RESEARCH ON PARAPHRASE IDENTIFICATION

Prachi Kalbhor[1], Gautami Patil[2], Somil Agarwal[3], Anshit Singh Rajput[4], Professor. Prajakta Dhamdhere[5]

[1,2,3,4]*Student, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*
[5]*Professor, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*

---***---

Abstract - *In the natural language system, the identification of paraphrases plays a critical role. As a result of this research, we used an immersive representation to model the interaction between two sentences not just at the word level, but also at the expression and phrase level, by employing a convolutional neural network, recurrent neural and multihead attention neural network to conduct paraphrase detection using semantic characteristics at the same time. The most important factors are semantic equivalence and similarity. Paraphrasing methods find, create, or extract sentences that express nearly the same content. The identification of paraphrases will discern various worded sentences that have the same meaning. Textual statements that use different surface types to communicate the same context are known as paraphrases. Paraphrase identification is important since it helps with a variety of NLP activities, including text summarization, document clustering, query response, inference of natural language, knowledge retrieval, plagiarism detection, and text simplification. The aim of the paper was to compile a list of all the methods, techniques, and current developments for detecting paranormal activity. Not only can detection be used to address a text's identity and protect its context, but it can also be used to provide a metric for analysing a text's computer translations. The existing available requests fail to verify the authenticity of a text if it is paraphrased and fails to mark it as plagiarised. Text mining, text summarization, plagiarism identification, authorship verification, and question answering all include the ability to detect identical sentences written in natural language. The aim is to determine if two sentences are semantically similar. An significant takeaway from this research is that current parasystems function well when put to use. We will use already proven conventional algorithms to identify whether the content is a copy of an existing work, and we will use our application to determine whether the content has been paraphrased in some way.*

Key Words: Paraphrase detection, Sentence similarity, Deep learning, RNN, CNN, MAN, Semantic Similarity

1.INTRODUCTION

When we hear a sentence, we don't have to retrain our brains to comprehend it. Is it possible for an algorithm to take this into account? As a result, the neural network idea was born. The principle of neural networks emphasises the importance of learning. The degree to which linguistic words, such as documents or sentences, are semantically identical is known as semantic similarity. The calculation of resemblance between documents or language is called semantic textual similarity. Semantic textual similarity can be used for a variety of tasks, including identifying paraphrases. A paraphrase on how to say the same thing in a different way without missing the essence. For paraphrasing, two types of "Para Generation" and "Para Detection" are used. Parametric identity is defined in semantic terms. The process of deciding if two sentences have exactly the same meaning is known as paraphrase identification. Many natural language applications, such as text summarization, query answering, computer translation, natural language creation, and plagiarism detection, have been shown to benefit from this challenge. Simplify the comparison to detect semantic similarities between two texts written in the same language. As a result, we suggest a framework for detecting semantic similarities between two paraphrases of the same language that incorporates neural models. In conjunction with Convolution Neural Network, Recurrent Neural Network and Multi-head Attention Network mechanism will be proposed to aid the model in learning relevant information in different presentation subspaces for better architectural performance.

1.1 Problem Statement

Paraphrase Identification (PI) problem is to classify that whether or not two sentences are close enough in meaning to be termed as paraphrases. The problem we aim to address is to develop a model, which can reliably provide an unit of measure as to whether the two sentences are paraphrased or non-paraphrased.

In particular, we aim to at answering the following question : Whether the two texts are semantically similar & If so, is the system capable enough to tell the results ?

1.2 Motivation

Natural Language Processing (NLP) is the study of how computers can analyse, comprehend, and produce natural human languages. Paragraphs may be broken down into terms, phrases, and even sentences. Major difficulties faced in natural language processing is ambiguity where the same text has several possible interpretations. To work on a project that solves a real-world challenge while still addressing the issues listed above. Deep learning architectures, neural networks, and machine learning are all recent ideas in NLP.

1.3 Summary

The output of this study evaluated the actual relationship between neural networks as classifier. Overall, what can we learn from this project is, we can the prediction of probability for our classified model with coupled architectural model by comparing. The model can be used to apply on another system to get more accurate results.

2. RELATED WORK

2.1 Paraphrases for Pattern Learning

In framework of open-domain computing, paraphrase have arisen as a valuable method for learning IE patterns. Given that the majority of this is focused on the majority of current paraphrase learning methods are geared to studying paraphrases of two arguments. However, if only one statement is to be extracted, using these two-argument methods results in a very poor recall. Our system knows paraphrases with a single argument because our emphasis in this paper is on the one-argument case. Since two arguments have better restrictions on the contexts they fit and on the sequence boundaries, learning in our case is different and arguably harder. In our case, incorrect boundary identification means that our patterns would be either too broad or too narrow. Scaling the learning algorithm is also more difficult in our situation since there are far more possible phrases for one statement than there are with two. However, there has been some previous research on understanding paraphrases for one statement.

