# Exploring the Recent Trends of Paraphrase Detection

Mohamed I. El Desouki
Information Systems Department
Prince Sattam Bin Abdulaziz University

Wael H. Gomaa
Beni-Suef University
Shaqra University

## ABSTRACT
This study is to examine paraphrase detection (PD) for diagnostic purposes. Which is defined as the capability to find and discover the similarity between sentences that are written in a natural language? Where detecting similar sentences written in natural language is extreme importance and it is very essential for computer software used in plagiarism detection, Q and A automated systems, text mining, authorship authentication and text recapitulation. The goal of paraphrase detection is to detect whether two statements have the identical semantic or not. There is hundreds of empirical research in this direction. This study will focus on the discussion of recent studies of the PD methods and will categorize them in two categories, supervised learning and unsupervised learning. Also will give an idea about text similarity, machine learning and deep learning approaches. The performance of the selected researches is assessed by how accurate the F-measures are in detecting paraphrase in Microsoft Research Paraphrase Corpus (MSPR).

## General Terms
Artificial Intelligence, Natural Language Processing, Paraphrase Identification.

## Keywords
Paraphrase Detection, Text Similarity, Machine Learning, Deep Learning, Microsoft Research Paraphrase Corpus.

## 1. INTRODUCTION
Natural language processing has been the focus of technological contributions recently. This is evident in in many applications such as extracting information from big data on the web although the most of those data are in unstructured format. Much has been accomplished in studying sentiment analysis in social networks and in the field of grammatical analysis that is used heavily in easy marking and assessing student's answers.

Paraphrase detection (PD) has always been problematic in the study of natural language processing (NLP). The goal is to establish if a pair of sentences is conveying identical meaning or not. Two sentences are judged as paraphrase if they yield identical meaning. If the meaning is not identical they are considered non-paraphrase.

The majority of the existing systems such as Microsoft paraphrase corpus (MSRP) are efficient at cleaning text corpora [1]. (PD) is deemed a fundamental subtask in many NLP tasks , it obviously appear in for example question answering, finding similar relations between question needs paraphrases, also it's widely used in machine translation, document clustering and Information retrieval, etc. The survey at hand briefly investigates the current research done on PD by means of presenting them in two approaches. The first approach is unsupervised approach that heavily depends on text similarity. The second is the supervised approach that depends on machine learning (ML) and deep learning (DL).

In text similarity the first thing to start with is the Word similarity. Words might be alike either lexically or semantically. Words are lexically similar if they have the same character sequence. This is also true if they are used in the same context and give the same meaning and one of them is a type of the other. Lexical similarity is identified by means of dissimilar string based algorithms whereas semantic similarity is identified through Corpus-Based and knowledge-based algorithms. String metrics measures check sequences of characters and character composition to measures the similarity or dissimilarity between two strings for approximate string matching or comparison.

Corpus-Based similarity is a kind of semantic similarity measure which indicates the similarity between words based on information gathered from large corpora. On the other hand, Knowledge-Based similarity is a measure of semantic similarity that indicates the degree of similarity between words using information obtained from semantic networks [2].

Machine learning (ML) is an artificial intelligence (AI) application which makes it possible for systems to automatically learn and grow from experience without any being explicit programming. It revolves around the development of computer software that can access data and utilizes them to learn for themselves.

Machine learning approach treats the PD problem as a problem of normal text classification that employs the syntactic and/or linguistic features. ML is categorized into supervised learning and Unsupervised learning. Supervised learning is the ML task of inferring a function from labeled training data which contains a group of training examples in which each example represents a pair consisting of an input object and a favorable output value. In supervised learning, the algorithm analyzes the training data and produces an inferred function. That function can be further used for mapping new examples.

On the other side, unsupervised machine learning algorithms deduce patterns from a dataset without reference to known or labeled outcomes. What distinguishes the unsupervised learning form the supervised learning and reinforcement learning [3] is that, the examples that are given to the learner are unlabeled in unsupervised learning. Therefore, the relevant algorithm does not conduct evaluation of accuracy of the output structure.

The evaluation of structure of accuracy is nonexistent of the structure that is output by the relevant algorithm.

