A Survey on Semantic Similarity

1st SUNILKUMAR P
M Tech CSSE, Dept. of Computer Science & Engineering
Govt. Engineering College, Idukki
Kerala, India
sunilkumar.p.kochi@gmail.com

2nd ATHIRA P SHAJI
Asst. Professor, Dept. of Computer Science & Engineering
Govt. Engineering College, Idukki
Kerala, India

***Abstract*—This paper provides a survey of semantic similarity of text documents. Semantic Similarity is an important task in Natural Language Processing (NLP). It is widely used for information retrieval, text classification, question answering, and plagiarism detection. This survey will classify different types of semantic similarity approaches such as corpus-based, knowledge-based and string-based. Various papers are reviewed and prepared performance analysis in this survey.**

***Index Terms*—Semantic Similarity, Knowledge Graph Based, Corpus Based, String Based, WordNet, Cosine Similarity, Jaccard Similarity**

I. Introduction

Semantic similarity find out the degree of semantic equivalence between two items, which can be concepts, sentences, or documents. Semantic Similarity between sentences or documents is also known as Semantic Textual Similarity (STS). It is an important linguistic technique in Natural Language Processing (NLP). It is being applied in document classification, semantic search, information retrieval, question answering, sentiment analysis, plagiarism detection, etc. Accuracy is a critical issue in the process of semantic similarity. Several techniques have been identified to measure the similarity between text documents. This paper analyses various semantic similarity approaches based on text documents. The Fig. 1 shows different types of similarity approaches.

Text similarity can be checked either lexically or semantically. Lexical similarity is a string-based similarity and semantic similarity is a meaning-based similarity. The string-based or a lexical based similarity method deals with the sentence as a sequence of characters. Obtaining similarity depends on measuring the similarity between sequences of characters. Various methods have been proposed in this type of similarity measure.

Different approaches are also available for Semantic-based similarity. These approaches apply different techniques to compare two sentences semantically. The corpus-based approach finds the similarity of the words based on statistical analysis of big corpus. Deep learning methods can be used to analyze a large corpus to denote semantics of words. The knowledge-based approach depends on a handcrafted semantic net for words. The meaning of words and relations between words has been included in this semantic net. WordNet is one of the popular semantic net used with applications.

String-based similarity is the third approach. String similarity measures the likeness of a sequence of text string and characters. It determines similarity or dissimilarity (distance) between two text strings. Several methods and approaches are available for measuring the string similarity.

This survey gives valuable information about various methodologies in measuring text similarity. It also helps to build a solid background in this domain.

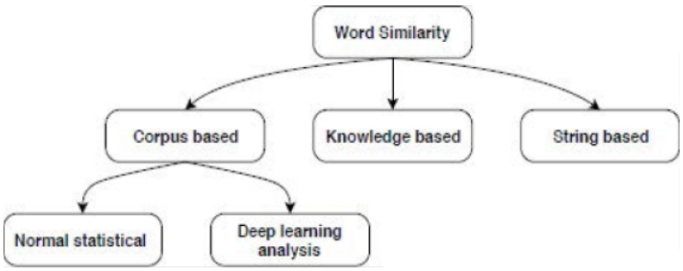

Fig. 1. Semantic Similarity

II. Related Works

A. Knowledge-Based Methods

Knowledge-based approaches measure the semantic similarity of concepts using Knowledge graphs. A knowledge graph is defined as a directed graph, $G = (V, E, \tau)$, where V is a set of nodes, E is a set of edges connecting those nodes; and τ is a function of $V \times \mathrm{V} \rightarrow E$ that defines all triples in G. The Fig. 2 shows the skeleton of Knowledge representation in a Knowledge graph.

Knowledge-based approaches measure the semantic similarity between concepts $c1, c2 \in V$, formally $sim(c1, c2)$, using semantic information contained in Knowledge graph. The most intuitive semantic information is the semantic distance between concepts, which is usually represented by the path connecting two concepts in knowledge graph. Intuitively, the similar concepts have shorter paths from one concept to

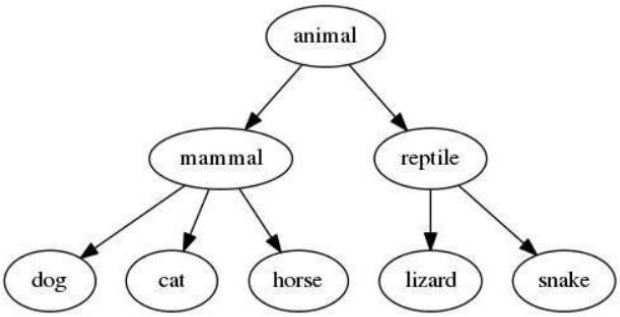

Fig. 2. Knowledge Representation

another.

