Raphael Zender et al. (Hrsg.): Die 18. Fachtagung Bildungstechnologien (DELFI),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 265

Automatic Feedback for Open Writing Tasks: Is this text appropriate for this lecture?

Sylvio Rüdian[1], Joachim Quandt[2], Kathrin Hahn[2] and Niels Pinkwart[3]

Abstract: Giving feedback for open writing tasks in online language learning courses is time-consuming and expensive, as it requires manpower. Existing tools can support tutors in various ways, e.g. by finding mistakes. However, whether a submission is appropriate to what was taught in the course section still has to be rated by experienced tutors. In this paper, we explore what kind of submission meta-data from texts of an online course can be extracted and used to predict tutor ratings. Our approach is generalizable, scalable and works with every online language course where the language is supported by the tools that we use. We applied a threshold-based approach and trained a neural network to compare the results. Both methods achieve an accuracy of 70% in 10-fold cross-validation. This approach also identifies "fake" submissions from automatic translators to enable more fine-granular feedback. It does not replace tutors, but instead provides them with a rating based on objective metrics and other submissions. This helps to standardize ratings on a scale, which could otherwise vary due to subjective evaluations.

Keywords: Feedback; online courses; language learning.

1 Introduction

In second language learning giving feedback to students is of high importance to enable them to reach the desired learning outcome. While closed tasks, e.g. filling blanks, can be easily checked for correctness due to existing sample solutions, open writing tasks often need to be assessed by tutors to provide appropriate feedback for students.

Research shows that language levels can be derived from written texts; the results show the language level (A1, A2, …) of the submitted student's text [Kl84]. When observing online language learning courses of individual language levels, all students generally learn the same words and grammatical structures as defined in the "Common European Framework of Reference for Languages" [Co01]. However, course book editors have a degree of freedom to decide which topic and which grammar will be taught within each course at any individual level. This results in a diversity of online courses with different structures. However, simply obtaining the information that a student's written text meets the requirements of a particular level is not enough to give automatic feedback, apart from in the

[1] Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Weizenbaum Institute for the Networked Society, ruediasy@informatik.hu-berlin.de, https://orcid.org/0000-0003-3943-4802
[2] Goethe-Institut e.V., Oskar-von-Miller-Ring 18, 80333 Munich
[3] Humboldt Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, pinkwart@hu-berlin.de

extreme scenario that a student's text does not fit into the current language level at all or when students are not aware, especially at the beginning of the course, that they are required to write texts that reach the language level of their current course.

Many automatic essay scoring systems only take samples of texts with existing ratings into account, but miss the relation to an online course where users have been taught other words, expressions, and grammar; i.e., learners may have extended their knowledge that should be applied in any particular practical open writing task. Thus, we aim to bridge the gap and focus on features that are related to taught contents in combination with already known submissions to get a more in-depth insight to predict teachers' ratings of student texts. In this paper, we follow the research question: which generalizable method could be used to predict tutor ratings of open writing tasks in online courses by combining the meta data collected from such tasks with the actual language content of any particular course section.

2 Related Work

Alikaniotis et al. [AYR16] introduced an approach of automatic text scoring in language learning using a long short-term memory (LSTM) neural network. Rating typically depends on a set of textual features such as grammar, vocabulary, style, relevance, and complexity of sentences. According to the authors, regression and ranking are still the state of the art approaches to obtain automatic ratings that are indistinguishable from human ratings. Using their architecture of an LSTM, they achieve comparable accuracies with fine-tuned models. With a feature importance analysis, they state that they were able to find features that have an impact on predictions, but there is no causality to suggest that these features cause different ratings.

Klebanov et al. [KF13] introduced new features using the co-existence of words in a text for automatic essay scoring; these features show a relationship between the quality of writing and word association profiles. Automatic essay scoring was also investigated by McNamara [Mc15] using a hierarchical classification and predict scores based on competency levels. They state that a hierarchical approach has the advantage of having the ability to inform by providing general formative feedback. Truscott [Tr96] shows that it is not necessary to give feedback for every single mistake in language learning as this kind of correction is unnecessary, ineffective and expensive. Teachers have to decide how fine-granular a correction should be; a detailed correction, for example, is always more time-consuming and time itself is a limiting factor. Having a method for giving a less-detailed, more general automatic feedback, could help to reduce tutor workload and also makes it possible to give immediate general feedback, without the need to wait for the teacher's rating. The usefulness of getting detailed or less-detailed feedback is a much debated topic amongst researchers, with little agreement being reached. Beuningen et al. [BJK08] notice that direct corrective feedback has a long-term effect, but short-term effects were observed

for both direct and indirect corrective feedback. There is no need to discuss whether feedback is necessary, but providing good feedback in online courses on a large scale requires either a considerable number of teachers or good technology-driven approaches that enable a feedback almost comparable to that of a tutor.

