English Correction Software

ToBIT Paper

Tobias Koller

University of Applied Science Northwestern Switzerland (FHNW),
4000 Basel, Switzerland

Abstract. This paper is based on literature review and capture the current state of Natural Language Processing (NLP) in the application of automatic grammar correction and writing evaluation. Focus is set on the tools that are available on the market and how useful they are in helping students to write better English texts and even improve their skills long-term. Some general NLP techniques will be presented to give insight on how these types of tools work. Eventually, a recommendation is given to the University of Applied Sciences and Arts Northwestern Switzerland (FHNW) on how to proceed if they want to let their students benefit from this technology.

Keywords: Grammarly® · Natural Language Processing · language correction software

1 Introduction

Numerous software solutions on the market promise to help in particular non-native English speakers with grammatical error detection and improving the style and structure of their writing. In the course of the module "Innovative Topics in Business Information Technology" (ToBIT), I am going to evaluate different products regarding their functions, ease of use and effectiveness in supporting

the writing process. The main goal is to use literature review to discover how effective the correction software is under real conditions.

One widely known and used writing tool for grammar checking, spell checking, and plagiarism is Grammarly®. Since it is one of the leading products in this field and many research papers are analysing its use, I focus on this particular tool.

After the evaluation of the tools, I will describe different natural language processing (NLP) techniques that are being applied. In the final part of the document, the findings will be discussed to show the current state of the art in this field. A recommendation to the University of Applied Sciences and Arts Northwestern Switzerland (FHNW) will be given on what products to consider or how to develop a new product internally.

2 Existing solutions

2.1 Grammarly®

Grammarly®, a company with its eponymous popular language correction software, aims to "help people improve their communication" ("Grammarly", n.d.). Max Lytvyn, Alex Shevchenko, and Dmytro Lider founded the company 2009 with a strong focus on supporting the student's writing process. In the meantime, they broadened their scope to businesses as well as private customers. Grammarly® @edu and Grammarly® business both meet the need of each of these customer segments. Their correction service is based on analysing the written content in real-time while showing suggestions in the form of correction cards that can be accepted or rejected. To promote learning, further grammatical information regarding the specific issue can be retrieved from the card which can help to decide on whether to accept the changes or not. ("Grammarly", n.d.)

A successful grammar checker needs to be available where people create texts. These places are as diverse and multi-faceted as the client groups. While academic and business end-users might use predominantly classical text processing software, most private users need the grammar support in their browser (web forums, social media) and increasingly on their mobile devices (chats, e-mails). Grammarly® responds to this with various products. Browser plugins for all major browsers (Chrome, Safari, Firefox, Edge) provide accompanying grammar service during your browsing experience. Input text fields are automatically detected and errors are highlighted with the option to see the correction card. Users of Windows and Mac can download a Grammarly® text editor, but similar functionality can also be used in their web application on app.grammarly.com. To also serve customers who write on mobile devices, the Grammarly® team developed a keyboard for Android and iOS. For both Microsoft Word and Outlook, Grammarly® provides a special plugin that integrates directly and seamlessly with the writing process.

Writers can expect three different types of support from the software. According to their webpage ("Grammarly", n.d.) support ranges from grammar checking, tone detector and for paying users even a plagiarism checker. Grammarly® claims that their grammar checker does not only find misspelt words but also recognises comma and other punctuation errors. Furthermore, their premium service includes more advanced suggestions to enhance the writing style. Assistance in appropriate language style is given by defining so-called goals on which the algorithm bases its recommendations. Parameters that can be predefined include target audience, formality, domain and tone of writing.

Use of the base functionality is free but requires registration. This may lower the entry barrier since all that is required is an e-mail address and a computer or smartphone with an internet connection. However, a premium service is offered against payment. Grammarly®'s website ("Grammarly", n.d.) shows the deviation from the premium plans to the free version. The plans "Premium"

and “Business” both comprise the same advanced features including grammar correction in the following categories.

- Fluency
- Readability
- Engagement
    1. Compelling vocabulary
    2. Lively sentence variety
- Delivery
    1. Confident language
    2. Politeness
    3. Formality level
    4. Inclusive language
- Plagiarism detection

“Business” plan only differs in the account management tools that are available to the organisation’s administrator and in business-oriented billing. The available software and plugins (browser, MS Word, MS Office, mobile keyboard) are equally available in the paid and free subscription model. Since Grammarly® has its focus on supporting students, they target educational institutions with their programme “grammarly@edu”. The grammar checking functionalities seem to be the same as in the “Premium” and “Business” plans but they offer specialised licenses and 24/7 support. Different versions of the same tool make the evaluation of the tool in the following chapter more difficult since some of the research found is based on the limited functionalities. Furthermore, the software developed greatly in the past years as can bee seen on print screens from Dembsey (2017). This needs to be taken into account when judging the results from this analysis.

2.2 Criterion®

The online writing evaluation service named Criterion® is specially tailored to be used in schools and universities and focuses on the tasks of automated essay scoring and improving student's writing skill by use of automated feedback but also facilitating the revisioning process online. The software is developed by Educational Testing Service (ETS) ("ETS Criterion writing evaluation service", n.d.).

This focus on institutions can be recognised when visiting their webpage ("ETS Criterion writing evaluation service", n.d.). To access the tool you need to login through your organisation. Private users are excluded from the potential customers.

According to Lim and Kahng (2012) using Criterion® usually starts with the teacher setting-up an assignment for his class. He as ample parameters to tailor the assignment to his needs, like a category (level of writing), different topic modes and an essay topic. Criterion® comes with a library of predefined essay topics for which there is a catalogue of past scripts written by former students and scored by human raters. This collection is later used to evaluate the student's work by comparing them to the already scored scripts. The teacher can run the assignment as a test and impose a time limit or decide to give unlimited time (for learning and practice). Furthermore, the teacher can set the number of possible re-submissions, the categories that should be displayed to the student and the deadline for submission. While the assignment is ongoing the teacher has its dashboard to supervise the progress of the class or each student. Additionally to the feedback provided by Criterion®, the teacher can add personal feedback manually. Teachers and students can discuss directly via the Criterion® web platform which allows them to directly comment on certain passages of the text. After the submission by a student, he is presented with an overall score and an overview of the errors detected including a detailed description of those.

