Survey on Sentence Similarity System

Nayan Joshi[1], Darshan Kadam[2], Sahil Kadu[3], Prof. Shubhangi Chavan[4]
[1,2,3] *Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*
[4]*Guide, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*

Abstract- **The measure of how similar are the given sentences are, is the Sentences similarity which plays an important role in text-related research and application in area as text-mining. In this system we Pre-process the given sentences in a bag of words using tokenization, stemming and other Natural language techniques. Then we apply syntax similarity techniques and semantics similarity techniques. The syntax similarity technique finds the grammatical syntax similarity between sentences. The semantic similarity technique finds the semantic similarity between words, it creates a relationship between words and sentences through the meanings of the words. The technique used to calculate semantic similarities are cosine similarity, word order similarity and feature based similarity. In cosine similarity we find the cosine similarities between sentences, in word ordered similarity we find the Intersection word set of them which contains common words between the sentences and in feature based. The sentence similarity is used in plagiarism detection system, Question-Answering system, etc.**

Index terms- **Semantic similarity, syntactic similarity, Word Net, Hindi similarity, Text similarity**

I.INTRODUCTION

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Most NLP techniques rely on machine learning to derive meaning from human languages. NLP is commonly used in many applications such as.

- Language translation applications such as Google Translate
- Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.
- And many more...

NLP entails applying algorithms to identify and extract the natural language rules such that the unstructured language data is converted into a form that computers can understand. When the text has been provided, the computer will utilize algorithms to extract meaning associated with every sentence and collect the essential data from them. Sometimes, the computer may fail to understand the meaning of a sentence well, leading to obscure results. The current times the Sentence Similarity measures are used more and needed in the Text-based Research and other areas. Some similarity measure calculates the similarity between 2 sentences, thoroughly using Word-net semantic dictionary. The Sentence Similarity is the one of the core functions in NLP tasks such as Paraphrase detection, etc. Given the 2 sentences, the task of calculating the similarity is defined how similar the meaning of 2 sentences are. The higher the similarity, the more similar the meaning of 2 sentences are.

II. LITERATURE SURVEY

The studies approaches to calculate the text similarity are done in previous years but there are few which give similarity based on sentences. This is due to the ambiguity and complexity of grammar. In this section we will review the previous system and approaches to find their strengths and their limitations

Pantulkar Sravanthi, Dr. B. Srinivasu suggested a system to check similarity between two sentences [1]. The approach uses the Word Net corpus. The two sentences are preprocessed in the preprocessor, the NLP techniques Tokenization, Tagging and lemanization are used. The preprocessed data is directly sent to find its cosine similarity between the sentences which gives the syntactic similarity. The semantic similarity uses the meaning behind* the words and syntax of sentences to find similarity. If structure of two similarity they are more probably have same meaning. The POS tagging is used and labelling them. This data is forwarded to lemmatizer to get lemmas, which is used to generate synset of Word Net corpus. The (WU and Palmer) measure is used to find the similarity between words. The first word of both sentences are compared, if they match then the similarity score is calculated based on syntactic level. If not then the synset of the word is extracted and compared, if this is also not similar then its comparison can be done on basis of semantics and the semantic similarity is calculated.

In 2014, Mohamed Elkhdir, Mohannad N. Ibrahim, Tarig A. Khalid, Shawgi Ibrahim, Mohamed Awadalla suggested a system to detect plagiarism using Free Text Fingerprint Analysis[2]. It uses Winnowing algorithm to analyse the data Then N-grams of length N are created and hash values are generated for each gram. These hash values are used as fingerprints and compared to check the similarity between two documents. The similar fingerprints are stored in an array and are given output in GUI.

In 2014, Deepa Gupta, Vani K, Charan Kumar Singh proposed a system to detect external plagiarism using NLP techniques and Fuzzy Similarity techniques[3]. This system compares a source collection of data against a suspicious document. At first, Basic preprocessing is done. Then two approaches are followed i) POS tagging is done using Stanford Tagger. Words which belong to Noun, adjective, verb, adverb classes are retained and other are pruned. ii) Stop word removal followed by lemantisation is done. Then for both approaches, N-grams are created and Fuzzy semantic similarity is used for comparisons. All matched N-grams are stored and passages are formed on the basis of passage boundary which depends on total word length. At evaluation stage, Recall, Precision and Granularity is calculated for ranking purposes.

In 2015, Vasilei Hatzivassiloglow, Judith L. Klavans, Eleazar Eskin proposed a system to detect text similarity over short passages using machine learning[4]. This system proposes a approach in which it computes a feature vector over a pair of textual units. It calculates primitive features like word co-occurrence, matching noun phrases, WordNet synonyms, common semantic classes for verbs, shared proper nouns for given textual units. Then it computes composite features like ordering, distance over pairs of primitive features. Then a feature vector is computed over primitive and composite feature values. Finally, a machine learning algorithm, RIPPER is used and trained over documents to get desired output of similar textual units.

