ISSN 2047-3338

Situation Modeling Using Sentence Correlation and Importance

Shahzeb Patel, Shreyas Talele, Abhijay Patne, Radhika Simant and Saurabh Karwa
College of Engineering, Pune – 411005
{patelsn10.comp, talelesr09.comp, patneav09.comp, simantrr09.comp, karwass09.comp} @coep.ac.in

***Abstract*—Biomedical abstracts are abundant with information on medical entities. Large numbers of these abstracts are easily accessible and are used in the field of medical research. Extracting useful information out of these abstracts is a field of major interest among text mining community. Biomedical corpus forms a major dataset of research in this area. This paper aims at exploring the development of a Situation Model (SM) on a set of such medical abstracts. It aims at finding, gathering, aggregating and analyzing information from these abstracts which is of utmost importance to the user, made readily available so that the user can see what the abstract aims at establishing. The work also tracks the relations within the text, protein interactions (PPI) to be precise which helps in developing a separate model of interest for the user. This paper details the steps to transform textual resources to a structured SM which covers integrated and exile representation of the available abstracts.**

***Index Terms*—Situation Modeling, Coherence, WordNet, Sentence Similarity and Word Importance Value**

I. Introduction

SITUATION models are mental representations of the state of affairs described in a text rather than of the text itself. People not only care for what it expresses, but also care for the most important idea of it. Situation modeling extracts the important part of the text and presents it to the user in a way the user would like to perceive [1].

Researchers proposed that understanding any text, involves more than merely constructing a mental representation of the text itself. Comprehension is first and foremost the construction of a mental representation of what that text is about: a situation model.

Thus, situation models are mental representations of the people, objects, locations, events, and actions described in a text, not of the words, phrases, clauses, sentences, and paragraphs of a text. The situation-model view predicts that comprehenders are influenced by the nature of the situation that is described in a text, rather than merely by the structure of the text itself.

As an illustration, consider the following sentences: Mary baked cookies but no cake versus Mary baked cookies and cake. Both sentences mention the word cake explicitly, but only the second sentence refers to a situation in which a cake is actually present. If comprehenders construct situation models, the concept of cake should be more available to them when the cake is in the narrated situation than when it is not, despite the fact that the word cake appears in both sentences.

II. Previous Work

When people comprehend text, they not only construct a mental representation of its words and sentences but also the situation conveyed by these. These representations have become known as situation models (van Dijk and Kintsch, 1983) [2], or mental models (Johnson-Laird, 1983) [3]. According to van Dijk and Kintsch, situation models are responsible for processes like domain-expertise, translation, learning from multiple sources or completely understanding situations just by reading about them.

Text comprehension can also be improved by rewriting poorly written texts in order to make them more coherent and to provide the reader with all the information needed for reading. Text coherence refers to the extent to which a reader is able to understand the relations between ideas in a text. This is generally dependent on whether these relations are explicit in the text. The general approach to increasing text coherence is to add surface-level indicators of relations between ideas in the text. Such modifications range from adding low-level information, such as identifying anaphoric referents, synonymous terms, or connective ties, to supplying background information left unstated in the text.

However, increasing text coherence is not necessarily the best condition for learning. Making readers participate more actively in the comprehension process can help memory and learning. In many research domains it has been shown that learning can be improved by making the learners task more difficult (Mannes and W. Kintsch, 1987; Healy and Sinclair, 1996 for skill acquisition; McDaniel et al., 1995 for text recall) [4], [5], [6].

There is evidence that readers construct a situation models while reading texts under particular conditions (e.g. Anderson, Garrod, and Sanford, 1983) ; Ehrlich and Johnson Lard, 1982; Fletcher and Chrysler, 1990; Franklin and Tversky, 1990;

Glenberg, Meyer, and Lindem. 1987; Mandler, 2004; Mani and Johnson-lard, 1982; Morrow, Bower, and Greenspan, 1989; Morrow, Greenspan, and Bower, 1987). Each of these studies has focused on a single aspect of situation models, such as temporal order of events, the spatial layout of the situation, or the causal relations among the described events. [7], [8], [9]

For this reason several researchers wrote about the importance of fore-grounding important information. The other important issue about situation models is the multidimensionality. Here the important question is how the different dimensions are related and what their weight for constructing the model is. Some researchers claim that the weight of the dimensions shifts according to the situation which is described.

III. Background

Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain [3, 9]. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics. There is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications stored in databases such as PubMed. Because of the amount of electronic literature now available, it is challenging for biologists to search biomedical corpuses for any kind of desired information beyond simple text retrieval. Several tools have been developed to make text mining easier for them. Some of these tools like Jsre focus on extracting biomedical terms; such as protein names and biological processes, given any input text.