A path $P(c_i, c_j)$ between $c_i, c_j \in V$ through G is a sequence of nodes and edges $P(c_i, c_j) = \{c_i, e_i ..., v_k, e_k, v_{k+1}, e_{k+1}, ..., c_j\}$ connecting the concepts c_i and c_j with cardinality or size n. For every two consecutive nodes $v_k, v_{k+1} \in V$ in $P(c_i, c_j)$, there exists an edge $e_k \in E$.

The similarity of two concepts can be obtained from the shortest path length between two concepts. Following equation can be used to find the similarity between two concepts.

$$sim_{path}(c_i, c_j) = \frac{1}{1 + length(c_i, c_j)} \quad (1)$$

Various ontologies are being used to find the similarity of text using the knowledge graph based approach. WordNet is a large lexical database of English language widely used for finding the similarity of texts. It groups Nouns, verbs, adjectives, and adverbs into sets of cognitive synonyms, which is called synsets. Synsets record several relations among these synonym sets or their members. WordNet is a combination of dictionary and thesaurus. These synsets are organized into a hierarchy constructing a semantic network in which semantic relations between synsets can be extracted easily. WordNet like word dictionaries are available for other popular languages also.

Ganggao Zhu, et al. [1] This paper proposes a method named wpath for combining knowledge-based and corpus-based semantic similarity approaches. Conventional corpus-based information content is computed from the concepts over textual corpus, which requires high computational cost. Since the instances are already extracted from textual corpus and annotated by concepts in knowledge graph-based IC, the wpath semantic similarity method shows significant improvement over other semantic similarity methods. The wpath method also shows excellent performance in a real category classification evaluation in terms of accuracy and F score.

Hai Jin, et al. [3] This paper presents ComQA - a three-phase knowledge-based question-answer framework by which users can ask questions and get answers. In ComQA, a question is split into several triple patterns. subsequently, it retrieves candidate subgraphs matching the triple patterns from the knowledge base and evaluates the semantic similarity between the subgraphs and the triple patterns to find the answer. It is a deep-rooted problem to evaluate the semantic similarity between the question and the heterogeneous subgraph containing the answer. Several testing over a series of QALD challenges confirm that the performance of ComQA is par excellence with other state-of-the-art approaches in terms of precision, recall, and F1-score

Nilima Sandip Gite [6] In this paper, a comparison based approach is being adopted for valuing the answer books. candidates' descriptive answers are compared with a specific standard descriptive answer stored on the server machine. The approach is mainly based on text mining technique which involves keyword matching as well as sequence matching. WordNet is being used for the keyword matching.

Riya Goswami, et al. [10] This paper proposes an approach to descriptive answer book evaluation using lexical and semantic similarity techniques. The goal of this system was to evaluate descriptive answer books programmatically and reduce the time and effort required for the valuation. Various tests show that the system can give moderate accuracy while comparing with the human valuation. In the next phase, the Semantic similarity approach is used with the WordNet dictionary for the answer book evaluation and obtained more accurate results than the previous one.

Marek Kubis [11] This paper introduces a new framework for computing the semantic similarity of words and concepts using WordNet-like databases. The key advantage of the proposed approach is the ability to implement similarity measures as concise expressions in the embedded query language. The framework was engaged to model the semantic similarity of nouns derived from Polish wordnets. Prospective results are obtained for this model in the testing process. Authors are planning to extend the framework with additional measures and to cover the content of PolNet more extensively as their future work.

