# JU-Evora: A Graph Based Cross-Level Semantic Similarity Analysis using Discourse Information

**Swarnendu Ghosh**
Dept. of Computer Science and Engineering
Jadavpur University, Kolkata, India
swarbir@gmail.com

**Nibaran Das**
Dept. of Computer Science and Engineering
Jadavpur University, Kolkata, India
nibaran@ieee.org

**Teresa Gonçalves**
Dept. of Informática
University of Évora, Évora, Portugal
tcg@evora.pt

**Paulo Quaresma**
Dept. of Informática
University of Évora, Évora, Portugal
pq@evora.pt

## Abstract

Text Analytics using semantic information is the latest trend of research due to its potential to represent better the texts content compared with the bag-of-words approaches. On the contrary, representation of semantics through graphs has several advantages over the traditional representation of feature vector. Therefore, error tolerant graph matching techniques can be used for text comparison. Nevertheless, not many methodologies exist in the literature which expresses semantic representations through graphs. The present system is designed to deal with cross level semantic similarity analysis as proposed in the SemEval-2014 : Semantic Evaluation, International Workshop on Semantic Evaluation, Dublin, Ireland.

## 1 Introduction

Text Analytics has been the focus of much research work in the last years. State of the art approaches typically represent documents as vectors (bag-of-words) and use a machine learning algorithm, such as k-NN or SVM, to create a model and to compare and classify new documents. However, and in spite of being able to obtain good results, these approaches fail to represent the semantic content of the documents, losing much information and limiting the tasks that can be implemented over the document representation structures. To overcome these shortcomings some research has been done aiming to use and evaluate more complex knowledge representation structures. In this paper, a new approach which integrates a deep linguistic analysis of the documents with graph-based classification algorithms and metrics has been proposed.

## 2 Overview of the Task

This task provides an evaluation for semantic similarity across different sizes of text, which we refer to as lexical levels. Specifically, this task encompasses four semantic similarity comparisons:

- paragraph to sentence(P2S),
- sentence to phrase(S2Ph),
- phrase to word(Ph2W), and
- word to sense(W2S).

Task participants were provided with pairs of each comparison type and asked to rate the pair according to the semantic similarity of the smaller item to the larger item. As an example, given a sentence and a paragraph, a system would assess how similar is the meaning of the sentence to the meaning of the paragraph. Ideally, a high-similarity sentence would reflect overall meaning of the paragraph. The participants were expected to assign a score between [0,4] to each pairs of sentences, where 0 shows no similarity in concept while 4 shows complete similarity in concept.

## 3 Theoretical Concepts

### 3.1 Discourse Representation Structures

Extraction and representation of the information conveyed by texts can be performed through several approaches, starting from statistical analysis to deep linguistic techniques. In this paper

*Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 375–379,
Dublin, Ireland, August 23-24, 2014.

we will use a deep linguistic processing sequence: lexical, syntactic, and semantic analysis.

One of the most prominent research work on semantic analysis is the Discourse Representation Theory (DRT)(Kamp & Reyle, 1993). In DRT, we aim to associate sentences with expressions in a logical language, which indicate their meaning. In DRT, each sentence is viewed as an update of an existing context, having as result a new context.

DRT provides a very powerful platform for the representation of semantic structures of documents including complex relations like implications, propositions and negations. It is also able to separately analyse almost all kinds of events and find out their agent and patient.

The main component of DRT is the Discourse Representation Structure (DRS These expressions have two main parts: a) a set of referents, which refer to entities present in the context and b) a set of conditions, which are the relations that exist between the entities. An example of a DRS representation for the sentence "He throws a ball." is shown below.

```
[
    x1, x2, x3:
    male(x1),
    ball(x2),
    throw(x3),
    event(x3),
    agent(x3,x1),
    patient(x3, x2)
]
```

### 3.2 GML Structure

Graph Modelling Language (GML)(Himsolt & Passau, 1996) is a simple and efficient way to represent weighted directed graphs. A GML file is basically a 7-bit ASCII file, and, as such, can be easily read, parsed, and written. Several open source applications[1] are available that enable viewing and editing GML files.