E-rater

The evaluation engine behind Criterion® is named e-rater®. As detailed by Lim and Kahng (2012) e-rater® bases the scoring on a statistical approach by analysing previously graded essays. An overall of 11 features are being detected and compared to the evaluated dataset. By use of linear regression, the score is then predicted. It is important to notice that e-rater® is trying to mimic human scorer and can, therefore, by definition, never overtake the scoring quality of human beings.

3 Evaluation

3.1 Grammarly®

Grammarly®'s web presence ("Grammarly", n.d.) makes strong claims about the effectiveness and usefulness of the system. Over the past years, researchers attempted to measure the impact of using Grammarly® on student's written performance and determine the overall quality of the feedback produced (Dembsey, 2017; Nova, 2018; Ventayen & Orlanda-Ventayen, 2018). Cavaleri and Dianati (2016) aimed to determine the perceived ease of use and usefulness to answer the question whether the technology will be accepted or not. The methods used include comparing Grammarly®'s machine-generated feedback to the feedback provided by online writing consultants (Dembsey, 2017), comparing student performance in language tests before and after being exposed to the software (Qassemzadeh & Soleimani, 2016) and different kinds of surveys and questionnaires (Cavaleri & Dianati, 2016; Nova, 2018; Ventayen & Orlanda-Ventayen, 2018).

In the following paragraphs we will focus on the quality of feedback, the number of errors found accuracy and see how students perceive the usefulness and ease of use.

Quality of feedback

Feedback provided by a grammar correction software should be understandable by the end-user to promote learning. Simply accepting suggestions of the software blindly does not guarantee avoiding the mistake in the future. Additionally, the correction algorithms are not flawless and the proposed changes should always be questioned. This is only possible if the writers understands the issue that was found in their writing (Dembsey, 2017).

A clear understanding of the problem described and the terminology used in the feedback is required for a student to derive learning gains. In the field of linguistics specialised terminology is being used to describe an error. Especially for the students of English as a foreign language (EFL), those terms might be ambiguous and need further explanation. Dembsey (2017) found that Grammarly® used 52 different terms from the linguistic vocabulary in the correction of three essays while, on the other hand, 10 online writing consultants used only 32 terms (all consultants combined), or 10 terms on average, for the same documents. Furthermore, the consultants used much more accessible language in their comments and even attempted to use the student's language to give more comprehensible feedback. Giving feedback in an appropriate format for the receiver can be achieved by humans better than by algorithms. In general, advanced terminology does not help to learn. Simple language should be used whenever possible (Dembsey, 2017).

In the best case, grammar correction feedback encourages the user to scrutinise the passage containing an issue and gives some valuable recommendations to improve it. However, if the issues are not correctly detected, the feedback can mislead and confuse the writer. In an interview conducted with Indonesian EFL postgraduate students (Nova, 2018), multiple participants reported that Grammarly® changed the sentence's intentional meaning, thus leading to confusion. As long as students are aware of the software's mistake, they can simply ignore it and proceed. More harmful are those incidents when the student is

at a beginner's level in English. Students lacking language experience are more tempted to accept proposed corrections without spotting the erroneous suggestions. As Vojak, Kline, Cope, McCarthey, and Kalantzis (2011) point out, such a situation of uncertainty is counterproductive to the author's development of confidence in his/her writing. The notion of something being wrong discourages creative writing and instead motivates one to comply with standard language patterns and generic phrases. Instead of promoting better writing style, Vojak et al. "fear that the persistent underlying urge towards conformity may stifle individual creativity" (Vojak et al., 2011).

Grammar correction on a sentence level follows rather clear rules whereas connections within a paragraph or even the logical structure of the whole document are much more sophisticated tasks. It is not surprising automated correction software like Grammarly® has difficulties. Students experienced the lack of context-aware checkings such as examining the text for coherence and cohesiveness. Those who needed the software only to check the grammar did not find this an issue (Nova, 2018). The same results were found by Dembsey (2017) who observes that Grammarly® treats each word and sentence individually and does not make any connections between them, therefore drastically reducing the learning opportunity, compared to expert feedback.

Numbers of errors found

The quantitative figure of detected errors is not sufficient to demonstrate the superiority of a correction algorithm. As a result of the comparison between Grammarly®'s and writing consultant's analysis of three student essays, Dembsey (2017) observed a total of 118 issues whereas the cumulative average of the 10 consultants only identified 51. Repetition of the same issues was the main driver for such a high number of detected issues. A human proofreader could encourage the student to look for additional instances of the same mistake by themselves, leaving more room in the coaching session to address other issues.

To get a better view of the issues discovered, all issues were categorised, which led to a total of 16 categories to which every issue could be assigned. In all the essays combined, Grammarly®'s correction cards could be assigned to only six types of issues. This again shows the rather narrow range of recommendations. Altogether, 10 consultants addressed 15 issue categories and even on average they addressed more (eight) diverse topics than Grammarly® did.

Despite having found more issues than a human proofreader, Grammarly®'s issue detection was highly repetitive and only addressed a narrow range of issues. The consultants used fewer comments but gave a more in-depth explanation and could even connect sentence-level issues to general (thesis) level issues. Furthermore, a high number of issues is often not beneficial to learning as students might become intimidated and demotivated (Dembsey, 2017).

