Future Computing and Informatics Journal

Volume 4
Issue 1 (2020) *Volume 4 issue 1*

Article 5

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
ama35@fayoum.edu.eg

Special issue

Proceeding of 3^{rd}. International Conference for Computing and Informatics, *Ubiquitous Computing and Engineering* (ICCI, 2018) July 3-4 2018, organized by Future University in Egypt & Helwan University

Digital Commons
ELSEVIER
FCIJ FUTURE COMPUTING AND INFORMATICS JOURNAL

Similarity Evaluation Based on Contextual Modelling

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

[a] *Faculty of computers and information systems, Helwan University, Cairo, Egypt*
[b] *Ministry 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 sentences1 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

.

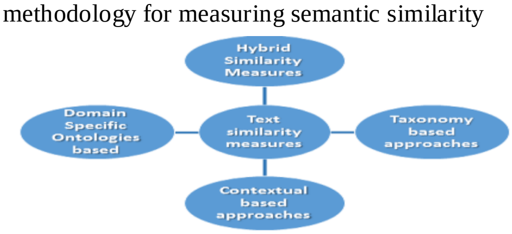
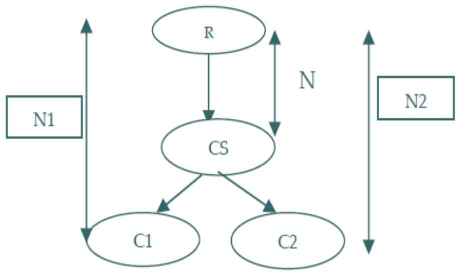


Figure 1: Semantic Measures Categories

3. Related work

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

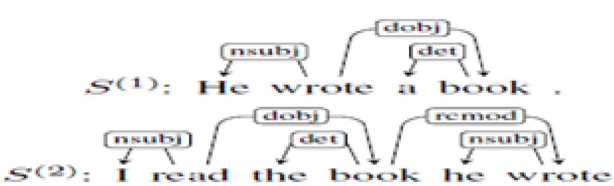


Wu and Palmer[5]:- simple, and gives good performance, its disadvantage that it does not consider how far the concepts are semantically. The semantic similarity can be formulated as the next equation

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

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



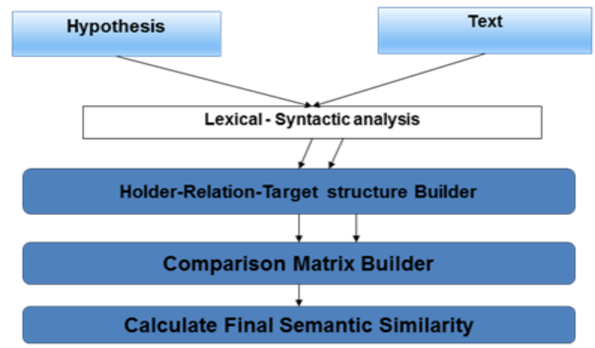
according to information gained from large corpora in specific domain .

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



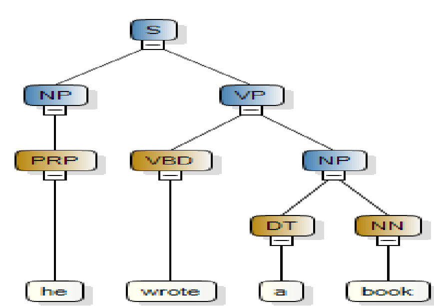
- **Lexical- Syntactic analysis**:- also referred to lexical-syntactic parsing. It has two processes:

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.

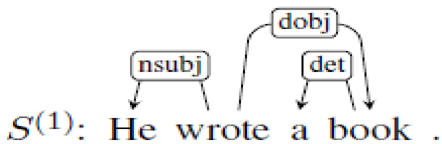
II. **Fully syntactic parsing** is building a hierarchical syntactic structure from lexical items to the whole sentence.

