Review on NLP Paraphrase Detection Approaches

Swati Meshram
P.G. Department of Computer Science
SNDT Women's University
Mumbai India

Abstract:- Paraphrasing methods identify, generate, or extract phrases, sentences that convey almost the same information. Different worded sentences may bear the similar meaning and can be identified by paraphrase identification. Paraphrase detection has importance as it contributes to various NLP tasks like Text summarization, Document Clustering, Question Answering, natural language inference, information Retrieval, Plagiarism Detection, Text Simplification. The motivation of the paper was to summaries all available approaches for paraphrase detection, resources and recent trends.

***Keywords**:- NLP, Sentence Encoding, Statistical Similarity, Semantic Similarity,Attention, Machine Translation, Classification.*

I. INTRODUCTION

A paraphrase is a variation in the content in the same language representing the semantic coherence. Paraphrase detection task is modeling a pair of sentences, by comparing two sentences and identifying the relationship between them. For instance one would like to detect two sentences are paraphrases:

- Sentence1: Every year thousands of people visit Taj Mahal.
- Sentence2: Taj Mahal is visited by thousands of tourists every year.

Paraphrasing has applications in information retrieval system[1], text summarization, machine translation, plagiarism detection etc. Paraphrase detection is a fundamental technique which can be beneficial to other natural language processing task like :

II. APPLICATIONS

- **Information retrieval by** automatic generation of variation in query to retrieve answers from databases [1].
- **Semantic textual similarity** measures the degree of similarity in the contextual meaning of two sentences[2].
- **Question Answering** where questions are turned into query and answers are retrieved by applying the query to an existing knowledge base. Further select the correct sentences answering a factual question from a set of candidate sentences. It gives rank to each selected answer sentence based on the measuring semantic similarity between question and answer[2][3].
- **Natural language inference** determines whether a hypothesis can be inferred from a premise based on the examination of context of the premise and the hypothesis.[2]

III. STATISTICAL TECHNIQUES

Some popular statistical techniques exploit the lexical properties of sentences are

- Cosine Similarity: Each sentence is represented using word vector based on the frequency of words in the sentences. It aims to find correlation between two vectors[4]. The cosine similarity score is given by

$$\textbf{Cosine Similarity}(\boldsymbol{S1},\boldsymbol{S2}) = \frac{\mathbf{S1.S2}}{|\mathbf{S1}|.|\mathbf{S2}|}$$

- Jaccard Similarity : The proportion of number of similar words to the number of distinct words in given two sentences[4].

$$\textbf{Jaccard Similarity}(\boldsymbol{S1},\boldsymbol{S2}) = \frac{\boldsymbol{S1} \cap \boldsymbol{S2}}{\boldsymbol{S1} \cup \boldsymbol{S2}}$$

- Resnik similarity : The res metric of similarity (Resnik, 1995) depends on the amount of information shared by two concepts

$$\boldsymbol{Sim_{res}} = \boldsymbol{IC}(\boldsymbol{LCS}(\boldsymbol{C1},\boldsymbol{C2}))$$

- **Lesk measure:** The lesk metric measures relatedness considering hypernymy based on is-a relationship.(Pedersen and Banerjee, 2003). It measures the overlap between two concepts and also considers relation using hypernyms and meronyms.

- **Lch measure :** The lch metric (Leacock and Chodorow, 1998) determines the similarity of two nodes by finding the shortest path length between two concepts in the is-a hierarchy.

$$\boldsymbol{Sim_{lch}}(\boldsymbol{C1},\boldsymbol{C2}) = \mathbf{max}(-\mathbf{log}\ \frac{\boldsymbol{shortestLen}(\boldsymbol{C1},\boldsymbol{C2})}{\mathbf{2} * \boldsymbol{TaxonomyDepth}})$$

- **Wu and Palmer Simliarity :** Wu and Palmer's semantic similarity measure wup (Wu and Palmer, 1994) is based on the path length between concepts organized in a hierarchical structure of lexicalized conceptual domain. It determines path length from the least common superconcept (LCS) of the nodes.

$$\boldsymbol{Sim_{wup}} = \frac{\mathbf{2} * \boldsymbol{depth}(\boldsymbol{LCS}(\boldsymbol{C1},\boldsymbol{C2}))}{\boldsymbol{depth}(\boldsymbol{C1}) + \boldsymbol{depth}(\boldsymbol{C2})}$$

Where depth(C) is the pathlength of concept C in the WordNet hierarchy.