Accuracy

A more crucial measure of value provided by the feedback than the number of issues detected is the accuracy of the results. False positives are reported issues that are no problems at all. Dembsey (2017) also considered incorrect use of term or incorrect explanation as an inaccuracy. 41% of Grammarly®'s correction cards were inaccurate, either representing false positives or using wrong terms for the specific issue. At the same time, consultants only had an average inaccuracy of 10% which originated mostly from using wrong terms. (Dembsey, 2017)

The decision whether an issue should be raised or not is also dependent on the type of writing. This puts automated correction software in a disadvantageous position, since detecting the type of writing as well as the target audience is generally difficult. Cavaleri and Dianati (2016) tested Grammarly®'s premium version and could indicate the type of writing. For "essay", "dissertation", "presentation", "blog", "business document" or "creative writing" different rules of

raising issues would be applied. This improved the accuracy of the feedback profoundly.

At the moment of writing, Grammarly® also allows adding some meta information to the document allowing for increased accuracy. Audience (general, knowledgeable, expert), formality (informal, neutral, formal), tone (neutral, confident, joyful, optimistic, friendly, urgent, analytical, respectful) and intent (inform, describe, convince, tell a story) are available in the free version. The latter two are marked as experimental. Only the domain (academic, business, general, technical, causal, creative) is available in the premium version. The fact that they describe these features as experimental, shows that Grammarly® has already detected the necessity to increase accuracy by employing better discourse analysis.

Perceived ease of use

For correction software to be used, it must be simple and intuitive. These are non-functional requirements and therefore more difficult to measure. The literature found focuses on the perceived ease of use reported by students using the tool.

In the Pangasinan State University, Ventayen and Orlanda-Ventayen (2018) conducted a usability study employing a SUS (System Usability Scale) and detected an average usability score of 86.04%. Students found it very easy to use the system and even thought most people would learn to interact with the system very quickly. In a survey (Cavaleri & Dianati, 2016) conducted in an Australian college, 94.4% of the students rated the ease of use of Grammarly® with 4 or 5 with 5 being 'extremely easy'. Only one out of 18 students reported having technical issues using the system. Negative statements about the ease of use were made about the automatic detection of Australian or American grammar and spelling. The tool did not allow for manual language selection and the detection

did not always work. Furthermore, some students found it difficult to navigate the page.

Easiness of access to the tool was also be considered. This includes especially the barriers that need to be overcome before the actual usage of the system can take place, such as registration, download and installation. The only requirement to start using Grammarly® is the registration with an e-mail address and password. Technically, the installation of any software is not required since the system can be used immediately through the browser which gives access to the Grammarly® application ("Grammarly", n.d.). The features are better integrated into the writing process when the plugins are used. The students interviewed in Nova's study found no barriers in the download and setup process.

Perceived Usefulness

According to the Technology Acceptance Model (TAM) (Davis, Bagozzi, & Warshaw, 1989), beside ease of use, perceived usefulness is a key factor that influences the intention to use computer systems. In the survey conducted by Cavaleri and Dianati "most students reported that they found the suggestions helpful for improving the particular paper they had submitted to Grammarly® and half felt it helped them achieve a better mark" (Cavaleri and Dianati, 2016). Effects were not only short term, as students felt the feedback helped them in understanding issues better and in improving their writing skills long-term. Therefore, usefulness is not limited to the current piece of writing but rather supports the whole learning experience of each user and self-directed learning. Besides the direct improvements on grammar, 77.8% of respondents felt an increase in confidence after using Grammarly®.

These results corroborate the conclusions of students interviewed by Nova (2018). They mention the positive impact of feedback cards on their self-revision. Increased reflection on the detected issues helped them to improve the quality

and avoid the repetition of errors. Especially the example sentences helped them to understand the issues better and apply a correction.

However, some students reported disadvantages that reduced the overall usefulness of Grammarly®. Both Ventayen and Orlanda-Ventayen (2018) and Nova (2018) claim that some parts of the document should be excluded from grammar checking such as the bibliography that follows certain standards. Checking on a reference list does not yield any benefit and only distracts (Nova, 2018; Ventayen & Orlanda-Ventayen, 2018). Another limitation found by Cavaleri and Dianati (2016) was the complex language used in some of the recommendations. Deciding on whether to accept the change or not required some deeper understanding of the problem at hand. When students were not able to understand the issue and the underlying grammar rule, they were not able to make those decisions. Therefore, advanced English writers could benefit more than others. The complex language used in the feedback cards can be seen as a barrier for beginner-level students.

3.2 Criterion®

As the underlying engine of Criterion® uses human evaluated texts as a basis, the evaluation of its fit should be determined by the degree it resembles human raters. Weigle (2010) validated that the scoring of a human evaluator and e-rater® do correlate. Lim and Kahng state "that the correlations between e-rater and human ratings were indistinguishable from those between two humans' ratings, suggesting that the two different rating procedures are measuring writing equally well." (Lim and Kahng, 2012). However Lim and Kahng (2012) also detect the limitations of the tool. The statistical method chosen puts a high focus on the quality and the linguistical accuracy but fails to value the strong argumentation and coherence.

3.3 AWE application in classrooms

In this section, we look at automatic writing evaluation from three different perspectives: the teacher's, the student's and the usage in classrooms.

Automatic Writing Evaluation

Automatic Writing Evaluation (AWE) (Grimes & Warschauer, 2010) programmes are specially designed for the application in the classrooms where students write reports and essays. They use "artificial intelligence (AI) to score stu- dent essays and support revision" (Grimes and Warschauer, 2010). The features of such an AWE are usually tailored to the use case of a class, providing the student with the option of submitting the paper for grading. Before final submission, the student usually has the opportunity to go through several revisions and receive automated feedback by the AWE. Eventually, the grading can also be done by the AWE or support the lecturer in this task. Such a system can accomplish more revisions of a student's document than a human can do because of capacity restrictions (Warschauer & Ware, 2006). Grimes and Warschauer (2010) state the limited capacity of teachers is the main bottleneck in the feedbacks they can provide to students and consequently impacts the development of writing skills. AWE is often seen as the silver bullet that solves all these problems. Removing this bottleneck would allow for more revisions, more writing practice, and as a result, improved motivation to write and revise.

