Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7716462

DETECTION AND CORRECTION OF SEMANTIC ERRORS - A SURVEY

Swati Gangwani
Department of Computer Science and Engineering, Jabalpur Engineering College, Jabalpur
swati.gangwani5@gmail.com

Satvika Shrivastava
Department of Computer Science and Engineering, Jabalpur Engineering College, Jabalpur
satvika144@gmail.com

Dr Jitendra Singh Thakur
Department of Computer Science and Engineering, Jabalpur Engineering College, Jabalpur
jsthakur@jecjabalpur.ac.in

ABSTRACT

Detecting semantic errors in a text is still a challenging area of investigation. A lot of research has been done on lexical and syntactic errors while fewer studies have tackled semantic errors, as they are more difficult to treat. The objective of this survey is to examine the existing studies, highlighting the current issues and suggesting the potential directions of future research. This survey is a result of analyzing different research papers and studying different approaches used for EDC. We also present a possible scheme for the classification of Semantic errors. This survey article also traces the evolution of different methods, categorizing them based on their underlying principles as knowledge-based, corpus-based, deep neural network-based methods, and hybrid methods. Discussing the strengths and weaknesses of each method.

Among the main observations, we found that not many studies have been done in the field of semantic errors and also there is lack of proper datasets for semantic errors which can help in training and testing of the machine learning models. We present a table that summarizes these approaches along different dimensions such as target error types, linguistic dataset used, strengths and focus of each approach. This facilitates better understandability, comparison and evaluation of previous research. This survey provides a comprehensive view of existing systems in place, for new researchers to experiment and develop innovative ideas to address the issue of semantic error correction and detection.

1.INTRODUCTION

English is a West Germanic language which has become a global *lingua franca*. Over 600 million users use English as a second language. English as second language learners are more likely to make semantic errors as compared to other syntactic errors because syntactic errors are governed by some grammar rules which is not the case with semantic errors. The problem of semantic error is a more complex one. Usually, such error disturbs the syntax and semantics of the whole sentence, which requires a human-being to detect it.

The difficulty in tackling semantic errors lies in the fact that analyzing “language beyond sentence level gets a prominent role in the study of language with the basic tenet that the study of language in context will offer a deeper insight into how meaning is attached to utterances than the study of language in isolated sentences”(Obeidat, 1986: 74).Examining and studying semantic errors is an interesting and challenging area of investigation which is still fertile requiring much more research. Compared to phonological and syntactic errors, there are relatively few studies which have tackled semanticerrors. An automatic syntactic/semantic analysis of a 'correct' sentence itself is a difficult task and the analysis of an 'erroneous' sentence is almost impossible in most cases. To the best of our knowledge, the problem of semantic error detection and correction is still at the research and development stage.

In this paper, we have given the survey to provide a comprehensive account of the various methods used in the field of semantic error detection and correction including the most recent advancements using deep neural network-based methods. This survey traces the evolution of techniques over the past decades and also suggest the recent techniques which could be used as a future research work in this field.

This paper is organized as follows. Section 2 defines what exactly are semantic errors. Section 3 describes the classification of Semantic errors. In section 4 various available datasets have been listed. Section 5 and section 6 highlights the various approaches existing till date and Concluding remarks are given in section 7. This survey provides a deep and wide knowledge of existing techniques for new researchers who venture to explore one of the most challenging NLP tasks, Semantic Errors detection and correction.

2.SEMANTIC ERRORS

The word ‘Semantic’ relates to ‘**meaning in language’.** Semantic errors occur when statements are not meaningful. Semantic refers to the set of rules which give the meaning of a statement. Semantic errors result in morphologically and syntactically valid words whose use in context is senseless or absurd.

