The Effect of Grammatical Correctness for Stack Overflow Question Success

Russell Woodcock
Adelaide University
a1706856@student.adelaide.edu.au

ABSTRACT

Online forums such as StackExchange and StackOverflow provide a platform for uses to share knowledge. Questions are answered at the discretion of other users, and many Stack Overflow questions remain unanswered. Many relevant papers seek to identify to factors that determine the quality of both questions and answers on Stack Overflow. One such factor that has been neglected in such works is grammatical correctness. This research paper seeks to determine the effect of correct grammar usage on the success of Stack Overflow questions. To answer this, an empirical study was implemented which analysed a large dataset of Stack Overflow questions for grammar errors using a high quality grammar checking library, and these results were analysed to determine the correlation between correct grammar and question success. As a secondary contribution, the distribution of correct grammar usage among Stack Overflow questions was identified. It was demonstrated that grammatical correctness was very weakly and positively correlated to the success of the question.

1. Introduction

Online forums such as StackExchange and StackOverflow provide a platform for uses to share knowledge. Those seeking answers for questions must rely on other users to provide relevant and effective answers, and there is a very real risk of a question not being answered entirely.

Stack Overflows provides a means for question answers to indicate satisfactory answers based solely on the asker's discretion. Any user with an account can vote on the quality of all questions, comments and answers, which increments or decrements its score. This score directly affects the rank of questions in the website. Higher scoring questions are more visible to general users on the website, and higher scoring questions have been clearly shown to receive more comments and answers [1].

Stack Overflow has implemented various strategies aimed at maintaining a high level of content quality, which including allowing users with high reputations to directly edit questions, however, there is no guarantee that questions that poorly communicate to the community will be edited. Unsuccessful questions can remain stuck in a position in which they lack visibility, which is a major predictor of questions remaining unanswered [1].

Analysing how the grammatical correctness of Stack Overflow questions contribute to a lack of response is of focus in this paper. As such, an empirical study was employed which analysed data from Stack Overflow's latest data dump for grammatical correctness in user question bodies, and the correlation of this to the success of Stack Overflow questions was assessed.

The use of **rule-based** grammar checking is proposed in this paper to perform grammar analysis of Stack Overflow questions. Using a rule based method here allows this analysis to commence without the need for an appropriate training data set, which for the context of online communication forums such as Stack Overflow, lacks development. The *LanguageTool* python wrapper is used to perform this grammar analysis.

Linear regression had been used in the past by *A. Baltadzhieva* and *G.Chrupała* to predict the success of questions based on shallow attributes, and this approach is applicable for grammar checking. In this paper, linear regression is used to assess the correlation between grammatical correctness and question success variables, including the number of answers, existence of an accepted answer, and the score awarded to a given question.

2. Motivation

While relevant papers do assess the effect of question quality attributes on the success of Stack Overflow questions, the task of investigating the effect of grammar on this success has not been performed in depth. Important contributions of this paper are twofold. Firstly, the correlation of grammar error frequency in Stack Overflow questions to the success of the question is identified, and secondly, raw data is compiled that describes the grammatical quality of the dataset. This includes the total number of errors and the frequency of errors in a question. Compiling this intermediate dataset is a computationally time consuming process, and is therefore of value.

The research topic is of interest to non-English speakers and those who speak English as a second langue, who are commonly involved in Stack Overflow [2]. While, many foreign versions of Stack Overflow exist, the English version has the largest database of answered questions and the highest number of users by far. This poses a challenge for Non-English fluent speakers who may write questions in non-correct English or use online translation services which may not provide perfect translations. If these people and more general users are not communicating effectively in their answers, then it can be expected that these questions are of limited success, and the questions are ultimately unanswered. For

Stack Overflow users, quality writing is generally considered necessary to engage potential question answerers, who may not wish to give their time. It would be important to those who lack proper writing skills to know exactly what effects a correctly written questions has on its outcomes.

3. Background

Over the past few decades, text editors such as Microsoft Word have become the norm for writing and publishing texts of all kinds. With these digital text editors, also came the introduction of live spelling and grammar checking features, where grammatically incorrect passages of text are flagged as incorrect, the category of error is identified, and suggestions to rectify the error are made.

