English Correction Software

ToBIT Paper

Tobias Koller

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

Abstract. This paper forms ....

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 there are innumerable research papers to support, I focus on this particular tool.

After the evaluation of the tools, I will also 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 clearly the current state of art in this field. A recommendation to 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 improve the communication of people. 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 personal writers. With Grammarly @edu and Grammarly business, they have their dedicated divisions for those clients. Their correction service is based on analysing the written content in real-time while showing suggestions in the form of correction cards. With a simple click, the changes can be applied or rejected. To promote learning of the writer 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. ("Write your best with Grammarly." n.d.)

Being a successful grammar checker requires to be available where people create texts. With such a diverse client group, those places are multi-faceted. While academic and business end-users might use classical text processing software, most personal use cases are not so easy to unify. Web forums, chat messages, e-mails or social-media platforms are some of the areas users compose content. 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.

Which type of help can writers expect from the software? According to their webpage ("Write your best with Grammarly." n.d.) support ranges from grammar checking, tone detector and for paying users even a plagiarism checker. They claim that their grammar checker does not only find misspelt words but also recognises comma and other punctuation misuses. Furthermore, their premium service includes more advanced suggestions to enhance the writing style. To help writing an appropriate language one can define so-called goals on which the algorithm bases its recommendations on. Parameters that can be predefined include target-audience, formality, domain and tone of writing.

Users can immediately start using Grammerly with just a simple registration. The usage of the base functionality is available for free for an unlimited time

span. This lowers the barrier of entry immensly, since all a user need to have is an e-mail address and a computer or smartphone with an internet connection. However, a premium service is offered against payment. Grammerly's website ("Write your best with 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 suggestions in the following categories.

- Fluency
- Readability
- Engagement
  - Compelling vocabulary
  - Lively sentence variety
- Delivery
  - Confident language
  - Politeness
  - Formality level
  - Inclusive language
- Plagiarism detection

"Business" plan only differs in the account managament tools that is available to the organisation's administrator as well as business-oriented billing. The available software and plugins (browser, MS Word, MS Office, mobile keyboard) are equally available for paid and free users. Since Grammerly has its focus on supporting students they target educational institutions with their programm "grammerly@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 printscreens from Dembsey (2017). This need to be taken into account when judging the results from this research.

some notes to the author

3 Evaluation

3.1 Grammarly

The Grammerly's web presentation ("Write your best with Grammarly." n.d.) makes strong claims about the effectiveness and usefulness of the system. Over the past years several researcher attempted to measure the impact of using

Grammarly on the student's written performances and determine the overall quality of the feedback produced. Others (Cavaleri & Dianati, 2016) aimed to determine the perceived ease of use and perceived usefulness to answer the question if this technology will be accepted or not. The methods used include comparing Grammerly's feedback to the one provided by online writing consultants (Dembsey, 2017), comparing student's performance in language tests before and after being exposed to the software (Qassemzadeh & Soleimani, 2016) and different kinds of surveys and questionnaires (Nova, 2018) (Cavaleri & Dianati, 2016) (Ventayen & Orlanda-Ventayen, 2018).

Quality of feedback Feedback provided by a grammar correction software should be understandable by the end user in order to promote learning. Simply accepting suggestions blindly without questioning support in avoiding the same error in the future. Additionally, the correction algorithms are not flawless and their proposed changes should always be questioned. This is only possible if the writer understands the issue that was found in his writing (Dembsey, 2017).

The understanding of the problem and thereof the possible learning gain is dependent on whether the student understands the terminology used in the feedback. In grammar situations and constellations often can be described very precisely by a specialised term. Especially for students in English as a foreign language (EFL), those terms might be ambiguous and need further explanation. Dembsey (2017) found that Grammarly used 52 different terms in the correction of three essays while on the other hand 10 online writing consultants used only 32 terms (all conultants combined), or 10 terms on average, for the same documents. Furthermore the consultants used much more accesible language in their comments' explanation 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 way better than by algorithms. In general avanced terminology is not supportive for the learning process of the student. Simple language should be used whenever possible (Dembsey, 2017).

In the best case the the grammer correction software's feedback encourages the user to scrutinise the passage of attention and give some valuable recommendation to improve it. However, when not detecting the issues correctly, misleading feedback can lead to the author's confusion. In an interview conducted with Indonesian EFL postgraduate students (Nova, 2018) multiple participants reported that Grammarly changed the sentence's intentional meaning and therefore led to their 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-level in English. He might be more tempted to accept any changes proposed without the lack of experience spotting those erroneous suggestions. As Vojak, Kline, Cope, McCarthey, and Kalantzis (2011) **(!!! isn't that a bit too long of an author's list??!!!)** points out such a situation of uncertainty is counterproductive to the author's development of confidence in

his writing. The notion of something being wrong with his writing motivates to comply with general phrases and standard structure in the future. Instead of promoting better writing style, they "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. Unsurprising that also automated correction software like Grammarly have their difficulties. Students experienced the lack of context aware checking like coherency and cohesiveness within a text. 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 not making any connections between them, therefore drastically reducing the learning opportunity compared to expert feedback.

Amount of errors found On first sight high quantitative figure of detected errors seems to demonstrate the superiority of a correction algorithm. We will see why this appearance is deceptive.