Muhidin Mohamed, et al. [12] This paper proposes a hybrid approach for sentence paraphrase identification. The proposal addresses the issue of examining sentence-to-sentence semantic similarity when the sentences contain a set of named-entities. The crux of the proposal is to differentiate the computation of the semantic similarity of named-entity tokens from the rest of the sentence text. It is mainly based on the integration of word semantic similarity derived from WordNet and Wikipedia. This approach has been validated using two different datasets; Microsoft Research Paraphrase Corpus (MSRPC) and TREC-9 Question Variants. The empirical

evaluation reveals that this system outperforms baselines and most of the related state-of-the-art systems for paraphrase detection.

Mariam Hassanein, et al. [17] This paper presents an approach for discovering personality traits based on text semantic analysis. Various representations of user text combined with several semantic-based measures are proposed to predict users' personalities through their Facebook status updates. The proposed approach is tested and validated on data released by the myPersonality project for the Workshop on Computational Personality Recognition. Promising results are obtained in the testing and it proves that the information content-based measure achieves the best average personality trait prediction with an accuracy of 64%.

Kunal Khadilkar, et al. [19] This paper proposes a new method to detect plagiarism of documents using semantic knowledge graphs. The method use Named Entity Recognition and semantic similarity between sentences to detect possible cases of plagiarism. The doubtful cases are visualized using semantic Knowledge Graphs for a thorough analysis of authenticity. Usage of text contents that convey the same meaning can be detected using the proposed method. It also detects the use of Synonyms/phrases conveying the same meaning and overcome the drawback faced by most of the plagiarism checking software.

Svitlana Petrasova, et al. [22] This paper proposes a model for searching similar collocations in English texts to determine semantically connected text fragments for social network data streams analysis. The logical-linguistic model uses semantic and grammatical features of words to obtain a sequence of semantically related text fragments from different actors of a social network. To implement the model, it uses Universal Dependencies parser and Natural Language Toolkit with the lexical database WordNet. The experiment achieved a precision of 0.92 based on the Blog Authorship Corpus.

Anutharsha Selvarasa, et al. [24] This paper presents a system to measure semantic similarity for Tamil short phrases using a hybrid approach that makes use of knowledge-based and corpus-based techniques. We tested this system with 2000 general sentence pairs and 100 mathematical sentence pairs. For the dataset of 2000 sentence pairs, this approach achieved a Mean Squared Error of 0.195 and a Pearson Correlation factor of 0.815. For the 100 mathematical sentence pairs, this approach achieved an 85% of accuracy.

B. Corpus-Based Methods

A large number of the proposed approaches in word similarity are corpus-based. In this type, valuable information is extracted from analyzing a big corpus. A large corpus helps to check words co-occurrence to estimate similarity between words accurately. Furthermore, two different ways for statistical analysis of a corpus exist. The first is using normal statistical analysis such as Latent Semantic Analysis and the second is using deep learning. In the first approach, a big corpus is statistically analyzed by counting words in the corpus and documents. Calculating Term frequency-inverse document frequency(TF-IDF), which is used as a word weighting, is an important objective in the corpus analysis.

Latent Semantic Analysis (LSA): One of the most popular approaches in corpus-based analysis is LSA. In LSA each word is represented by a vector based on statistical computations. To construct these vectors a big text is analyzed and word matrix is constructed. In this matrix words and paragraphs are represented in rows and columns respectively. Besides, the Singular Value Decomposition (SVD) is applied to reduce dimensionality. It is a popular mathematical technique applied to LSA. Based on the constructed word vector, the similarity of the words are calculated using cosine similarity.

Word Embedding: Deep learning is another method used to represent words semantically. A very large corpus is used for training to find word representation in a semantic space. The generated word representation depends on the co-occurrence of words in the corpus. The idea of using deep learning is to train the model to guess a word given the surrounding words. Using this model a vector representation for words can be learned. The deep learning methods have shown good results in representing words semantically. Cosine similarity between words' vectors is used to measure word similarity.

Popular word embedding models used for finding the semantic similarity are :

- Word2Vec (by Google)
- GloVe (by Stanford)
- fastText (by Facebook)

Word2Vec: This model is provided by Google and is trained on Google News data. This model has 300 dimensions and is trained on 3 million words from google news data. It used skip-gram and negative sampling to build this model.