How is this technology being applied in these days in classrooms and how effectively is it supporting the learning goals? In a multi-year study, Grimes and Warschauer (2010) observed the attitudes of students and teachers towards this new technology.

Teacher's attitude towards AWE

Incorporating an AWE system comes with a change in the structure of writing classes and the role of the teacher. When working with AWE, the teacher became more of a supervisor who was present to help the students with the usage of the system or to answer questions. Their role shifted from judge to a supportive coach with whom the students wanted to collaborate. This was only possible since the judgement was offloaded to a machine which distanced the teacher from his role as a rater and the students sought advice for improvement from a third party (Grimes & Warschauer, 2010). This made the management of a class much easier. Students tended to be more autonomous and self-motivated when working with AWE and their reluctance to write decreased significantly. Teachers saved a lot of time they would otherwise have spent on low-level issues. This allowed the teacher to put their focus on high-level concerns like the style and overall structure since low-level grammatical errors were taken care of by the system.

The participating schools in the study by Grimes and Warschauer (2010) were using the AWE named "My Access" (MA) which offers an automatic scoring feature. The score will be visible to both the teacher and the student and students can do further revisions by working on the feedback given by MA. The final grade was still determined by the teacher but influenced by the score given by MA. Teachers indicated that 18% of the students' grade was determined by the score MA gave. This number is relatively low since most teachers did not put much confidence in the accuracy and fairness of the automated scoring. On average, they treated the fairness of the system slightly lower than neutral. Knowing the limitations of the automated scoring, it is not surprising that teachers still read the students work very thoroughly. 41% reads as thoroughly as when they would not use MA.

Teachers observed different reactions to the automated scoring feature (Grimes & Warschauer, 2010). While some students were increasingly motivated to write

a high-quality text for the immediate reward, others were highly distracted by the score and could no longer focus on their task. Some teachers even disabled the automated scoring and only showed students the score after submission. Some of the high-performing students that reached a very high score on their first submission were no longer motivated to revise whereas if they had not known the score, they could have still found parts to improve. Another observation was that students tried to learn how the scoring algorithm works and then submit text that would simply lead to a higher score but does not make sense in the context of the paper. From those reports, it can be concluded that teachers are advised to tightly observe the usage of the AWE. Only if they support their students and prevent misuse, the automated scoring can provide real value by allowing the writers to assess and motivate themselves.

Student's attitude towards AWE

Increased motivation towards writing and revising was found by Grimes and Warschauer (2010). Reasons identified were the immediate feedback by the AWE instead of week-long waiting time for human feedback. For some students, the automatic score seemed to have the characteristics of gamification. As a result, they tried to outperform each other which increased motivation even further. They were also able to use the time after the first submission for further improvements since the feedback is available immediately.

Students also did not rate the fairness of the automatic grades as critical as the teachers. They rated the fairness with 3.4 (on a 5-point scale) whereas the teacher's rating was only 2.8.

Regarding the number of revisions done by students, the first year did not show any increase and only 12% of essays had more than one revision (Grimes & Warschauer, 2010). In the following year, this changed to 53%. On the one hand, it can be reasoned that teachers allocated more time for the revision process but

also the students learned how to properly use the system and make the best use of its features. Students who revise their writing first focus on low-level issues like spelling and punctuation before moving to feedback about text organisation and argument development. It seems to be a natural behaviour to focus on the low-hanging fruits that can be fixed with lower effort. Improving on the text structure takes much more time and often requires reading large parts again to come up with a strategy to re-arrange it.

AWE usage

After looking at both the teacher's and the student's side it can be concluded that AWE usage can remarkably improve the learning process. More time is available to focus on higher-level concerns like text organisation and argument development since issues in spelling, punctuation, grammar and word choice are taken care of by the software. Overall student motivation significantly increased. According to Grimes and Warschauer (2010) this needs to be taken with care. The higher motivation observed was mainly based on the goal to reach a high score, not mainly to write better texts and learn from it. This shift from internal to external motivators is not beneficial to the students. Furthermore, students must be closely observed when using AWE and teachers need to take appropriate action when they see problems. Not all students interact equally with this new support. Some might be distracted by the scoring while others lose motivation after receiving a good initial score. Grimes and Warschauer conclude that there is a need for "sensible teachers who integrate AWE into a broader writing program emphasizing authentic communication, and who can help students recognize and compensate for the limitations of software that appears more intelligent at first than on deeper inspection." (Grimes and Warschauer, 2010).

4 Techniques of natural language processing

Grammatical error detection (GED) is the task of identifying errors within texts. Algorithms applied for this task belong to the category of supervised learning and require a labelled dataset. Whereas grammatical error correction (GEC) is concerned about fixing the issue detected Bell, Yannakoudakis, and Rei, 2019. The next paragraphs give an overview of some techniques used in NLP and describe their advantages and limitations.

4.1 Part of speech tagging

Different techniques used for GED and GEC make use of part-of-speech (POS) tagging. During this process, every word of a sentence is assigned a tag of its grammatical category ("POS tags and part-of-speech tagging — Sketch Engine", 2018). The different tags used are defined in the tagset. They can vary by the degree of detail. Basic tags only distinguish between noun, verb, adjective, etc. while others distinguish male/female, plural/singular, tense and person. This tagging alone does not yet give any insight into the correctness of a sentence, it simply gives insight about the structure of a sentence and the role of the words in it. This then allows automated text processing software to do further analysis.

4.2 Syntactic parser

The rules that describe the structure that a sentence has to conform with, are referred to as the syntax of a language. To detect ill-formed sentences the syntactic parsing can be applied. As a first step POS tagging needs to be applied to know about the structure of the sentence. In the next step, the sentence is parsed using predefined constraints which resemble the rules of the specific language. According to Manchanda, Athavale, and kumar Sharma (2016), the sentence is matched with the given tree structures. If the sentence can be matched with

the constraints available the parsing succeeds and the sentence is syntactically correct. Failure of the parsing process simply tells that there is some syntactical error present, but does not reveal any further information.