In 2016, Sandip Sarkar, Saurav Saha, Jereemi Bentham, Partha Pakray, Dipankar Das, Alexander Gelbukh proposed a paraphrase detection system which was language independent[5]. This system proposes a system which is focused in sentence-level paraphrase identification for Indian languages.It is based on three language independent techniques i) Jaccard Similarity ii) Levenshtein Ratio iii) Cosine Similarity. Output of these techniques is fed into Probabilistic Neural Network (PNN) for paraphrase classification.

Zhao Jingling, Zhang Huiyun, Cui Baojiang proposed a system to check sentence similarity based on a Semantic Vector Model[6]. This system proposes the similarity of sentences using the semantic and structural information of the compared sentences.In this to compute the similarity we have to follow three steps: i) We calculate the semantic similarity of words. ii) Semantic similarity between the sentences. iii) Word order similarity and combine word order and semantic similarity as final similarity. Word similarity uses a corpus for a specific language to calculate. The corpus used is How-Net. The semantic similarity is calculated using the mutual vectors in both the sentences computing the cosine similarity of sentence.In word order similarity a word occurs with two keywords in order to both the sentences.

Haipeng Ruan, Yuan Li, Qinglin Wang, Yu Liu proposed a system to check system similarity for a Question Answering System which is based on Multi-Feature Fusion[7]. It is a system used for question answering system in chinese language. It compares the sentences and finds the similarity of sentences. The system uses four feature checks to find the similarity. The four feature checks are: i) the sentence similarity as a whole sentence. ii) The sentence similarity using word to vector. iii) Morphological similarity: - The mutual words between sentences. iv)Order of word similarity: - Bigram* is created and checked in with Bigram* another sentence. These four features are used to train the Neural Network Model. The algorithm used to train the Neural Network is back propagation.The output given by the model is the total similarity.

Asad Abdi, Norisma Idris, Rasim M. Alguliyev, Ramiz M. Aliguliyev proposed a system to detect plagiarism using linguistic knowledge[8]. In this paper the system is developed to detect plagiarised documents. The system has 3 steps, i) Pre-processing: In this step the source document and suspicious document are prepared for further processing. This step consists of 3 function stopword removal, stemming and sentence segmentation. ii) Detailed Comparison: In this both the documents are compared to find the plagiarized sentences in suspicious documents with reference to the source documents. This step includes the use of word net, Semantic similarity between words, Semantic similarity between sentences. Then the similarity is calculated using an algebraic equation that combines the semantic and word order similarity. iii) Post-processing: The sentences classified as plagiarized sentences are displayed.

In 2015, Md Arafat Sultan, Steven Bethard, Tamara Summer proposed a system to check sentence similarity using Word Alignment and Semantic Vector Composition[9]. This paper proposes a system which checks similarity between two given sentences. At first Word Alignment is done using a monolingual word aligner. An extra alignment-based feature is used in which content words aligned only if they have a contextual similarity. Then using word2vec toolkit, vector representation of each sentence is done and take cosine similarity between two sentences. Finally, outputs of cosine similarity and word alignment are combined using a ridge regression module and final output is given.

In 2016, Sneha Bagde, Mohit Dua, Zorawar Singh Virk proposed a system to compare different similarity techniques on a Hindi QA System[10]. This system discusses a comparative analysis of different similarity measures for Hindi question answering system. Different similarity functions used are : i) N-gram approach ii) Euclidean approach iii) Cosine similarity iv) Jaccard coefficient v) Jaro-Wrinkler. Sentences containing multiphrase words or misspelled words are passed through each function and results are obtained respectively.

Issa Atoum, Ahmed Otoom proposed an efficient system for Hybrid Text Similarity using Wordnet and a Corpus[11]. In this paper, the proposed approach uses a sentence to sentence similarity and word 2 word similarity .The paper uses the 2 sentences and compare both sentences as a whole to each other and then compares the word of each sentence to every word of the other sentence. Then the both similarities are summed up to give the total similarity of the sentences.