GloVe: GloVe stands for global vectors for word representation. It is an unsupervised learning algorithm developed by Stanford for generating word embeddings by aggregating global word-to-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the word in vector space.

fastText: fastText is a library for learning of word embeddings and text classification created by Facebook's AI Research (FAIR) lab. The model allows to create an unsupervised learning or supervised learning algorithm for obtaining vector representations for words. Facebook makes available pretrained models for 294 languages. fastText uses a neural

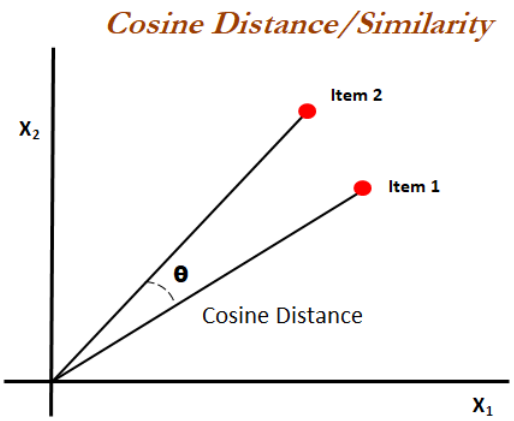

Fig. 3. Cosine Similarity

network for word embedding.

Cosine Similarity:

Cosine similarity is a mathematical method used to measure the similarity of documents irrespective of their size. It measures the cosine of an angle between two vectors projected in a multi-dimensional space.

$$similarity = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

where A_i and B_i are components of vector A and B respectively.

The similarity measure values range from -1 to 1, -1 shows the exact opposite, 1 shows exact similarity and 0 indicating orthogonality or decorrelation, while in-between values indicate intermediate similarity or dissimilarity. The vectors A and B are usually the term frequency vectors of the documents. Cosine similarity is one of the popular similarity checking methods and it can be seen as a method of normalizing document length during the comparison. The cosine similarity of two documents will range from 0 to 1 since the term frequencies (tf–idf weights) cannot be a negative value. The angle between two term frequency vectors cannot be greater than 90°. Fig. 3 shows the cosine similarity.

Atish Pawar, et al. [2] This paper proposes a methodology that can be applied in multiple areas by including corpora-based statistics into a standardized semantic similarity algorithm. The proposed method follows an edge-based approach using a lexical database for calculating semantic similarity between words and sentences. It has been tested on benchmark standards as well as the mean human similarity dataset and the methodology shows a high correlation value for the word and sentence similarity concerning Rubenstein. It shows remarkable performance in the SICK dataset and performs better than other unsupervised models.

Taihua Shao, et al. [4] In this paper, a Transformer-based neural network is proposed for answer selection. It set up a bidirectional long short-term memory (BiLSTM) behind the Transformer to acquire global information and sequential features in the question or answer. Apart from the original Transformer, this Transformer-based network focuses on sentence embedding rather than the seq2seq task. Besides, BiLSTM is being utilized to incorporate sequential features. This method is evaluated on a popular QA dataset WikiQA and the experimental results show that this Transformer-based answer selection model can achieve a comfortable performance compared with several competitive methods. Experiments reveal that this model is par excellence with other state-of-the-art methods in terms of MAP, MRR, and accuracy.

Sijimol P J, et al. [5] This paper, Handwritten Short Answer Evaluation System (HSAES) is an automated short answer evaluation system that can identify texts in answer papers and evaluates marks for each short answer based on knowledge acquired by the model through training. In the proposed system, OCR tools are used to extract handwritten texts. NLP is used to extract the keywords from the human evaluated sample dataset of handwritten answer papers and answer key. The proposed model evaluates scores based on cosine similarity measures. Marks will be given to each of the sentences in the evaluated answer paper.

Alla Defallah Alrehily, et al. [7] This paper proposes an automated assessment system for subjective questions. It finds the matching ratio for the keywords in original answers with students' answers. The marks are given based on semantic and document similarity. The assessment system contains four modules such as preprocessing, keyword expansion, matching, and grading. Various tests are conducted and obtained prospective results. Moderate values are obtained for testing parameters recall, precision, accuracy and f measure. This system helps to increase the efficiency and productivity of the examination system. It also helps to save time, reduce cost and minimize usage of resources.