In response to the unbridled growth of information in literature and biomedical databases, researchers require efficient means of handling and extracting information. As well as providing background information for research, scientific publications can be processed to transform textual information into database content or complex networks and can be integrated with existing knowledge resources to suggest novel hypotheses. Information extraction and text data analysis can be particularly relevant and helpful in genetics and biomedical research, in which up-to-date information about complex processes involving genes, proteins and phenotypes is crucial.

A. Situation Model

Situation models are mental representations of the state of an affairs described in a text rather than of the text itself. Comprehension is first and foremost the construction of a mental representation of what that text is about: a situation model. Situation models are mental representations of the people, objects, locations, events, and actions described in a text, not of the words, phrases, clauses, sentences, and paragraphs of a text. The situation-model view predicts that comprehenders are influenced by the nature of the situation that is described in a text, rather than merely by the structure of the text itself.

B. Coherence

Coherence in linguistics is what makes a text semantically meaningful. It is especially dealt with in text linguistics. Coherence is achieved through syntactical features such as the use of deictic, anaphoric and cataphoric elements or a logical tense structure, as well as presuppositions and implications connected to general world knowledge [10]. In this section we find the coherence values between sentences.

C. WordNet

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet database (version 2.1) contains 155,287 words organized in 117,659 synsets for a total of 206,941 wordsense pairs; in compressed form, it is about 12 megabytes in size. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules.

D. Types of Similarities

Four types of similarities are taken into considerations as follows:

1). Syntactic Similarity: Syntactically same words are given a similarity value of 1. Sentence similarity is influenced to a great extent, if the sentences contain same words.

2). Semantic Similarity: Words may not be same (syntactically) but may be semantically similar. We also include the similarity of such semantically related words in the sentence similarity.

3). Similarity by Co-occurrence: In literature, it is said that words which co-occur repeatedly with similar words are possibly similar. In our approach we also apply this co-occurrence theory by increasing the similarity value of such words by some factor and hence increasing the sentence similarity.

4). Similarity of Grammatical Relation: In literature, it is seen that words which occur repeatedly with the similar Grammatical Relations are possibly similar. In our approach we also apply this Grammatical Relation theory by increasing the similarity value of such words by some factor and hence increasing the sentence similarity.

IV. Motivation and Problem Statement

As discussed above a lot of work has been done on SM and also on BioNLP, but in our research we have not found any work done which accomplishes the task of integrating the conventional situation modeling and the ever advancing work being done in the field of biomedical Text Mining. Situation Modeling has so far been studied on worldly situations consisting of multidimensional facets. While BioNLP has its focus on extracting important entities and identifying relations.

Our motivation behind this work is to assimilate the information provided by these two largely different fields of

Text Engineering and form a clearer, larger and a more descriptive picture of the textual resources available at hand.

The main aim of the project is to represent the textual abstracts in a way that the user can gain maximum knowledge of what the abstracts says while also adding some helpful insights as to add to the information provided by the text. The result should be relevant to the text and should assume that the information provided as input is error free. Text comprehension forms the core motive behind this project.

General objectives of the project are:

- Analysis and preprocessing of Biomedical abstract
- Identify the important entities of text which gives insights into SM.
- Identify relations among tagged entities in the text
- Enhance the extracted relations depending on domain knowledge.
- Proposing an algorithm for building the SM on given text
- Provide visualization to set of extracted relations

V. IMPLEMENTATION

The complete relation extraction pipeline deals with Jsre. The verified relations are one dimension of the user results. The visualization module adds to the insight of PPI. The situation generation part where sentence similarity value is considered and sentence importance is calculated to use the algorithm to generate a situation representation.

The outline of the algorithm is shown below:

I. Find the coherence value (similarity value) between consecutive sentences.
II. Divide the text into chunks of sentences depending on the coherence value and the threshold.
III. From each chunk found above infer the situation using our algorithm

A. Sentence Similarity

We have used the following approach for finding sentence similarity:

I. Take the text as an input.
II. Split the text in sentence (Sentence Splitting).
III. Tokenize the sentences into words (Word Tokenization).
IV. Remove stopwords and punctuation marks from list of words.
V. Find similarity between each pair of words in consecutive sentences.
VI. Find the sentence similarity using word similarity calculated above.

Depending upon the sentence similarity, divide text into chunks of coherent sentences.

B. Algorithm 1(calculating similarity)

```
1: procedure GetCosine(vector1; vector2)
2:   for all keys in vector1 do
3:     for all keys in vector2 do
4:       if keys are equal then
5:         sim = 1:0
6:       else
7:         sim = (Similarity return by WordNet)
8:         if High correspondence in local scope then
9:           sim = sim + ϕ
10:        end if
11:        if both keys are in same grammatical relation with same word then
12:          sim = sim +χ
13:        end if
14:      end if
15:      numerator = numerator + (freq (key of vector1) * freq (key of vector2) * sim)
16:    end for
17:    average the numerator
18:    final_numerator = final_numerator + numerator
19:  end for
20:  denominator = length (vecotr1) * length (vector2)
21:  Sentence Similarity =(final_numerator)/denominator
22: end procedure
```

Where the score given to a word is the frequency of the word.