Deep learning is a machine learning technique which instructs computers to achieve what the people acquire naturally: learn by example. In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Learning can be supervised, semi-supervised or unsupervised. Recently, deep learning has been the focus of research for a number of reasons. For one thing, the results accomplished by it are impressive. Models of deep learning

are more or less associated with information processing and communication patterns in a biological nervous system such as neural coding which seeks to identify a relationship between miscellaneous stimuli and neuronal responses associated to them in the brain.

Types of deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been implemented in a variety of fields including NLP, computer vision, speech recognition, audio recognition, social network filtering, machine translation, bioinformatics and drug design.

Research findings have shown that some models of deep learning were able to achieve tremendous accuracy, sometimes exceeding human-level performance [4].

The rest of the current survey would be organized as follows. Sections 2, 3 will introduce the previous PD work based on unsupervised and supervised approaches respectively. Section 4 will present the discussion and conclusions of the Survey..

## 2. PARAPHRASE DETECTION BASED ON UNSUPERVISED APPROACH

As mentioned above, the unsupervised learning on PD relies on text similarity features. The authors [5] introduced techniques for measuring the semantic similarity of texts, by means of corpus-based and knowledge-based measures of similarity. The researchers showed that some algorithms like PMI_IR, LSA, Lesk, Lin and Resnik were tested on MSRP. The researches prove that adding semantic information to measures of text similarity reveals an obvious increase in the probability of recognition compared to the random baseline and the vector-based cosine similarity baseline in paraphrase recognition task. In [6] an unsupervised method for semantic relatedness that generates a semantic profile for words by using salient conceptual features gathered from encyclopedic knowledge was proposed. The basic idea behind the model comes from the thought that a word meaning can be signified by the remarkable concepts that are found in its direct context. the efficiency and maturity of this model was proven by applying the standard word-to-word and text-to-text relatedness benchmarks. The performance of the model seems to be independent of the distance metric used in the evaluation (cosine or SOCPMI). And this leads to additional support for the underlying assumption about profiling words using strong unambiguous word-concept associations. In [7] a software system that utilizes a recently developed lexico-syntactic method to the task of paraphrase identification was introduced. The approach offered competitive results against the results achieved by other approaches on MSRP data set. The usefulness of this approach comes from the fact that it mostly uses few resources when compared to similar systems while producing results similar if not better than other methods to paraphrase identification. In [8], the authors proposed a method for measuring the semantic similarity of texts using a corpus based measure of semantic word similarity and a normalized and modified versions of the Longest Common Subsequence (LCS) string matching algorithm. They focus of the research was on measuring the similarity between two sentences or between two short paragraphs. And the method was evaluated on MSRP data set and the results showed that this method is obviously perform better than several competing methods. In [9] a comparative study was provided between neural word representations and traditional vector spaces based on co-occurrence counts, in a number of compositional tasks Different semantic spaces and tensor-based compositional models were implemented and tested. The authors conducted two small-scale tasks (word sense disambiguation and sentence similarity) and two large-scale tasks (paraphrase detection and dialogue act tagging). A WordNet-based lexical similarity measures were applied in [10]. Various refinements to the matrix similarity approach were presented. The work was focused on the use of lexical similarity, essentially using a bag-of-words model. The system was evaluated on MSRP and found to outperform previously text similarity reported approaches.

**Table 1 summarizes the PD unsupervised learning approach results on MSRP corpus.**

| Reference | Methods | Accuracy | F-measure |
|---|---|---|---|
| [5] | - Cosine similarity with tf-idf weighting<br>- Combination of several word similarity measures | 64.5 %<br>70.3 % | 75.3 %<br>81.3 % |
| [6] | - Explicit semantic space<br>- Salient semantic space | 67 %<br>72.5 % | 79.3 %<br>81.4 % |
| [7] | Graph subsumption | 70.6 % | 80.5 % |
| [8] | Combination of semantic and string similarity | 72.6 % | 81.3 % |
| [9] | Additive composition of vectors and cosine distance | 73 % | 82 % |
| [10] | JCN WordNet similarity with matrix | 74.1 % | 82.4 % |