In open writing tasks, computer-generated submissions are a problem where learners send a text that they claim to have written themselves. Lavoie et al. [LK10] show that texts can be classified relying on keyword-based features. This is an important step to identify "fake" submissions, instead of rewarding a user for a good text. The authors recognize that there is still considerable room for improvement in detecting fakes.

In this paper, we propose a methodology to give language learning students feedback for the appropriateness of written texts for every section of an online course to learn German as a foreign language. This approach is generalizable and can be applied for every language learning course for languages that are supported by our used libraries. Additionally, it can detect fake submissions.

3 Methodology

Our prediction combines different layers of language learning. We explore the use of grammar and wording separately, and also combine them to find the most practicable accuracy.

3.1 Grammar & Mistakes

The first focus is on the appropriateness of the applied grammatical structures that students were taught during the lecture. A "trivial", but very comprehensive and expensive method would be to define, which grammatical structures the learner should be able to use, for each section of a language learning course. As this method requires language experts, we do not want to limit ourselves to their availability. Instead, we propose a greedy algorithm by observing the receptive parts of the online course; this includes texts, where learners have to solve tasks. We follow the assumption that learners know how to use at least some of the grammatical structures that have been taught in the particular section involved.

To determine whether grammatical structures were used in open writing tasks, we first extract all receptive texts of a particular section. We transfer all texts to their Part-of-Speech (POS) Tags representation [He99]. Each word will be transferred to the corresponding part of speech (e.g. NN for a noun). This allows us to have access to a generalized version, thus obtaining a list of word- and punctuation classes. Subsequently, we define n-grams and set n to three, named trigrams. These trigrams consist of three connected sequential POS tags. Consecutive parts of speech tags represent possible grammatical structures that the user was confronted with. The same is done with previous submissions of users that have previously been marked by tutors as appropriate for the section.

The next step is to find overlapping POS trigrams of the section and the user's submission. This set contains grammatical structures that are part of the online course section itself and those that the learner used within the submitted text. For every existent acceptable submission of our observed section, we are able to obtain a set of trigrams. If we count the number of overlapping trigrams as values, we can put them into a sorted list, that represents the range of acceptance according to the grammatical structure. By finding the overlap with the POS trigrams of the section and a new user's submission, it is possible to see whether the resulting length of the intersection will be in our learned range.

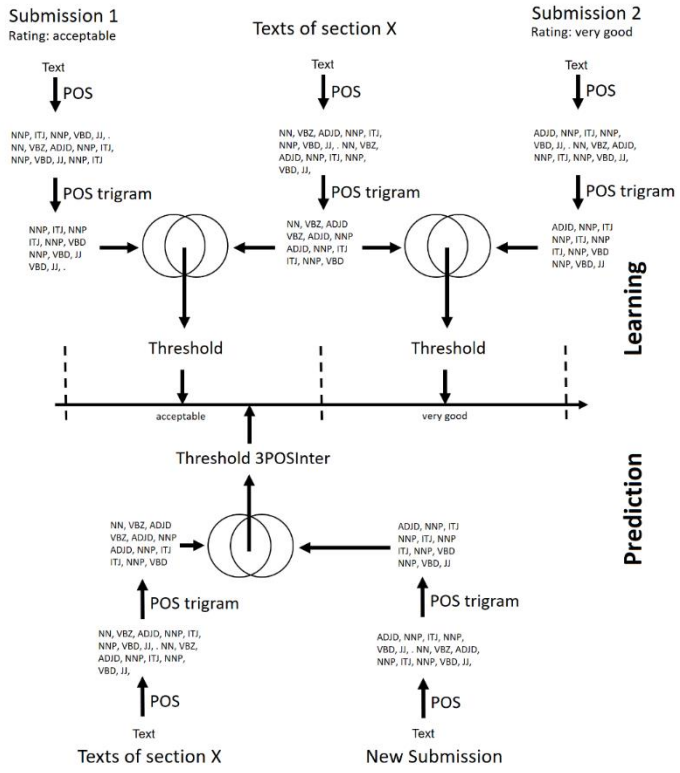

Fig. 1: Methodology for using thresholds of POS trigrams intersections.