M. Syamala Devi, et al. [8] In this paper computerized evaluation of subjective answers is done with machine learning techniques such as Latent Semantic Analysis (LSA) and Bilingual Evaluation Understudy (BLEU) and Maximum Entropy. The above techniques are implemented with ontology and tested with input data consisting of subjective answers in computer science. Detailed analysis has been carried out and obtained accurate results and a high correlation with human performance.

Avani Sakhapara, et al. [9] In this paper, a machine learning-based subjective answer grader system (SAGS) is designed and implemented using latent semantic analysis (LSA) and information gain (IG) algorithms. It proposes an enhancement of these algorithms through synonym replacement using WordNet. Accuracy of this algorithm is tested by comparing the generated scores with the scores given by evaluators.

Pongsakorn Saipech, et al. [13] This paper presents the automation of subjective examination for the Thai language based on semantic similarity, applying cosine similarity techniques together with the applicable synonym. Students' answers and reference answers are input into the system in short Thai sentences. The input answers which are reference answers provided by faculties, and student answers. Using the longest matching algorithm and dictionary (Lexitron). These answers are converted into vectors using TF-IDF, which is a frequency computation technique. Cosine similarity is used to measure the similarity of the answers. The results generated by the proposed system is compared with the scores prepared by expert teachers and found that the scores produced from the system using cosine similarity with synonyms were similar to those obtained from the expert.

Ruchindramalee Chandrathlake, et al. [15] This project performs validation of news posts in social media. The proposed system extracts the content of the news item, searches the Internet to find similar articles in reliable online news sources, matches the extracted content with the content of the news sites and generates an accuracy level. Several techniques have been used in developing this system such as web scraping techniques, web crawling techniques, URL ranking methodologies, automatic text summarization techniques, Word2vec, and cosine similarity techniques. This system has obtained an accuracy of 70% for the news posts on social media compared to the reliable online news sources.

Gokul P. P, et al. [16] This paper uses individual words synonyms to find the similarity between two sentences. It presents the observations on sentence similarity between two Malayalam sentences using cosine similarity method. It used test data of 900 and 1400 sentence pairs of FIRE 2016 Malayalam corpus that used in two iterations to present. This semantic similarity method got an accuracy of 0.8 and 0.59 respectively.

Xiaolin Jin, et al. [18] This paper proposes a semantic similarity computation method based on Word2vec. This method improves Chinese dictionary-based approaches such as HowNet and Tongyici Cilin, and also adds the word vector model as a weighing parameter to calculate the word similarity, after comparing the similarity of the words by assigning different weights to the three methods. In the analysis, it shows a high Pearson correlation coefficient and the method can cover most words so that it can effectively solve the problem of the similarity of the word calculation in the dictionary.

Nishant Nikhil, et al. [20] In this paper, a deep learning-based supervised approach is used to recommend similar documents based on the similarity of content. It combines the Convolutional Deep Structured Semantic Models (C-DSSM) with Word2Vec distributed representations of words to create a novel model to classify a document pair as relevant/irrelevant by assigning a score to it. Using this model retrieval of documents can be made in reduced time and less memory complexity. From the experiments it is seen that learn the embeddings of documents using neural network are performed with reduced complexity and increased accuracy.

Gerardo Orellana, et al. [21] This paper proposes an approach to semantically compare the syllabus contents through text similarity methods. Such methods have been widely used in different domains. In this work, the syllabus of higher education institutions is extracted using Text mining and compared their semantic contents. It uses various approaches such as pre-processing techniques, Latent Semantic Analysis for dimensionality reduction, text enrichment through the Wikipedia API and Google Engine, Support Vector Machine as classifier, and cosine similarity as similarity metric. Test results show that this method successfully measures similarity among syllabuses. It achieved promising values like accuracy, precision, and recall for the comparison of selected course syllabuses.

Qiufeng Ren, et al. [23] This paper proposes a study to determine a resource recommendation scheme based on the semantic similarity and sentiment analysis of the review text. The process to obtain personalized recommendation involves Extracting the semantic and sentiment information of the resources, filling user rating matrix, and calculating users' similarity with adjusted cosine measures. Test results reveal that the proposed algorithm could characterize user preference in a better way by obtaining information-in-depth.

C. String-Based Methods

String-based word similarity depends on the comparison between two sequences of characters. There are different methods to measure the similarity between words based on string matching such as jaccard distance, levenshtein distance, and n-gram.