This paper is a comprehensive survey of recent paraphrase detection approaches developed in the recent years exploiting lexical properties and other models using deep neural network.

IV. APPROACHES TO PARAPHRASE DETECTION

➢ *Models based on lexical properities :*

- [5]The authors project the word relation for the linguistic representation of knowledge unit and its various semantically equivalent (SE) forms of its expert description. There work is based on joint estimation of the coupling strength of its word combinations found in the phrases of a text analyzed, and further decomposed these words into classes with value of the TF-IDF metric relevant to the corpus texts. They proposed, to select target text-corpus phrases which are either mutually equivalent or have to be semantically complementary to each other and represent the same image.

Further, the selected phrases are ranked by the degree of closeness to a semantic pattern (i.e. sense standard). The coupling strength is estimation is done on the basis of prepositions and conjunctions and excluding them. They have used text information representing a selected knowledge unit compressing at least two times preserving its meaning.

- [6] here the Abstract Meaning Representation (AMR) parsing framework converts sentences into a canonical form as AMR graphs due to its ability of abstracting away from syntax and representing the core concepts expressed in the sentence. Their research based on latent semantic analysis for paraphrase detection. Here the similarity between two graphs of corresponding sentences is found along with a score function. They have used kernel based SVM classifier and have demonstrated LSA with different reweighting scheme based on PageRank and TF-IDF. The JAMR parser classifier accuracy is 86.6%.

- [7] This paper presents an algorithm for paraphrase identification which relies on word similarity information derived from WordNet. The proposed method is based on evaluating semantic similarity metrics to find the similarity of two text segments by considering all word-to-word similarities, not just the maximal similarities between the sentences. Here, each sentence is represented as a binary vector a and b. The similarity between the pair of sentences is computed using the following formula:

$$sim(\vec{a},\vec{b}) = \frac{\vec{a}W\vec{b}^{T}}{|\vec{a}||\vec{b}|}$$

With W as a semantic similarity matrix containing information about the similarity of word pairs. The model is evaluated on Microsoft Research Paraphrase Corpus (MSRP) freely available for English.

- [8] The author's perspective is to view paraphrase detection as a classification problem. Two sentences are classified into completely equivalent, roughly equivalent and not quivalent classes based on multinomial logistic regression classification technique. Cosine similarity, word overlap measure, stemmed overlap measure, bigram based similarity and semantic seminlairty has been used in the approach. Paraphrase detection system has been experimented in four Indian Languages-Hindi, Punjabi, Tamil and Malayalam. Their system achieves the highest f-measure of 0.95 in Punjabi language.The performance of our system in Hindi language is f-measure of 0.90.

Recently, current models in NLP are built using the deep neural networks focusing on implicit learning vector representations of sentence meaning which have shown effectiveness in this task.

➢ *Models based on Neural network :*

- [9] has proposed a matching aggregation framework which is a multiway attention network. Their model is built to learn word representation for two sentences using a bi-directional RNN. The activation function used is gated recurrent unit. They have designed four attention functions to match words in corresponding sentences viz concat, bilinear, dot, minus. Then the aggregation of matching information from multiway attention functions is computed.

They have used 300-dimensional GloVe embeddings for training, zero vectors to represent all out-of-vocabulary words, hidden vector length which is set to 150 for all layers. They apply dropout rate 0.2 between layers, with dropout. The model is optimized using AdaDelta with initial learning rate of 1.0. The dataset is Quora Question Pairs with over 400,000 question pairs. Their experiments demonstrate 89.12% model accuracy.

- [10] The authors propose a bilateral multi-perspective matching (BiMPM) model which is belongs to the "matching-aggregation" framework with five layers. Given two sentences P and Q, their model first encodes them with a BiLSTM encoder. Further, they estimate the similarity between two encoded sentences in two directions ($P \rightarrow Q$ and $P \leftarrow Q$) with the probability distribution $Pr(y|P, Q)$. They have employed LSTM with full, maxpooling, attentive, max attentive matching functions and softmax function at the output layer. Their model performance attended the accuracy of 88.8%.

- [11] have represented tree structured neural network for encoding of sentences. Their model Stack augmented Parser-Interpreter Neural Network (SPINN) combines parsing tree structured sentence interpretation into a

ISSN No:-2456-2165

shift-reduce parser. Their model takes two sentences, a premise and a hypothesis as input. And the outcome is a decision of entailment, contradiction, or neutral which reflects the relationship between the senses of the two sentences. A Tree LSTM composition function with active state and memory have been projected. The activation functions used for Shift parsing is softmax classifier, for reduce are sigmoid and tanh. The corpus SNLI used had 570k human labeled pairs of scene description. The finding of the paper is that parse trees make easy to identify any instance of negation and separate from the rest of the sentence. Their model has shown improvement by achieving 83% accuracy as compared to LSTM RNN encoders with 80%.