Two general types of grammar checking libraries re most common: rule based and statistical natural language processing. For rule based checking, a large number of hand crafted rules are required. The disadvantage of this approach is that it is language specific and sometimes an exhaustive number of rules have to be developed manually, which will not be applicable to other languages [3]. The main advantage of this technique is that the developed rules can be developed over time, achieving a high degree of perfection [3] [4]. New rules can be added and existing rules can be deleted and existing knowledge can be easily applied.

In the late-1980s and mid-1990s, automated statistical analysis became much more commonplace, and since then natural language processing research has focused on machine learning [5]. The machine-learning method of grammar checking uses statistical inference to automatically learn grammar rules through the analysis of large corpora of typical real-world examples. A corpus (plural, "corpora") is a set of documents. Systems based on machine-learning algorithms have several advantages over hand-produced rules:

Firstly, the learning procedures used during machine learning automatically focus on the cases which are most statistically relevant, whereas manually produced rules can vary in the actual statistical relevance of the rule, and so much of the effort in the process of developing these hand written rules can be inefficiently directed [5].

Secondly, automatic learning procedures can make use of statistical-inference algorithms to produce models that are applicable to new input types that have not been seen before. It is also able to handle erroneous input such as misspelled words or words accidentally omitted. In general, handling these cases with hand-written rules is very time-consuming and difficult, as there are many cases in implement [5].

However, statistical inference based grammar checkers rely on the existence of training data. This data can be very difficult to produce depending on the context and to construct a corpus or corpora for a new context is not a trivial task [6]. Often, sub-optimal training data sets rather than perfect training sets must be used which can negate the benefit of following the machine learning method [6]. The training process to create the statistical model can be a time consuming process computationally, and must be repeated for new corpora [6].

4. Research Method

4.1 Research Questions

A number of research questions to be answered in the paper are listed below. These questions reflect the need to evaluate both the influence of grammar correctness for question success, and the current state of grammar correctness in Stack Overflow questions.

R1. What effect does incorrect use of grammar or language have on the outcomes of these questions, in terms of popularity and success?

R2. How is the distribution of correct or incorrect use of grammar among Stack Overflow questions?

4.2 Data Collection

Stack Overflow releases a comprehensive data dump of all public activities every few years. Relevant to this paper, the full texts and attributes of every question on Stack Overflow is available in a raw data format. This data is freely available and is hosted by many sources. The data sourced for all purposes in this paper were downloaded from *archive.org*.

The dataset used contains questions for the year leading up to the 3rd of June, 2019. From this dataset, 500,000 questions were included in data analysis. Out of the total number of questions analysed, 51851 remained unanswered (10.37%). The dataset contains information regarding the question title, body, up votes, down votes, and others. The question owner, e.g. registration status, reputation, name, id are also available. The variables as described below are used in data analysis:

- Body - String containing the text of the question body.
- Question Score - Reputation awarded by users to the question.
- Number of answers -Number of answers to the question (not including comments)
- Accepted Answer ID - ID of accepted answer of a question. This data value is not present in questions which have no accepted answer.

In order to aggregate the quality of a question's title and body, the text of each question body from the dataset was extracted and analysed using *LanguageTool*. Both the number of grammar errors, as well as the frequency of grammar errors were recorded. A Python scrip was used to extract question body and title texts from the dataset and to then analyse these texts for grammatical correctness.

A Python wrapper of the online grammar checker *LanguageTool* was used to perform grammar checking operations. This library operates based on a rule based grammar checking methodology, rather than a statistical - NLP based one. This tool has maintained a high level of support and as such can be considered to produce quality results compared to competing solutions.

On Stack Overflow, users can style their questions using generic HTML mark-up language. Users must include mark-up tags which specify blocks of text which are to be styled as determined by the tag. To apply a tag, an opening and closing tag must be written by the user, and all code within the opening and closing tags is modified. For example, it is common practice to include code snippets within the <code> tag, which applies a styling to the text which identifies the text visually to readers as code. These HTML tags are included in the dataset and must be removed.

