International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Comparison of Free Text Semantic Similarity Measures Using Dependency Relations

Richa Dhagat[1], Arpana Rawal[2], Sunita Soni[2]

[1]Mtech Scholar, Dept. of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India
[2]Professor, Dept. of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India

---

Abstract - *Very recently, NLP research has observed much inclined shift from syntactic based techniques to dependency structure-based techniques of exploring NLP semantics. A precise set of text semantic similarities based on dependency relation structures have been explored that achieve effective retrieval of semantic closeness of text pairs and justify the significant role of dependency structures (grammatical relations).*

Key Words: Natural Language Processing, Text Similarity, Semantic Similarity Metric, Dependency Relations, Text-based Corpora

1.INTRODUCTION

Natural language processing (NLP) is a field of artificial intelligence in which real-time application driven tools and software's analyze, understand, and derive meaning from human languages (whether English or regional) in a machine-understandable and machine-processing formats. Advances in NLP, preferably combining the field of computational linguistics, along with a large number of text-based resources nowadays publicly available in digital formats (e.g., online encyclopedias) allow newer NLP prototype or commercial tools to become more reliable and robust in accomplishing many natural language applications such as semantic search, summarization, question answering, document classification, and sentiment analysis, plagiarism detection tasks. All these are possible only if syntax-semantic based intelligent techniques are employed to capture the meaning, concept and idea revealed within the text documents.

Further, the recent research carried out in NLP using sentential similarity approach of text mining have invaded information retrieval effectiveness in application realms like web page retrieval, image retrieval from web, web page ranking and document summarization from the topic-relevant web-pages at hand. In this paper, a comparative approach is adopted to investigate the effectiveness of state-of-the-art free text semantic similarity measures devised and used by researchers who worked upon semantic similarity metrics in accomplishing their own undertaken NLP tasks.

2.LITERATURE SURVEY

In text-mining research, sentences can be detected either as lexically or semantically similar. This is evident from abundant of research carried out by text miners from the first two decades of 21st century when numerous lexical similarity measures using document level, paragraph level and sentence level syntactic structures and also lexical variations like synonyms, antonyms and hypernyms. It was Li et. al (2006) who devised semantic based text similarity measures based on syntactic structures, semantic ontology and corpus statistics and Lee (2011) crafted the metrics based on noun and verb vector semantic spaces [1,2].

The first approach involves the use of corpus-based measures that finds the word-to-word semantic similarity based on information exclusively derived from large corpora. In this direction, the two popular preliminary metrics was devised namely, pointwise mutual information (Turney 2001), and latent semantic analysis (Landauer, Foltz, &Laham 1998) [3,4]. In the second knowledge-based approach, the similarity metrics depend on handcrafted semantic net for arriving at word-to-word similarity computations, of which WordNet is the most widely used semantic net (Ontology) by NLP researchers [5]. The third is the structured based approach where usually, the sentential structures are exploited to arrive at contextual relationships among various word groupings within a sentence or in surrounding sentences. In this category, various popularly used metrics are 'path', 'lch', 'wup', 'res', 'lin', 'jcn', 'lesk' and 'hso' metrics. Six metric measures of semantic closeness have been explored by the work group till date among which three of the metrics based on information content are 'res' Resnik (1995),'lin' Lin(1999) and 'jcn'Jiang and Conrath(1997), while the rest of the three measures depend on path length: 'lch'Leacock and Chodorow(1998),'wup' Wu and Palmer(1994) and Path Length(path)[6,7,8,9,10] .Metrics based on measures of relatedness are 'hso' Hirst and St-Onge(1998),'lesk' Banerjee and Pedersen(2003),'vector' Patwardhan(2006)[11,12,13].

Recently, Vakare has devised a novel metric to compute sentence similarity using Dependency Parsing [14]. Here, the documents were represented as sentences which were converted to dependency tree structures. The semantic similarity measure was the result of adding two similarity components, one owing to matching of head and tail arguments denoted by wordsim(d^i_A,d^j_B) and wordsim(h^i_A,h^j_B) and the other owing to matching of dependency relation names (titled as tags by Vakare) and denoted by tagsim(t^i_A,t^j_B)[14].It may be noted that similarity between

corresponding head and dependent nodes is obtained using path, lch and wup similarity metric expressions. Their work attempts to learn grammatical tag relations by training the similarity scores on pre-defined datasets. This leaves a query as to what will be the learnt weights, if at all, the background corpus gets changed and indicates an element of uncertainty aroused due to domain-dependency nature of the undertaken problem objective.