Jaccard Similarity: It is also known as the Jaccard similarity coefficient, which is a statistical method used to find the similarity and diversity between two finite sets. It is defined as follows.

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{3}$$

Levenshtein distance: Distance between two strings is the minimal number of basic operations (insert, delete, or replace) needed to convert one string to another. The ratio between the distance and length of longer string is considered as the similarity between these strings.

n-gram: n-gram is a method of checking 'n' continuous words or sounds from a given sequence of text or speech. This model helps to predict the next item in a sequence. Unigram refers to n-gram of size 1, Bigram refers to n-gram of size 2, Trigram refers to n-gram of size 3.

Mubbashir Ayub, et al. [14] In this paper, the Jaccard similarity approach is used for improving the performance of collaborative filtering (CF) based recommender systems. This model works in the same manner as Jaccard similarity works. But Jaccard similarity does not consider the absolute value of rating and only considers the ratio of co-rated items. It takes into account the ratio of absolute rating values which are equal in value, to the total no of co-rated items. An additional argument it takes into account is the average rating value of users. The performance of the proposed model has been tested with many state-of-the-art similarity measures. The results show that the proposed measure has some performance improvement in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

Seiya Temma, et al. [25] This paper proposes a new index of similarity for the classification of emails into ham and spam ones with the Jaccard index. It takes advantage of the co-occurrence value of all pairs of two words in emails. The proposed method classified emails into hams or spams with a high accuracy rate than the present filtering system using the appearance frequency of the word. This method could extract patterns of word usage reflecting the context of emails.

D. Summary

In this survey, 25 papers have been analyzed in various domains out of which 8 papers from Knowledge-based methods, 13 papers from Corpus-based methods, 2 papers from String-based methods and 2 papers used hybrid methods. Table 1 shows the summary of survey.

Analysis reveals that the Knowledge-based and Corpus-based methods are performing well. Knowledge-based methods have less computational cost while compared with the corpus-based methods. Knowledge-based methods are not restricted to language comparing and it can be used to compare the similarity of any ontology-based systems. The availability of the ontology is one of the demerits of knowledge-based methods.

Corpus-based methods are used in several systems that Works with unstructured or semi-structured texts. High Computational cost and difficulty in preparing domain corpus are the disadvantages of corpus-based methods.

String-based similarity measures are simple and easy to use. String-based similarity measures only be used for dissimilarity or distance measures.

III. CONCLUSION

This paper performed a survey of semantic similarity methods and surveyed papers in three groups such as Knowledge-based methods, Corpus-based methods, and String-based methods. Detailed study and analysis are conducted for 25 papers, which use various approaches and methods for measuring semantic similarity. The analysis shows that Knowledge-based and Corpus-based methods are widely used for measuring semantic similarity and these methods show promising results.

