation triplets* (R) *mined from text documents*
Output: *Semantic net - a directed graph* (G)
Steps:
1. Initialize G *with* ϕ
2. For every relation triplet $\langle E_i, R_a, E_j \rangle \in R$ *do*
3. If $E_i, E_j \notin G$ *then*
  a. Create separate nodes for both left and right entities
  b. Draw a directed edge from E_i *to* E_j *and label with* R_a
4. If E_i *(or* E_j*) is a complex entity then*
  a. Break E_i *(or* E_j*) into a set* $S = \{E_{i1}, E_{i2}, ..., E_{in}\}$ *of atomic entities on the basis of connectors and prepositions*
  b. Create a connected sub-graph G_s *as follows:*

- *Create a separate node for each atomic entity* $E_{ik} \in S$ *if* $E_{ik} \notin G$, $1<k<n$
- *Draw a directed edge from left entity to right entity - with edge marked by the connector between the entities*

  c. Replace node E_i *(or* E_j*) of* G *with sub-graph* G_s
5. Stop

4 Results

We elucidate the proposed approach through results generated from querying the PubMed database for the query term *"Alzheimer's disease"*. A total of 100 documents consisting 1047 sentences were filtered out of 38135 sentences as likely to conatin valid relations triplets. Using *typedDependenciesCollapsed* option, the documents were parsed by Stanford Parser and a total of 751 triplets were extracted from them by applying the RT extraction rules mentioned in section 2.2. A partial list of such triplets is shown in table 2. To initiate the generation of semantic net, we first identified those triplets that contained the query string within the left or right entities as the candidate concept nodes. After applying the linguistic rules over these entity sets the list of concepts is extended to contain the atomic entities also. The semantic net generation algorithm is applied over these elements. Due to limitation of space, a partial view of the generated semantic net is shown in figure 1.

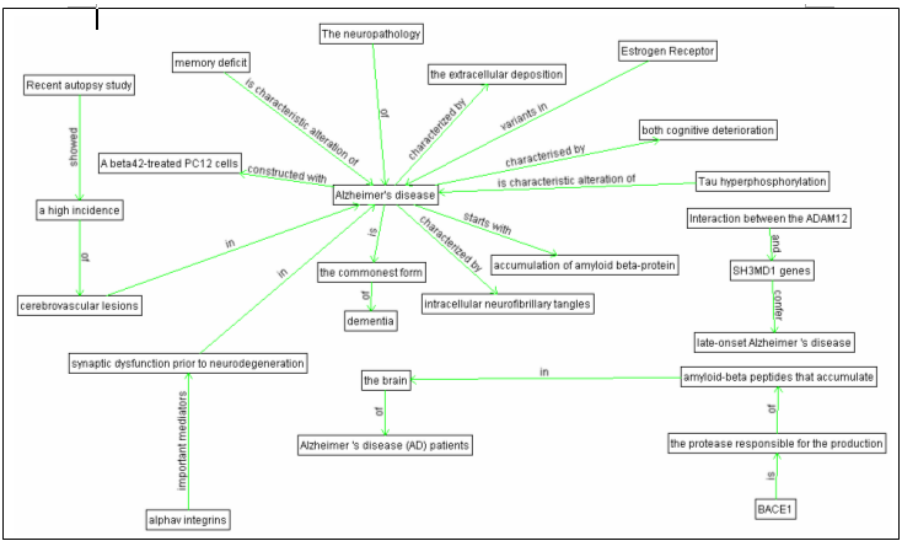

Fig. 1. Semantic Net generated around the biological concept *"Alzheimer's disease"*

5 Related Work

Visualization is a key element for effective consumption of information. Semantic Nets provide a consolidated view of domain concepts and can aid in the process. Wagner *et al.* [6] have suggested building a semantic net using the Wiki technology for making e-governance easier through easy visualization of information. In [5], similar approaches have also been proposed for integrating and annotating multimedia information. In [3], a soft-computing based technique is proposed to integrate information mined from biological text documents with the help of biological databases. [1] proposes building a semantic net for visualization of relevant information with respect to use cases like the *nutrigenomics* use case, wherein the relevant entities around which the semantic net is built are predefined.

The proposed method differs from all these approaches predominantly in its use of pure linguistic techniques rather than using any pre-existing collection of entities. Though biological relation mining [2] have gained attention of researchers for unraveling the mysteries of biological reactions, their use in biological information visualization is still limited.

6 Conclusion and Future Work

In this paper, we have presented a scheme for extracting relevant information from text documents and their semantic integration. The extraction principle is based on semantic analysis of text from which entities and relations are extracted

using Natural Language Processing techniques. 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. The system is being integrated to work as a front-end visualizer for a search engine which can enable quick comprehension of information. As the graph shows, the semantic net highlights the role of a single entity in various contexts which are useful both for a researcher as well as a layman. The limitations of the currently used graph drawing software restrict the appropriate representation of n-ary relations. We are also working towards a proper graph-based visualizer in which we shall also add a method to point to the original documents where concepts occur.

References

1. Castro, A.G., Rocca-Serra1, P., Stevens, R., Taylor1, C., Nashar, K., Ragan, M.A., Sansone, S.-A.: The use of concept maps during knowledge elicitation in ontology development processes - the nutrigenomics use case, BMC Bioinformatics (May 25, 2006)
2. Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology On-tology. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), pp. 659–664 (2005)
3. Cox, E.: A Hybrid Technology Approach to Free-Form Text Data Mining, http://scianta.com/pubs/AR-PA-007.htm
4. Fensel, D., Horrocks, I., van Harmelen, F., McGuinness, D.L., Patel-Schneider, P.: OIL: Ontology Infrastructure to Enable the Semantic Web. IEEE Intelligent Systems 16(2), 38–45 (2001)
5. García, R., Celma, O.: Semantic Integration and Retrieval of Multimedia Metadata. In: Knowledge Mark-up and Semantic Annotation Workshop, Semannot 2005. CEUR (2005)
6. Wagner, C., Cheung, K.S.K., Rachael, K.F.: Building Semantic Webs for e-government with Wiki technology. Electronic Government 3(1) (2006)