2019

**Similarity Evaluation Based on Contextual Modelling**

Mohamed H. Haggag  
*Helwan University*, mohamed.haggag@fci.helwan.edu.eg

Marwa M. A. ELFattah  
marwa_8_80@yahoo.com

Ahmed Mohammed Ahmed  
am35@fayoum.edu.eg

Follow this and additional works at: [https://digitalcommons.aaru.edu.jo/fcij](https://digitalcommons.aaru.edu.jo/fcij)

Part of the Computer Engineering Commons

**Special issue**


**Recommended Citation**

Available at: [https://digitalcommons.aaru.edu.jo/fcij/vol4/iss1/5](https://digitalcommons.aaru.edu.jo/fcij/vol4/iss1/5)

This Article is brought to you for free and open access by Arab Journals Platform. It has been accepted for inclusion in Future Computing and Informatics Journal by an authorized editor. The journal is hosted on Digital Commons, an Elsevier platform. For more information, please contact rakan@aaru.edu.jo, marah@aaru.edu.jo, u.murad@aaru.edu.jo.
Similarity Evaluation Based on Contextual Modelling

Mohamed H. Haggag a*, Marwa M. A. ELFattah a*, and Ahmed Mohammed Ahmed b

aFaculty of computers and information systems, Helwan University, Cairo, Egypt
bMinistry of Communication and information Technology, Telecom Egypt Company, Giza, Egypt

* mohamed.haggag@fci.helwan.edu.eg

Abstract

Measuring text similarity problem still one of opened fields for research area in natural language processing and text related research such as text mining, Web page retrieval, information retrieval and textual entailment. Several measures have been developed for measuring similarity between two texts: such as Wu and Palmer, Leacock and Chodorow measure and others. But these measures do not take into consideration the contextual information of the text. This paper introduces new model for measuring semantic similarity between two text segments. This model is based on building new contextual structure for extracting semantic similarity. This approach can contribute in solving many NLP problems such as text entailment and information retrieval fields.

Keywords: Text Similarity, Word Net, Semantic Similarity Measures.

1. Introduction

Text semantic similarity measures play important role in text related research and applications in tasks such as
• Information retrieval,
• Text Classification,
• Document Clustering,
• Topic Detection
• Question answering.

Semantic similarity between concepts is a method to measure the semantic similarity, or the semantic distance between two concepts (texts) according to a given ontology.

Measurements of semantic similarity between a pair of sentences provide fundamental function in natural language understanding, machine translation, information retrieval and voice based automation tasks, among many other applications. In machine translation, for example, one would like to quantitatively measure the quality of the translation output by measuring the effect that translation had in the conveyed message.

Current approaches to semantic similarity measurement include techniques that are specific or custom to the task at hand. For example, in machine translation, the BLEU metric [1] is used in measuring similarity of the MT output. In call routing, vector based methods (e.g., [2, 3]) are used to compare the input utterance against a set of template categories.

Semantic similarity and semantic relatedness are two related words, but semantic similarity is more specific than relatedness and can be considered as a type of semantic relatedness. For example ‘Student’ and ‘Professor’ are the related terms, which are not similar. All the similar concepts are related and the vice versa is not always true.

Semantic similarity and semantic distance are defined conversely. Let be C1 and C2 two concepts that belong to two different nodes n1 and n2 in a given ontology, the distance between the nodes (n1 and n2) determines the similarity between these two concepts C1 and C2. Both n1 and n2 can be considered as an ontology (also called concept nodes) that contains a set of terms synonymous and consequently. Two terms are synonymous if they are in the same node and their semantic similarity is maximized [4].

The use of ontologies to represent the concepts or terms (humans or computers) characterizing different communicating sources are useful to make knowledge commonly understandable. Additionally, it is possible to use different ontologies to represent the concepts of each knowledge source.