Text within some tags, such as the <code> or <a href> tags must be omitted from each question's body before being analyse for grammar errors, as the grammar checking libraries used are not applicable to code or urls, and similar such cases. Some tags only provide an estimate of the text that lies inside the tag. For example, code is usually present in the <code> tag, but this is not always the case as it is up to the Stack Overflow user to include the <code> formatting tag. Text within some tags should not be omitted, such as those which format lists or apply generic styles like font size or colour. There are a large number of these tags and each must be considered if appropriate to include for grammar checking. A blacklist describing common HTML tags which are not applicable text for grammar checking was created. Text within tags that are in this list are omitted from grammar checking. Self-contained tags do not need to be included in this list as not text lies within these tags. This list is displayed below:

Tag	Description
<code>	Commonly used to format code.
<time>	Used to format times.
<tbody>	Used to format tables.
	Creates a listed item.

Table 1. Contains HTML tags present in Stack Overflow question bodies which must be omitted in grammar checking.

A Sample of 300 questions bodies that were analysed for grammar errors were manually checking whether any omittion of HTML content broke the grammatical flow of the question. It was found that in 97% cases, omitting html content did not cause the introduction of invalid grammar errors.

The values for the number of answers, score of question and answered status of each question were considered as quality attributed which measure the success of a question. This values were extracted using a python script directly from the downloaded dataset.

4.3 Linear Regression Analysis

Linear regression models were used to determine correlation between question success variables – *question score, number of questions & accepted answer status,* and question quality variables – number of grammar errors & grammar errors per character.

A possible factor in determining the effect of grammar misuse in question bodies is the rate at which mistakes are noticed by the reader. For this reason, the number of grammar mistakes per length of the question body is also taken into consideration.

Coefficient of determination (r^2) values were calculated for each pair of question success variable set and question grammar quality set. An outline of this mathematical process is described. The expected relationship in a linear model and the general definition of the coefficient of correlation used are given below:

$$y_i = ax_i + b$$

Formula 1 – describes the general linear regression model, where x_i denotes the independent variable of question i and y_i the dependent variable for question i. a and b are unknown constants.

$$\boldsymbol{r^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}}$$

Formula 2 – describes the general definition of the coefficient of determination. SSres denotes the regression sum of squares, which mathematically describes the how well a linear regression model represents the data being modelled. SStot denotes the residual sum of squares, which describes the discrepancy between the data and a linear regression model.

All data produced is stored by the python script in an intermediate CSV file.Using this intermediate data format, NumPy is used to perform distribution analysis. In order to answer R2, histograms were constructed and distribution attributes such as the mean, median were also recorded. These are presented in the results section.

4.4 Python Scripting Structure

A python script was used to perform all grammar analysis. This script performed this task used default python libraries, the grammar check library for *LanguageTool*, SciPy and Numpy. The grammar checking library contains necessary function declarations necessary to perform the required operations. The

python wrapper for *LanguageTool* itself operates by calling on a local client-side server which is provided as a downloadable on the *LanguageTool* website. This *langaugeTool* server must be started before any operations can commence, and this adds additional overhead to the python script runtime.

Some features of *LanguageTool* are open source; however the main engine and other important features are closed. *LanguageTool* is commercial and this has some ramifications for the downloadable server. This server, although ran locally places some restrictions on the number of words that can be processed in a single batch. In order to get around this problem, the server has to be restarted periodically. To achieve this, a bash script was used to run the python script multiple times for different batches of data values from the dataset. 1000 blocks of 500 paragraphs were used in each batch, which was low enough to keep the word count limit under the required threshold in all batches. Running the program in batches also alleviates the risk of fatal bugs or crashes. The server which language tool operates on is commonly at risk of losing connection for unknown reasons. This is a commonly reported bug for the python wrapper version of LanguageTool, and there is not current solution that fixes it.

In practice, linear regression is performed using the *SciPy and* Numpy python library and all supporting operations were automated using python.

