SEMANTIC WORD ANALYSIS FOR THE MARATHI PARAPHRASE SENTENCES

Dattatray Solanke and C Namrata Mahender

Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University Aurangabad, (MH). India

solanked031@gmail.com and cnamrata.csit@bamu.ac.in

ABSTRACT

Many time word communicate the same meaning or intention or even different meaning or intention in different sentence which makes paraphrase detection a difficult task. Humans are capable of extracting appropriate meaning of the word according to the context of sentence. Understanding intention of words or a complete sentence is termed as Intention paraphrasing. It is very important in text summarization, machine translation, question answering, natural language understanding, word net building, ontology, plagiarism detection, author identification. Corpuses for Marathi paraphrase is developed as such dataset for Marathi are not available. Cosine similarity is applied for calculating semantic similarity among the paraphrase sentences.

Keyword: *Paraphrase, Marathi, levenstein, cosine, sentences*

DOI: 10.48047/ecb/2023.12.6.278

INTRODUCTION

Language being a tool of communication reflects the thoughts, needs, or expression of a person to others. The way we communicate actually many times determines the different meaning of the same sentence (uttered) or different sentence making the same sense. It is that makes a sentence analysis a difficult or complex task when texts are lexically or syntactically modified to differ in appearance but retain the same intention is called “paraphrasing”. Paraphrasing helps to present a single concept in multiple ways making sense of using various permutation combinations or synonyms of words, paraphrase can be generated, extracted, or identified based on the application where it is applied. Paraphrase generation involves the use of synonyms of a word including changing the word sequence and grammatical structure also without changing the meaning whereas paraphrase extraction is a process of compilation of various words or phrases, which have the same meaning or same intention. Paraphrase identification is a technique of detecting the variety of expressions that communicates the same meaning or intention. Paraphrasing is very important in text summarization, machine translation, question answering, natural language understanding, word net building, ontology identification. This paper tries to highlight the importance of understanding semantic similarity in given Marathi sentence to better analyse paraphrasing. As very negligible work on Marathi semantic similarity sentences, thus the present work tries to put emphasizes on semantic similarity of sentences for the Marathi language.

2 LITERATURE REVIEW

1. Rada & Corley et al [1]in their research focused on measuring the semantic similarity of short texts. Through experiments performed on a paraphrase data set, the semantic similarity method outperforms methods based on simple lexical matching, resulting in up to 13% error rate reduction with respect to the traditional vector-based similarity. The best performance is achieved using a method that combines several similarity metrics into one, for an overall accuracy of 70.3%, representing a significant 13.8% error rate reduction with respect to the vector-based cosine similarity baseline. Moreover, if considered into account the upper bound, 83% was established by the inter-annotator agreement achieved on data set the error rate reduction over the baseline appeared even more significant.
2. Lee & Keely et al focus [2]. on measuring the semantic similarity between sentences or very short texts, based on semantic and word order information. First, semantic similarity is derived from a lexical knowledge base and a corpus. The lexical knowledge base models is a common human knowledge about words in a natural language; this knowledge is usually stable across a wide range of language application

areas. A corpus reflects the actual usage of language and words. Thus, our semantic similarity not only captures common human knowledge, but it is also able to adapt to an application area using a corpus specific to that application. Second, the proposed method considers the impact of word order on sentence meaning. The derived word order similarity measures the number of different words as well as the number of word

3. Gunasinghe et al [3]. focuses on the sentence similarity measure algorithm which is developed based on both syntactic and semantic similarity measures. This algorithm is based on measuring the sentence similarity by adhering to a vector space model generated for the word nodes in the sentences. In this implementation they consider two types of relationships. One of them is relationship between verbs in the sentence pairs while the other one is the relationship between nouns in the sentence pairs. One of the major advantages of this method is, it can be used for variable length sentences. They focus only on measuring the sentence similarity in a well-structured manner. Rather than focusing on sentence similarities of two sentences.

4. Sazianti et al. in [4], introduced the metaphor for the semantic similarity of two sentences with the synonyms matching from WordNet while applying the vector to the words. The result is calculated after concerning the output as 3 similarity measures comparisons as better in measuring the semantic similarity between sentences. However, they suggested that more testing and evaluation work is needed for real test data and human experts.

5. Issa & Ahemad et al [5]. In their article evaluates previous word similarity measures on benchmark datasets and then used a hybrid word similarity in a novel text similarity measure(TSM). TSM is based on information content and WordNet semantic relations. TSM includes exact word match, the length of both sentences in a pair, and the maximum similarity between one word and the compared text. The similarity measure outperforms much of the compared similarity measures and is significant at the 0.05 level. The reason behind the high achievement of the method is due to the employment of additional information (corpus and information content) and the effectiveness of the borrowed word similarity measure.