2. Background

Textual semantic similarity measures are varied to reach to best results in text similarity research. Several methods of determining semantic measures have been proposed according to its methodology for measuring semantic similarity.
3. Related work

**Taxonomy based approaches (Structure-based measures):**
It is based on edge counting in taxonomy like WorldNet or SENSUS or Ontology

![Diagram of taxonomy](image)

Wu and Palmer[5]:- simple, and gives good performance, its disadvantage that it does not consider how far the concepts are semantically. The semantic similarity can be formulated as the next equation

$$Sim_{wup}(C1, C2) = \frac{2*N}{N1*N2+2*N}$$

**Contextual based measures:** these approaches are based on Dependency-based contextual similarity defines the context for the pair (w(1) i , w(2) j ) using the syntactic dependencies of w(1) i and w(2) j . The two dependencies are either identical or Semantically “equivalent” according to the equivalence table provided by Sultan et al[6]

**Domain Specific Ontologies based Similarity measures:**-his category determines the similarity between sentences according to information gained from large corpora in specific domain .

A Corpus is a large collection of written or spoken texts that is used for language research. It is tagged by humans [8]

**Hybrid Similarity Measures:**- Hybrid methods use multiple similarity measures. Many researches trend to this area to achieve better results

4. Proposed model

The proposed model can be categorized as hybrid similarity model, as it combines taxonomy based approach with contextual based approach

![Diagram of proposed model](image)

- **Lexical- Syntactic analysis:**- also referred to lexical-syntactic parsing. It has two processes:
  1. Lexical-parsing: dividing the input sequence of tokens in order to produce its grammatical structure.
  2. Syntactic parsing: syntactic parsing might be divided into shallow parsing and fully syntactic parsing.

  I. **Shallow parsing** is the analysis process of the sentence which identifies the Constituents, or linguistic phrases, but does not specify their internal structure, or their role in the sentence, i.e. producing non-hierarchical syntactic structure.

  II. **Fully syntactic parsing** is building a hierarchical syntactic structure from lexical items to the whole sentence.

Lexical- Syntactic analysis uses link parser[11] to generate output as the following figure

![Dependency tree](image)

Dependency tree is generated after link parser finished as the following:-

https://digitalcommons.aaru.edu.jo/fcij/vol4/iss1/5
[nsubj(wrote-2, he-1), root(ROOT-0, wrote-2), det(book-4, a-3), dobj(wrote-2, book-4)]

\[ S^{(1)}: \text{He wrote a book} . \]

**Holder-Relation-Target structure**

The proposed structure is composed of four components; holder, target, relation and complements. Each component is a sentence entity having a role and described with a set of attributes. The new semantic role labeling structure for Sentence, this structure is constructed based on Link Parser system and Word Sense Disambiguation technique (WordNet). This model uses link Parser to parse the sentence and return all the sentence components (NOUN, Verb, Propositions...) and links in the sentence. The link parser generates two kinds of syntactic parsers are considered to parse a sentence conforming two formalisms of grammar: context-free syntactic parsers and dependency parser. Correspondingly, there are two kinds of syntactic parsing representations: context-free grammar parsed trees and dependency grammar parsed trees.

The basic extracted components are (holder- relation- target- other objects)

- Holder: - holder is Event Initiator and It similar to the subject of the sentence.
- Relation: - it is the object that links the holder with the target or the action which happened to reach to the target.
- Target: - it is Event Recipient or it receives the action of the holder .the Target can be Word or Sentence.
- Complement: - it is object is a complement for the sentence such as Adjective or adverb.

**Comparison Matrix Builder**

The semantic similarity between two elements is calculated and fills the comparison matrix

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Holder Text</th>
<th>Relation Text</th>
<th>Target Text</th>
<th>Complement Objects Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holder hypothesis</td>
<td>X11</td>
<td>X12</td>
<td>X13</td>
<td>X14</td>
</tr>
<tr>
<td>Relation hypothesis</td>
<td>X21</td>
<td>X22</td>
<td>X23</td>
<td>X24</td>
</tr>
<tr>
<td>Target hypothesis</td>
<td>X31</td>
<td>X32</td>
<td>X33</td>
<td>X34</td>
</tr>
<tr>
<td>Complement Objects hypothesis</td>
<td>X41</td>
<td>X42</td>
<td>X43</td>
<td>X44</td>
</tr>
</tbody>
</table>

**Calculate Final Semantic Similarity**

The last step is calculating final semantic similarity value. This step will compute the final semantic similarity value depending on the priority matrix