## 3. PARAPHRASE DETECTION BASED ON SUPERVISED LEARNING

### 3.1 Classic Machine Learning Approach

In contrast to most PD systems that concentrates on sentence similarity, in [11] a supervised two-phase framework was presented to address the problem of PD through detecting dissimilarities between sentences and made its paraphrase judgment based on the how far they are dissimilar. The capability to detect differences between significant dissimilarities reveals what makes two sentences a non-paraphrase. In addition, it helps to introduce additional paraphrases that contain extra but insignificant information. The Experimental results showed that, the implemented system was accurate at distinctive non-paraphrasing dissimilarities and also was able to achieve higher paraphrase recall compared to other alternatives. In [12], an approach was introduced based on enhanced pre-processing and semantic heuristics which relied on enhanced features set. The system produced comparable or even better results than the state of the art systems in this category. Another important part of this work was misclassification analysis which highlighted the pros and cons of semantic heuristics based features used in this study and in addition it showed also some criticisable annotations of sentence pairs included in benchmark corpus like MSRP. In [13] both of Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Maximum Entropy

(MaxEnt) algorithms were tested to determine which one of them is the most appropriate for the PD task. And the SVM algorithm showed the best performance. The authors in [14] showed that some features can be derived and then used to decide if similar sentences are paraphrases of each other from methods currently utilized to automatically evaluate machine translation systems. The experiments showed that using features that encode the distribution over the Part Of Speech (POS) tag set of both matching words and non-matching words can significantly enhance the performance of a Position independent word error rate (PER) based system on PD task. The authors in [15] employed a generative model that generates a paraphrase of a given sentence, and they used probabilistic inference to reason about whether two sentences share the paraphrase relationship. The model used quasi-synchronous dependency grammars to cleanly incorporates both syntax and lexical semantics. In addition, they combined the model with a complementary logistic regression model based on state-of-the-art lexical overlap features depending on a product of experts. In [16], the authors showed that the final accuracy of PD can be enhanced by using dependency-based features in addition to bigram features. Furthermore, the results showed that using weighted dependency overlap seemed to provide promise, yielding the best F-Measure for False Paraphrase classification seen so far. They concluded that dependency features may thus be useful in more accurately classifying cases of False Paraphrase. In [17], the authors proposed to re-examine the hypothesis that automated metrics developed for machine translation evaluation can prove useful for PD in light of the significant work on the development of new machine translation metrics. They showed that a meta-classifier trained using nothing but recent machine translation metrics outperforms all previous PD approaches on the MSRP corpus. They conducted extensive error analysis and uncover the top systematic sources of error for a PD approach relying solely on machine translation metrics. The key idea behind the work done in [18] was that the existence of similarity in the latent space suggests the existence of semantic relatedness. The authors described three ways in which labeled data can enhance the accuracy of these approaches on PD. First, they designed a discriminative term-weighting metric called TF-KLD, which outperforms TF-IDF. Next, they showed that using the latent representation from matrix factorization as features in a classification algorithm substantially improves accuracy. Finally, they combined latent features with fine-grained n-gram overlap features, yielding a high accuracy on MSRP corpus. The authors in [19] studied the use of structural representations for learning relations between pairs of short texts. Their work mainly focused on defining syntactic and semantic structures to represent the text pairs and then applying graph and tree kernels to them for automatically engineering features in SVM. They did a lot of comparative analysis among the state-of-the-art models of this type of relational learning. A summary of the PD supervised learning approaches and their results on MSRP corpus using classical machine learning methods mentioned below in Table 2.

**Table 2: Paraphrase Detection Based On Classical Machine Learning Methods**

| Reference | Methods | Accuracy | F-measure |
|---|---|---|---|
| [11] | Sentence dissimilarity classification | 72 % | 81.6 % |
| [12] | PI using semantic heuristic features | 74.4 % | 81.8 % |
| [13] | Combination of lexical and semantic features | 76.6 % | 79.6 % |
| [14] | Combination of MT evaluation measures as features | 75 % | 82.7 % |
| [15] | Product of experts | 76.1 % | 82.7 % |
| [16] | Dependency-based features | 75.6 % | 83 % |
| [17] | Combination of eight machine translation metrics | 77.4 % | 84.1 % |
| [18] | Matrix factorization with supervised reweighting | 80.4 % | 85.9 % |
| [19] | Combination of Convolution Kernels and similarity scores | 79.1 % | % |