Sl. No.	Paper	Journal	Date	Base Method	Method Used	Application	Acc%	Advantages	Disadvantages
1	Computing Semantic Similarity of Concepts in Knowledge Graphs	IEEE	Jan-2017	Knowledge Based	Wpath	Semantic Similarity	80	Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
2	Challenging the Boundaries of Unsupervised Learning for Semantic Similarity	IEEE Access	Feb-2019	Corpus Based	Corpus	Semantic Similarity	87	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
3	ComQA: Question Answering Over Knowledge Base via Semantic Matching	IEEE Access	Jun-2019	Knowledge Based	Knowledge Graph	Question Answering	74	Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
4	Transformer-Based Neural Network for Answer Selection in Question Answering	IEEE Access	Mar-2019	Corpus Based	Transformer Based NN	Question Answering	73	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
5	Handwritten Short Answer Evaluation System (HSAES)	IJSRST	Feb-2018	Corpus Based	Cosine Similarity, NLP	Short Answer Evaluation		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
6	Implementation of Descriptive Examination and Assessment System	IJARSE	Mar-2018	Knowledge Based	WordNet	Evaluation of Descriptive Answers		Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
7	INTELLIGENT ELECTRONIC ASSESSMENT FOR SUBJECTIVE EXAMS	CSIT	May-2018	Corpus Based	Cosine Similarity	Evaluation of Subjective Exam	89	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
8	MACHINE LEARNING TECHNIQUES WITH ONTOLOGY FOR SUBJECTIVE ANSWER EVALUATION	IJNLC	Apr-2016	Corpus Based	LSA and Ontology	Evaluation of Subjective Exam		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
9	Subjective Answer Grader System Based on ML	SPRINGER	Feb-2019	Corpus Based	LSA, Cosine Similarity	Evaluation of Subjective Exam	83	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
10	A Knowledge based Approach for Long Answer Evaluation	IEEE CONF	May-2017	Knowledge Based	WordNet	Evaluation of Descriptive Answers	83	Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
11	A Semantic Similarity Measurement Tool for WordNet-Like Databases	SPRINGER	Jun-2018	Knowledge Based	WordNet	Semantic Similarity		Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
12	A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics	SPRINGER	Apr-2019	Hybrid Method	WordNet	Semantic Similarity	76	nan	Knowledge Based Requires Ontology
13	Automatic Thai Subjective Examination using Cosine Similarity	IEEE CONF	Nov-2018	Corpus Based	Cosine Similarity	Evaluation of Subjective Exam		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
14	A Jaccard base similarity measure to improve performance of CF based recommender systems	IEEE CONF	Jan-2018	String Based	Jaccard Similarity	Recommender System		Simple Similarity Measure	Used for dissimilarity or distance measure
15	A Semantic Similarity Measure Based News Posts Validation on Social Media	IEEE CONF	Jun-2019	Corpus Based	Word2vec, Cosine Similarity	News Posts Validation	70	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
16	Sentence Similarity Detection in Malayalam Language using cosine similarity	IEEE CONF	May-2017	Corpus Based	Cosine Similarity	Sentence Similarity Detection		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
17	Predicting personality traits from social media using text semantics	IEEE CONF	Feb-2019	Knowledge Based	Knowledge Based, WordNet	Predicting Personality	64	Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
18	Word Semantic Similarity Calculation Based on Word2vec	IEEE CONF	Oct-2018	Corpus Based	Word2vec	Semantic Similarity		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
19	Plagiarism Detection Using Semantic Knowledge Graphs	IEEE CONF	Apr-2019	Knowledge Based	Knowledge Graph	Plagiarism Detection		Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
20	Content Based Document Recommender using Deep Learning	IEEE CONF	May-2018	Corpus Based	Word2vec	Document Recommender		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
21	A text mining methodology to discovery syllabi similarities among Higher Education Institutions	IEEE CONF	Dec-2018	Corpus Based	LSA & Cosine Similarity	Syllabus Similarity	83	Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
22	Building the Semantic Similarity Model for Social Network Data Streams	IEEE CONF	Oct-2018	Knowledge Based	WordNet	Semantic Similarity		Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
23	Resource Recommendation Algorithm Based on Text Semantics and Sentiment Analysis	IEEE CONF	Feb-2019	Corpus Based	Cosine Similarity	Resource Recommendation		Works with unstructured or semi-structured texts	High Computational cost Difficulty in preparing domain corpus
24	Short Tamil Sentence Similarity Calculation using Knowledge-Based and Corpus-Based Similarity Measures	IEEE CONF	Jul-2017	Corpus & Knowledge Based	WordNet	Sentence Similarity	85	Less Computational Cost Not Restricted to Language Comparing	Requires Ontology
25	The Document Similarity Index based on the Jaccard Distance for Mail Filtering	IEEE CONF	Aug-2019	String Based	Jaccard Similarity	Document Similarity		Simple Similarity Measure	Used for dissimilarity or distance measure

TABLE I
SUMMARY OF SURVEY

REFERENCES

[1] Ganggao Zhu and Carlos A. Iglesias "Computing Semantic Similarity of Concepts in Knowledge Graphs," *IEEE Trans, Jan-2017*

[2] Atish Pawar and Vijay Mago "Challenging the Boundaries of Unsupervised Learning for Semantic Similarity," *IEEE Access, Feb-2019*

[3] Hai Jin , (Fellow, IEEE), Yi Luo, Chenjing Gao, Xunzhu Tang, and Pingpeng Yuan, (Member, IEEE) "ComQA: Question Answering Over Knowledge Base via Semantic Matching," *IEEE Access, Jun-2019*