During the comparison of Grammarly with writing consultants in analysing three student essays Dembsey (2017) observed a total of 118 issues whereas the cummulative average of the 10 consultants only brought up 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 time for different issues. In order 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 show the rather narrow range of recommendations. Cummulated all 10 consultants addressed 15 issue categories and even on average they addressed more (8) diverse topics than Grammarly.

Despite having found more issues than human proofreader, Grammarly's issue detection was highly repetitive and only addressed a narrow range of issues. The consultants used less comments but gave 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 for the learning rate of students, as they might become intimidated and demotivated. (Dembsey, 2017)

Accuracy A more crucial measure of value provided by feedback than 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 inaccuracy. 41% of Grammarly's correction

cards where inaccurate, either being 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 if an issue should be raised or not is also dependent on the type of writing. This puts an automated correction software in a disadvantagous position, since detecting type of writing as well as target audience is generally difficult. Cavaleri and Dianati (2016) tested Grammerly'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 setting some meta information to the document allowing for increased accuracy. Audience (general, knowledgeable, expert), fromality (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 only available in the premium version. Seing those features, specially the one being experimental, shows that Grammarly has already detected the necessity to increase accuracy by means of better contextual issue detection.

Perceived Ease of use In order for a correction software to be used in the student's writing process it is crucial that the usage is 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 by means of 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 that 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 to have technical issues using the system. Negative statements about the ease of use were made about the automatic detection of Australian or American grammar spelling. The tool did not allow the manual selection of language and the detection did not always work. Furthermore, some students found it difficult to navigate the page.

We also consider how easy it si to access the tool. This includes specially the barriers that need to be overcome before the actual usage of the system can take place like 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 serves the main correction features ("Write your best with Grammarly." n.d.). The features are better integrated in the writing process when the provided plugins are used. The interviewed students in Nova's study found no barriers in the download and setup process.

Perceived Usefulness According to the Technology Acceptance Model (TAM), besides ease of use, perceived usefulness is a key factor that influences the people's intention to use computer systems (Davis, Bagozzi, & Warshaw, 1989). 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, students felt the card's feedback helped them in understanding issues better and improve their writing skills also long-term. Therefore, usefulness is not limited to the current piece of writing but rather on the whole learning experience of each individual user and supports self-directed learning. Besides the direct improvements on the correctness of the grammer, 77.8% of respondents felt an increase in their confidence level after using Grammerly.

These results cover with the conclusion by students interviewed by (Nova, 2018). They mention the positive impact of feedback cards on their self-revision. Inreased refelction on the issues found helped them to improve the quality and avoid the repetition of errors. Specially the indication of example sentences helped them to understand the issues better and apply a correction.

However, some students detected also disadvantages that reduced the overall usefulnes of Grammarly. Both Ventayen and Orlanda-Ventayen and Nova claimed that some parts of the document should be excluded from grammar checking like 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 Application in classrooms

Automatic Writing Evaluation Automatic Writing Evaluation (AWE) programmes are specially designed for the application in the classrooms where students write reports and essays. They use "artificial intelligence (AI) to score student 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. It is obvious that such a system can serve much 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 a teacher in English language arts is the main bottleneck on the feedback he can provide to his students and consequently their 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, writing practice as a result more motivation by students to write and revise.

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

Teacher's attitude towards AWE Incorporating a 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 that was around 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 of the writing was offloaded to a machine which distanced the teacher from his role as a rater and the students seeked 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 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 by both the teacher and the student and students have the ability to 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 the grade given was influenced by MA by an average of 18%. 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 of the limitations of the automated scoring it is not surprising that teachers still read the students work very thoroughly. 41% reads them even 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 teacher even disabled the automated scoring and only showed their students after submission. Some of the high-performing students that reached a very high score on their first submission were no longer motivated on revisioning whereas if they would not have known they could have still found parts to improve. Another development observed was students that 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 we can draw the conclusion that teachers are advised to tightly observe the usage of the AWE by teir students. 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 a human feedback. For them the automatic score took the characteristics of gamification and they tried to 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 amount of revisions done by students, the first year did not show any increase and only 12% of essays had more then 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 who learned how to properly use the system and make best use of its features. Students who revision their writings first focus on the low-level issues like spelling and punctuation before moving to feedback about organisation and development. This seems to be a natural behaviour to focus

on the low-hanging fruits than can be fixed with lower efforts. Improving on the structure takes much more time and often requires reading large parts again in order to come up with a strategy to re-arrange the structure.

AWE usage After looking at both the teacher's and the student's side we can conclude that AWE usage can remarkably inprove learning process. More time is available to focus on higher-level concerns like organisation and development since issues in spelling, punctuation, grammar and word choice were taken care of by the software. Overall student motivation significantly increased. According to Grimes and Warschauer (2010) this need 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 actions 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 concludes 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).

- (Warschauer & Ware, 2006) page 15
- better over the revisions, lower errror rate, more discourse elements
- 16/17 limit revision process to spelling

Potentional on data ?

- (Warschauer & Ware, 2006)
- page 16: valuable data: nr revisions, improvements

teacher have time to focus on other ghings

Li, Link, and Hegelheimer, 2015

- instructors view and usage of AWE
- students view highly depend on the instructors approach

4 Techniques of natural language processing

5 Self experiment

6 Discussion and recommendation

text quote “word for word ” (Dembsey, 2017)

7 Bibliography used

The list of references is not final and will be extended during the process of writing the ToBIT paper.

References

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

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

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. doi:10.1016/j.jslw.2014.10.004

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

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

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

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