English subordinate clause connective correction using NLP and hybrid model

[1]Ms. Kiran R. Borade, [2]Prof. N. M. Shahane
[1]PG student, [2]Associate Professor
[1]Department of Computer Engineering,
[1]K. K. Wagh Institute of Engineering Education & Research, Nashik.

Abstract : In this paper we proposed a novel method for English subordinate clause connective correction using NLP. New English learners make many grammatical mistakes in English subordinate clause connectives. It is not possible for checker to check each sentence for detecting and correcting connective errors. Hence for English subordinate clause connective error correction, an automatic error correction model is implemented. At first, syntactic and semantic features are extracted based on sentence context and then feature selection is done using metaheuristic algorithm and machine learning algorithm for classification. Technologies are combined and most fitted feature set is selected. This feature set is used by machine learning algorithm for classifying the connective. The proposed system is tested on Tatoeba dataset which is collection of sentences.

Index Terms **- Machine Learning; feature extraction; feature selection; connective correction; NLP.**

I. INTRODUCTION

English learners make many grammatical errors while writing. Mostly English as a second language (ESL) learners make many mistakes in grammar articles prepositions, determiners, spelling mistake and many more. Each error domain plays important role in grammar error correction. In such domains different strategies are used for grammar correction.
English as second language (ESL) learners frequently make errors in subordinate clause connectives due to lack of understanding correct use of connectives. Hence clause becomes an important grammar project in English.To correct this connective errors an automatic correction model is proposed which is based on metaheuristic algorithm and machine learning algorithm.Syntactic and semantic analysis is done using Stanford core NLP. It is a java pipeline framework which provides core natural language processing steps from tokenization to other annotations. It helps to identify the feature set that is helpful for feature extraction. Features to be extracted are predefined based on the study of type of error. Features are extracted according to surrounding context of a connective. Feature subset is selected using metaheuristic algorithm and then using machine learning algorithm classification is done.

In this paper, main focus is on clause connective correction, studying and implementing automatic connective error correction model based on metaheuristic algorithm and machine learning algorithm. Firstly using Stanford core NLP toolkit used for syntactic and semantic analysis of sentences, syntactic and semantic features are extracted according to sentence connective context and builds subordinate clause connective automatic correction model. The model adds an automatic feature selection module before training and testing data. This reduces the dimension of the data and improves error correction accuracy of connectives. Feature selection is done using heuristic algorithm and most fitted feature set is selected. This feature set is used by machine learning algorithm for classifying the connective. Finally comparison is done with existing algorithms. The error correction effect of proposed model can be optimal.
The paper is organized as follows: Section II represents review on grammar error correction. Section III represents algorithm used in system. Section IV represents Result and discussion and section V concludes the paper.

II. LITERATURE SURVEY

Stanford parser is employed which consists of a competitive phrase structure parser. Preposition selection is prejudiced by parse features which have a strong hold on selection in well formatted text. The performance of preposition error detection system using parse features is improved, even though errors of learner's text, parse features make small and non significant effect. The input sentence is split into chunks before and after the preposition and both is parsed separately. A preposition model is amplified with tokenization and parse feature. After examining the output of parser shows that parse features can be extracted from ESL data [1].

To correct errors of preposition and determiners, pipeline of confidence-weighted linear classifiers in system are used. In this system determiner and preposition correction is considered as classification problem. From possible correction based on confusion set confidence weighted linear classifiers are used to predict the correct word from set. Separate classifiers are built for correction of determiner errors, preposition replace errors, and preposition insert and deletion errors [1]. To form an error correction system the classifiers are combined into a pipeline of correction steps. System consists of pipeline of sequential steps where the output of one step serves as the input to the next step. Feature extraction analyzes the syntactic structure of the input sentences part-of-speech (POS) tagging, chunking, and parsing and relevant instances are identified for correction all noun phrases (NP) for correction of determiner. Determiner error correction is treated as a multiclass classification problem. A classifier is trained to predict the correct article from a set of possible article choices (a, the, an) as per the sentence context [2].

In feature extraction words surrounding the context of article are taken considered as features. Instead of considering features relying on human skill and prior knowledge in NLP, this approach simply helps in system automation. Depending on this approach both an error annotated corpus and an error non-annotated corpus is trained. It is possible to learn a strong statistical model on sufficient examples of error type. This system focus on the article error correction using Neural network CNN. The preprocessing module extract surrounding context of an article including that the article is not used while representing it is considered as X in features. Post processing Module, in English, there are rules to use a, an or the considering properties of word immediately after the or a/an. These rules implemented are employed to revise the output of our CNN module [3].