Example 1: Constraint: subject-verb agreement for number and person. Sentence: "He (3rd person sg.) is tall". Result: Constraint matches.

Example 2: Constraint: subject-verb agreement for number and person. Sentence: "He (3rd person sg.) am tall". Result: Constraint does **not** match.

Constraint relaxation

In its basic form syntactic parsing only fulfils the role of grammar error detection but fails to provide insightful advice to the author. To do so, a diagnosis technique can be applied. The most widely used, constraint relaxation, is described here in more detail. In syntactic parsing, the matching of partial structures is only allowed if the constraints are met (Vandeventer, 2001). In constraint relaxation some of the constraints present are relaxed, meaning that a partial structure is allowed to match even if this constraint is violated. To give useful information to the user the relaxed constraint needs to be labelled while defining them.

Example: Constraint: subject-verb agreement for number and person. Sentence: "He (3rd person sg.) am tall". Result: Constraint matches because it is relaxed. **Label:** Verb does not correspond to the subject.

The main advantage of constraint relaxation in syntactic parsing is that it provides the full analysis of the structure and at the same time diagnoses the violations to the underlying syntactic constraints (Vandeventer, 2001). The problem with this approach is memory-intensive computing. This is because every correct sentence does not only match with the correct structure but also

with all the relaxed versions of that structure. Additionally, the constraints of parsing are manually assembled which represents a labour intense task.

4.3 Statistical technique

The statistical approach also makes use of the POS representation of a sentence. Other than the syntactical approach it does not depend on predefined rules, instead, it learns from POS-annotated corpus (Manchanda et al., 2016). The decision of whether a sentence is well-formed or not depends on the statistical measures of the frequency of those POS sequences in the corpus. It is therefore crucial that corpus of language is well chosen for the task. This means that different types of writing may need different text corpora. Different language differ widely in the sequences of POS considered as correct, which requires also separate annotated corpora.

The great advantage in this procedure is the irrelevancy of any hand-crafted rules. A large enough POS annotated corpus is sufficient to achieve the desired results. However, the occurrences considered as incorrect based on non-occurrence in the corpus cannot be explained. The author is left uninformed about the reason for the detected error. This disadvantage can be mitigated by building the model not only with the POS tagged features but add a lexical feature-set (Gamon et al., 2009).

4.4 Rule based technique

The rule-based technique requires the composing of a collection of error rules (Manchanda et al., 2016). These rules describe errors by use of POS tags that are expected to occur in the text. For every possible error, an individual rule must be manually constructed. It is self-evident that this technique will never lead to a complete set of rules since the types of erroneous sentences are innumerable.

As the rules are specifically constructed for a type of error it is simple to add a detailed description with a suggestion on how to fix it supplemented with the affected grammar rules.

One variation of this rule-based technique does not require the definition of each error but allows the matching of whole error types by usage of regular expressions in n-gram templates (Method for rule-based correction of spelling and grammar errors, 2003). An n-gram is a sequence of n words in a sentence that is matched against the set of rules. The regular expressions are used so that not all the characters of a word must match the rule. This invention allows the replacement of the illegal n-gram with a legal one as the following examples illustrate:

Example 1:
Rule: fuly$ → fully
illegal n-gram: hopefu**l**ey
replacement: hopefu**ll**y
($ represents the end of the word)

Example 2:
Rule: their seem → there seem
illegal n-gram: the**ir** seems to be...
replacement: the**re** seems to be...

This kind of matching allows to cover numerous common errors with just one rule. The replacement of the error is then done in a context-specific fashion since only errors in the specified n-gram are being replaced. This approach does not find all errors but if an issue was detected has a high certainty for a replacement and does not introduce new errors. Therefore, this technique is especially useful for automated replacements.

5 Self experiment

In a self-conducted experiment, I want to find out first-hand how helpful the automatic grammar correction is. Grammarly® is designated for this experiment since it is easily accessible in its free version and they use advanced machine learning and deep learning techniques to improve their products ("Grammarly", n.d.). The experiment should answer the following questions:

- How is the user experience when revising issues with Grammarly® web?
- Is the feedback provided by Grammarly® understandable?
- Does Grammarly® help build new knowledge that later can be applied?

5.1 Method

To get a more objective result, a survey paper of a fellow student is used, instead of any product that was written by the author of this document. The paper chosen is in an early draft stage with almost no revision made. This should yield a broad detection of different issues. The Grammarly® web-based editor is used to conduct the test. The goals of the writing can be manually configured, the following settings have been made:

- **Audience:** Expert
- **Formality:** Formal
- **Domain:** Normal (cannot be changed in the free version)
- **Tone:** Analytical
- **Intent:** Inform, Describe

Only a subsection (2 chapters) of a total length of 2667 words have been submitted for evaluation.

5.2 Result

Within less than a minute after pasting the content to the browser the calculation has already been finished. The sidebar on the right gives an overview of all the issues found. It prominently mentions that 68 alerts have been detected, but there are 90 more issues present which are only accessible in the premium version of Grammarly®. Those 68 alerts are subdivided into the two categories **Correctness** and **Clarity**. A progress bar also indicates the achievement of the two other categories **Engagement** and **Delivery**. If one wants to see how to improve and selects the category, they also mention buying premium service first. Additionally, an overall score of 48 (out of 100) is displayed.

The revision screen of Grammarly® is divided into three columns. The first column (from left to right) shows the text that is being edited at the moment. Issues are highlighted by underlining the sections in a specific colour. Each of the before mentioned issue categories is assigned to one colour. The second column shows a feedback card for each issue. They appear on the same height as the is located in the first column. While scrolling through the page they are synchronised. The last column is the sidebar that was already described before.