## 3.2 Deep Learning Approach

The authors in [20] introduced a method for PD based on recursive auto encoders (RAE). RAEs are based on a totally new objective and learn feature vectors for phrases in syntactic trees. These features were used to measure the word-wise and phrase-wise similarity between two sentences. Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size. They introduced a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation was then used as input to a classifier. In [21] the idea behind the work done by the authors was that, they systematically compared three types of distributional representation to test how they affect the semantic composition. The comparisons involved a simple distributional semantic space, word embedding computed with a neural language model and a representation based on weighted word-link-word tuples arranged into a third-order tensor. These representations vary in many respects: the amount of preprocessing and linguistic information involved, whether the semantic space is the by-product of a learning process, and data requirements. These representations served as input to three composition methods involving addition, multiplication and a deep recursive auto encoder. They experimented with several possible combinations of representation and composition, exhibiting varying degrees of sophistication. Some are shallow while others operate over syntactic structure, rely on parameter learning, or require access to very large corpora. They found that shallow approaches are as good as more computationally intensive on PD task. In [22], the authors proposed a framework for comparing sentences that uses a multiplicity of perspectives.

First, the authors used a Convolutional Neural Network (CNN) that derives characteristics at different levels of granularity to model each sentence and used several types of pooling. After that they compared their sentence representations at several granularities using multiple similarity metrics. And the results showed a high strong performance on PD task, the performance exceeded the state of the art although they didn't use any external resources such as WordNet or parsers. In [23] the authors prepared a compositional distributional framework that is built on a rich form of word embedding that aims to simplify the relationships between words in the context of any given sentence. Embedding and composition layers were jointly learned against a generic objective that enhanced the vectors with syntactic information from the surrounding context. Additionally, each word is associated with a number of senses and the system dynamically selects the most feasible in the course of the process of composition. They evaluated the produced vectors qualitatively and quantitatively with positive results. At the sentence level, the effectiveness of the framework was demonstrated on the MSRP corpus, and the results appeared within the state-of-the-art range. They also demonstrated the benefits of a Siamese architecture in the context of a PD task. While the architectures tested in this work were limited to a recursive neural network (RecNN) and a Recurrent Neural Network (RNN), the ideas they presented were in principle directly applicable to any kind of deep network. In [24], a model proposed to consider the resemblance and dissemblance between sentences by decomposing and composing lexical semantics over sentences. The model represented each word as a vector, and calculated a semantic matching vector for each word based on all words in the other sentence. Following that, each word vector was decomposed into a similar component and a dissimilar component based on the semantic matching vector. After this, a two-channel Convolution Neural Network (CNN) model was used to capture features by composing the similar and dissimilar components. Finally, a similarity score was estimated over the composed feature vectors. Experimental results show that this model achieved a comparable result on the PD task. Table 3 summarizes the PD supervised learning approach results on MSRP corpus using deep learning methods.

**Table 3: Paraphrase Detection Based On Deep Learning Methods**

| Reference | Methods | Accuracy | F-measure |
|---|---|---|---|
| [20] | Recursive autoencoder with dynamic pooling | 76.8 % | 83.6 % |
| [21] | Simple distributional semantic space | 73 % | 82.3 % |
| [22] | Multi-perspective Convolutional NNs and structured similarity layer | 78.6 % | 84.7 % |
| [23] | Recursive NNs using syntax-aware multi-sense word embeddings | 78.6 % | 85.3 % |
| [24] | Sentence Similarity Learning by Lexical Decomposition and Composition | 78.4 % | 84.7 % |

# 4. DISCUSSION AND CONCLUSIONS

This survey investigates paraphrase detection (PD). Given two sentences, the objective is to detect whether they are semantically identical. We go through the articles in PD and we categorize them into two categories, unsupervised and supervised learning. The former type of learning is the learning task of inferring a function to describe hidden structure from "unlabeled" data. The unsupervised learning on PD task depends on the various approaches of text similarity such as string-based, corpus-based and knowledge-based. The later type of learning is a machine learning task of inferring a function from labeled training data. We classified the supervised learning approach into two categories classical ML and deep learning. The execution of the selected articles is evaluated based on the accuracy and F-measures in identifying paraphrase in Microsoft Research Paraphrase Corpus (MSPR). The best accuracy and F-Measure on the unsupervised category were obtained by authors in [10]; they implemented a system that based on JCN knowledge-based similarity algorithm. This algorithm relies on a combination of using edge counts in the WordNet 'is-a' hierarchy and utilizing the information content values of the WordNet concepts. The authors in [18] achieved the best results on the classical machine learning category. The accuracy and F-Measure values were 80.4 % and 85.9 % respectively. They presented three ways in which labeled data can improve distributional measures of semantic similarity at the sentence level. The main innovation was TF-KLD, which discriminatively reweights the distributional features before factorization, so that discriminability impacts the induction of the latent representation. They then transform the latent representation into a sample vector for supervised learning, obtaining results that strongly outperform the prior state-of-the-art; adding fine-grained lexical features further increased performance. The authors in [23] achieved the best results on the deep learning category. The accuracy and F-Measure values were 78.6.4 % and 84.7 % respectively. The main contribution of [23] was a deep compositional distributional model acting on linguistically motivated word embeddings. The effectiveness of the syntax-aware, multi-sense word vectors and the dynamic compositional disambiguation framework in which they were used was demonstrated by appropriate tasks at the lexical and sentence level, respectively, with very positive results. As an aside, they also demonstrated the benefits of a Siamese , RecNN and RNN architectures. Paraphrase detection is an open research area, especially in deep learning. Therefore, advancement in the field of natural language processing requires enhancing performance by models that require a magnitude of data and do not need much linguistic expertise to train and operate.