A maximum entropy classifier is trained to select among a/an, the, or zero i.e no article for noun phrases (NPs), based on a set of features extracted from local context of each article. The system uses features based on local context in form of words and POS tags to compute the probability that the NP will have a, an, the, or 0 article. The system's performance is evaluated in 2 ways: On held-out data from the same corpus as the training set, and on essays written for the Test of English as a Foreign Language by native speakers of Japanese and Russian[4].

Grammatical error correction methods that employ statistical machine translation (SMT) have been proposed for dealing with many grammatical errors. An SMT system generates instances with scores for all sentences and selects the sentence with the highest score as the correction result. In SMT system 1-best result is not always the best result. Ranking approach is proposed for grammatical error correction. Also to re-score N-best results of the SMT and reorder the results reranking approach is used. When we use the discriminative reranking with features, both precision and recall increases [5].

Number of discriminate features matter for higher classification performance. The smaller number of features reduces the problem's dimensionality and hence improves the performance. In DWFS (Dynamic wrapper feature selection) project is developed, which is a tool that allows selection of features for a variety of problems. DWFS follows the wrapper approach and applies search strategy based on Genetic Algorithms (GA).

A parallel GA implementation evaluates simultaneously large number of collections of features. DWFS integrates various filter methods that are applied as a pre-processing step in the feature selection step. According to the application requirements weights and parameters in the fitness function of GA can be adjusted. Experiments done using heterogeneous datasets from biomedical applications demonstrate that DWFS is fast and efficient in reduction of the number of features without sacrificing performance as compared to several existing methods [6].

In many domains the problem of selecting most useful features from lots of features in a low sample size data set arises. In such classification tasks feature subset selection is key problem. It is common to use filter methods. Because of correlation between genes due to which complexity of new feature techniques are in research hence ignorance in the correlations between genes which are prevalent in gene expression data and standard wrapper algorithms cannot be applied.

Additionally, existing methods are not especially able to handle the small sample size data which is one of the main cause's instability of feature selection. In order to deal with these issues, a new hybrid, filter and wrapper, based approach is proposed based on instance learning. Convert the problem to a tool that allows choosing only a few subsets of features in filter step. A cooperative subset search (CSS), is used with a classifier algorithm to represent an evaluation system of wrappers. Comparison results show that existing approach is better than other methods in terms of accuracy and stability of the subset[7].

In this paper an approach is proposed to solve problem of English article error correction, which is as important for a large proportion in grammatical errors. Genetic algorithm is employed for feature selection with confidence tuning for error correction. Machine learning based approach is considered which works on error annotated corpus and proposed a strategy to solve correction problem of article. At first, large numbers of related syntactic and semantic features are extracted from the context of connectives. With the help of genetic algorithm, a best feature subset is selected out which reduces feature dimensionality. For each testing instance, according to the predicted scores generated by the classifier, given approach measures the difference between scores in order to enhance the precision to a certain category[8].

The Stanford CoreNLP toolkit is designed, which is an extensible pipeline that provides core natural language analysis. It is used for semantic and syntactic analysis of sentences. It is easy for users to get started with framework, and to keep framework small, so it is easily comprehensible, and can easily be used as a component within the large system that a user is developing. It helps in analyzing the syntax and semantics of text data or sentences. It helps in feature extraction process it helps by letting know the meaning of the sentence according to which we can extract the features. For developers coding in Java Stanford NLP is used and for working in python use NLTK toolkit[9].

Text classification isto assign the documents to one of the predefined classes according to their contents. Text classification consists of many features which is a pattern recognition problem, where important step is feature selection. Still researches are going in which researchers are proposing new feature selection methods for text classification. Two-stage based feature selection methods are constituted by combining filter based feature selection methods along feature transformation methods and wrapper based feature selection method.

The main objective is to do two-stage feature selection methods analysis for text classification considering different views. Filter-based feature selection methods and feature set construction methods are considered in the first stage of two-stage feature selection and in second stage feature transformation methods PCA and wrapper-based feature selection method such as genetic algorithms (GA) used. kNN classifier is used for text classification [10].

In this paper wrapper based evolutionary algorithm is used for feature selection and kNN is used as machine learning algorithm. An automatic error correction model is built that makes English subordinate clause correction. [11]

III. PROPOSED SYSTEM

3.1 Architecture

English subordinate clause connective is a multiclass classification problem that involves number of word classes.
The connectives considered in error correction are what, who, when, where, that, which and if and &. Where & means there should be no connective.
English writing includes common mistakes like misuse of connectives and omission of connectives (or no connective required). System is able to correct these two errors. Figure 1 shows general architecture of proposed system with system module and their output.