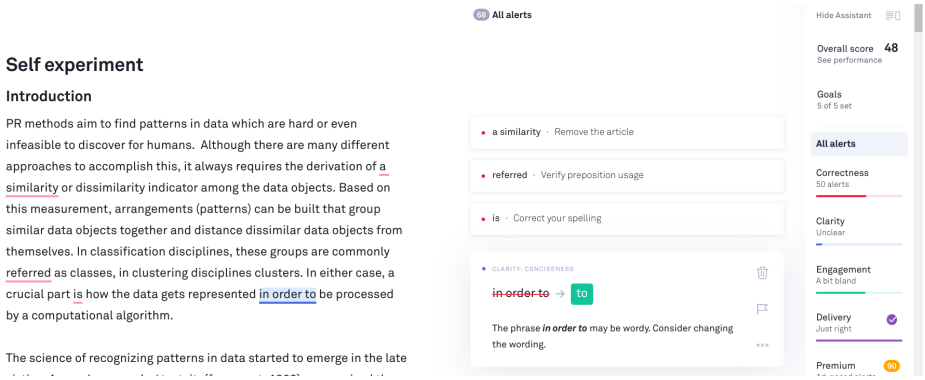

Fig. 1. Revision view.

Each feedback card can be in three different states. In the overview, all the cards are in the collapsed state to not use up much space. Only the text containing an issue and a quick recommendation are displayed.

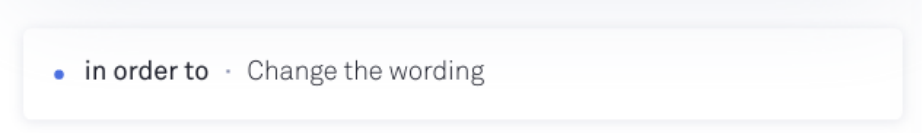

Fig. 2. Collapsed state.

When clicking into the underlined text section or by clicking the feedback card directly, the card expands and gives a full-sentence explanation of the issue. With a simple click on the green correction, the proposed solution can be accepted or declined by clicking the bin button.

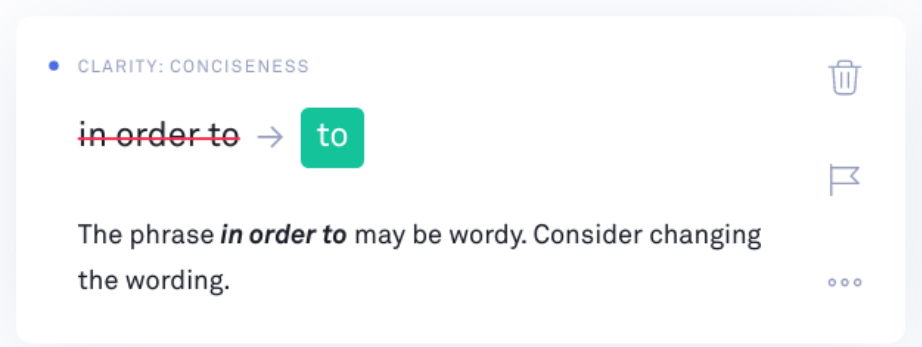

Fig. 3. Extended state.

If more information is still needed to make the final decision, a detailed explanation of the rule, including examples, can be displayed by clicking the three dots.

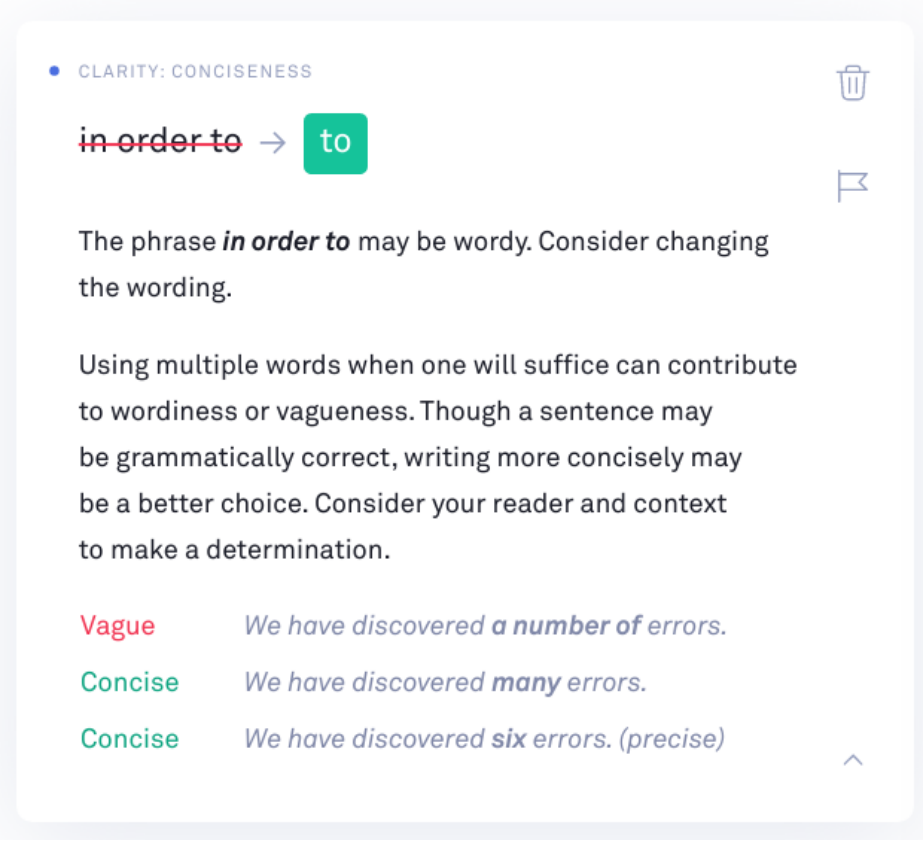

Fig. 4. Detailed rule.

This three stages of information help to keep the overall view organised and still show maximum information when needed. The synchronous display of the cards with the floating text make it easy to associate them together.