Graphs are represented by the keys viz. graph, node and edge. The basic structure is modelled with the node's id and the edge's source and target at-tributes. The id attributes assign numbers to nodes, which are then referenced by source and target. Weights can be represented by the label attribute.

### 3.3 Similarity Metrics for Graphs

It has already been mentioned that the objective of the present work is to generate similarity scores among documents of different lexical levels using an approach which integrates a deep linguistic analysis of the documents with graph-based classification algorithms and metrics. Here, five different distance metrics taken from (Bunke, 2010) are utilized for this purpose. They are popularly used in object recognition task, but for text similarity measure they have not yet been used.

For two graphs $G_1$ and $G_2$, if $d(G_1, G_2)$ is the dissimilarity/similarity measure, then this measure would be a distance if $d$ has the following properties:
1. $d(G_1, G_2) = 0$, iff $G_1 = G_2$
2. $d(G_1, G_2) = g(G_2, G_1)$
3. $d(G_1, G_2) + d(G_2, G_3) \geq d(G_1, G_3)$

The measures used in the present work follow the above rules and the corresponding equations are

$$d_{mcs}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{max(|G_1|, |G_2|)} \qquad \dots(1)$$

$$d_{wgu}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|G_1| + |G_2| - |mcs(G_1, G_2)|} \qquad \dots(2)$$

$$d_{ugu}(G_1, G_2) = |G_1| + |G_2| - 2 * |mcs(G_1, G_2)| \qquad \dots(3)$$

$$d_{MMCS}(G_1, G_2) = |MCS(G_1, G_2)| - |mcs(G_1, G_2)| \qquad \dots(4)$$

$$d_{MMCSN}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|MCS(G_1, G_2)|} \qquad \dots(5)$$

In the equations $mcs(G_1, G_2)$ and $MCS(G_1, G_2)$ denote maximal common subgraph and minimum common super graphs of two graphs $G_1$ and $G_2$. Theoretically $mcs(G_1, G_2)$ is the largest graph in terms of edges that is isomorphic to a subgraph of $G_1$ and $G_2$. The $mcs(G_1, G_2)$ has been formally defined in a work of Horst Bunke (Bunke, Foggia, Guidobaldi, Sansone, & Vento, 2002). As stated earlier, it is a NP complete problem and actually, the method of finding the $mcs()$ is a brute force method which finds all the subgraphs of both the graphs and select the maximum graph which is common to both. To make the program computationally faster, the program is modified to an approxi-

[1] http://en.wikipedia.org/wiki/Graph_Modelling_Language

mate version of $mcs(G_1, G_2)$ on the fact that the vertices which exhibit greater similarity in their local structures among the two graphs have a greater probability of inclusion in the $mcs()$. The two pass approach used in the present work to form the approximate $mcs(G_1, G_2)$ is as follows:

- All the node pairs (one from each graph) are ranked according the number of matching self-loops.
- The *mcs* is built by including each node pair (starting with the one with the highest number of matching self-loops) and considering it as a common node; and then include the rest of the edges (i.e. non-self-loop edges) which occur in the same fashion in both the graphs.

In this way it ensures that the approximation version exhibits most of the properties of a *mcs*, while keeping the complexity in a polynomial time.

The minimum common supergraph $(MCS)$(Angelova & Weikum, 2006) is formed using the union of two graphs, i.e. $MCS(G_1, G_2) = G_1 \cup G_2$.

The distance metrics of Equations 1-3 were used directly without any modifications; the ones of Equations 3-4 were divided by $(|G_1| + |G_2|)$ and $|MCS(G_1, G_2) + mcs(G_1, G_2)|$ respectively to make them normalized, keeping the value of distance metrics within the range $[0, 1]$.

It is worthy to note that label matching that is performed during the above mentioned step may not necessarily be exact matching. Rather in this case we have used the WordNet to find an approximate conceptual similarity between two labels. For our experiment we have used the Wu and Palmer's conceptual similarity (Wu & Palmer, 1994).