All of the methods decode the text with a dependence parser and learn similar paths to learn syntactic paras or entailments. Our first tool, on the other hand, knows surface-level paraphrases, making it easily scalable to broad corpora and allowing us to use various stages of post-processing. Our second approach uses shallow parsing to practise lexico-syntactic paraphrases, which is many times faster than absolute parsing. No one, to the best of our experience, has ever taught paraphrases in the way we do.

2.2 Paraphrases for Information Extraction

Although pattern-based knowledge extraction has been prominent since the early 1990s, the early work concentrated on manual pattern construction. However, the focus soon turned to automatically learning patterns from domain-specific corpora. Annotated teaching corpora were used in one line of research to learn these patterns. However, since this requires a lot of repetitive manual annotation, weakly-supervised methods that need very little annotation are becoming more common. These and other domain-specific IE methods use domain-specific corpora to learn patterns. Their reliance on domain-specific corpora makes it difficult for these approaches to be easily transferred to new domains. However, unlike previous methods, the pattern-learning framework we propose here does not include a domain-specific corpus to learn patterns; instead, it learns patterns from a general broad-coverage corpus. Furthermore, our system only extracts patterns from a large corpus once. Using only a few seed patterns, patterns can be created for any (new) domain.

3. PROPOSING SYSTEM

The presented System Architecture is using the semantic approach with classifier:

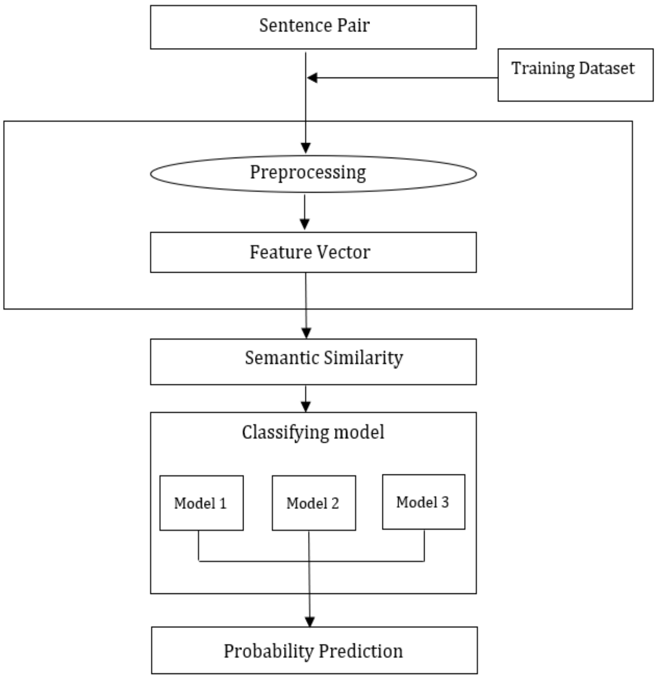

Fig -1: Architectural Program

Semantic Similarity: Orthogonality is one of the standard space vector's properties, which means that all of the dimensions are perpendicular to one another. To put it another way, we believe that each word in our language is distinct from the others. Despite the existence of synonyms or closely related terms in languages, this statement is clearly incorrect. This problem is thought to be caused by a flaw in the standard vector space model's ability to compute semantic similarity when the text has been paraphrased. The concept of semantic similarity, as opposed to lexicographical similarity, is a metric specified over a collection of documents or words, where the idea of distance between objects is centred on the resemblance of their context or semantic material.

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

Manhattan similarity: The system checks one similarity technique for semantic level of similarity calculation. This makes the system not so accurate on the level of Semantic similarity. The Manhattan distance is a metric that calculates the distance between two points as the sum of their Cartesian coordinates' absolute differences. To put it another way, it is the complete number of the differences between the x- and y-coordinates.