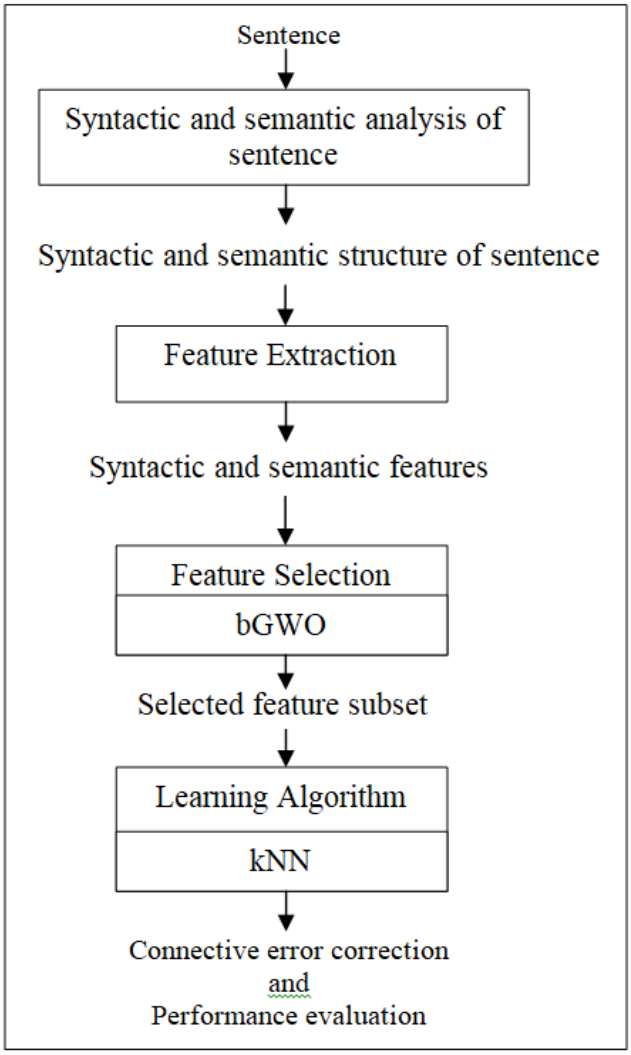


Figure 1: System Architecture

Dataset in given as input to the system and after that syntactic and semantic analysis of the sentences is done using the Stanford core NLP [9] libraries. Text sentences are taken as input into an Annotation object and then information is added by annotators in anpipeline of analysis. Sequences of annotators are:

1) Tokenization
2) Sentence splitting
3) Part of speech(POS)
4) Syntactic parsing

Dependency relationship, etc

After this feature selection algorithm is applied to extract the features from the sentences. For each sentence features are separated by a comma. Syntax and semantic features are extracted according to sentence connective (if, which, what, etc) context (local context).

Feature Extraction Algorithm:

(a) Feature extraction function(tok-sent,tok-pos)

(b) If a sentence contain connective:

Feature extraction based on the position of the connective.

(c) Return feature else,

Insert X as unknown connective based on study (connective appears before words containing PRP,VBZ, NN, VBD).

Feature extraction based on the position of the X.

(d) return feature.

where,

tok-sent and tok-pos is tokenization and part of speech of sentence, PRP is pronoun, VBZ is Verb 3rd person singular, NN is noun singular or mass, VBD is verb past tense.

Once the features are extracted, feature selection algorithm is implemented according to evaluation to select the features. Feature selection is divided into two approaches: Filter and Wrapper. In the filter feature selection, evaluation criterion of feature selection is independent of the learning algorithm and is obtained directly from dataset itself. The core of the Wrapper feature selection algorithm is that the evaluation performance of the selected feature subset based on the learning algorithm is better than filter

approach. Therefore, performance of the learning algorithms is considered to be the feature selection evaluation criterion in Wrapper feature selection is better.

Feature Selection:

The central outline of the proposed feature selection algorithm is as shown infigure 2: binary grey wolf optimization (bGWO)as search agent in feature selection. Before giving feature vector to bGWO values are binarized along with their feature index.