The other part i.e., the score of the clause is calculated using the aggregated score of the words in all the trigrams containing w.

C. Situation Extraction

This section aims at exploring the approach used for finding important parts from the chunks of text available from above algorithm (Algorithm 1). Here we describe the methods used for finding sentence importance. We then describe an algorithm which uses the importance values of sentences and outputs the comprehension of the text.

We call the score a word gets from its functional role the local importance and the score from the word significance the global importance. We calculate the importance value of each sentence from the local and global importance scores.

I. Calculation of the Local Importance Score:

The local importance score of a word LI (w) is defined as:

LI(w)=(the_score_given_to_w)+(the_score_given_to_the_clause_in_which_w_appears)

Where the score given to a word is the frequency of the word. The other part i.e., the score of the clause is calculated using the aggregated score of the words in all the trigrams containing w.

II. Calculation of the Global Importance Score:

The global importance score of a word GI (w) is calculated by:

$$GI(w) = \sum_{w'} (LI(w') \times Similarity(w, w'))$$

Where the set w' is the set of all the words in the abstract, for which the similarity values of w and the word are greater than a preset threshold. We use WordNet path similarity to know the similarity values.

III. Importance Value of Sentence:

The importance value of a sentence is calculated by the local importance and global importance scores. However, we do not calculate it simply by adding the importance values of words in the sentence. If we do this, a long sentence will get higher importance value as it contains more words in it. To avoid this partiality, we normalize the importance value using the sentence length as a parameter. The importance value of each sentence IV is defined as:

$$IV(s) = \sum_{w\_in\_s} \sqrt{LI(w) \times GI(w)}$$

D. Algorithm 2 (Calculating Situation Extraction)

```
1: procedure GetSituation()
2:    Select the first sentence from the list of sentences
3:      if the sentence selected contains a connective word or phrase then
4:         if it is an elaborating connective then
5:            make the order of importance to be 0
6:            move the sentence to the end of list
7:            go back to 1
8:          else if the sentence just in front of this is already chosen for the situation then
9:            include the sentence selected to the situation
10:    else
11:         lower the order of importance by 1
12:         move the sentence to the 2nd position in the list
13:         go back to 1
14:    end if
15:    include it to the situation
16:    end if
17:      Remove from the list the sentence included into the situation
18:    If the amount of sentences taken does not exceed the situation rate, then go back to 1
19: end procedure
```

Algorithm 2 explores a methodology for finding the important sentences from the chunks formed. But here we do not take the more important sentences as a part of the situation, but we also consider the occurrence of words like and which may speak about the things already spoken about and may be found unnecessary.

The sentence list is sorted in decreasing order of their importance. Sentence with the highest impotence is selected from the list. If the sentence is a simple there is a high possibility of it speaking about a topic not spoke about before (as it is an important member of the chunk). Hence it is included straight away in the situation. If the sentence starts with an elaborating connective, it is sent back to the list decreasing its impotence to 0.If it does not start with an elaborating connective but has a connection to the previous sentence, the fate of the sentence depends on the previous sentence. If even that's not the case, the sentence has lower importance and is moved back. The sentence considered is removed from the list of sentences to be considered. This process is repeated till we achieve the required number of sentences in the situation. Thus by applying this process on every chunk we form situations of incoherent chunks, thereby forming the complete situation.

VI. Results

We experimented on the system with the following features.

- OS : Fedora 17
- Processor: Intel i5 x86-64
- Python 2.7, Java 1.7.0

The experimentation was done on many input abstracts from Aimed corpus and Ohsumed corpus out of which we present here three case studies.

A. Case Study 1 : AIMed corpus

Situation Number	Situation Coherence	Chunk Coherence	Situation Score
1	0.379568582396	0.321171620005	31.9118069663
2	NA	NA	65.6335726517
3	NA	NA	53.5371702936
4	NA	NA	33.6700687841
5	NA	NA	28.1388655105
6	0.23481943429	0.227957057301	32.2300553143
7	0.302757212374	0.284265806245	27.516558884
8	0.13561095375	0.312939247279	25.887904135
9	0.273256358295	0.251228407347	24.7239296225
10	0.265080669139	0.312473434119	28.2660055454
11	NA	NA	36.4425410103
12	NA	NA	26.6514120665
13	0.138126692514	0.207396478997	27.9025122484
14	NA	NA	26.3071613688
15	NA	NA	0.0

The end result observed:

Full Positive	Full Negative	Half Positive	Total
10	1	4	15

Analysis:

1) *Situation Coherence and Average coherence:* Coherence value describes the average similarity value between the consecutive sentences. The situation coherence gives the value for the sentences in the generated situation, while average coherence gives it for all the sentences that are a part of that chunk. The situation should consist of minimum sentences with high similarity. Hence for a good result the average coherence value should be more than the situation coherence.