2.1 Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs) are neural networks with one or more convolutional layers that are primarily used for image recognition, detection, segmentation, and other auto-correlated data. Convolution is the process of sliding a filter over an input signal. What is the greatest benefit of using CNN? There is less reliance on pre-processing, which reduces the amount of human work required to create the functionalities. It is easy to comprehend and execute. It forecasts images with the greatest precision of any algorithm. Because of their high precision, CNNs are used in image detection and identification. The CNN uses a hierarchical model that builds a network in the shape of a funnel and then outputs a fully-connected layer where all the neurons are connected to one another and the data is stored. CNNs are mainly used for image detection and identification, as you can see. The potential to convolutionalize is a CNN's specialisation. The scope for further applications of CNNs is infinite, and it must be pursued and driven to new limits to learn all that this sophisticated machinery is capable of. Convolutional Neural Networks (CNN) are commonly used in computer vision and have been effective in tasks such as facial recognition, self-driving vehicles, and handwriting recognition. Recently, they've been used in a number of natural language processing tasks, especially classification problems, such as author profiling, personality recognition, paraphrase recognition, and sentiment analysis.

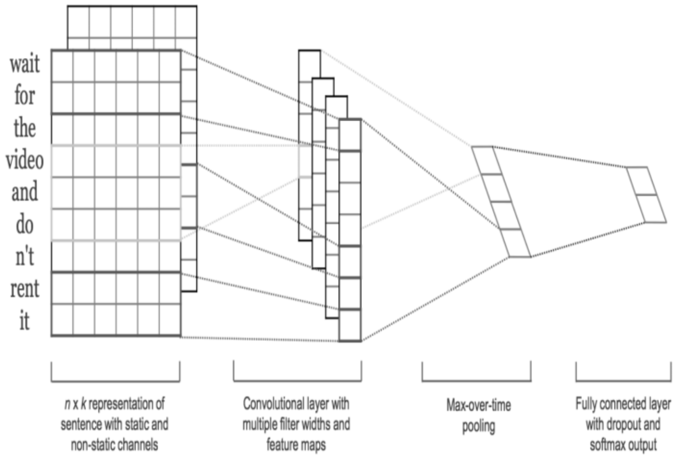
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Fig -2: Convolutional Neural Network

2.2 Recurrent Neural Network (RNN)

RNNs(recurrent neural networks) are a type of neural network that can be used to model sequence data. RNNs, which are derived from feedforward networks, behave similarly to human brains. Simply put, recurrent neural networks can model sequential data in a way that other algorithms can't. RNNs are a kind of artificial neural network in which nodes' connections form a directed graph in a sequential order. Since text is naturally sequential, this architecture allows RNN to display temporal actions and capture sequential data, making it a more "simple" solution when working with textual data. When dealing with sequential data such as text, audio, or video, recurrent neural networks (RNN) or sequence models in general are extremely helpful. Speech processing, sentiment classification, and machine translation are only a few of the issues that have been successfully solved using these models. Since they perform the same series of operations on any part of the chain, these neural networks are called "recurrent." RNNs' key feature is their ability to remember things they've seen before in the series. Part-of-Speech tagging models, for example, output a tag for each input term (many-to-many), sentiment analysis returns a class given a sequence of words (many-to-one), a text generator generates a sequence from an input word (one-to-may), and computer translation translates a sequence of words after the first has been completely processed (second type of many-to-many).

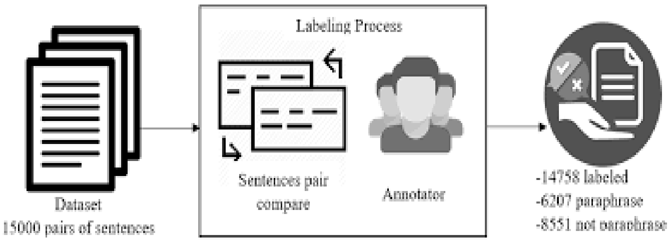

Fig -3: Recurrent Neural Network

2.3 Multihead Attention Neural Network (MAN)

In terms of mathematics, it entails paying attention not only to the individual words in the sentence, but also to different segments of the words. The words vectors are divided into a fixed number of chunks (h, number of heads), and then self-attention is applied to the corresponding chunks, yielding h background vectors for each word. Concatenating all of the context vector results in the final context variable. Multi-head attention focus to help the model to jointly attend to data from several representation subspaces at different locations. Averaging prevents this with a single focus attention head. Attention takes two sentences, converts them into a matrix in which one sentence's changes shape the columns and another sentence's changes form the rows, and then allows connections to search similar background. This is particularly useful when it comes to machine translation. You don't have to use only your attention to match meanings of sentences in two different languages. You may also place the same sentence in the columns and rows at the same time to see how many elements of the sentence apply to each other. As a result, studying helps you to see at the end of a sentence and draw connections between any given word and its meaning. This differs significantly from the small-memory, upstream-focused RNNs, as well as the proximity-focused convolutional networks.