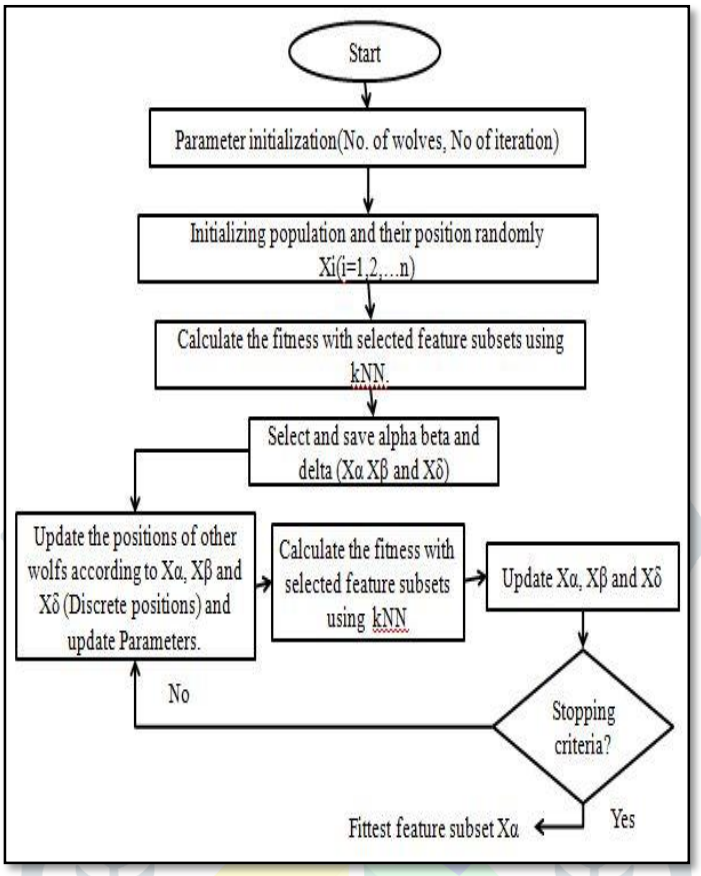


Figure 2. Wrapper feature selection using bGWO(search) and kNN(check hypothesis)

To update position of grey wolf according to fittest wolfs(Xα, Xβ, Xδ)following rule and parameters are considered:

(a) foreach Wolf_i∈pack do:

(b) X_i^{t+1} ←crossover among x1,x2,x3 using equation (3)(7)(11).

where, x1={x_1^1,....x_1^d}. Similarly,x2 and x3 are feature vectors generated.

Consider,

$\overrightarrow{A}$ = 2a $\overrightarrow{r1}$-a (1)

$\overrightarrow{C}$ = 2$\overrightarrow{r2}$ (2)

where, a is linearly decreased from 2 to 0 over the course of iteration and $\overrightarrow{r1}$,$\overrightarrow{r2}$are random vectors in [0,1].

x1,x2,x3arecalculatedasfollows:

$$x_1^d = \begin{cases} 1 \; if(X_\alpha^d + bstep_\alpha^d) \geq 1 \\ 0 \qquad otherwise \end{cases} \quad (3)$$

where, X_α^d is position vector of alpha wolf in d dimension and $bstep_\alpha^d$is a binary step in dimension d that can be calculated usig following equation:

$$bstep_\alpha^d = \begin{cases} 1 \; if(cstep_\alpha^d) \geq rand \\ 0 \qquad otherwise \end{cases} \quad (4)$$

where, rand is a random number drawn from uniform distribution ∈[0,1], and $cstep_\alpha^d$ is the continuous valued step size for dimension d and can be calculated by function given below:

$$cstep_\alpha^d = \frac{1}{1+e^{-10(A_1^d D_\alpha^d - 0.5)}} \quad (5)$$

where, A_1^d is calculated using equation 1, D_α^d is calculated using equations in the d dimension as below:

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1 X_\alpha} - \vec{X}\right| \quad (6)$$

where, $\overrightarrow{C_1}$iscalculatedasgiveninequation2andXisthecurrentpositionofwolf.

$$x_2^d = \begin{cases} 1 \; if\left(X_\beta^d + bstep_\beta^d\right) \geq 1 \\ 0 \qquad otherwise \end{cases} \quad (7)$$

where, X_β^d is position vector of alpha wolf in d dimension and $bstep_\beta^d$is a binary step in dimension d that can be calculated usig following equation:

$$bstep_\beta^d = \begin{cases} 1 \; if\left(cstep_\beta^d\right) \geq rand \\ 0 \qquad otherwise \end{cases} \quad (8)$$

where, rand is a random number drawn from uniform distribution ∈[0,1], and $cstep_\beta^d$ is the continuous valued step size for dimension d and can be calculated by function given below:

$$cstep_\beta^d = \frac{1}{1+e^{-10(A_1^d D_\beta^d - 0.5)}} \quad (9)$$

where, A_1^d is calculated using equation 1, D_β^d is calculatedusing equations in the d dimension as below:

$$\overrightarrow{D_\beta} = \left|\overrightarrow{C_1 X_\beta} - \vec{X}\right| \quad (10)$$

where, $\overrightarrow{C_1}$iscalculatedasgiveninequation2andXisthecurrentpositionofwolf.