All code can be found on the Github repository for this research paper. The files of importance in the github repository are the following:

- results.csv – contains all results compiled in this research project
- grammar_check.py – performs a batch of grammar checks. Question bodies are read in from the *results.csv* file and results are written to the results.csv file.
- analyse.py – contains all code used to perform linear regression analysis on values recorded in results.csv.
- run_grammar_check.sh – bash script that runs the overarching grammar checking process

5. Results

The results of the linear regression analysis is shown in the following two tables:

	r	r^2	β
no. answers	-0.11776	0.01386	-0.12441
score	-0.19447	0.03782	-0.01646
accepted answer	-0.03512	0.001234	-1.3442

Table 2. Displays linear regression analysis results for the **total number of errors** against the three success variables: the number of answers to a question, the awarded score of a question and whether the question has an accepted answer (which is true or false value). r denotes the correlation coefficient, r^2 denotes the coefficient of determination and β denotes the slope of the best fit model.

	r	r^2	β
no. answers	-0.10235	0.01048	-0.05624
score	-0.12508	0.01564	-0.03253
Accepted answer	-0.12486	0.01559	-0.05433

Table 3. Displays linear regression analysis results for the **error density** against the three success variables: the number of answers to a question, the awarded score of a question and whether the question has an accepted answer (which is true or false value). The parameters used in this table continue from table 2.

These results suggest that correlation between the grammar correctness of question bodies, as determined by the grammar checking method used, is very low. In regards to the research question R1 - *What effect does incorrect use of grammar or language have on the outcomes of these questions, in terms of popularity and success?* There seems to be a very low negative correlation between the number and rate of grammar errors in these questions. Furthermore, the slopes formed from these linear regression models are flat, which indicates there would only be a minimal decrease in question success attributes for larger numbers of grammar errors present.

Given the intermediate dataset compiled in this study, the distribution of the frequency and number of grammar errors is determined. A histogram showing this distribution is displayed below:

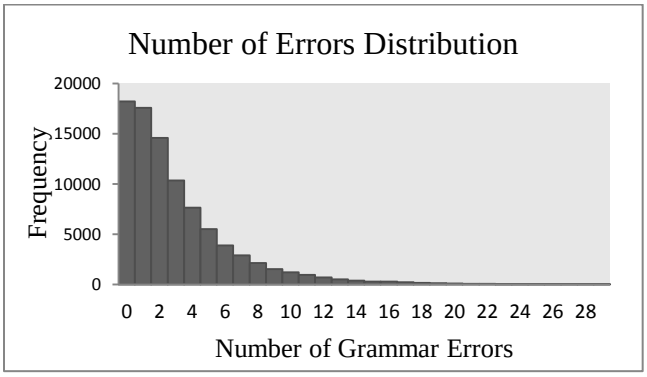

Chart 1. This histogram shows the distribution of the **number of errors** variable in the data produced. A bin size of 1 was used. This chart can be interpreted as describing the distribution of correct use of grammar in Stack Overflow questions. The mean number of errors for all Stack Overflow questions is 3.2578 and the median is 3 errors. 95% of all questions contain 10 grammar errors or less.

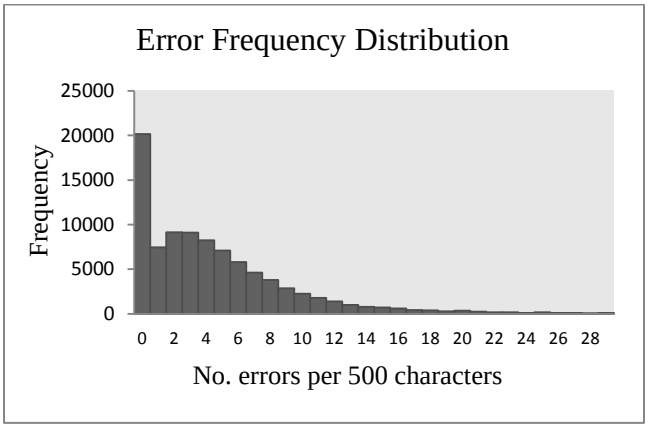

Chart 2. This histogram shows the distribution of the **error density** variable in the data produced. A bin size of 1 was used. This chart can be interpreted as describing the distribution of correct use of grammar in Stack Overflow questions. The mean frequency for all Stack Overflow questions is 4.9058 errors/500 characters and the median is 5 errors/500 characters. 95% of all questions contain less than 15 errors/500 characters.

These distributions demonstrate the apparent grammar correctness of Stack Overflow questions. To answer R2 - *How is the distribution of correct or incorrect use of grammar among Stack Overflow questions?* Questions on Stack Overflow follow a right skewed distribution with questions containing 0 grammar errors being most common with 50% of all questions containing 3 or fewer errors. This distribution is also seen when the number of errors is adjusted by the length of the question, with 95% of question containing 5 or less grammar errors per 500 characters.