- [12] The approach is to represent each word as a low dimensional vector. Then and computes a semantic matching vector for every word based on all words in the other sentence. Further, split each word vector into a similar and dissimilar component on the basis of word-by-word semantic matching vector. Then, a two-channel CNN model with multiple types of ngram filters is employed to extract the features for sentence similarity calculation. Finally, a similarity score is computed on the composed feature vectors. The performance of the model is measured in MAP on QASent as 77% , MRR is 84%, whereas on WikiQA dataset with mean average precision (MAP) as 70%, mean reciprocal rank (MRR) as 72%.

- [13] is a model for answer sentence selection incorporating distributed representation of sentence to understand the semantic encoding. Where questions and answers are transformed into vectors and learned a semantic matching function between QA pairs. They found improvement in performance by using a complex sentence model based on a convolutional neural network and the convolution layer encodes every bigram into feature. The average pooling layer to combine all bigrams. The activation function used is hyperbolic tan. They have trained their model with unigram, bigram, unigram+count and bigram+count. Where the MAP was found as 70% and MRR was 78%.

- [14] their research work is a model based on a Recursive Autoencoder(RAE) for unsupervised feature learning and dynamic pooling. RAE projects vector representation of phrases as nodes in a parse tree. Feature representation for each node in the tree are learnt by recursively constructing word vector. Authors have reported the dynamic pooling layer is powerful as it captures the global structure of similarity matrix for measuring similarity of sentences. The model achieved accuracy upto 75.9% with standard RAE method and 76.8% with unfolding RAE. Further, adding 1 and 2 hidden layers had lowered the accuracy by 0.2% and training became slower.

- [15] This paper, presents recursive autoencoder architecture to learn representation of phrases which is unsupervised learning. These representations help to extract features for classification algorithms. The authors find these representations are decent by extensively examining their nearest neighbors. The tree structure constructed by autoencoders is built using greedy approach and CKY algorithm to find global optimal tree structure. Have further worked on parsed tree. They have also adopted SimMatrix method to calculate the distance between a pair of representation of sentences. Here a sigmoid function is used for the computing the similarity. The Microsoft Research Paraphrase Corpus is used. Accuracy achieved Greedy aggregate – 68.57%, CKY aggregate – 68.75%, Parsed tree – aggregate – 70.55%, SimMatrix – 73.33%

- [16]Here authors propose a Collaborative and Adversarial Network (CAN), to model the common features between given two sentences to enhance the sentence similarity modeling. They have introduced a common feature extractor in the CAN model, which consist of a generator and a discriminator via both collaborative and adversarial learning. They have measured similarity with manhattan distance. The model activation function are hyperbolic tangent and softmax at the output layer. The datasets, namely TREC-QA and WikiQA for answer selection and MSRP for paraphrase identification are used. The authors claim their proposed model is effective to improve the performance of sentence similarity modeling by outperforms the state-of-the-art approaches on TREC-QA without use of any external resource and pre-training. For the other two datasets, they have found the model is comparable but not better than the recent neural network approaches. The performance of the model measured in MAP on TREC-QA was 84% , MRR is 91%, whereas on WikiQA dataset with MAP as 73%, MRR as 74%. MSRP dataset MAP was 77%, MRR is 84%.

- [17] The authors have presented a study on the effectiveness of subword (character and character n-gram) level models in sentence pair modeling without using pretrained word embeddings. Their research proposes that subword models can benefit from multi-task learning with simple language modeling. The model is pair-wise word interaction model which encodes word context and d sequence order through bidirectional LSTMs. A 19-layer-deep CNN is designed to aggregate the word interaction features and the softmax layer to predicate classification probabilities Embedding of subwords is based upon char C2W, char CNN. A multi-task structure for sequential tagging is adapted, to further improve the subword representations in sentence pair modeling and language modeling to predicts the next word and previous word using softmax over the hidden states of Bi-LSTM. Their experiment is based on including 19-layer-deep CNN and without 19-layer-deep CNN model. The finding is, in most cases the 19-layer CNN helps to achieve better or comparable performance, at the expense of increase of training time by 25%.