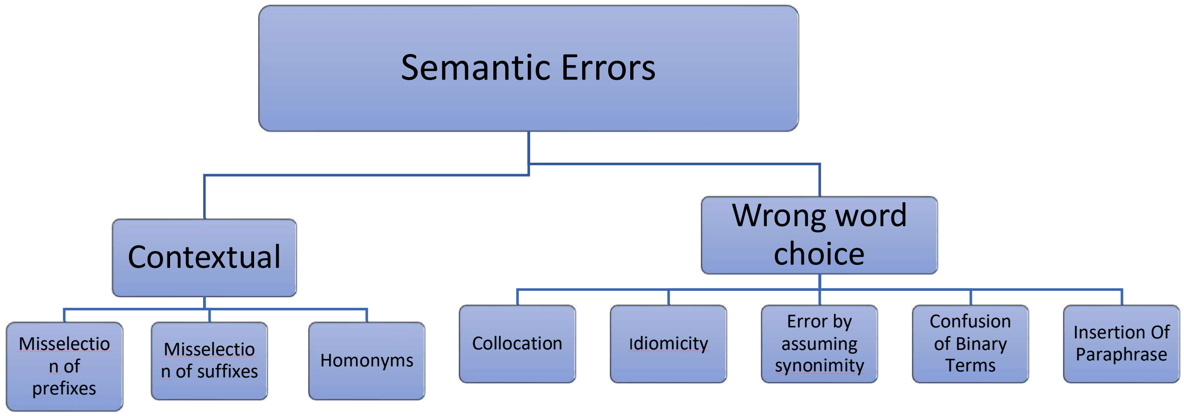
These are the errors that do not violate English grammar rules, but make the sentence senseless or absurd. However, a semantic error as used in this study can be defined as a violation of the rules of semantic system particular to English language. A semantic error can be a contextual error wrong word choice error.

When the writer’s knowledge of the meanings of words is imprecise, he or she may choose a word whose placement and syntax might seem appropriate but which is in fact incorrect. These types of sentences are correct in grammar but wrong semantically. The following example will give the correct view of these errors:

*The contractor was not happy with the progress at the construction **sight**.*

Here, the word sight seems correct in terms of syntax and grammar but it makes the meaning of the sentence unambiguous. The meaning of the word sight is ‘the ability to see something, or a thing one sees, but it does not fit properly with the meaning of the sentence, hence it is a semantic error. If we replace the word ‘sight’ with the word ‘site’ then the sentence would become more meaningful.

3.CLASSIFICATION



A semantic error can be a contextual error or wrong word choice error.

- **Contextual Error-** When a wrongly typed word is a real word in the language, it is not detected as a spelling error, yet it does not fit in the given context; such errors are called as contextual errors.

Example- Her mother prepared a delicious **desert**. (**dessert**)
He brought everything **accept** the book. (**except**)

- **Wrong word choice Error-** is using a rare word (possibly due to limited knowledge of vocabulary) which is often not used in the given context.

Example- We spent our afternoon **looking** movies. (**watching**)
The teacher asked us to meet him when he is **empty**. (**free**)

Semantic errors usually occur in language due to lack of knowledge. So, we have further classified these errors into following types:

Contextual errors

1. Homonyms:

Example: The tea was **two** hot to drink. (too)
Her mother prepared a delicious **desert**. (dessert)

2. MISSELECTION OF PREFIXES:

Example: This question is **disclear** for me. (unclear)
I am **nonhappy** in my studies. (unhappy)

3. MISSELECTION OF SUFFIXES:

Example: I am **interesting** in reading books. (interested)
I am an **ambitionable** person in my life. (ambitious)

Wrong word-choice errors

1. Collocation Error

Example: I am going to the tennis**field**. (court)

The teacher gives us **expensive** advice. (valuable)

2. Error by Assuming Synonymity:

Example: The teacher asked us to meet him when he is**empty**. (free)

I will **communicate** with you through email. (contact)

3. Confusion of Binary Terms (Relational Opposites):

Example: But I have to return the books I **lent** from the library. (borrowed)

Everyday I**come** to school on foot.(go)

4. Idiomacity

Example: I always get up at 6 o'clock. (wake up)

She **heard** English music today. (listened to)

5. Insertion of Paraphrase:

Example: The women who are **carrying babies** should stay at home. (pregnant)

Tomorrow, I have a party of my **day I was born**. (Birthday)

4.DATASETS

In this section, we discuss some of the popular datasets used to evaluate the performance of Grammatical Errors involving all types of errors including Semantic Errors also. As such there is no dataset prepared till date which is purely based on Semantic errors. We are listing here some of the datasets which have been used for evaluation of Grammatical as well as Collocation errors.