6. Discussion

It is notable that the r^2 values that were calculated indicate very low correlation. A more significant correlation was expected.
From these results, a cause and effect relationship cannot be identified of the frequency or density of grammar mistakes to the success of the question as there is no significant correlation.

It is also notable that for most question bodies' analysed, no or few errors were present. Specifically, it was found that half of question bodies contained three or fewer grammar errors with the highest frequency of question bodies containing no grammar errors. Furthermore, over half contained less than 5 grammar errors per 500 characters. This suggests for the most part, that the Stack Overflow community maintains a high quality use of grammar.

A consideration in these results is that the effect of poor grammar can be expected to vary for different contexts. Online forums such as Stack Overflow have been observed to contain answers with minimal length compared with the actual amount of information they convey [7]. This can be expected to be relevant factor in the effect of poor grammar in Stack Overflow questions, as some grammar mistakes may not diminish the questioner's credibility or perceived level of effort so long as the question contains all informational requirements that allow the question to be answered. This measure would not be reflected by simply tallying the number of grammar mistakes in each question's body and could be a reason for the low correlation found.

Furthermore, a promoted activity on Stack Overflow is the editing of questions which are not clear or don't follow community guidelines. It is common for such questions to be edited by other users who possess a high reputation score on Stack Overflow, rather than site administrators or staff [7]. The dataset used for this research study was compiled in a data dump, and as such it reflects the visible question bodies at the time. Meaning some edited questions will be in the dataset, and some questions in the dataset are yet to be edited in the future. This is certainly a large contributor to the high quality and consistent use of grammar in the dataset.

Another possible effect contributing to the high quality of grammar in Stack Overflow questions in the quality of translators. It would be expected that if non English speakers utilised online translators, then the quality of the questions written would depend on the translation service. This topic lacks research development however, so the extent of this particular factor is not known.

7. Threats to validity

7.1 HTML tag issues

The Stack Overflow text editor can be considered powerful as users to access some basic HTML elements to format text. Due to the large number of possible HTML tags that could be used by a user, a blacklist was created containing only known html tags with text that should not be omitted from grammar checking. All other text was omitted to minimise the amount of text that was not applicable to grammar checking was not included in grammar checking tasks. However, the introduction of HTML content in the dataset introduces the possibility of erroneous input, where unexpected grammar cases arise. It is clear from manually viewing a large number of question bodies that these erroneous cases do exist, although they are rare. In some cases, HTML styles are important to the flow of the grammatical structure of a question. For example, there exists more than one question which uses hyperlinks as nouns in a regular paragraph, which in this study has to be omitted. This is one issue of using more traditionally focused grammar checking solutions on a unique context such as online content. In practice, using url links as nouns is valid communication in contexts such as Stack Overflow, so some erroneous errors should be considered valid or invalid. Although through manual verification, it was concluded that only a minimal number of questions did not omit sections that were vital to the flow of the question, there dataset still contains many erroneous cases which will be analysed as including some grammar errors according to the grammar checking, but not so to the Stack Overflow community.

Another key consideration regarding HTML tags is that the removal of sections of text in between these tags is a process of estimation. Users in Stack Overflow are expected to use proper HTML tags to style the body of their question, and there are provisions in place to ensure this. Questions that are not correctly formatted may be edited by other users with high reputation. The results of this study rely on pieces of questions that must be omitted form grammar checking correctly being removed. Manual checks suggest that the overwhelming number of pieces of information that are expected to be contained within the appropriate HTML tags, are indeed contained within the HTML tag, so the impact of this risk would be minimal.

7.2 Quality of rule based grammar checkers.

There have been some great challenges in the development of grammar checking software. While grammar checkers of this category do not require training, they rely on the quality of manual rules that govern correct grammar usage. *LanguageTool* is in limited open source [8], with a total of 88 individual contributors, and has been extensively documented. However as noted in [8], rule based grammar checkers risk lacking some more convoluted grammar error cases as rules are based on theoretical understandings of grammar but these generally do not cover all cases, although have been shown to cover the vast majority of them[8]. The results compiled in this study are dependent on the quality of the *LanguageTool* python wrapper. While the tool is generally considered of high competitive quality, it cannot be ensured that all cases are covered in the grammar checking rule set.