Revising suggestions

While going through the various feedback cards almost all of the 68 issues could be resolved as proposed by Grammarly®. The decision if the suggested change should be applied could be done fairly quick since the context of the sentence

was always visible and could be reread quickly. The issues were so obvious that the extended feedback card was always sufficient and no further grammatical explanation was required to detect the problem. There was only one card that reported an error where there was none.

The following issue is a false negative. The rule shown does not apply to the situation. The word “similarity” belongs to the term “similarity indicator” which is not an uncountable noun.

“(. . .) it always requires the derivation of a similarity or dissimilarity indicator (. . .)”

Feedback: The indefinite article, **a**, may be redundant when used with the uncountable noun **similarity** in your sentence. Consider removing it.

After all, issues have been handled, by either accepting or ignoring the suggestion the view now shows how many additional writing issues are present but can only be reviewed with a premium account.

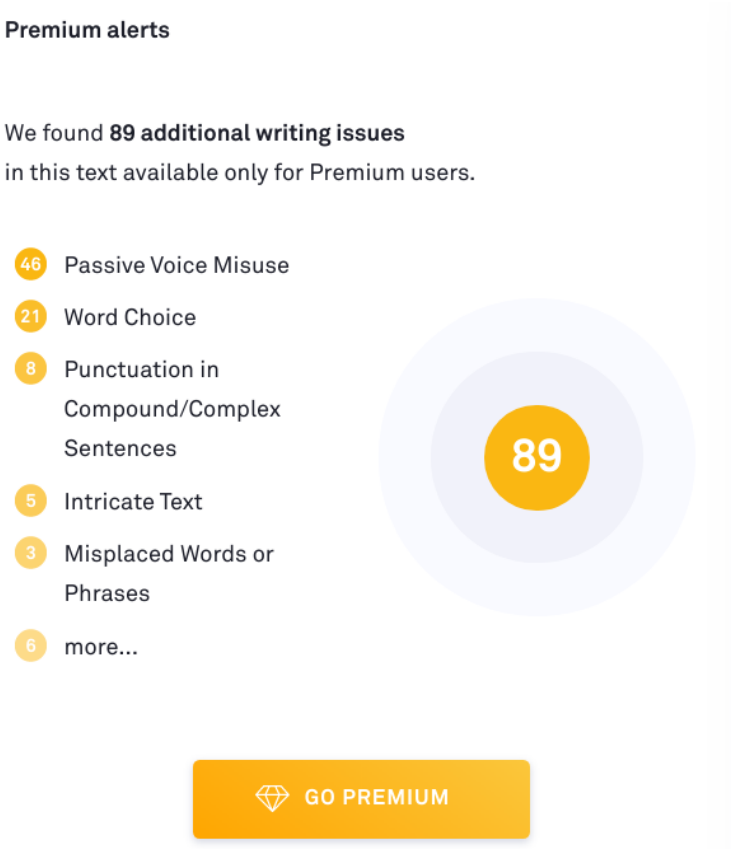

Fig. 5. Further issues.

5.3 Discussion

Experience

The whole revision process was smooth and the user interface was intuitive and could be managed without further documentation or explanation. Since most of the corrections could be applied as recommended this was done with the simplicity of a click. No further typing with the keyboard had to be done.

The synchronous scrolling of the actual text with the cards helped to keep the overview.

The progress bars in the sidebar always showed the current state of the text. While more issues were resolved those bars started to present better results which motivated, even more, to fix as many issues as possible. Visual components are cleverly utilised to boost the user's motivation.

The user gets constantly reminded of plenty issues still present but not marked because of the lack of a premium account. While also this is a very smartly placed advertisement for the paid service, it can daunt the author when he cannot make use of the full scope of improvements available.

Feedback understandable

All the issues presented in the free version of Grammarly® were easily understandable. I most cases the extended feedback card was sufficient to fully recognise the problem at hand. If I doubted the appropriateness of the suggestion the detailed rule helped to clarify. Especially the correct and incorrect examples made it clear how to apply a specific rule. The terminology used was understandable and did not lead to any confusion.

Building new knowledge

Especially with repeating issues that needed to be solved in the same manner over and over again, a learning effect could be detected. The iterative revisioning of the same error made this rule stick to the mind. In the further writing process, the author remembers this and can think of alternative ways to construct the phrase.

However, most of the errors seemed to be spelling or typing errors and inadvertent mistakes. Surely, Grammarly® helps to detect them and suggests cor-

rections but it does not help to avoid them in the future. The author is usually aware of those

The most potential for building new knowledge would have been in the areas of sentence structure, development and argumentation. However, those enhancements are not available in the free version and therefore the learning effect of the author is very limited.

6 Discussion and recommendation

6.1 Overview of studies

It can be concluded that the application of automated evaluation and correction software has its place in the process of writing English texts. On a word and sentence level, the feedbacks help to correct not only spelling and grammar issues but also advises in rewording, applying stylistic rules and writing in an appropriate tone. The immediate feedback acts as a motivator for the writer (Grimes & Warschauer, 2010), encouraging repeated revision cycles. Additional information retrieved about the issue helps the language learner to understand the grammar rules better and thus allows the improvement of writing skills for the future.

In an educational setting, such systems can take over the discovery of surface-level (affecting individual sentences) issues and free up time for teachers to focus on content development and stylistic concerns (Dembsey, 2017). An automated score was found to be useful for students to self-assess their product in an early phase of writing (Grimes & Warschauer, 2010).

Nevertheless, human proofreading and assessment is indispensable. Machine-generated scores tend to focus on strict rules and lack the recognition of a well-developed structure. Close supervision of the students' usage is recommended to

be able to mitigate negative effects such as frustration, cheating the system or stifled creativity through conformity with rules.