1. **CoNLL-2014 Shared Task:**
   The CoNLL-2014 shared task test set is the most widely used dataset to benchmark GEC systems. The test set contains 1,312 English sentences with error annotations by 2 expert annotators. Models are evaluated with MaxMatch scorer (Dahlmeier and Ng, 2012) which computes a span-based Fβ-score (β set to 0.5 to weight precision twice as recall).

2. **Lang-8 Corpus of Learner English:**
   Lang-8 is an online language learning website which encourages users to correct each other's grammar. The Lang-8 Corpus of Learner English is a somewhat-clean, English subsection of this website (Mizumoto et al., 2011; Tajiri et al., 2012).
   This corpus contains English learners texts extracted from Lang-8. It has 100,051 English entries written by 29,012 active users. We also include automatic tense/aspect annotation used in our ACL 2012 paper.

3. **NUCLE:**
   The National University of Singapore Corpus of Learner English (NUCLE) consists of 1,400 essays written by mainly Asian undergraduate students at the National University of Singapore (Dahlmeier et al., 2013).

4. **THE CAMBRIDGE LEARNER CORPUS (CLC FCE DATASET):**

1. THE CLC FCE DATASET IS A SET OF 1,244 EXAM SCRIPTS WRITTEN BY CANDIDATES SITTING THE CAMBRIDGE ESOL FIRST CERTIFICATE IN ENGLISH (FCE) EXAMINATION IN 2000 AND 2001.THE SCRIPTS ARE EXTRACTED FROM THE CAMBRIDGE LEARNER CORPUS (CLC), DEVELOPED AS A COLLABORATIVE EFFORT BETWEEN CAMBRIDGE UNIVERSITY PRESS AND CAMBRIDGE ASSESSMENT.

   For each exam script, the CLC FCE Dataset includes the original text written by the candidate (transcribed and anonymized, but otherwise unmodified) as well as marks, error annotation and essential demographic details including the candidate's first language and age bracket.

5. **Project Gutenberg:**

   The Standardized Project Gutenberg Corpus contains the files of all books that were part of Project Gutenberg (PG) which is used in the paper,Samanta, P., & Chaudhuri, B. B. (2013) A simple real word error detection and correction using local word bigram and trigram.

   In this paper, real word error has been simulated synthetically and this erroneous document is subjected to the proposed n-gram method in the paper. To make such a corrupted document, one in every 20 words is chosen. Suppose this current word is W. Then W is converted into a set of strings by one edit operation (insertion, deletion, substitution) at one-character position. If W contains n characters then n substitutions, n deletions and n+1 additions will create 3n+1 strings. From all the generated strings those words are found which are valid words. One of these valid words is chosen at random and W is replaced by this word. In this way, 100/20 = 5% real-word errors are introduced in the corpus. Here we have considered real-word error generated by single operation like substitution, deletion or insertion.

   The main problem in working with the problem of Semantic Errors detection and correction is the lack of availability of required dataset for training and testing. All the datasets are mainly based on GEC which includes very few semantic errors. So, there is further need to work on developing a dataset which is purely based on semantic errors.
   This would help in training the models especially based on Supervised learning models.

   In our research paper, the classification of Semantic Errors is based on the studies from different research papers and also **ESL Learner testset** mentioned in paper **Choosing the right word: Using bidirectional LSTM tagger for writing support systems.**

5.APPROACHES FOR THE DETECTION AND CORRECTION OF SEMANTIC ERROR

In the recent decades different approaches have been proposed by a number of people all around the globe and in this section, we have given a brief view of the approaches used for the above task.