2) *Situation Importance and Average importance:* The Importance value describes the average importance value for the sentences in that part. The situation importance gives the value for the sentences in the generated situation, while average importance gives it for all the sentences that are a part of that chunk. The situation should consist of sentences with higher importance value. Hence for a good result the average importance

3) *Aggregate Analysis:* The accuracy of the model could be seen by the number of half positive and full positive instances. The full positive instances of situations are the ones which satisfy both the above requirements, half positive are those which satisfy at least one of the requirement, while full negative satisfy none. The table at the end gives the total analysis of the situations.

B. Case Study 2 : Ohsumed corpus

The abstract supplied as input to the system is a merger of 5 abstracts from the Ohsumed medical corpus. As is known the corpus does not contain protein names or interactions. This is reflected in the output. The pass 2 performs well as is suggested by the situation statistics.

Situation Number	Situation Coherence	Chunk Coherence	Situation Score
1	NA	NA	59.3965396059
2	0.247355603488	0.185860025221	57.9242351139
3	NA	NA	103.211870762
4	0.218072257723	0.175190828859	63.1797156834
5	0.204586190409	0.170965508421	46.1691743689
6	0.154076320918	0.238010990708	101.358488543
7	NA	NA	86.0014059911
8	NA	NA	85.2364113006
9	NA	NA	68.5355348771
10	NA	NA	64.1459117028
11	0.117307156798	0.166781650532	55.6880188844
12	0.247355603488	0.185860025221	57.9242351139
13	NA	NA	103.211870762
14	0.218072257723	0.175190828859	63.1797156834
15	0.204586190409	0.170965508421	46.1691743689
16	0.154076320918	0.238010990708	101.358488543
17	NA	NA	86.0014059911
18	NA	NA	85.2364113006
19	NA	NA	68.5355348771
20	NA	NA	64.1459117028
21	NA	NA	51.9794981628
22	NA	NA	39.0939010744
23	NA	NA	44.4192015386
24	0.0660318534627	0.127455495698	44.1269026678
25	NA	NA	39.0939010744
26	0.179401851299	0.163877942094	41.1010190508
27	0.0660318534627	0.0976675360472	44.1269026678

The end result observed:

Full Positive	Full Negative	Half Positive	Total
20	0	7	27

C. Case Study 3 :

The abstract is a merger of both the corpuses. The model still performs well even though there is no specific relation between the abstracts of the two corpora

Situation Number	Situation Coherence	Chunk Coherence	Situation Score
1	0.364215062675	0.321171620005	23.7027158488
2	NA	NA	57.6668385827
3	NA	NA	50.8373846899
4	NA	NA	28.8000557986
5	NA	NA	24.9115234493
6	NA	NA	44.1336842902
7	0.199357306384	0.138787452967	22.5174448136
8	NA	NA	31.3268223529
9	NA	NA	27.6889480371
10	0.238431221307	0.171528085617	28.325755074
11	NA	NA	24.7303790011
12	NA	NA	19.8649867019
13	0.344117085028	0.266451724605	26.2412817958
14	NA	NA	18.7166423222

The end result observed:

Full Positive	Full Negative	Half Positive	Total
9	0	5	14

VII. Conclusion and Future Work

In this paper, we have proposed the system which deals with medical abstract and output a user friendly interpretation of the same. The result is a combination of three sources which we have describe in the previous chapter.

We see that the number of true positives form 70 % of the total chunks. Hence we can say that our approach provide for higher number of cases wherein it delivers maximum of variety (minimum similarity) and higher importance level in the summary (situation).

This work is not limited to biomedical domain but could also be extended for other domain. Also WordNet can be used in a broader sense.

References

[1] Situation Modeling: http://www.comp.lancs.ac.uk/computing/research/cseg/projects/dirc/papers/situation%20modeling.docI. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[2] Kintsch and van Dijk, Model of text comprehension, 1983.

[3] Johnson-Laird, The history of mental models, 1983.

[4] Mannes and W. Kintsch, The construction-integration model, 1987.

[5] Healy and Sinclair, Memory: Handbook of perception and cognition, 1996.

[6] McDaniel, Strategic and automated prospective memory retrieval, 1995.

[7] Anderson, Garrod, and Sanford, Memory and attention in text comprehension : The problem of reference, 1983.

[8] Mandler J M, The foundations of mind: The origins of conceptual thought, 2004.

[9] Mani and Johnson-lard, The effect of mental retardation on mental representation, 1982.

[10] Angelika Storrer, Coherence in text and hypertext, 2002.