# 5. REFERENCES

[1] Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

[2] Gomaa, W. H., & Fahmy, A. A. (2011). Tapping into the power of automatic scoring. In The Eleventh International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC).

[3] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[5] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In AAAI (Vol. 6, pp. 775-780).

[6] Hassan, S. (2011). Measuring semantic relatedness using salient encyclopedic concepts. University of North Texas.

[7] Rus, V., McCarthy, P. M., Lintean, M. C., McNamara, D. S., & Graesser, A. C. (2008, May). Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In FLAIRS conference(pp. 201-206).

[8] Islam, A., & Inkpen, D. (2009). Semantic similarity of short texts. Recent Advances in Natural Language Processing V, 309, 227-236.

[9] Milajevs, D., Kartsaklis, D., Sadrzadeh, M., & Purver, M. (2014). Evaluating neural word representations in tensor-based compositional settings. arXiv preprint arXiv:1408.6179.

[10] Fernando, S., & Stevenson, M. (2008, March). A semantic similarity approach to paraphrase detection. In Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics (pp. 45-52).

[11] Qiu, L., Kan, M. Y., & Chua, T. S. (2006, July). Paraphrase recognition via dissimilarity significance classification. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 18-26). Association for Computational Linguistics.

[12] Ul-Qayyum, Z., & Altaf, W. (2012). Paraphrase identification using semantic heuristic features. Research Journal of Applied Sciences, Engineering and Technology, 4(22), 4894-4904.

[13] Kozareva, Z., & Montoyo, A. (2006). Paraphrase identification on the basis of supervised machine learning techniques. In Advances in natural language processing (pp. 524-533). Springer, Berlin, Heidelberg.

[14] Finch, A., Hwang, Y. S., & Sumita, E. (2005). Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

[15] Das, D., & Smith, N. A. (2009, August). Paraphrase identification as probabilistic quasi-synchronous recognition. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (pp. 468-476). Association for Computational Linguistics.

[16] Wan, S., Dras, M., Dale, R., & Paris, C. (2006). Using dependency-based features to take the'para-farce'out of paraphrase. In Proceedings of the Australasian Language Technology Workshop 2006 (pp. 131-138).

[17] Madnani, N., Tetreault, J., & Chodorow, M. (2012, June). Re-examining machine translation metrics for paraphrase identification. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 182-190). Association for Computational Linguistics.

[18] Ji, Y., & Eisenstein, J. (2013). Discriminative improvements to distributional sentence similarity. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 891-896).

[19] Filice, S., Da San Martino, G., & Moschitti, A. (2015). Structural representations for learning relations between pairs of texts. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Vol. 1, pp. 1003-1013).

[20] Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Advances in neural information processing systems (pp. 801-809).

[21] Blacoe, W., & Lapata, M. (2012, July). A comparison of vector-based representations for semantic composition. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 546-556). Association for Computational Linguistics.

[22] He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1576-1586).

[23] Cheng, J., & Kartsaklis, D. (2015). Syntax-aware multi-sense word embeddings for deep compositional models of meaning. arXiv preprint arXiv:1508.02354.

[24] Wang, Z., Mi, H., & Ittycheriah, A. (2016). Sentence similarity learning by lexical decomposition and composition. arXiv preprint arXiv:1602.07019.