Users of automated grammar correction software and administrators at schools need to be aware of the limitations of the technology. A machine cannot address the needs of an individual student as this is the case with human coaching sessions. Issues detected are usually on a low-level and rather technical. Advising on structure and development requires a broader understanding of the topic and the ability to logically connect paragraphs over the course of a document. Grimes and Warschauer (2010) claim this field of natural language understanding (NLU) is not yet sufficiently developed and only has been successfully applied in narrow domains of medicine and engineering.

6.2 Benefits to FHNW students

Before giving any concrete consultancy on which system the University of Applied Sciences and Arts Northwestern Switzerland (FHNW) should acquire or produce it is crucial to define the intended use cases and requirements. This chapter gives an overview of possible cases from a student's perspective.

At FHNW various degree programmes are being taught in English ("FHNW", n.d.) but since English is not a national language in Switzerland, most students learned English as a second language. While listening and talking was trained in high school, students have little experience in the writing of academic papers or essays. Providing students with a service such as Grammarly® which they can integrate in the process of creating all kinds of writing and presentations could help them find mistakes and learn from them, thereof improving the quality of their writing. Alternatively, writing courses focusing on academic writing skills could make use of specialised AWE software. Lecturers would be supported in their task of giving feedback, correcting and grading the submitted papers.

The first approach targets at the students' autonomy and willingness to improve their writing skills, whereas the second use case utilizes AWE to make writing courses more effective.

6.3 Recommendation to FHNW

After declaring the use cases that should be covered by the system, FHNW should check if one of the existing systems on the market (candidate systems) fulfils some or all of the criteria. For the first use case Grammarly®, for the second use case, Criterion® and My Access should be considered.

Once a product is found that can cover the use cases, a pilot project should be started. Licenses for a limited set of students (or classes) should be bought for one year allowing the extensive testing of the software product under real-world conditions. During and after the trial period surveys and questionnaires should be conducted to test the benefits provided.

Under certain circumstances, the development of its product can be recommended to FHNW. FHNW can avoid being dependent on a third-party service provider and would also evade the yearly licenses. Furthermore, possession of the source code and control of the NLP algorithms used are a benefit. Nevertheless, the long development time and effort must be taken into account. This might only be justified if students' and researchers' resources can be used to accomplish the project. While in the beginning, this undertaking would consume many resources and requires investing money, FHNW could consider monetise the product in the future.

7 Bibliography

References

Bell, S., Yannakoudakis, H., & Rei, M. (2019). Context is Key: Grammatical Error Detection with Contextual Word Representations. *arXiv:1906.06593 [cs]*. arXiv: 1906.06593. Retrieved October 14, 2019, from http://arxiv.org/abs/1906.06593

Cavaleri, M. R., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. *Journal of Academic Language and Learning*, *10*(1), A223–A236.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, *35*(8), 982–1003. doi:10.1287/mnsc.35.8.982

Dembsey, J. M. (2017). Closing the Grammarly® Gaps: A Study of Claims and Feedback from an Online Grammar Program. *The Writing Center Journal*, *36*(1), 63–100. Retrieved October 14, 2019, from https://www.jstor.org/stable/44252638

ETS Criterion writing evaluation service. (n.d.). Retrieved December 10, 2019, from https://criterion.ets.org/criterion/default.aspx

FHNW. (n.d.). Retrieved December 12, 2019, from https://www.fhnw.ch/en

Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J., Belenko, D., & Klementiev, A. (2009). Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal*, *26*(3), 491–511. Retrieved December 9, 2019, from www.jstor.org/stable/calicojournal.26.3.491

Grammarly. (n.d.). Retrieved November 27, 2019, from https://www.grammarly.com/

Grimes, D., & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *The Journal of Technology, Learning and Assessment*, *8*(6). Retrieved October 14, 2019, from https://ejournals.bc.edu/index.php/jtla/article/view/1625

Kantrowitz, M., & Baluja, S. (2003). *Method for rule-based correction of spelling and grammar errors*. US6618697B1. Retrieved December 9, 2019, from https://patents.google.com/patent/US6618697B1/en

Lim, H., & Kahng, J. (2012). Review of Criterion for English Language Learning. *Language Learning & Technology*, *16*(2), 38–45.

Manchanda, B., Athavale, V. A., & kumar Sharma, S. (2016). Various techniques used for grammar checking. *International Journal of Computer Applications & Information Technology*, *9*(1), 177.

Nova, M. (2018). UTILIZING GRAMMARLY IN EVALUATING ACADEMIC WRITING: A NARRATIVE RESEARCH ON EFL STUDENTS' EXPERIENCE. *Premise: Journal of English Education*, *7*(1), 80–97. doi:10.24127/pj.v7i1.1332

POS tags and part-of-speech tagging — Sketch Engine. (2018). Retrieved December 9, 2019, from https://www.sketchengine.eu/pos-tags/

Qassemzadeh, A., & Soleimani, H. (2016). The Impact of Feedback Provision by Grammarly Software and Teachers on Learning Passive Structures by Iranian EFL Learners. *Theory and Practice in Language Studies*, *6*(9), 1884–1894. doi:10.17507/tpls.0609.23

Vandeventer, A. (2001). Creating a grammar checker for CALL by constraint relaxation: A feasibility study. *ReCALL*, *13*(1), 110–120. doi:10.1017/S095834400100101X

Ventayen, R. J. M., & Orlanda-Ventayen, C. C. (2018). *Graduate Students' Perspective on the Usability of Grammarly® in One ASEAN State University* (SSRN Scholarly Paper No. ID 3310702). Social Science Research Network. Rochester, NY. Retrieved October 23, 2019, from https://papers.ssrn.com/abstract=3310702

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New Spaces and Old Places: An Analysis of Writing Assessment Software. *Computers and Composition*, *28*(2), 97–111. doi:10.1016/j.compcom.2011.04.004

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157–180. doi:10.1191/1362168806lr190oa

Weigle, S. C. (2010). Validation of Automated Scores of TOEFL iBT Tasks against Non-Test Indicators of Writing Ability. *Language Testing*, *27*(3), 335–353. doi:10.1177/0265532210364406