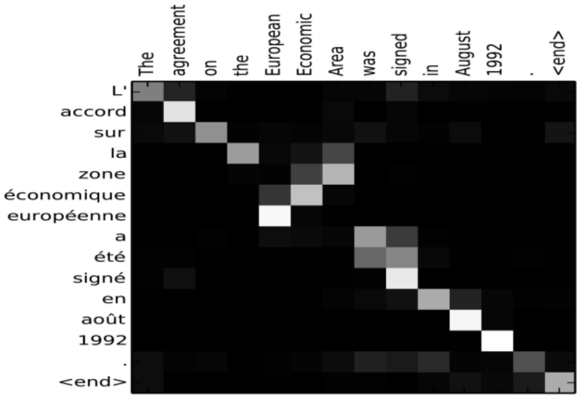

Fig -3: Multihead Attention Neural Network

3. DATASET

We conducted experiment on Quora Question Pair dataset ,each comprising a large set of instances in the form of qid1, qid2, where qid1 and qid2 are two questions, and question1, question2 indicating the full text of each question and is_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise. Table I shows a brief description of these datasets.

- Quora Question Pair dataset is an open-domain English dataset derived from Quora.com.

The aim of this dataset is to figure out which of the given pairs of questions contains two identical questions. Kaggle also added computer-generated query pairs to the test collection as an anti-cheating tool. Many of the questions in the training package are real Quora questions. Quora required an automatic means of identifying redundant query pairs at a scale of millions to minimise the inefficiency of redundant posts. About 400,000 lines of possible query repeat pairs make up the dataset. Each line includes the IDs for each question in the pair, as well as the full text of each question and a binary value indicating if the line contains a repeat pair.

Table -1: Dataset used

Dataset	Source	Items	Type
QQP	Quora Question Pairs	4,00,000 sentence pairs	Corpa

Corpus - A corpus is a vast group of texts. Texts in a single language (monolingual corpus) or text data in different languages may be included in a corpus (multilingual corpus). A linguistic research is focused on a corpus of written or spoken content.

3.1 Risk Analysis

Following are the details for each risk.

Table -2: Risk Analysis 1

Risk ID	1
Risk Description	Large set of dataset
Source	Hardware requirement
Probability	High
Response	Accept
Strategy	Sufficient dataset taken in consideration
Risk status	Identified

Table -3: Risk Analysis 2

Risk ID	2
Risk Description	Virtual Environment
Source	Technology
Probability	High
Response	Mitigate
Strategy	Proper libraries must be included
Risk status	Identified

4. RESULTS

The comparative findings for the legitimate achievement of the target on QQP dataset is given below. All of the networks we evaluated were real-time networks with a probability to the respective model. To conclude our results multihead attention network performs better than the convolutional neural network and recurrent neural network in all sequences. These results are based on probability factors via Training of dataset. Not only the perfomance of the MAN network is higher than CNN and RNN networks on the QQP dataset, but also the MAN was trained in a 00:03:05.27 epoch time, while the CNN AND RNN network was trained in 00:00:18.68 and 00:01:23.85 epoch time respectively. There is an obvious benefit to employing MAN networks for general-purpose applications. In addition, we present a comparison of running time of the most recent deep network work performing in a static dataset. We provide this comparison to show that although these networks perform extremely well on a specific dataset, we in future can conclude on various datasets too. The data presented in public reports and data gathered informally on the best outcomes produced in a given dataset.

MAN reported 0.803 test accuracy as compared to CNN and RNN with 0.705 and 0.786 respectively. We estimated probablities based on the speed of our networks with average results in ratios.

5. CONCLUSIONS

Finally, we do an in-depth review of state-of-the-art Deep Learning techniques. We examine their results by comparing training and validation accuracy, as well as adding missing data to their models. We discussed the difficulties and problems that Deep Learning methods for paraphrase recognition face. Integrating paraphrase detection into a plagiarism detection scheme and using our proposed genetic algorithm to refine the deep learning model's hyper-parameters is one of the areas of potential research. In order to train state-of-the-art models on larger datasets and apply transfer learning to the particular task at hand, it is necessary to train state-of-the-art models on larger datasets. Combining neural network models with knowledge base approaches is another avenue to investigate in order to reduce the need for comprehensive databases to cover all imaginable scenarios. We'll also assess combinations of proposed models and investigate their classification errors in depth. This observation leads to the conclusion that acquiring a knowledge-base containing paraphrases is an effective general methodology for information extraction. It also points to the need of building more such knowledge resources for other NLP applications.

This paper presents a sustainable approach to the problem of paraphrase identification. Our method makes use of similarity measures applied differently from previous approaches. The system was evaluated on the Quora Question Pairs Corpus and found to outperform reported approaches.