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



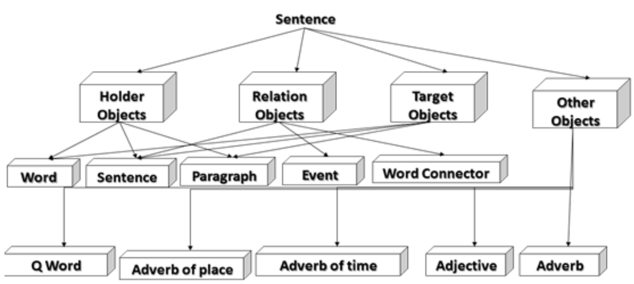
Dependency tree is generated after link parser finished as the following:-

[nsubj(wrote-2, he-1), root(ROOT-0, wrote-2), det(book-4, a-3), dobj(wrote-2, book-4)]



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 (Word



Net). 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 from text and hypothesis structure is calculated. The semantic similarity between two elements is equal to the average of summation of three values which are

1. Shortest Path algorithm

$$sim(C1,C2) = 2 * Max(C1,C2) - sp$$

2. Wu and Palmer algorithm

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

3. Leacock and Chodorow algorithm

$$Sim_{LC}(C1,C2) = -\log\left(\frac{length}{2.D}\right)$$

Where length is the length of the shortest path between the two concepts (using node-counting) and D is the maximum depth of the taxonomy.

Comparison Matrix Builder

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

Element Name	*Holder Text*	*Relation Text*	*Target Text*	*Complement Objects Text*
Holder hypothesis	X11	X12	X13	X14
Relation hypothesis	X21	X22	X23	X24
Target hypothesis	X31	X32	X33	X34
Complement Objects hypothesis	X41	X42	X43	X44

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

Element Name	*Holder Text*	*Relation Text*	*Target Text*	*Complement Objects Text*
Holder hypothesis	P11	P12	P13	P14
Relation hypothesis	P21	P22	P23	P24
Target hypothesis	P31	P23	P33	P34
Compleme nt Objects hypothesis	P41	P24	P43	P44

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), conj:and(prices-7, orders-10), nmod:in(seen-5, inventories-12), conj:and(prices-7, inventories-12), cc(prices-7, and-13), nmod:in(seen-5, exports-14), conj:and(prices-7, exports-14)]

Tyntax parser tree : (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

No Holder
-------------------- **Relation Data**----------------------
relation text ----> [seen]
relation type ----> [Event]
linguistic Description ----> [Verb]
relation Tag ----> [VBN]
-------------------- **Target Data**-------------------------
Target text ----> [gains]
Target type ----> [Word]
linguistic Description ----> [NOUN]
Target Tag ----> [NNS]
--------------------- **Complements**---------------------
text --> [prices, new orders, inventories , exports]

----------------------**Holder** ----------------------
text --> [prices, new orders, inventories , exports]
Holder type ----> [MANY Words]
linguistic Description ----> [NOUN]
Holder Tag ----> [NNS]
---------- ----------**Relation Data**-----------------
relation text ----> [increased]
relation type ----> [Event]
linguistic Description ----> [Verb]
relation Tag ----> [VBD]
-------------------- **Target Data**-----------------
No Target
------------------- **Complements**------------------
No other object

3- Comparison Matrix Builder. The result of this step for the pervious example is

Element Name	*Holder Text*	*Relation Text*	*Target Text*	*Complement Objects Text*
Holder hypothesis	0	.3	.6	1
Relation hypothesis	0	.2	.6	.16
Target hypothesis	0	0	0	0
Complement Objects hypothesis	0	0	0	0

Calculating the final value will be .572

6. Results Evaluation

Sentences Corpus Dataset Size: Microsoft Research Paraphrase Corpus

The Proposed Model results Evaluation

Data set	**True positive**	**False positive**	**True negative**	**False negative**	**accuracy**	**Recall**	**Precision**
1650	921	177	402	150	80.1%	85.9%	83.8%

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

Measure	**Data Sources**	**Semantics**	**Using syntactic analysis**
Shortest Path	Ontology	Distance	No
Wu and Palmer	Ontology	Similarity	No
Leacock and Chodorow	Ontology	Similarity	No
Proposed model	Ontology	Distance +Similarity	yes

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.