KNOWLEDGE-BASED SEMANTIC DETECTION AND CORRECTION METHODS

The earlier models for the detection and correction of semantic errors are based on the knowledge-based methods that utilizes knowledge from different sources like the language corpuses or lexical databases and use this existing knowledge about the sequences and combinations of words to decide whether a given sentence is semantically correct or not. Most of the pertinent knowledge used by these models helps the system to decide what would be the most probable word that can be used in the sentence in place of some erroneous word.
Some knowledge-based methods for correction of the semantic error are-

Correction of Semantic Errors in Natural Language Texts with a Dictionary of Literal Paronyms

One of the earliest works in the knowledge based category is by Alexander Gelbukh and Igor A. Bolshakov in which they tended to detect and correct the malapropisms, [i.e., semantic errors replacing a word by another existing word similar in letter composition and/or sound but semantically incompatible with the context] by relying on a generator of correction candidates—paronyms [i.e., real words similar to the suspicious one encountered in the text and having the same grammatical properties] and arguing that a dictionary of literal paronyms [words at the distant of few editing operations from a given word] should be compiled beforehand and that its units should be grammeme names.

N-gram probabilistic model

The n-gram probabilistic model is one of the basic ways to find the real word semantic error and replace it with the most probable word from a set of similar words.
In the n-gram based real word semantic error detection and correction method, the scores of n-grams generated by the neighbours of the candidate word are combined. Mostly a single character position error model is assumed so that if a word W is erroneous then the correct word belongs to the set of real words c(W) generated by some character edit operation on W. In this model, the observed word W is assumed to be correct with probability or degree of belief α. Hence any member of c(W) is equally likely to be a correction candidate with constant probability $(1- \alpha) / n$ where n is the cardinality of c(W). The above combined score is calculated also on all members of s in the form of probability. These words are ranked in the decreasing order of the score. The member for which the sentence probability/score is maximum is the correction word. By observing the rank and using a rule-based approach, the error decision and correction candidates are simultaneously selected.
Fossati, D., & Di Eugenio, B. (2007), addressed the problem of real-word spell checking by proposing a methodology based on a mixed trigrams language model. Their experiments showed promising results with respect to the hit rates of both detection and correction, even though the false positive rate were still high.

P. Samanta and Bidyut B. Chaudhuri [2013] proposed another way to deal with the real word errors is based on bigram and trigram model. The method tries to detect an error by noting bigrams and trigram constituted by immediate left and right neighbour of candidate word and then generate some suggestions according to ranks/score calculated for the correction set of words. They used the BYU corpus ofbigram and trigram and tested the method on text from Project Gutenberg. Their proposed algorithm initially chose a confusion set for each candidate word using Levenshtein distanceequal to one from the dictionary words. Then it calculated the ranks of the elements of the confusion set based on which it detected an error and suggested some words against the detected error.

The n-gram models have also been used to correct the semantic errors in other languages other than English. K. M. Azharul Hasan, Muhammad Hozaifa and Sanjoy Dutta applied the same approach as P. Samanta and Bidyut B. Chaudhuri to identify and correct semantic errors in Bangla sentences. Shailza Kanwar, Manoj Sachan and Gurpreet Singh [2017] used the n-gram model forHindi sentences. Aqil M. Azmi, Manal N. Almutery, and Hatim A. Aboalsamh [2019] came up another hybrid approach combining the n-gram model with machine learning to avoid the use of predefined confusion sets. They use the pre-processing and feature extraction to classify the sentences in the detection stage and the n-gram model for the correction stage.

UNSUPERVISED SEMANTIC ERROR DETECTION AND CORRECTION METHODS

These methods tend to detect the patterns in which the words occur together in the sentences. It allows the model to work on its own to discover patterns and information that was previously undetected and use these patterns to identify the erroneous words in the sentences and replace them with some other word that fit into the detected pattern.

The unsupervised methods don’t need large scale labelled datasets with the correct and incorrect sentences but just a huge corpus of sentences that can be used to train the models for the detection of pattern in the sentences. With the unavailability of labelled datasets for the semantic errors these methods have been successful in finding a good replacement of the erroneous word in a sentence.