$$x_3^d = \begin{cases} 1 \; if\left(X_\delta^d + bstep_\delta^d\right) \geq 1 \\ 0 \qquad otherwise \end{cases} \quad (11)$$

where, X_δ^d is position vector of alpha wolf in d dimension and $bstep_\delta^d$is a binary step in dimension d that can be calculated usig following equation:

$$bstep_\delta^d = \begin{cases} 1 \; if\left(cstep_\delta^d\right) \geq rand \\ 0 \qquad otherwise \end{cases} \quad (12)$$

where, rand is a random number drawn from uniform distribution ∈[0,1], and $cstep_\delta^d$ is the continuous valued step size for dimension d and can be calculated by function given below:

$$cstep_\delta^d = \frac{1}{1+e^{-10(A_1^d D_\delta^d - 0.5)}} \quad (13)$$

where, A_1^d is calculated using equation 1, D_δ^d is calculatedusing equations in the d dimension as below:

$$\overrightarrow{D_\delta} = \left|\overrightarrow{C_1 X_\delta} - \vec{X}\right| \quad (14)$$

where, $\overrightarrow{C_1}$iscalculatedasgiveninequation2andXisthecurrentpositionofwolf.

KNN classification algorithm is applied for classification of the result. features considered are the feature subset ones that are selected using feature selection algorithm. Features of test vector are matched with features of each instance in training dataset and assign the class of most matched vector using matching criteria as hamming distance in kNN.

Binarization of Data

Features extracted are represented as binary vector using binarization technique which is given as input to feature selection algorithm.Selected features are given to machine learning algorithm for connective error correction.

It is also called one-hot encoding. Its only needs to be performed with the categorical variable. The feature has categorical values hence before using feature selection all features are encoded into binary string. That is a sequence consisting of 0 and 1. Where 0 represents feature is not present 1 indicates feature is present.

Classification algorithm used is kNN in proposed system.

IV. RESULTS

4.1 System Configuration:

The system is going to be implemented in java using jdk 1.8.0 along with Netbeans-8.0 IDE. The test results are generated on Intel core i5 Processor and 4GB RAM on windows 10.

4.2 *Dataset:* Tatoeba[12] is free online database that collects foreign language learners including English, German, French and Chinese, etc. The Tatoeba Corpus consists of multilingual sentences out of which English corpus consisting of English sentences is selected as experimental data.Total 1000 English sentences are selected out of which 400 are English subordinate clause sentences on which proposed system works.

4.3 *Experimental setup:*Experiments are going to be conducted using Holdout method. Hence out of 1000 sentences work is done on 400 sentences which are having connectives 300 are taken for training and 100 for testing.

4.4 *Performance Metric:*

There are four performance evaluation indexes of the performance of English subordinate clause connective correction: accuracy rate (Accuracy), precision rate P (Precision), recalls rate R (Recall) and F value (F1), and their formulas are as follows:

Accuracy = TP+TN/TP+FN+FP+TN

Precision = TP/TP+FP

Recall = TP/TP+FN

F1-measure= 2 x Precision x Recall / Precision x Recall where,

TP = Number of sentences with correct connective that were classified as correct connective.

TN = Number of sentences with incorrect connective that were classified as incorrect connective.

FP = Number of sentences with incorrect connective that were misclassified as having correct connective.

FN = Number of sentences with correct connective that were misclassified as having incorrect connective.

The performance is going to be measured automatically after classification.