[4] Taihua Shao , Yupu Guo, Honghui Chen, and Zepeng Hao "Transformer-Based Neural Network for Answer Selection in Question Answering," *IEEE Access, Mar-2019*

[5] Sijimol P J and Surekha Mariam Varghese "Handwritten Short Answer Evaluation System (HSAES)," *IJSRST, Feb-2018*

[6] Nilima Sandip Gite "Implementation of Descriptive Examination and Assessment System," *IJARSE, Mar-2018*

[7] Alla Defallah Alrehily, Muazzam Ahmed Siddiqui, Seyed M Buhari "Intelligent Electronic Assessment for Subjective Exams," *CSIT, May-2018*

[8] M. Syamala Devi and Himani Mittal "MACHINE LEARNING Techniques with Ontology for Subjective Answer Evaluation," *IJNLC, Apr-2016*

[9] Avani Sakhapara, Dipti Pawade, Bhakti Chaudhari, Rishabh Gada, Aakash Mishra and Shweta Bhanushali "Subjective Answer Grader System Based on Machine Learning," *SPRINGER, Feb-2019*

[10] Riya Goswami, Somik Karmakar, Avik Bisai, Alok Ranjan Pal "A Knowledge based Approach for Long Answer Evaluation," *IEEE Conf, May-2017*

[11] Marek Kubis "A Semantic Similarity Measurement Tool for WordNet-Like Databases," *SPRINGER, Jun-2018*

[12] Muhidin Mohamed and Mourad Oussalah "A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics," *SPRINGER, Apr-2019*

[13] Pongsakorn Saipech and Pusadee Seresangtakul 'Automatic Thai Subjective Examination using Cosine Similarity," *IEEE Conf, Nov-2018*

[14] Mubbashir Ayub, Mustansar Ali Ghazanfar, Muazzam Maqsood, Asjad Saleem "A Jaccard base similarity measure to improve performance of CF based recommender systems," *IEEE Conf, Jan-2017*

[15] Ruchindramalee Chandrathlake, Lochandaka Ranathunga, Sumudu Wijethunge, Prabhath Wijerathne, Dilki Ishara "A Semantic Similarity Measure Based News Posts Validation on Social Media," *IEEE Conf, Jan-2018*

[16] Gokul P.P, Akhil BK and Shiva Kumar K.M "Sentence Similarity Detection in Malayalam Language using cosine similarity," *IEEE Conf, Jun-2019*

[17] Mariam Hassanein, Wedad Hussein, Sherine Rady and Tarek F. Gharib "Predicting personality traits from social media using text semantics," *IEEE Conf, Feb-2019*

[18] Xiaolin Jin, Shuwu Zhang and Jie Liu "Word Semantic Similarity Calculation Based on Word2vec," *IEEE Conf, Oct-2018*

[19] Kunal Khadilkar, Dr. Siddhivinayak Kulkarni and Poojarani Bone "Plagiarism Detection Using Semantic Knowledge Graphs," *IEEE Conf, Apr-2019*

[20] Nishant Nikhil and Muktabh Mayank Srivastava "Content Based Document Recommender using Deep Learning," *IEEE Conf, May-2018*

[21] Gerardo Orellana, Marcos Orellana, Victor Saquicela and Fernando Baculima "Computing Semantic Similarity of Concepts in Knowledge Graphs," *IEEE Conf, Dec-2018*
[22] Svitlana Petrasova, Nina Khairova and Włodzimierz Lewoniewski "Building the Semantic Similarity Model for Social Network Data Streams," *IEEE Conf, Oct-2018*
[23] Qiufeng Ren, Yue Zheng, Guisuo Guo and Yating Hu "Resource Recommendation Algorithm Based on Text Semantics and Sentiment Analysis," *IEEE Conf, Feb-2019*
[24] Anutharsha Selvarasa,Nilasini Thirunavukkarasu, Niveathika Rajendran, Chinthoorie Yogalingam, Surangika Ranathunga and Gihan Dias "Short Tamil Sentence Similarity Calculation using Knowledge-Based and Corpus-Based Similarity Measures," *IEEE Conf, Jul-2017*
[25] Seiya Temma, Manabu Sugii, and Hiroshi Matsuno "The Document Similarity Index based on the Jaccard Distance for Mail Filtering," *IEEE Conf, Aug-2019*