A more general criticism towards rule based grammar checkers for analysing grammar in a new context is that it lacks the ability to adapt to new grammar assessment criteria, and this is certainly true for Stack Overflow. One relevant consideration here is that trivial grammar errors that break rule based grammar checks are not as influential as other more convoluted forms of errors that cause the communication in a question body to break down, and that these other errors may be specific to the Stack Overflow platform, which rule based grammar checking libraries will not be able to adapt to.

7.3 Timeframe limitations

It has been observed that there is very high standard deviation in the score given to questions with high grammar error densities. While the sample size was very high, the frequency of grammar error density values was found to be highly skewed in favour of low values (most questions have none or next to none grammar errors), so the sample size considering only questions with high density of errors relatively low compared to the standard deviation, even though the total sample size was high at 500,000 questions. This sample size is limited by time constraints regarding computations. The task of checking the grammar of each question's body is a bottleneck in this process. Increasing the sample size would be ideal, but any meaningful increase n sample size would increase the computational time by several days. An increase in sample size can be expected to yield a smaller standard deviation, and therefore a more accurate linear regression model and correlation. This requirements would require a much larger project timeframe however.

7.4 Confounding factors

There is the possibility of unidentified confounding factors altering the perceived influence of grammar errors. Question body length is an example of this, and this was controlled for in linear regression, however because the convoluted nature of the dataset, in which there is a very large number of factors influencing the outcome of Stack Overflow questions and answers [9][10], it is not clear if all confounding factors have been identified. For this reason, demonstrating a cause and effect relationship is difficult using the supplied Stack Overflow dataset, however demonstrating correlation between variables is useful in constructing a prediction model for the success of questions, where many variables are taken into consideration.

7.5 Known technical issues in *LanguageTool*

There are some known bugs in *LanguageTool*. Although this software has been in development for a long time and is considered to be completed in its intended scope or features, there are still some current bugs in circulation as identified in the tool's GitHub repository. Developers have described all active issues as of minimal or negligible impact, however there is no 100% guarantee of this.

8. Related work

There are numerous studies that aim to predict the quality of both answers and questions regarding Stack Overflow. *A Baltadzhieva & G. Chrupała* [*10*] used linear regression to predict Stack Overflow question quality based on a small number of available features. The quality they measured was an aggregate of the score of the question and the number of replies to the question's thread. They found that questions that are considered not worthy of receiving an answer are generally questions that include typos or that are found to be off-topic. Questions that contained typos were also least successful. Questions with relatively large body sizes negatively affected question quality, though this effect was small.

R. Saha, A. Saha* & D. Perry [11] performed an automated analysis to understand the reasons why Stack Overflow questions remained unanswered. A number of shallow factors were identified, including the question's number of tags, body length, score and other. These features were ranked by influence by the statistical influence on the question's score. They found that predicting whether a question will be answered is largely dependent on quality attributes such as number of views, favourites, scores, and questioners' reputation.

J. Lin, T. Lin & P. Schaedler [12] performed an empirical study to determine the most important influential factors when determining an answer's success for Stack Overflow questions, and developed a prediction model for an answer's reputation score. Random forest and logistic regression machine learning models were used to predict the best answers for questions on Stack Overflow based on shallow features, including body length, hyperlink count, length of code snippets, etc and semantics based features, such as the similarity of the question to the set of received answers. They found that the most influential factors in predicting high scoring answers were the chronological order of the answers (answers posted sooner received higher scores) and the answerer's account reputation. While their paper is focused on Stack Overflow answers, not questions, the design of the prediction model and approach is applicable to the design of this paper.

M. Choetkiertikul, D. Avery, H. Dam, T. Tran and A. Ghose [13] employed machine learning models to predict whether a given user will answer a new question from the characteristics of the question. They produced a model that achieved a 44% success rate, which is somewhat higher than competing models. The paper as opposed to other relevant papers, uses factors relating to both feature-based prediction (i.e. topic, length, code fragment, readability score, etc.) and social network based prediction (i.e. sharing the same interesting tags, staying in the same time zones).