REFERENCES

[1] E. W. Frank, and M. a. Hall. "Data Mining: Practical Machine Learning Tools and Techniques", 3rd Edition Textbook, 54(2), 2011.

[2] V. Ilamathi, "Preprocessing Techniques for Text Mining - an Overview.", International Journal on Computer Science and Communication Networks, 5(1):7-16, 2015.

[3] G. Sahaj. "Stock Market Prediction Using Data Mining", in 2017 International Journal on Intelligent Sustainable Systems 2(2):2780–2784, 2017.

[4] M. F. Patrick Uhr, and J. Zenkert. " Sentiment analysis in Financial Markets", IEEE International Conference on System Management, 912–917, 2014.

[5] A. Søgaard, "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies, 5(1):1-167, 2012.

[6] K. S. Loke. "Impact of Financial Ratios and Technical analysis on Stock Price Prediction Using Random Forests", 2017 IEEE International Conference on Computer and Drone Applications , 8–42, 2017.

[7] L. I. Bing, C.Chan and C. Ou. "Public Sentiment analysis in Twitter Data for Prediction of A Company`s Stock Price Movements", IEEE 11th International Conference on E-business, Engineering, 232-239, 2014.

[8] A. Assaf, and E. Alnagi" Predicting Stock Prices Using Data Mining Techniques.", International Arab Conference for. Information Technology, 1–8, 2013.

[9] S. Kannan, P. S. Sekar, "Financial Stock Market Forecast using Data Mining Techniques", International Multi Conference Engineering. Computer Science, 4-8, 2010.

[10] M. K. =Alkhatib, H. Najadat, and I. Hmeidi,"Stock Price Prediction Using K -Nearest Neighbor (k NN) Algorithm", International Journal of Business And Humanities Technology., 3(3):32–44, 2013.

[11] S. M. Price, J. Shriwas, and S. Farzana. "Using Text Mining and Rule Based Technique for Prediction of Stock Market Price", International Journal Emerging. Technology and Advanced Engineering, 4(1):246-250, 2014,

[12] V. S. Pagolu, K. N. R. Challa, G. Panda, and B. Majhi., "Sentiment analysis of Twitter Data for Predicting Stock Market Movements", International conference on Signal Processing, Communication, Power and Embedded System, 1345-1350, 2016

[13] N. Sharma and A. Juneja. "Combining of random forest estimates using LSboost for Stock Market Index Prediction" , 2017 2nd International Confernece on Convergence Technology, 1199-1202, 2017

[14] M. N. Elagamy, C. Stanier, and B. Sharp, "Stock Market Random Forest-Text Mining System Mining Critical Indicators of Stock Market Movements", International. Conference Natural Language and Speech Processing , 1-8, 2018.

[15] A. E. Khedr, S. E. Salama, and N. Yaseen, " Predicting Stock Market Behavior using Data Mining Technique and News Sentiment analysis", International Journal of Intelligent Systems and Applications, 9(7):22-30, 2017.

[16] M. Granik and V. Mesyura, "Fake News Detection Using Naive Bayes Classifier", First International Conference on Electrical and Computer Engineering , Ukrain , 900–903, 2017.

[17] S. Kogan, T. J. Moskowitz, and M. Niessner., "Fake News in Financial Markets", 2017.

[18] R. Desai. "Stock Market Prediction Using Data Mining", 2017 International Conference on Intelligent Sustainable Systems 2780–2784, 2017.

[19] D. Lyon and B. Cedex., "N-grams based feature selection and Text Representation for Chinese Text Classification.", International Journal on Computer Intelligent Systems, 2(4):365–374, 2009.

[20] Y. Shynkevichl, T. M. Mcginnityl, S. Colemanl, and A. Belatrechel., "Stock Price Prediction based on Stock-Specific and Sub-Industry-Specific News Articles", 2015 International Joint Conference on Neural networks, 1-8, 2015.

[21] United States and European Commission, http://ec.europa.eu/trade/policy/countries-and-regions/countries/united-states/, 2014.

[22] B. D. Trisedya, and Y. E. Cakra. "Stock Price Prediction using Linear Regression based on Sentiment analysis", International Conference on Advancement of Computer Science and Information Systems, 147–154, 2015