III. PROPOSED SYSTEM

Existing System Architecture

The presented System Architecture is using the syntactic approach and semantic with only one syntactic technique:

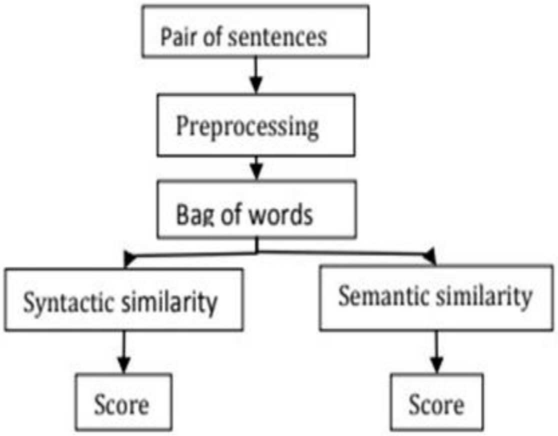

Figure 1 Existing system architecture

This approach uses only both syntactic and semantic way of the similarity measure. In the semantic approach the techniques used:

1. Wu-Palmer similarity
2. Short Path distance

The syntactic approach only uses a single technique for measurement between text:

1. Cosine Similarity
2. Jaccard similarity
3. Word-order similarity

The system checks the only uses one similarity technique for syntactic level of similarity calculation. This makes the system not so accurate on the level of syntactic similarity.

Proposed System Architecture

As discussed above, architecture does not have the more accurate syntactic score and all the scores aren’t calculated in a single score for the system.

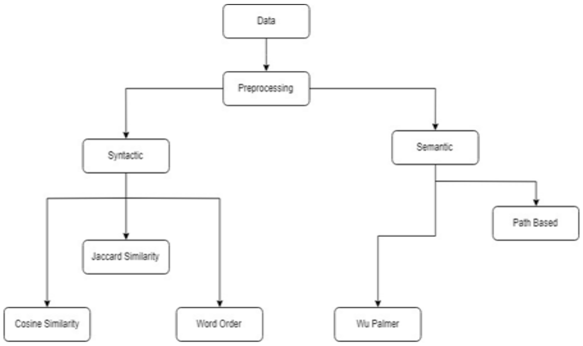

Figure 2 Proposed system architecture

In the proposed system, the data is first passed through the Pre-Processing stage where the data get processed using the NLP (Natural Language Processing) Techniques. Then the processed data is created in BOW (Bag of words). This processed data is used in different Syntactic and Semantic techniques.

The techniques give a score as per the algorithm presented. This score is calculated as per a semantic level and syntactic level as a single score for each level. This score is then presented to the user.

3.2 Hardware and Software Specifications

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 1 and Table 2 respectively.

Table 1 Hardware details

Processor	1.6Ghz Intel/Amd
Graphics	512Mb
RAM	4 GB

Table 2 Software details

Operating System	Windows 7 and up.
Programming Language	Python 3.6, Html, Css
IDE	IDLE/Pycharm
Database	Mysql
Browser	Chrome

REFERENCES

[1] Pantulkar Sravanthi, Dr. B. Srinivasu, “Semantic Similarity between Sentences”. International Research Journal of Engineering and Technology (IRJET), 2017.

[2] Atoum, Issa & Otoom, Ahmed. (2016). Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.070917.

[3] H. Ruan, Y. Li, Q. Wang and Y. Liu, "A Research on Sentence Similarity for Question Answering System Based on Multi-feature Fusion," 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 507-510.

[4] Sarkar, Sandip & Saha, Saurav & Bentham, Jereemi & Pakray, Dr. Partha & Gelbukh, Alexander. (2016). NLP-NITMZ@DPIL-FIRE2016: Language Independent Paraphrases Detection.

[5] Sneha B., Mohit D., Zorawar Singh V. (2016) Comparison of Different Similarity Functions on Hindi QA System. In: Satapathy S., Joshi A., Modi N., Pathak N. (eds) Proceedings of International Conference on ICT for Sustainable Development. Advances in Intelligent Systems and Computing, vol 408. Springer, Singapore

[6] Sultan, M.A., Bethard, S. and Sumner, T., 2015. DLS $@ $ CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 148-153).

[7] Hatzlvassiloglou, V., Klavans, J.L. and Eskin, E., 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora.

[8] Elkhidir, M., Ibrahim, M.M., Khalid, T.A., Ibrahim, S. and Awadalla, M., 2015, September. Plagiarism detection using free-text fingerprint analysis. In 2015 World Symposium on Computer Networks and Information Security (WSCNIS) (pp. 1-4). IEEE.

[9] Jingling, Z., Huiyun, Z. and Baojiang, C., 2014, November. Sentence similarity based on semantic vector model. In 2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (pp. 499-503). IEEE.

[10] Gupta, D., Vani, K. and Singh, C.K., 2014, September. Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In 2014

International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2694-2699). IEEE.

[11] Abdi, A., Idris, N., Alguliyev, R.M. and Aliguliyev, R.M., 2015. PDLK: Plagiarism detection using linguistic knowledge. Expert Systems with Applications, 42(22), pp.8936-8946.