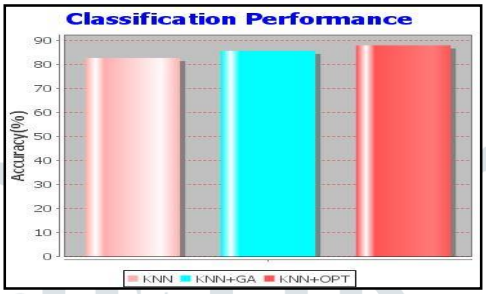
4.5 *System Result comparision:*

Algorithm	Accuracy	Precision	Recall	F-Measure
KNN	70.90899999999999	0.301	0.409	0.343
NB	65.932	0.316	0.369	0.329
DS	90.773	0.301	0.398	0.313
KNN+GA	76.75	0.277	0.438	0.337
NB+GA	72.455	0.279	0.455	0.343

Algorithm	Accuracy	Precision	Recall	F-Measure
KNN	70.90899999...	0.301	0.409	0.343
KNN+GA	76.75	0.277	0.438	0.337
KNN+Opt	91.75	0.769	0.638	0.833

Figure4. Comparison of proposed bGWO + kNN system with other existing systems.

Figure.4 shows the comparison of proposed system (bGWO+kNN) with other existing systems. It shows that proposed system enhances or improves performance of English subordinate clause connective correction. Following are graphs that show analysis of system performance.



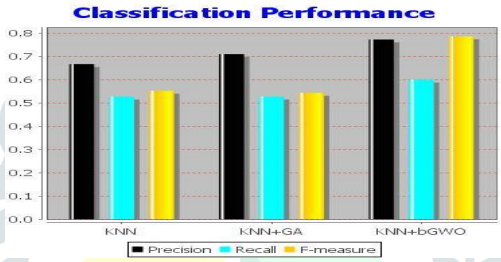


Figure 5. Graphs showing variation in performances of existing and proposed system(bGWO+knn).

V. CONCLUSION

In this paper we proposed a novel method for English subordinate clause connective correction and to detect the clause connective error. Feature extraction is done by taking into account system configuration. bGWO feature selection algorithm is used to enhance system performance. The system measures the execution performance with respect to accuracy under various execution criteria such as features to be extracted, feature selection technique and learning algorithm used. Proposed system is tested on tatoeba dataset and the results depicts our proposed system performs well as compare to the state of art systems studied in literature survey.System performance is compared with other techniques. In future, different metaheuristic algorithms can be used and also techniques used in proposed system can be used for speech recognition system.

REFERENCES

[1] Tetreault, Joel, Jennifer Foster, and Martin Chodorow. "Using parse features for preposition selection and error detection." In Proceedings of the acl 2010 conference short papers, pp. 353-358. Association for Computational Linguistics, 2010.

[2] Dahlmeier, Daniel, HweeTou Ng, and Eric Jun Feng Ng. "NUS at the HOO 2012 Shared Task." In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 216-224. Association for Computational Linguistics, 2012.

[3] Sun, Chengjie, Xiaoqiang Jin, Lei Lin, Yuming Zhao, and Xiaolong Wang. "Convolutional neural networks for correcting English article errors." In Natural Language Processing and Chinese Computing, pp. 102-110. Springer, Cham, 2015.

[4] Han, Na-Rae, Martin Chodorow, and Claudia Leacock. "Detecting errors in English article usage by non-native speakers." Natural Language Engineering 12, no. 2 (2006): 115-129.

[5] Mizumoto, Tomoya, and Yuji Matsumoto. "Discriminative reranking for grammatical error correction with statistical machine translation." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1133-1138. 2016.

[6] Soufan, Othman, DimitriosKleftogiannis, PanosKalnis, and Vladimir B. Bajic. "DWFS: a wrapper feature selection tool based on a parallel genetic algorithm." PloS one 10, no. 2 (2015): e0117988.

[7] Brahim, Afef Ben, and Mohamed Limam. "A hybrid feature selection method based on instance learning and cooperative subset search." Pattern Recognition Letters 69 (2016): 28-34.

[8] Xiang, Yang, Yaoyun Zhang, Xiaolong Wang, Chongqiang Wei, Wen Zheng, Xiaoqiang Zhou, Yuxiu Hu, and Yang Qin. "Grammatical error correction using feature selection and confidence tuning." In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1067-1071. 2013.

[9] Manning, Christopher, MihaiSurdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP natural language processing toolkit." In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60. 2014.

[10] Uysal, AlperKursat. "On Two-Stage Feature Selection Methods for Text Classification." IEEE Access 6 (2018): 43233-43251.

[11] Huang, Guimin, Chuang Wu, Sirui Huang, Hongtao Zhu, Ruyu Mo, and Ya Zhou. "An english subordinate clause connective correction model based on genetic algorithm and k-nearest neighbor algorithm." In Progress in Informatics and Computing (PIC), 2017 International Conference on, pp. 302-306. IEEE, 2017.

[12] https://tatoeba.org/eng/stats/sentences_by_language