ACKNOWLEDGEMENT

First and foremost, we would like to thank our superiors, Dr. Rekha Sugandhi and Prof. Prajakta Dhamdhere, for their invaluable assistance in this project. We appreciate all of their efforts to our study and for sharing their vast expertise. We are eternally grateful for their unwavering support for any mission and partnership we have undertaken.

REFERENCES

[1] Berant, Jonathan, and Percy Liang. "Semantic parsing via paraphrasing." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Fernando, Samuel, and Mark Stevenson. "A semantic similarity approach to paraphrase detection." Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics. 2008.

[3] Vrbanec, Tedo, and Ana Meštrović. "Corpus-Based Paraphrase Detection Experiments and Review." Information 11.5 (2020): 241.

[4] Wahle, Jan Philip, et al. "Are neural language models good plagiarists? a benchmark for neural paraphrase detection." arXiv preprint arXiv:2103.12450 (2021).

[5] Anchiêta, Rafael Torres, and Thiago Alexandre Salgueiro Pardo. "Exploring the Potentiality of Semantic Features for Paraphrase Detection." International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2020.

[6] Chi, Xiaoqiang, Yang Xiang, and Ruchao Shen. "Paraphrase Detection with Dependency Embedding." 2020 4th International Conference on Computer Science and Artificial Intelligence. 2020.

[7] Shakeel, Muhammad Haroon, Asim Karim, and Imdadullah Khan. "A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts." Information Processing & Management 57.3 (2020): 102204.

[8] Gangadharan, Veena, et al. "Paraphrase Detection Using Deep Neural Network Based Word Embedding Techniques." 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE, 2020.

[9] Agarwal, Basant, et al. "A deep network model for paraphrase detection in short text messages." Information Processing & Management 54.6 (2018): 922-937.

[10] Bhargava, Rupal, Gargi Sharma, and Yashvardhan Sharma. "Deep paraphrase detection in indian languages." Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. 2017.

[11] El Desouki, Mohamed I., and Wael H. Gomaa. "Exploring the recent trends of paraphrase detection." International Journal of Computer Applications 975 (2019): 8887.

[12] Aziz, Achmad Abdul, Esmeralda C. Djamal, and Ridwan Ilyas. "Paraphrase Detection Using Manhattan's Recurrent Neural Networks and Long Short-Term Memory." 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). IEEE, 2019.

[13] Qin, Yujia, et al. "Improving sequence modeling ability of recurrent neural networks via sememes." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2364-2373.

[14] Duong, Phuc H., et al. "A hybrid approach to paraphrase detection." 2018 5th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, 2018.

[15] Jang, Myeongjun, Seungwan Seo, and Pilsung Kang. "Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning." Information Sciences 490 (2019): 59-73.

[16] Mahmoud, Adnen, and Mounir Zrigui. "Sentence embedding and convolutional neural network for semantic textual similarity detection in Arabic language." Arabian Journal for Science and Engineering 44.11 (2019): 9263-9274.

[17] Yuan, Zhao, and Sun Jun. "Siamese Network cooperating with Multi-head Attention for semantic sentence matching." 2020 19th International Symposium on

Distributed Computing and Applications for Business Engineering and Science (DCABES). IEEE, 2020.

[18] Huang, Po-Yao, Xiaojun Chang, and Alexander Hauptmann. "Multi-head attention with diversity for learning grounded multilingual multimodal representations." arXiv preprint arXiv:1910.00058 (2019).

[19] Meshram, Ms Swati. "Survey on Attention Neural Network Models for Natural Language Processing."

[20] Srivastava, Shruti, and Sharvari Govilkar. "A survey on paraphrase detection techniques for Indian regional languages." International Journal of Computer Applications 163.9 (2017): 0975-8887.

[21] Issa, Fuad, et al. "Abstract meaning representation for paraphrase detection." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.

[22] Text corpus. https://en.wikipedia.org/wiki/Text_corpus

[23] Victor U Thompson & Chris Bowerman, (2018) "Methods for Detecting Paraphrase Plagiarism".

[24] Vila, Marta, M. Antònia Martí, and Horacio Rodríguez. "Is this a paraphrase? What kind? Paraphrase boundaries and typology." *Open Journal of Modern Linguistics* 4.01 (2014): 205.

[25] Lopez, Marc Moreno, and Jugal Kalita. "Deep Learning applied to NLP." *arXiv preprint arXiv:1703.03091* (2017).

BIOGRAPHIES

Student,
MIT School of Engineering, Pune

Student,
MIT School of Engineering, Pune

Student,
MIT School of Engineering, Pune

Student,
MIT School of Engineering, Pune

Professor,
MIT School of Engineering, Pune