Ozates (2016) used dependency grammatical structures in bigram formats in order to compute sentence semantic similarity15. Each bigram structure comprised a dependent word, a head word, and a typed dependency tag expressing the type of relationship between them. The consequent semantic similarity measure was expressed in terms of Simple Approximate Bigram Kernel (SABK) and its variants.

After undertaking an exhaustive survey of the above mentioned metrics, it was found that when NLP researchers used typed dependency grammar for text representations of document sentences, the results obtained from sentence similarity computations outperformed those similarity computations that did not use typed dependency structures.

3. METHODOLOGY

Any successful NLP task is initiated with machine-readable document representation task. In order to extract topical knowledge, concept and crux of any text document, NLP researchers have been performing broadly two types of text processing tasks, namely, shallow NLP and deep NLP tasks. Unlike, shallow NLP tasks explored for implementing automatic summarization, machine translation and named-entity recognition applications, extracting dependency relations is an inevitable step involved in Deep NLP tasks .Some of the shallow NLP tasks are: Part-of-speech tagging ,Noun phrase chunking, Syntax parsing(Dependency parsing and phrase-constituency parsing. While, tasks like semantic role labeling, Spatial role labeling, semantic dependency parsing, pragmatic word sense disambiguation, Named-Entity recognition, entity linking, co-reference resolution, Information extraction, Discourse parsing, Topic modeling, Sentence similarity, sentiment analysis, speech recognition, and topic segmentation comprise deep NLP tasks.

Manning (2008) pioneered the concept of dependency relation structures that came to be popularly known as Stanford typed dependencies [17]. This robust document representation format was designed to provide linguistically defined grammatical relationships between all possible pairs of sentences from corresponding candidate documents. In particular, rather than the phrase structure representations that have long dominated in the computational linguistic community, these structures represent all sentence relationships uniformly as typed dependency relations. The dependencies are all binary relations: a grammatical relation holds between a governor (also known as a regent or a head) and a dependent (also known as tail). As a result, many other Universities constructed variant versions of dependency parsers based on variety of POS tag definitions as per variety of linguistic corpora resources.

Broadly, the comparison criteria for evaluating the effectiveness of Text Semantic similarity measures have been either the use of reliable linguistic corpora having complete synsets or it can be the repositioning of words in the sentences. The methodology adopted in the current piece of work merely computes the text semantic similarity measures devised by competent researchers namely Miranda (2013) and Vakare (2019) [16,14].

4. EXPERIMENTS AND RESULTS

Having known the wide variety of semantic similarity metrics used, the current experimental setting simply attempts to compare the effectiveness of semantic similarities as articulated in some of the mentioned competent literary works. It is also subsumed that the documents have been already pre-processed and decomposed to sentential units. Miranda (2013) introduced a similarity metric based on the concept of dependency relation extraction [16]. Each sentence pair (A, B) were expressed as a set of dependency relations. These relations were compared in all combinations to check for dependency overlaps between the respective pairs of 1-gram word phrases using an expression of overlap coefficient shown in Equation 1.

$$Sim_{Dependency}(S_A, S_B) = \frac{|S(A,n) \cap S(B,n)|}{\min(|S(A,n)|, |S(B,n)|)} \quad (1)$$

Here S(A, n) and S(B, n) be the unique dependency relations contained in the sentence pairs respectively. The number of overlapping relations is normalized by the smaller set of S(A, n) or S(B, n).

In the experimental setup, the second comparable semantic similarity metric used was predicate extraction metric that exploits the matching of unique verb forms (without using the VerbNet generalization step) [16].

$$DTS(A,B) = \sigma\left(\sum_{i=1}^{m}\sum_{j=1}^{n} sim(A_b^i, B_b^j)\right) \quad (2)$$

These metrics introduced by Miranda was compared with Vakare's Dependency Tree Similarity (DTS) metric given in Equation 2. In nutshell comparative evaluation of the above semantic sentential similarity metrics have been assimilated in Table-1. All these three metrics were validated for computational effectiveness using human assessed scores. Although the sentence pairs under taken for metric comparisons are extracted from datasets, few sentence pairs were manually crafted to give a wide spectrum of dataset combinations [18].

5. CONCLUSION

The results obtained from conventional text similarity metrics devised by Miranda (2013) were compared with Vakare's dependency tree similarity scores computed on randomly selected sentences from SemEval Semantic Similarity Task 1. The difference in semantic similarity scores

is due to the fact that the semantic similarity expression used by Vakare et. al. (2019) considered the dependency tag relationship say, for instance, prep-pobj, nsubj-dobj, det-compound and amod-acl relations. Moreover, it was observed that Miranda did not consider the synset extraction from WordNet dictionaries to arrive at his dependency relation similarity scores as well as predicate generalization scores. This was the major reason of greater offsets between Miranda's scores and human assessed scores (gold standard responses considered in our research).