Error detection in content word combinations

Ekaterina Kochmar [2016] proposed a method to identify semantic errors through the content word combinations (adjective–noun (AN) and verb–object (VO) combinations, as they cover a substantial portion of learner errors in the use of content words). They implemented compositional distributional semantic models, and demonstrated how they can be applied to the learner data to detect errors in the choice of content words. They used the output of these models and derived “semantically informed” features which were used with a machine learning (ML) classifier. They designed their model on the idea that “words with similar meanings will occur with similar neighbours if enough text material is available” (Sch¨utze and Pedersen, 1995); “a representation that captures much of how words are used in natural context will capture much of what we mean by meaning” (Landauer and Dumais, 1997); and “words that occur in the same contexts tend to have similar meanings” (Pantel, 2005). Also, the distributional hypothesis claims that the words that co-occur with the target word and the

contexts in which the target word occurs implicitly describe the meaning of the word. Therefore, they assumed that the meaning of the word can be "accessed" through the observed examples of the word's use in context and can be represented with its co-occurrence counts with the other words in its context.
Their proposed algorithm performs EDC by comparing the original combinations to their possible alternatives and selecting the most fluent one according to the chosen measure of collocational strength. We show that an ML classifier that uses a small number of semantic features

Choosing the right word: Using bidirectional LSTM tagger for writing support systems
V. Makarenkov, L. Rokach, B. Shapira (2019) proposed the semantic error correction method through lexical substitution using the bidirectional LSTM tagger whichwas language independent because it was purely unsupervised and based on text corpora only, and does not involve an ensemble training. The model relied solely on an unlabelled data, without any human annotation and the application does not require a training of a specific classifier for each type of grammatical error or lexical substitution. The bidirectional LSTM tagger was applied to the task of text classification and each word was transformed into its dense embedding representation and then was fed into two LSTM networks: 1) left-to-right LSTM network, and 2) right-to-left LSTM network. The outputs of the two networks (the LSTM hidden states) were further concatenated and the classification was performed. When this approach was applied to text, model learned from target word's prefix (left-to-right) and suffix (right-to-left) contexts and performed the final classification based on the jointly learned representation of this context. The main limitation of this work is that their model does not attempt to address the detection of incorrect word usage, but only provide a better alternative to the targeted word and therefore the results are not directly comparable to other mentioned approaches or the existing GEC methods, which handles detection and correction and uses the F-measure as its evaluation metric for the complete test set.

SUPERVISED DEEP NEURAL NETWORK-BASED METHODS AND NEURAL MACHINE TRANSLATION

Error detection and correction methods have exploited the recent developments in neural networks to enhance performance. The most widely used techniques include Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM),Recursive Tree LSTM, encoder decoders, transformers and encoder decoder transformers. Encoder-decoder models are most widely used for machine translation from a source language to a target language. Similarly, an encoder-decoder model can be employed for GEC, where the encoder network is used to encode the potentially erroneous source sentence in vector space and a decoder network generates the corrected output sentence by using the source encoding.
Shamil Chollampatt and Hwee Tou Ng [2018] proposed automatic correction of grammatical, orthographic, and collocation errors in text using a multilayer convolutional encoder-decoder neural network. They mentioned that their model could find the semantic collocation errors along with the other grammatical errors but the explicit result regarding the performance of the model on the semantic errors was not calculated.

Other GEC methods based on the deep neural networks have been developed that attempt to correct not only grammatical errors, but also spelling and collocation errors like the GECToR – Grammatical Error Correction: Tag, Not Rewrite-GEC sequence tagger [2020] which uses a Transformer encoder and is pre-trained on synthetic data and then fine-tuned in two stages: first on errorful corpora, and second on a combination of errorful and error-free parallel corpora. But the performance results for the errors independently are not present foe these GEC models and hence the performance of these models on the semantic errors is unknown.

6.ANALYSIS

ERRORS		APPROACHES		
		KNOWLEDGE BASED (n-gram)*	SUPERVISED (CNN Model)**	UNSUPERVISED (Bi-LSTM Tagger)**
Contextual	Mis-selection of prefixes	✓	✓	✓
	Mis-selection of suffixes	✓	✓	✓
	Homonyms	✓	✓	✓
Wrong Word Choice Errors	Collocation	✓	✓	✓
	Idiomicity	×	✓	✓
	Error by assuming synonymity	✓	✓	✓
	Confusion of Binary Terms	×	×	✓
	Insertion Of Paraphrase	×	×	×

* Samanta, P., & Chaudhuri, B. B. (2013). A simple real word error detection and correction using local word bigram and trigram.
** Shamil Chollampatt and Hwee Tou Ng. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction.
*** Choosing the right word: Using bidirectional LSTM tagger for writing support systems.