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Holder Text</th>
<th>Relation Text</th>
<th>Target Text</th>
<th>Complement Objects Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holder hypothesis</td>
<td>P11</td>
<td>P12</td>
<td>P13</td>
<td>P14</td>
</tr>
<tr>
<td>Relation hypothesis</td>
<td>P21</td>
<td>P22</td>
<td>P23</td>
<td>P24</td>
</tr>
<tr>
<td>Target hypothesis</td>
<td>P31</td>
<td>P23</td>
<td>P33</td>
<td>P34</td>
</tr>
<tr>
<td>Complement Objects hypothesis</td>
<td>P41</td>
<td>P24</td>
<td>P43</td>
<td>P44</td>
</tr>
</tbody>
</table>

This matrix shows how the relationship between each element from the text with each element from hypothesis will impact in the final semantic similarity value.
5. Research Analysis and Discussion

By applying the proposed approach on the next example

**Text:** The largest gains were seen in prices, new orders, inventories and exports.

**Hypothesis:** Sub-indexes measuring prices, new orders, inventories and exports increased.

1- The output of the Lexical-Syntactic analysis for the text will be:

- The dependency tree:
  
  det(gains-3, The-1), amod(gains-3, largest-2), nsubjpass(seen-5, gains-3), auxpass(seen-5, were-4), root(ROOT-0, seen-5), case(prices-7, in-6), nmod:in(seen-5, prices-7), amod(orders-10, new-9), nmod:in(seen-5, orders-10), conjunction(prices-7, orders-10), nmod:in(seen-5, inventories-12), conjunction(prices-7, inventories-12), conjunction(prices-7, and-13), nmod:in(seen-5, exports-14), conjunction(prices-7, exports-14)

- Tyntax parser tree:

  ![Syntax Tree Image]

  (ROOT (S (NP (DT The) (JJS largest) (NNS gains)) (VP (VBD were) (VP (VBN seen) (PP (IN in) (NP (NP (NNS prices)) (,) (NP (JJ new) (NNS orders)) (,) (NP (NNS inventories)) (CC and) (NP (NNS exports))))) ())).

The output of Holder-Relation-Target is

<table>
<thead>
<tr>
<th>Element Name</th>
<th>Holder Text</th>
<th>Relation Text</th>
<th>Target Text</th>
<th>Complement Objects Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holder hypothesis</td>
<td>0</td>
<td>.3</td>
<td>.6</td>
<td>1</td>
</tr>
<tr>
<td>Relation hypothesis</td>
<td>0</td>
<td>.2</td>
<td>.6</td>
<td>.16</td>
</tr>
<tr>
<td>Target hypothesis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Complement Objects</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculating the final value will be .572

6. Results Evaluation

**Sentences Corpus Dataset Size: Microsoft Research Paraphrase Corpus**

The Proposed Model results Evaluation

<table>
<thead>
<tr>
<th>Data set</th>
<th>True positive</th>
<th>False positive</th>
<th>True negative</th>
<th>False negative</th>
<th>accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1650</td>
<td>921</td>
<td>177</td>
<td>402</td>
<td>150</td>
<td>80.1%</td>
<td>85.9%</td>
<td>83.8%</td>
</tr>
</tbody>
</table>

Calculate Final Semantic Similarity. To calculate final result we should multiply the comparison matrix by the priority matrix which is

<table>
<thead>
<tr>
<th>Measure</th>
<th>Data Sources</th>
<th>Semantics Using syntactic analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shortest Path</td>
<td>Ontology</td>
<td>Distance</td>
</tr>
<tr>
<td>Wu and Palmer</td>
<td>Ontology</td>
<td>Similarity</td>
</tr>
<tr>
<td>Leacock and Chodorow</td>
<td>Ontology</td>
<td>Similarity</td>
</tr>
<tr>
<td>Proposed model</td>
<td>Ontology</td>
<td>Distance +Similarity</td>
</tr>
</tbody>
</table>

50
REFERENCES


7. Conclusions

Semantic similarity evaluation is a good factor included in many applications enclosed in the artificial intelligence research area. Based on the theoretical principles and the way in which ontologies are investigated to compute similarity, different kinds of methods can be identified. The proposed model produced improved results in measuring textual semantic similarity compared to other models. It introduces contextual approach with taxonomy based semantic similarity method for measuring textual semantic similarity. The proposed model uses contextual structure to store syntactic information and semantic information of the input text.