A Proposed Model to Identify Paraphrasing in Marathi Text

[1]Ramesh R. Naik, [2]Maheshkumar B. Landge, [3]C. Namrata Mahender

[1,2]Research scholar, [3]Assistant professor, Department of CS and IT, Dr.B.A.M.University, Aurangabad (MS), INDIA.

[1]ramesh.naik31@yahoo.com,[2]maheshkumar.landge@gmail.com [3]nam.mah@gmail.com

Abstract— **Paraphrasing is an essential resources of linguistic and literature, as it provides the power of expression such as poems, stories. It also becomes confusing or difficult in some context like proverb those have dual means, moral of the stories. Even it is noteworthy to understand reinterpretation of same sentence may be different by different people. Thus to convey desirable semantics of a word/sentence. Paraphrasing is very important. When working with Marathi language many difficulties comes due to the linguistic aspect of language: 1) Marathi language is agglutinative, 2) Need of Dependency parser as it is object based, 3) Contextually some words can change meaning in a sentence, 4) Adjectives do not inflect under they end in Long /a/, in which case they agree with nouns in gender, number and case. In this paper we are proposing a doc2vec and word2vec based model for identifying the regions of paraphrasing in a given Marathi text.**

Keywords— word2vec; doc2vec; paraphrasing; plagiarism detection.

I. INTRODUCTION

The world of words, sometimes are so magical that it attracts the listener for eg which takes us in an rthymic journey, sometimes to complex to understand the real sense behind it or even difficult to reinterpret the same like proverbs.in real time expression we need to say same things to people, the sentences have same meaning but vary in their form. The understand it just consider

Many a times when things has to be explained to an adolescent will be different than adult not compulsory.

Thus need of paraphrasing is always there. How to do paraphrasing and why it is so difficult to work in automatic processing of a language.

A. *paraphrase an abstract view*

Paraphrase is a statement of the meaning of a text or passage using other words. The term itself is derived via Latin paraphrasis from Greek, meaning “additional manner of expression” and the process of paraphrasing is called “paraphrasis”

Rahul Bhagat and Edword Hovy (2013) [1] have identified 25 types of paraphrases with each class having its own specific way of retaning the requirement of strict semantic equivalent.

TABLE I. 25 CLASSES OF PARAPHRASE.

Sr No	Type	Method	Example
1	Synonym substitution :	Replacing a word/phrasal idiom by a synonymous word/set idiom , in the appropriate linguistic context of use , resolution in a paramusical musical idiom of the archetype prison term /musical phrasal idiom Aries the	Ram is fat. ⇔ Ram is chubby.
2	Antonym substitution	Replacing a word/phrase by its antonym accompanied by a negation or by negating some other word, in the appropriate context, consequence in a paraphrasis of the original sentence/phrase.	I am sad ⇔ I am not happy
3	Converse substitution :	Replacing a word/phrase with its converse and inverting the relationship between the element of a	Google buy YouTube. ⇔ YouTube was sell to Google.