In this section we try to summarize the advantages and disadvantages of the above methods. The papers onknowledge-based methods have reported the F1 scores ranging from 80-90 % which is good but the major drawback in these methods is that they solely rely on the knowledge bases and cannot encounter any uncommon sentence. Like for the n-gram approach by P. Samanta and Bidyut B. Chaudhuri [2013] the confusion set created for the word 'his' is {his, him, this, is, has} and the system can replace the word with only these alternatives and cannot go beyond this.

Other than this, the sentences used in these methods to check the precision and recall were created by introducing the error in the sentences using edit operations (insertion, deletion, substitution) on the words and replacing the correct word with the edited real word.

The unsupervised machine learning based methods detect the pattern of the sentences and can learn to fit a correct word in the sentence but the drawback these methods face is that they are trained on unlabelled data so the choice of correct word is not appropriate in many cases. The bidirectional LSTM tagger by V. Makarenkov, L. Rokach, B. Shapira (2019) could provide a better alternative to a word but cannot detect the target word or an erroneous word in a sentence, it needs the target word to be specified explicitly in order to provide an alternative to that word. The deep neural network-based approaches have been in more demand recently due to their ability of converting the bad grammar to good grammar and for providing a wide range of input. The models are rigorously trained on large annotated datasets and hence learn to translate the grammar accordingly and with the availability of large annotated datasets and corpuses these methods are achieving new progress day by day. But in case of semantic error the explicit performance of these GEC models is unknown due to the unavailability of datasets and test-sets on semantic errors. Also, till now no deep neural based model has been solely trained for the task of detection and correction of semantic errors hence the we cannot comment on the ability of neural machine translation to curb the problem of semantic errors.

APPROACH	DATA	FOCUS	REFER. CORPORA	DETECTION/CORRECTION	RESULTS	YEAR
Daniel Dahlmeier and Hwee Tou Ng, Correcting Semantic Collocation Errors with L1-induced Paraphrases	**NUS Corpus of Learner English (NUCLE)**	**error correction in learner English based on paraphrases extracted from parallel corpora.**	**an L1-English parallel corpus.**	**Correction of collocation errors**	**17.21MRR**	**2011**

Samanta, P., & Chaudhuri, B. B. (2013). A simple real word error detection and correction using local word bigram and trigram.	**Project Gutenberg**	**employs only two bigrams and one trigram around the test word** **Stemming based method**	**BYU corpus**	**both**	**Precision-71%-79%** **Recall-81-88%** **Accuracy-85%-93%**	**2013**
Ekaterina Kochmar. Error detection in content word combinations, Technical reports published by the University of Cambridge Computer Laboratory	**Cambridge Learner Corpus CLC-FCE dataset**	**3 models of compositional distributional semantics** **2content word combinations – (AN) and (VO) combinations.** **ML classifier**	**BNC & ukWaC**	**Main focus on detection**	**65%(approx.)**	**2016**
Shamil Chollampatt and Hwee Tou Ng. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction	**Concatenation of Reference corpora**	**convolutional encoder-decoder architecture with multiple layers of convolutions and attention** **larger English corpora to pre-train word embeddings and to train an N-gram language model**	**Lang-8 (Mizumoto et al. 2011) and** **NUCLE (Dahlmeier, Ng, and Wu 2013**	**grammatical errors, spelling and collocation errors correction**	**Prec. Recall F0.5** **Random 51.90 12.59 31.96** **Word2vec 52.80 12.80 32.49** **fastText 51.08**	**2018**

		to be used as a feature.			**13.63** **32.97**	
Detection and Correction of Real-word Errors in Bangla Language	**Self-prepared**	**N-gram, Markov model.**	**Self-developed corpora**	**both**	**Accuracy - 96%**	**2018**
Choosing the right word: Using bidirectional LSTM tagger for writing support systems	**ESL learner test-set**	**RNN based model**	**---**	**did not attempt to address the detection of incorrect word usage**	**0.25MRR**	**2019**
Real-Word Errors in Arabic Texts: A Better Algorithm for Detection and Correction	**SELF**	**FOR DETECTION- n-gram (n = 1–3) language model along with machine learning** **FOR CORRECTION-N-gram**	**KSU, ANCKACST, and JM corpus**	**both**	**Detection Precision 83.5% Recall 99.2%** **Correction- Accuracy- 98%**	**2019**