6. Paraphrase detection using semantic relatedness based shortest path [7] in this research semantic relatedness approach on the interlink considered from Wordnet for English language is used by author and overall result achieved from author is 71.1% .

3 METHODOLOGY DEPLOYED

For developing present system for Marathi the basic requirement was first to create Marathi corpus for paraphrase sentences. We have created total 40 sentences and 120 sentences which are paraphrased sentences based on the 40 sentences which are simple sentences then a synonym word dictionary is prepared. Then using cosine similarity, the Similarity is measured among the paraphrased sentences. The detail algorithm is given below.

PROPOSED METHODOLOGY

Original sentences = 40, paraphrased = 120 here for each original sentences, 3 paraphrased sentences.

Here, OS [] for original sentences, PS [] for paraphrased sentences.

Algo: similarity measurement in semantically paraphrased Marathi sentences

Step1: [initialize counter variable]

K = 1, N = 1, R = 1

Step2: Repeat Step 3 & 6 While K <= 120

While N ≤ 3

Step3: cos [R] = Measure between OS [K] & PS[N]

Step4: N = N+1, R = R+1

Step5: k = K+1

Step6: exit

Figure1: Procedure for similarity measurement in semantically paraphrased Marathi sentences

4. RESULT AND ANALYSIS:

In this experiment, we employed the cosine similarity method to determine semantic similarity between Marathi paraphrased sentences. We created our own dataset for the suggested system. A total of 160 sentences are used, of which 40 are original and 120 are paraphrased sentences derived from the original text. The values in the table below represent the cosine similarity of the original and paraphrased sentences. The cosine similarity result falls between 0 and 1. Here, 0 denotes sentence dissimilarity and 1 reflects total similarity to original sentences. And the value between 0 and 1 represents the degree of similarity between the paraphrased sentences and the original sentences. The Average of cosine similarity at paraphrased sentences (ps1, ps2, ps3) level is 0.35,0.35,0.35 respectively.

Table1: Result for Analysis of Paraphrased Sentences in Marathi

Sr. NO	Original sentences[OS]	Phrased sentence1	Phrased sentence2	Phrased sentence3	Cosine Result		
1	आपण सध्याच्या कामावर समाधानीआहात का	आपणास सध्याचेकाम आवडले का.	आपण सध्याच्या कामातआनंदी आहात का	आपण सुर आसलेल्या कामात संतोषीआहात का	0	0.5	0.5
2	तुमचीशेती बागायती आहे .	तुमचे रान पाण्याखालीआ हे	तुमची जमीनपाण्याखा ली आहे.	तुमचेक्षेत्रबागाय ती आहे.	0	0.308	0.154
3	महिला कार्यक्रमासउप स्थितहोत्या	महिला कार्यक्रमासस हभागी होत्या.	महिला कार्यक्रमास हजरहोत्या.	स्त्रीयाकार्यक्रमा ससहभागी झाल्या.	0.156	0	0.44
4	थोड्यावेळात कार्यक्रम सुरु होईल.	लवकरच कार्यक्रमाससुर वात होईल.	थोड्यावेळातका र्यक्रमप्रारंभहोई ल.	कार्यक्रमलवकर चचालूहोईल	0.22	0.5	0
5	आमच्यासंघाचा पराभव झाला.	आमच्या संघाला हारपत्करावा लागली .	आमच्या समूहाचा अपयश झाला.	आमचागटअपय शीठरला.	0	0.5477	0.5477
6	चार लोकांसमोर भाषण करायला धैर्य लागते.	चार लोकांसमोर भाषणकराय ला धाडस लागते.	चारलोकांसमोर वक्तृत्वासाठीमे हनतलागते.	चार माणसासमोर भाषणकरायला साहस लागते.	0.99	0.799	0.199
7	वडाचे झाड	वडाचे झाड	वडाचे झाड	वडाचे	0.79	0.19	0