$If\ L = lso(c_1, c_2)$ , where $c_1$and $c_2$ are a pair of concepts corresponding to two words and $lso(c_1, c_2)$ means the lowest super ordinate then,

$$sim_{WP}(c_1, c_2) = \frac{2 \times depth(L)}{len(c_1, L) + len(c_2, L) + 2 \times depth(L)}$$

### 3.4 Tools Used

In order to process texts C&C/Boxer (Bos, 2008; Curran, Clark, & Bos, 2007) a well-known open source tool available as a plugin to Natural Language Toolkit (NLTK) is used. The tool consists of a combinatory categorical grammar (CCG) (Curran et al., 2007) parser and outputs the semantic representations using discourse representation structures (DRS) of Discourse Representation Theory (DRT) (Kamp & Reyle, 1993).

## 4 System Description

The method described in the present work, is mainly divided into three major components. The first is the creation of the DRS of the semantic interpretation of the text. The second is the construction of graphs in GML from the obtained DRS using some predefined rules. The third one is the classification phase where the different graph distances are assessed using a k-NN classifier (Zhang, Li, Sun, & Nadee, 2013).

The algorithm semantic evaluation of text content may be described as follows.

- **NLTK Module** : For each pair of text, to

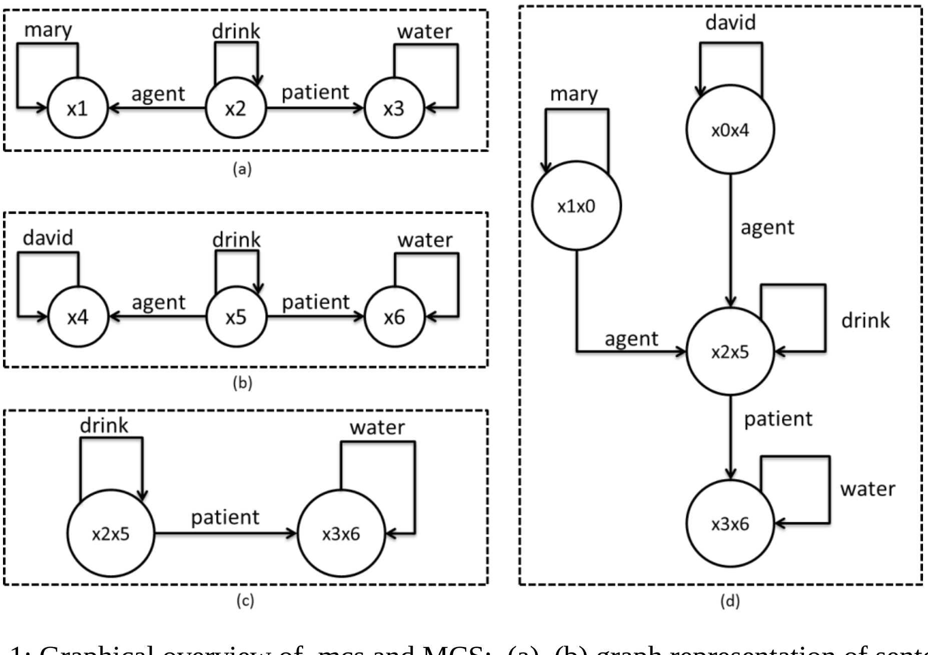

Figure 1: Graphical overview of mcs and MCS: (a), (b) graph representation of sentences meaning "Mary drinks water" and "David drinks water " respectively, (c) maximum common subgraph, (d) minimum common supergraph.

compare their similarity measure we need to find their DRS using the C&C/Boxer toolkit. The toolkit first uses the C&C Parser to find the combinatorial categorical grammar(CCG) of the text. Next the Boxer Module uses the CCG to find the discourse representation structures.

- **Graph building module** : In general Boxer represents a sentence through some discourse referents and conditions based on the semantic interpretation of the sentence. In the graph, the referent is represented by vertex after resolving the equity among different referents of the DRS; and a condition is represented by an edge value between two referents. The condition of a single referent is represented as a self-loop of the referent (source and destination referents are same). Special relationships such as proposition, implication etc. are treated as edge values between two referents; Agent and patient are also treated as conditions of discourse, hence represented by the edge values of two referents.
- **Calculating Similarity Index** : It has already been mentioned that the different distance metrics (see Equations 1-5) calculated based on the mcs() and MCS(). The values of mcs() and MCS() are represented by the number of similar edges. Thus, ten different distances are calculated based on Equations 1-5.
- **Learning** : We obtained 5 similarity scores for each pair of texts. Our task requires us to assign a score between 0-4 for each pair of text. Hence using the gold standard a K-NN Classifier have been trained to find the output score for a test sample. The value of K has been empirically adjusted using the cross validation technique to find the optimal value.