9. Future Work

An optimal method of analysing the grammatical correctness of each question's body would be to use a statistical natural language parser. This would involve a model that is trained on data specific to Stack Overflow. As described in this paper, some content of some question's had to be omitted from grammar checking such as code blocks or hyperlinks, but on some occasions these have been creatively used in sentence structures. By training a statistical NLP on some form of online grammar paradigms, this problem could be alleviated, and all components of the question, including code blocks could be included in grammar checking. However, this is not a trivial task. Generating trading data for this model would be challenging and time consuming due to the undeveloped set of corpora for this area. Such research would need to be analysed further to determine if it would be feasible.

Another useful extension would be to incorporate into this study, other relevant factors which have been identified in other studies at having some influence in question's success. Such factors that have been identified as important include the reputation of the user asking the question [13] and the frequency of spelling errors [10]. Although based on this study and other studies that include other factors, it does seem that grammar is not as influential as these other factors, the procedures from these other papers are not the same, and so making meaningful comparisons are difficult.

10. Conclusion

In this paper, the use of rule based grammar checking was proposed to assess the effect of grammatical correctness in Stack Overflow questions in regard to the success of these questions in terms of the number of answers received and the presence of an accepted answer. The *languageTool* library was used to generate an intermediate dataset containing grammar attributes of each question's body. Linear regression was used to assess the coefficient of correlation between a number of question grammar correctness variables and a number of question success variables. It was found that the correlation between the grammar and quality variables was low, and these models weekly and negatively correlated between an increase in presence of grammar errors and increase in score. It was also found that the distribution of grammar errors in Stack Overflow questions was right skewed with the most frequent number of grammar errors per question being 0 errors, and the median value for such being 3 errors.

R. Woodcock

GitHub

https://github.com/Russ8/Grammar-Analyser

REFERENCES

[1] Haifa Alharthi, Djedjiga Outioua, and Olga Baysal. 2016. Predicting questions' scores on stack overflow. DOI: http://dx.doi.org/10.1145/2897659.2897661

[2] Stack Overflow. (2018). Developer survey Results. https://insights.stackoverflow.com/survey/2018

[3] A. Fahda and A. Purwarianti, "A statistical and rule-based spelling and grammar checker for Indonesian text," 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, 2017, pp. 1-6. doi: 10.1109/ICODSE.2017.8285846

[4] S. P. Singh, A. Kumar, L. Singh, M. Bhargava, K. Goyal and B. Sharma, "Frequency based spell checking and rule based grammar checking," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 4435-4439. doi: 10.1109/ICEEOT.2016.7755557

[5] Lin, N.Y., Soe, K.M., & Thein, N.L. (2011). Developing a Chunk-based Grammar Checker for Translated English Sentences. PACLIC.

[6] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Stanford University, University of Arizona, IBM Research, Computer and Information Sciences U. of Alabama at Birmingham, Prismatic Inc.

[7] Shaowei Wang, Tse-Hsun Chen, and Ahmed E. Hassan. 2018. Understanding the factors for fast answers in technical Q&A websites: an empirical study of four stack exchange websites. ACM, New York, NY, USA, 884-884. DOI: https://doi.org/10.1145/3180155.3182521

[8] Naber, Daniel. (2003). A Rule-Based Style and Grammar Checker.

[9] Alton Y.K. Chua and Snehasish Banerjee. 2015. Answers or no answers. J. Inf. Sci. 41, 5 (October 2015), 720-731. DOI: https://doi.org/10.1177/0165551515590096

[10] Antoaneta Baltadzhieva, Grzegorz Chrupała, (2015). Predicting the Quality of Questions on Stackoverflow. Tilburg University

[11] Ripon K. Saha, Avigit K. Saha, and Dewayne E. Perry. 2013. Toward understanding the causes of unanswered questions in software information sites: a case study of stack overflow. DOI: https://doi.org/10.1145/2491411.2494585

[12] Jun-Wei Lin, Tzu-Chi Lin, Peter Schaedler. (2016). Predicting the Best Answers for Questions on Stack Overflow.

[13] Choetkiertikul, Morakot & Avery, Daniel & Dam, Hoa & Tran, Truyen & Ghose, Aditya. (2015). Who Will Answer My Question on Stack Overflow?. 10.1109/ASWEC.2015.28.