	आकाराने मोठे असते .	आकारेनेमहा काय असते.	आकाराने विशाल असते.	वृक्षप्रतीरूपाने अपार आहे.			
8	अवनीचे आपल्यावर फार उपकार आहेत	पृथ्वीचे आपल्यावर फारउपकार आहेत .	धरतीचेआपल्या वरखूपऋणआहे .	वसुंधराची आपण ऋणीआहोत.	0.59	0.59	0.59
9	तुमच्याकडे कोणत्या जातीचे पिक आहेत.	तुमच्याकडे कोणत्याप्रकार चे पिकअसते.	तुमच्याकडे कोणत्या प्रकारच्यावन स्पती आहेत.	तुमच्याकडे कोणत्याप्रकारचे वान आहेत.	0.59	0.833	0.46
10	वर्षातून कितीवेळ आपला व्यावसाय चालू असतो .	वर्षातून कितीकाळ आपलाव्याव साय चालूअसतो.	वर्षातूनकितीदि वसआपलाधंदा सुरुअसतो	वार्षिककितीअव धीआपलाधंदासु रुअसतो.	0.59	0.59	0.59
11	तुमचा मुख्यव्यवसाय कायआहे.	तुमचा प्रमुख धंदा काय आहे.	तुमचा प्रथम व्यापार काय आहे.	तुमचा मूळ उद्योग काय आहे.	0.66	0.66	0.66
12	अपुरे शासकिय अनुदान .	कमीशासकिय अनुदान.	अल्प शासकिय अनुदान.	अपूर्णसरकारी निधी.	0.39	0.19	0.18
13	त्यांनी खूप मायाजमवली आहे.	त्यांनी खूप धन जमविले आहे.	त्यांनी आतिशय पैसे जमवले आहे.	त्यांनी अत्याधिक रकम जमा केली आहे.	0.66	0.33	0.33
14	त्यांना भेट दया.	त्यांना नजरांना दया.	त्यांनी आतिशय पैसे जमवले आहे.	त्यांनी अत्याधिक रकम जमा केली आहे.	0	0	0
15	त्यांनाबशीसीद्या.	त्यांना भेट दया.	त्यांना नजरांना दया.	त्यांनाउपहार द्या.	0.25	0.32	0.56
16	घरदार नसणारा.	तो उपरा आहे	तो बेघर आहे.	निवासरहित आहे	0.59	0.23	0.22
17	पक्षाने त्यांनाअर्धचंद्र दिला	पक्षाने त्यांची हाकालपट्टी केली .	पक्षातून त्यांना काढून टाकले.	संघातून त्यांना हटवले.	0.63	0.66	0.66
18	फुलेवयानेछोटेअ	फूले ही अल्प	फूले ही कमी	फुले हि लहान	0.59	0.32	0.39

	सतात.	आयु असतात.	वयाचीअसतात.	असतात.			
19	धोनी हा विविध बाबीतप्राविण्य असलेला खेळाडू आहे.	धोनी हा आस्तपैलू खेळाडू आहे	धोनीहाविविधते नेनुपूनअसलेला खेळाडूआहे .	धोनीहाविविधक्षे त्रातीलप्रतीभाशा लीखेळाडूआहे.	0.24	0.33	0.054
20	त्यांनी पितांबरधारण केले होते.	त्यांनी पिवळे वस्त्र्य धारणकेले.	त्यांनीपिवळ्या वस्त्रपरिधानके ले.	त्यांनीपीतवर्णक पडेघातले.	0.39	0.39	0.39
21	कालपासून निरंतरपाऊस येतोय	कालपासून वर्षा चालू आहे	कालपासून पावसाची झाड लागली आहे.	कालपासून पाऊस सतत येत आहे	0.39	0.39	0.44
22	ती देखणी आहे.	तीसुंदर आहे.	ती छान आहे.	तीआकर्षकआहे.	0.66	0.66	0.66
23	मला जरा निजुदे.	मला जरा झोपुदे.	मलाथोडावेळआ रामकरुदे.	मला थोडी निद्रा घेऊ दे	0.66	0.25	0.25
24	सुर्य अस्ताला गेला .	सूर्य मावळतीला गेला.	सूर्यास्त झाला.	भास्कर दूर झाला.	0.33	0	0
25	तो रुबाबदारदाखवतो .	तो मिजास दाखवतो.	तो दिमाख दाखवतो.	तोताठा दाखवतो.	0.33	0.45	0.34
26	तुमची जागा प्रशस्त आहे.	तुमची जागा विस्तीर्ण आहे	तुमची जागा ऐस पैस आहे	तुमचीजागा अमर्याद आहे .	0.66	0.66	0.66
27	भारताचे रोड ओबडधोबड आहेत.	भारताचे रोड खडबडीत आहेत .	भरतेच रोड रांगडे आहेत.	भारताचे रोड बेडोल आहेत.	0.17	0.45	0.24
28	उसाचे खूप ओझे आहे.	उसाचे खूप वजन आहे.	उसाचा खूप बोजा आहे.	उसाचा खूप भार आहे.	0.28	0.28	0.28
29	माणसाचाशेवट होतो.	माणसाचाशेव ट होतो.	माणसाचामुर्त्यू होतो.	माणसाचामुर्त्यू होतो.	0.12	0.12	0.12
30	विस्तवावर पाय ठेऊ नका.	निखारायवर पाय ठेऊ नका .	अंगारायावर पाय ठेऊ नका.	इंगळ पाया खाली घेउनका.	0.45	0.23	0.34
31	प्रकाश अंधुक आहे.	प्रकाश अंधुक आहे.	प्रकाशमंद आहे.	प्रकाशमंद आहे.	0.32	0.31	0.39
32	प्रकाशमंद आहे.	प्रात: काल झाला.	सकाळ झाली .	उषा झाली.	0	0	0
33	काय उणीव आहे.	काय	कायन्यूनता	काय कमीपणा	0.147	0.16	0.146