7.CONCLUSION

The task of detection and correction of semantic error has been one of the most challenging tasks in the field of Natural Language Processing. Various methodologies have been proposed over the years to detect and correct the semantic errors. The survey discusses various approaches that have been proposed over the years and the advantages and disadvantages of various methods. This survey would serve as a good foundation for researchers who intend to find new methods to measure semantic similarity.

8.REFERENCES

1. Obeidat, H. (1986). *An Investigation of Syntactic and Semantic Errors by Arab Learners* (Doctoral dissertation). University of Illinois at Urbana-Champaign.
2. D. Dahlmeier, H. T. Ng, and S. M. Wu, “Building a large annotated corpus of learner English: The NUS corpus of learner English,” in Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 22–31.
3. T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto, “The effect of learner corpus size in grammatical error correction of ESL writings,” in Proceedings of COLING 2012: Posters. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 863–872.
4. S. Chollampatt and H. T. Ng, “A multilayer convolutional encoder-decoder neural network for grammatical error correction,” CoRR, 2018.
5. T. Tajiri, M. Komachi, and Y. Matsumoto, “Tense and aspect error correction for ESL learners using global context,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 198–202.
6. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English Daniel Dahlmeier1,2 and Hwee Tou Ng2,3 and Siew Mei Wu4.
7. Samanta, P., & Chaudhuri, B. B. (2013) A simple real word error detection and correction using local word bigram and trigram.
8. Gelbukh A., Bolshakov I.A. (2004) On Correction of Semantic Errors in Natural Language Texts with a Dictionary of Literal Paronyms. In: Favela J., Menasalvas E., Chávez E. (eds) Advances in Web Intelligence. AWIC 2004. Lecture Notes in Computer Science, vol 3034. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24681-7_13
9. Fossati D., Di Eugenio B. (2007) A Mixed Trigrams Approach for Context Sensitive Spell Checking. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2007. Lecture Notes in Computer Science, vol 4394. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70939-8_55
10. Hasan, K. M. & Moaj, Hozaifa & Dutta, Sanjoy. (2015). Detection of semantic errors from simple Bangla sentences. 2014 17th International Conference on Computer and Information Technology, ICCIT 2014. 296-299. 10.1109/ICCITechn.2014.7073156.
11. Kanwar, S., Sachan, M., & Singh, G. N-GRAMS SOLUTION FOR ERROR DETECTION AND CORRECTION IN HINDI LANGUAGE. International Journal of Advanced Research in Computer Science, 8(7), 2017.
12. A. M. Azmi, M. N. Almutery and H. A. Aboalsamh, "Real-Word Errors in Arabic Texts: A Better Algorithm for Detection and Correction," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1308-1320, Aug. 2019, doi: 10.1109/TASLP.2019.2918404.

13. Schfttze, Hinrich & Pedersen, Jan. (2002). Information Retrieval Based on Word Senses.
14. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.
15. Pantel, Patrick. (2005). Inducing Ontological Co-occurrence Vectors. 10.3115/1219840.1219856.
16. Makarenkov, Victor & Rokach, Lior & Shapira, Bracha. (2019). Choosing the Right Word: Using Bidirectional LSTM Tagger for Writing Support Systems.
17. GECToR -- Grammatical Error Correction: Tag, Not Rewrite Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, Oleksandr Skurzhanskyi . arXiv:2005.12592 **[cs.CL]**
18. Ekaterina Kochmar. Error detection in content word combinations, Technical reports published by the University of Cambridge Computer Laboratory