Our method works smoothly for the first two lexical levels. But for the last two levels i.e. phrase to word and word to sense it is not possible to find out DRS for a single word. Hence we have used the WordNet(Fellbaum, 1998) to extract the definition of the word in question and calculate its DRS and proceed with the method. When a word has multiple definitions, all the definitions are fused to a single sentence after conjugating them with the conjunction ‘or’.

## 5 Results and Discussions

The JU-Evora system performed fairly in the SemEval Competition 2014. All the correlation scores are not as good as the Baseline(LCS) scores, however it provides a better Pearson correlation score in case of Paragraph to Sentence. The other scores, though not higher, are in the vicinity of the baseline. All the scores are shown below in Table 1.

| **PEARSON’S CORRELATION** | | | | |
|---|---|---|---|---|
| | **P2S** | **S2Ph** | **Ph2W** | **W2S** |
| **JU-Evora** | 0.536 | 0.442 | 0.090 | 0.091 |
| **Baseline (LCS)** | 0.527 | 0.562 | 0.165 | 0.109 |
| **SPEARMAN CORRELATION** | | | | |
| **JU-Evora** | 0.533 | 0.440 | 0.096 | 0.075 |
| **Baseline (LCS)** | 0.613 | 0.626 | 0.162 | 0.130 |

Table 1: Performance of JU-Evora system with respect to Baseline.

## 6 Conclusion

In this paper a new approach has been proposed to the text comparison task which integrates a deep linguistic analysis of the documents with a graph-based comparison algorithm. In the linguistic analysis, discourse representation structures (DRS) are used to represent text semantic content and, afterwards, these structures are transformed into graphs. We have evaluated existent graph distance metrics and proposed some modifications, more adequate to calculate graph distances between graph-drs structures. Finally, we integrated graph-drs structures and the proposed graph distance metrics into a k-NN classifier for calculating the similarity between two documents. Future works in this area would be concentrated on the use of external knowledge sources to make the system more robust.

## References

Angelova, Ralitsa, & Weikum,Gerhard. (2006). Graph-based Text Classification: Learn from Your Neighbors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 485–492). New York, NY, USA: ACM.

Bos, Johan (2008). Wide-Coverage Semantic Analysis with Boxer. In J. Bos & R. Delmonte

(Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings* (pp. 277–286). College Publications.

Bunke, Horst (2010). *Graph Classification and Clustering Based on Vector Space Embedding* (Vol. Volume 77, pp. 15–34). WORLD SCIENTIFIC. doi:doi:10.1142/9789814304726_0002

Bunke, Horst, Foggia, Pasquale, Guidobaldi, Corrado, Sansone, Carlo, & Vento, Mario (2002). A Comparison of Algorithms for Maximum Common Subgraph on Randomly Connected Graphs. In T. Caelli, A. Amin, R. W. Duin, D. Ridder, & M. Kamel (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition SE - 12* (Vol. 2396, pp. 123–132). Springer Berlin Heidelberg.

Curran, James, Clark, Stephen, & Bos, Johan (2007). Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 33–36). Prague, Czech Republic: Association for Computational Linguistics.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. *British Journal Of Hospital Medicine London England 2005* (Vol. 71, p. 423).

Himsolt, Michael, & Passau, Universität (1996). GML : A portable Graph File Format. *Syntax*, 1–11.

Kamp, Hans, & Reyle, Uwe (1993). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*.

Wu, Zhibiao, & Palmer, Martha (1994). Verb semantics and lexical selection. *In 32nd Annual Meeting of the Association for Computational Linguistics,, 32*, 133–138.

Zhang, Libiao, Li, Yuefeng, Sun, Chao, & Nadee, Winai (2013). Rough Set Based Approach to Text Classification. *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*.