		कमतरताआहे.	आहे.	आहे.			
34	त्यंची चेष्टा करा.	त्यांचीथटटा करा.	त्यांची मस्करी करा.	त्यांचा उपहास करा.	0.13	0.12	0.17
35	उमेदीने काम करा.	हिमतीने काम करा.	उत्साहाणे काम करा.	धेर्याने काम करा.	0.16	0.56	0.59
36	अधीर मन झाले.	उत्कठिंत मन झाले .	आतुर मन झाले.	उत्सुक मन झाले.	0.12	0.18	0.19
37	खूप कर्ज आहे.	खूपऋण आहे.	खूप रीण आहे.	खूप देन आहे.	0.17	0.15	0.13
38	मन एकचीत ठेवा.	मन स्थिर ठेवा.	मन एकाग्र ठेवा.	मनऐकतान ठेवा.	0.16	0.15	0.9
39	माझ्यावर कृपा आहे.	माझ्यावर उपकार आहेत.	माझ्यावर एहसान आहे.	माझ्यावर दया आहे.	0.51	0.56	0.57
40	हा माझा बंदु आहे	हा माझा सखा आहे	हा माझा मित्र आहे.	हा माझा दोस्त आहे .	0.25	0.24	0.26

CONCLUSION

Paraphrasing is critical and important phase in text summarization, question answering systems. We have created Corpora for Marathi language due to the unavailability of Marathi datasets. Cosine similarity is used to analyse the semantic similarity of the paraphrased sentences. The cosine similarity score ranges from 0 to 1. In this case, 0 signifies different sentences and 1 denotes entire similarity to the original sentences. A value between 0 and 1 represents the degree of similarity between the related paraphrased sentences and the original sentences. We have obtained Average of cosine similarity at Paraphrased sentences (PS1, PS2, PS3) level is 0.35,0.35,0.35 respectively.

REFERENCES

1. Mihalcea, Rada, Courtney Corley, and Carlo Strapparava. "Corpus-based and knowledge-based measures of text semantic similarity." In *Aaai*, vol. 6, no. 2006, pp. 775-780. 2006
2. Li, Yuhua, David McLean, Zuhair A. Bandar, James D. O'shea, and Keeley Crockett. "Sentence similarity based on semantic nets and corpus statistics." *IEEE transactions on knowledge and data engineering* 18, no. 8 (2006): 1138-1150.
3. Gunasinghe, U. L. D. N., W. A. M. De Silva, N. H. N. D. de Silva, A. S. Perera, W. A. D. Sashika, and W. D. T. P. Premasiri. "Sentence similarity measuring by vector space model." In *2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 185-189. IEEE, 2014.
4. Saad, Sazianti Mohd, and SitiSakiraKamarudin. "Comparative analysis of similarity measures for sentence level semantic measurement of text." In *2013 IEEE international conference on control system, computing and engineering*, pp. 90-94. IEEE, 2013.
5. Atoum, Issa, and Ahmed Otoom. "Efficient hybrid semantic text similarity using WordNet and a corpus." *Int. J. Adv. Comput. Sci. Appl* 7, no. 9 (2016): 124-130.
6. Atoum, I., &Otoom, A. (2016). *Efficient hybrid semantic text similarity using WordNet and a corpus. Int. J. Adv. Comput. Sci. Appl*, *7*(9), 124-130
7. Lee, J. C., &Cheah, Y. N. (2016, August). Paraphrase detection using semantic relatedness based on Synset Shortest Path in WordNet. In *2016 International Conference On Advanced Informatics: Concepts, Theory and Application (ICAICTA)* (pp. 1-5). IEEE.