Table-1 : Comparative Evaluation Metrics of Semantic Similarities for Sentential Datasets

A	B	Gold Standard	$Sim_{Dependency}(A,B)$	$Sim_{Predicate}(A,B)$	DTS(A,B)
A woman supervisor is instructing the male workers.	A woman is working as a nurse.	.2	0	0	.65
A bike is next to a women.	A child is next to a bike.	.4	0	0	.85
There are dogs in the forest.	The dogs are alone in the forest.	.8	.4	0	.9
It is a container of juice.	It is a glass of cider.	.9	0	0	.99
Salt is crucial in cooking.	Salt is necessary while preparing food	.9	0	0	.97
A boy is at school taking a test.	The boy is taking a test at school.	1	.42	.33	.97
A blonde woman looks for medical supplies for work in a suitcase.	A blonde woman is searching for medical supplies in a bag.	1	.4	0	.98
A guy is sitting on the couch watching TV.	Some guy sitting on a couch watching television.	1	.42	.5	.75
A lady and her daughter look through a microscope.	A girl and a lady both looking through a microscope.	.92	.375	0	.86
Food is included in price of the accommodation.	The price of accommodation also includes the food.	.9	.285	0	.97
A man with tattoos is lounging on a couch and holding a pencil.	A tattooed man is on a sofa and is holding a pencil.	.88	.36	.4	.99

ACKNOWLEDGEMENT

The author(s) express their sincere gratitude to research and development cell of Bhilai Institute of Technology, Durg, Chhattisgarh for providing an excellent opportunity for sharing the preliminary milestones sought during the graduation level project tenure by the first (corresponding) author.

REFERENCES

[1] Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on Semantic Nets and Corpus Statistics. IEEE Trans. Knowledge Data Eng. 18(8), 1138–1150(2006).

[2] Lee M. C. (2010). A Novel Sentence Similarity Measure for Semantic-based Expert Systems. Expert Systems with Applications, 38(2011), 6392–6399.

[3] P. D. Turney (2002). Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July. pp. 417-424.

[4] Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. Discourse Processes. 25(2-3), 259-284.

[5] Miller, G. A. (1995) WordNet: A Lexical Database for English. Communications of ACM 38(11), 39-41.

[6] Resnik, P. (1995). Using Information Content to evaluate Semantic Similarity in a Taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, San Francisco, CA, USA, 14th Jul. pp. 448–453.

[7] Lin, D. (1998). An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conference on Machine Learning, Madison, WI, USA, 24–27 Jul. pp. 296–304.

[8] Jiang, J. J., & Conrath, D. W. (1997).Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.Proceedings of the 10th International Conference on Research in Computational Linguistics, Taipei ,Taiwan,August pp. 19–33.

[9] Leacock Claudia, Chodorow Martin and Miller George A. (1998). Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database. The MIT Press, Cambrige, MA,USA,pp. 265-283,ISBN: 9780262272551

[10] Wu, Z.& Palmer, M.(1994).Verb semantics and lexical selection. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 24th June ,pp. 133-138.

[11] Hirst and St-Onge(1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., editor,WordNet: An electronic lexical database. The MIT Press, Cambrige, MA,USA,pp. 305–332, ISBN: 026206197X

[12] Banerjee,S. and Pedersen,T.(2003). Extended gloss overlaps as a measure of semantic relatedness. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco , Mexico,9-15th Aug. pp.805–810.

[13] Patwardhan, S. , & Pedersen, T.(2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts.Proceedings of the EACL Workshop Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy,4thApril.pp. 1-8

[14] Vakare ,T. ,Verma,K.(2019).Sentence Semantic Similarity Using Dependency Parsing. Proceedings of the 10th International Conference on Computing, Communication, and Networking Technologies, Kanpur, India,6-8th July.

[15] Özateş, Ş. & Ozgur, A. & Radev, Dragomir.(2016) Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization. Proceedings of the Tenth International Conference on Language Resources and Evaluation , Portorož, Slovenia,23rd May .pp. 2833–2838.

[16] Man Yan Miranda Chong (2013). A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques, (doctoral dissertation). University of Wolverhampton. UK, pp 1 – 326, http://rgcl.wlv.ac.uk/papers/chong-thesis.pdf.

[17] Marie-Catherine de Marneffe and Christopher D. Manning.(2008). The Stanford typed dependencies representation. Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, Stroudsburg, PA, USA ,23rd August. pp. 1-8.

[18] Petr Baudis, Semantic Text Similarity Dataset Hub. https://github.com/brmson/dataset-sts. 13/08/2016