DOI : 10.18231/2454-9150.2018.0871

		sentence/phrase, in the appropriate context, results in a paraphrasis of the original sentence/phrase, presenting the situation from the converse Perspective This substitution may be accompanied by the addition/deletion of appropriate function Logos and sentence restructuring	
4	Change of voice:	Changing a verb from its active to passive form and frailty versa answer in a paraphrasis of the archetype time /phrasal idiom.	Geeta is esteemed by students⇔ students respect Geeta
5	Change of person:	Changing the grammatical somebody of a referenced object resultant role in a paraidiomatic expression of the pilot condemnation /phrasal idiom Rap said, "I like	Pat said, "I like football." ⇔ Dab said that he liked football.
6	Pronoun/Co-referent substitution :.	Replacement a pronoun by the noun phrasal idiom it co-refers with outcome in a paraphrasis of the original judgment of conviction /phrase	Pat the likes of Chris, because she is smartness. ⇔ Pat likes Chris, because Chris is smart.
7	Repetition/ Ellipsis:.	Ellipsis or elliptical construction results in a paraphrase of the original condemnation /phrase.	Pat can foot race fast and Chris can run fast, too. ⇔ Pat can run fast and Chris can, too.
8	Function word variations:	Changing the function Word of God in a time /phrase without affecting its semantics, in the appropriate context, context of use , results in a paraphrase of the original sentence/phrase	Pat showed a nice demo. ⇔ Pat's demo was nice.
9	Actor/Action substitution :	Replacing the name of an action by a word/phrase denoting the person doing the action (actor) and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase.	
10	Verb/"Semantic-role noun" substitution :	Replacing a verb by a noun corresponding to the agent of the action or the patient of the action or the instrument used for the action or the medium used for the action, in the appropriate context, results in a paraphrase of the original sentence/phrase.	
11	Manipulator/Device substitution :	Manipulator/Gimmick substitution: Replacing the name of a twist by a word/phrase denoting the person using the device (operator) and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase.	The pilot took off despite the stormy atmospheric condition condition . ⇔ The plane took off despite the stormy weather.
12	General/Specific substitution :	Replacement a word/set phrase by a more general or more specific word/phrase, in the appropriate circumstance , final result in a paraphrasis of the archetype time /phrase	Pat is flying in this weekend. ⇔ Pat is flying in this Saturday.
13	Metaphor substitution :	Replacement a noun by its standard metaphorical use and frailty versa, in the appropriate context , resultant in a paraphrasis of the master copy sentence/phrase.	Immigrants have used this network to send immediate payment . ⇔ Immigrants have used this network to send stashes of cash.
14	Part/Whole substitution :	Replacement a part by its corresponding whole and frailty versa, in the appropriate context, results in a paraphrase of the master sentence/phrase.	American plane pounded the Taliban defenses. ⇔ American airforce pounded the Taliban defenses.
15	Verb/Noun conversion:	Replacing a verb by its corresponding nominalized noun frame and frailty versa, in the appropriate context, results in a paraphrase of the original	The police interrogated the suspects. ⇔ The police subjected the suspects to an interrogation.

DOI : 10.18231/2454-9150.2018.0871

		sentence/phrase.	
16	Verb/Adjective conversion: .	Verb/Adjective conversion:. Replacing a verb by the corresponding adjective form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase	Pat loves Chris. ⇔ Chris is lovable to Pat.
17	Verb/Adverb conversion:	Replacing a verb by its corresponding adverb form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase.	Pat boasted about his work. ⇔ Pat wheel spoke boastfully about his work.
18	Noun/Adjective conversion:	Replacing a verb by its corresponding adjective form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase.	I'll fly by the end of June . ⇔ I'll fly late June.
19	Verb-preposition/ Noun substitution :	Replacing a verb and a preposition denoting location by a noun denoting the location and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase.	The finalists will play in Giants stadium. ⇔ Giants stadium will be the playground for the finalists.
20	Change of tense:	Changing the tense of a verb, in the appropriate context, results in a paraphrase of the original sentence/phrase.	
21	Change of aspect:	Changing the tense of a verb , in the appropriate context of use , result in a paraphrasal idiom of the archetype prison term /phrase . twenty-one Change of facial expression : Changing the aspect of a verb, in the appropriate context, results in a paraphrase of the original sentence/phrase. Pat solvent in a paraphrase of the master copy condemnation /phrase.	Pat is flying in today. ⇔ Pat flies in today.
22	Change of modality:	Add-on /deletion of a modal verb auxiliary or substitution of one modal by another, in the appropriate context, results in a paraidiomatic expression of the original sentence/phrase.	The government wants to boost the economy. ⇔ The government hopes to boost the economy.
23	Semantic implication:	Replacing a discussion /phrase denoting an action at law , event , and so forth, by a word/phrase denoting its possible future effect, in the appropriate context, answer in a paraphrase of the pilot sentence/phrase.	The Marines are fighting the terrorists. ⇔ The Marines are eliminating the terrorists.
24	Approximate numerical equivalences:	Replacement a numerical grammatical construction (a word/phrase denoting a telephone number , often with a building block of measurement) by an approximately equivalent numerical expression (even perhaps with change of unit), in the appropriate context, results in a paraphrase of the original time /phrase.	Disneyland is 32 miles from here. ⇔ Disneyland is around 30 minutes from here.
25	External knowledge:	Replacing a discussion /idiom by another word/phrasal idiom based on extra-linguistic (globe) knowledge, in the appropriate context, results in a paraphrase of the original sentence/phrase.	We must work hard to win this election. ⇔ The Democrats must work hard to win this election.