- [18] This paper is a study on paraphrase generation focusing on learning text representation using deep neural networks. The authors have constructed a model, bilateral LSTM (BiLSTM) using lexical features and transformers for paraphrase generation. SLING a neural transition-based semantic graph generator is used for paraphrase generation. The parser adds frames/roles to the graph and serves as a continuous internal representation of the input's incrementally constructed A feedforward Transition-Based Recurrent Unit (TBRU) to process the token vectors is also used. MSCOCO and WikiAnswers are the dataset for training and evaluation. Their finding is TRANSFORMER-PB (28.0%) performs over the basic TRANSFORMER (18.0%)—but both fall far short of CHIA (78.2%).

- [19] The authors incorporate a corpus to evaluate metaphor paraphrase detection. The authors have represented the metaphor interpretation task as supervised learning binary classification and gradient judgment prediction. A DNN architecture which includes combining a CNN, an LSTM RNN, and two densely connected neural layers is constructed for metaphor interpretation. The DNN architecture that authors have proposed encodes each sentence in a 10 dimensional vector representation. This is further merged through concatenation and fed to a series of densely connected layers.

V. CONCLUSION

In this article, various methods involving statistical, syntactical and semantic properties to represent sentence modeling in natural language text have been discussed. An overview of the behavior of each similarity measure is also presented. The recent trend of employing deep learning neural network models are also described with their performance. Neural network models give better performance as it involves extensive learning through training over data model.

REFERENCES

[1]. Nitin Madnani, Bonnie J. Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods, Computational Linguistics, Vol 36, no -3,341-387,2010.
[2]. Wuwei Lan and Wei Xu,Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering
[3]. L. Yu, K. M. Hermann, P. Blunsom, S. Pulman, Deep learning for answer sentence selection, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS) Workshop, 2014.
[4]. Mathew, D., Idicula, S.M.: Paraphrase identification of Malayalam sentences - an experience. IEEE (2013)
[5]. G. M. Emelyanova, D. V. Mikhailova, and A. P. Kozlova. Relevance of a Set of Topical Texts to a Knowledge Unit and the Estimation of the Closeness of Linguistic Forms of Its Expression to a Semantic Pattern ,ISSN 1054-6618, Pattern Recognition and Image Analysis, 2018, Vol. 28, No. 4, pp. 771–782. © Pleiades Publishing, Ltd., 2018.
[6]. F Issa, M Damonte, S.B.Cohen,X Yan, Y Chang. Abstract meaning representation for paraphrase detection. In Proceedings of NAACL-HLT18, 486–492.
[7]. S. Fernando and M. 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
[8]. Kamal Sarkar,Detecting Paraphrases in Indian Languages Using Multinomial Logistic Regression Model
[9]. Chuanqi Tan, Furu Wei, Wenhui Wang,Weifeng Lv, Ming Zhou.Multiway Attention Networks for Modeling Sentence Pairs, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)
[10]. Zhiguo Wang, Wael Hamza, Radu Florian, Bilateral Multi-Perspective Matching for Natural Language Sentences
[11]. R. Samuel Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, D. Christopher Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In ACL, pages 1466–1477. Association for Computational Linguistics, 2016.
[12]. Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. Sentence similarity learning by lexical decomposition and composition. In COLING, 2016.
[13]. Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. Proceedings of the Deep Learning and Representation Learning Workshop: NIPS, 2014.
[14]. R. Socher, E.H. Huang, J. Pennington, A.Y. Ng, and C.D. Manning. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In NIPS, 2011
[15]. Eric Huang. Paraphrase detection using recursive autoencoder. Source : http://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf,2011
[16]. Qin Chen, Qinmin Hu, Jimmy Xiangji Huang and Liang He. CAN: Enhancing Sentence Similarity Modeling with Collaborative and Adversarial Network, SIGIR'18, 2018, USA
[17]. Wuwei Lan and Wei Xu .Character-based Neural Networks for Sentence Pair Modeling ,arXiv:1805.08297, May. 2018.
[18]. Su Wang, Rahul Gupta, Nancy Chang, Jason Baldridge. A task in a suit and a tie: paraphrase generation with semantic augmentation, In AAAI, 7176–7183
[19]. Yuri Bizzoni and Shalom Lappin. Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks, Proceedings of The Workshop on Figurative Language Processing, NAACL 2018, New Orleans LA.
[20]. E. Altintas, E. Karsligil, and V. Coskun. A New Semantic Similarity Measure Evaluated In Word Sense Disambiguation, In Proceedings of the 15th NODALIDA conference, S. Werner (ed.), 2006, pp. 8–11 ISBN 952-458-771-8, ISSN 1796-1114.