Table1: 25 classes of Paraphrase by Bhagat, R., &Hovy, E. (2013)[1]

B. *Applications*

Under the preview of natural language processing many applications requires paraphrasing.

1) Natural language understanding

2) Machine Translation

3) Summarization

DOI : 10.18231/2454-9150.2018.0871

4) Plagiarism Detection

5) Grammer Checker

6) Data mining algorithm

7) Opinion Mining

8) Sentiment Analysis

II. Paraphrasing in Marathi Language

i) Marathi language is agglutinative

ii) Need of Dependency parser as it is object based

iii) Contextually some words can change meaning in a sentence

iv) Adjectives do not inflect under they end in Long /a/ , in which case they agree with nouns in gender, number and case.

III. Document to Vector and word to Vector

A. *Document to Vector*

Doc2Vec is an unsupervised learning algorithm, which aims to find the embeddings of documents. It is Similar to Word2Vec, there are two Doc2Vec models, namely, Distributed Memory (similar to CBOW) model and Distributed Bag of Words (similar to Skip-Gram) model. While the latter ignores word ordering, the former keeps it by concatenating the paragraph vector and word vectors in order to predict the next word in the given context. Doc2Vec algorithm has two advantages; i) it preserves word order and ii) it is an unsupervised learning algorithm.

B. *Word2vec*

Word2vec is widely used in natural Language processing.Word2vec is a word embedding method that takes a corpus of words as input and produces vectors as output.word2vec having two models which are continues bag of words and Skip-gram[3]. The difference between them is the word order, which is followed in Skip-gram and ignored in continues bag of words [4].In this paper , we have taken the result by using gensim python library[5]. We first build a dictionary from the whole training data then each word attaches a vector and it generates word vectors.

IV. Proposed Model

A. *Algorithm*

1. We have taken Marathi text document as an input.

2. Apply tokenization on input documents.

3. Word to Vector and Doc to Vector Generation.

4. Calculate the Similarity ratio in Documents by using Different Similarity Measures.

5. Identify Paraphrasing.

V. Conclusion

Paraphrasing is to state something written in different words. This paper covers the brief introduction of paraphrasing. There are 25 classes in paraphrasing. The paraphrasing can be used in different applications, like plagiarism detection, summarisation, opinion mining and grammar checker. Word to vector is a word embedding method that takes a document of words as input and produces vectors as output. Doc2vec is an extension to word2vec for learning document embeddings. The vector Produced in word to vector and doc to vector is used for identifying paraphrasing in Marathi text Documents.

Acknowledgment

We are thankful to CSRI DST Major Project sanctioned No.SR/CSRI/71/2015(G), Computational and Psycholinguistic Research Lab and Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (Maharashtra) for providing facility for carrying out the research.

References

[1] Bhagat, R., &Hovy, E. (2013) what is a paraphrase? Computational Linguistics, 39(3), 463- 472.

[2] Le and T. Mikolov. (2014). Distributed representa-tions of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014) , pages 1188–1196, Beijing, China.

[3] Mikolov, Tomas, et al. (2013)”Discovering word senses from text using 1random indexing. ”arXiv preprint arXiv: 1301.3781.

[4] Rong, Xin. (2014)”word2vec Parameter Learning Explained. ”arXiv preprint arXiv: 1411.2738.

[5] R. ehek, P. Sojka, Proceedings of the LREC (2010) Workshop on New Challenges for NLP Framework. ELRA, Valletta, Malta, 2010, pp.45-50.

DOI : 10.18231/2454-9150.2018.0871