This learned range can help to understand how good a submission is. We tag this value with "3POSInter". Fig. 1 visualizes the learning and prediction processes in general. The didactic approach follows the standard criteria for correction of written texts for German as a foreign language[4]:

[4] https://www.goethe.de/lrn/pro/gc1/C1_PrueferTraining_08.pdf

1. All parts of the task have been answered.

2. Content is presented in a coherent and well-structured way.

3. Vocabulary used is matching the current level of the course and according to the level of the Current European Framework of Reference for Languages [Co17].

4. The text is correct regarding morphology, syntax, orthography and punctuation.

While criteria 1 and 2 are implying content analysis and semantic understanding of the text, criteria 3 and 4 are orientated on formal aspects of language. As criteria 3 and 4 are easier to analyze with the technology in use, focus will be placed on these two criteria. We divided the acceptance range into three classes and added smileys for each class for visualization of general feedback: 😐=acceptable, 🙂=good, 😀=very good. If the POS trigram overlap is smaller than the known range, the tool gives ☹=bad as the feedback. The "very good" class is the range starting at the third of the range vector and is open-ended.

Practical usage has shown that learners who write more, are advantaged by this approach. Using grammar that the user should not be aware of at this stage could represent a fake submission, possibly achieved by using a translation tool, for example. To filter out these submissions, we used a superset of the sections' and submissions' trigrams; this superset contains all trigrams that occur in the submission of the user that are not part of the online course. We computed the length of this superset for all known submissions of this section to see whether the range is appropriate or not. This is labeled "3POSSuper". We use the end of this range as a fixed border. If the supersets length of a new submission is greater than this border, we have a solution that contains too many mistakes. This means either the learner uses grammatical structures that were not taught during the lecture or uses inadequate grammar that results in trigrams that do not exist.

To distinguish between "fake" submissions and an unreasonably high number of mistakes, we added the last step. We use all texts at every level of the online course to obtain the POS trigrams. We follow the assumption, that, if the trigram generated is not part of the online course (ranging from A1 to C1), it is seen as non-existent and should not be used. Computing the superset of the trigrams of the overall online course and the learner's submission gives the amount of possible non-existent grammar structures, which are then counted as mistakes. Fake submissions have very few of these mistakes; there are usually fewer trigrams that are non-existent in the overall online course compared to other submissions. Thus, the submission can be flagged as out of scope and should be revised by a tutor. Having numerous mistakes in a submission can be flagged as ☹=bad and should be revised by the learner.

The following paragraphs formalize our approach. Let $\{T_{i,lec}\}$ be the set of POS trigrams of the section lec which contains i texts T_i. Let $\{T_{j,lec}\}$ be the set of POS trigrams of the user's submission j for the specific open writing task in the section lec of the online course. The next step creates vector $\vec{V}$, ordered by value, that includes all lengths of the

POS trigrams overlaps between the online course section and each existing user submission:

$$\forall\, T_i \in lec\colon v_i = \left|\{T_{i,lec}\} \cap \{T_{j,lec}\}\right|$$

$$\vec{V} = sort[v_o, \ldots, v_i]$$

Vector $\vec{V}$ now contains acceptable lengths of intersections, ranging from low performing submissions to very good texts. To have a better orientation, we divide this vector into three parts of equal length: acceptable, medium, very good. Of these parts, we use the middle value and define our lower and upper thresholds: F = acceptable, L = good/very good:

$$F = \vec{V}\left[int\left(\frac{len(\vec{V})}{6}\right)\right], L = \vec{V}\left[int\left(\frac{5}{6}len(\vec{V})\right)\right]$$

By calculating the length of the POS trigrams overlap with a new submission and the section, we get a metric to compare the users' submissions among each other; this metric can also be used to compare future submissions.

3.2 Wording

Next, we used stem words that occur for the first time in the section of the online course and count the length of the intersection with the stem words of the submissions. This is labeled "WORDS". We limit used words to stem words as declined words could be misspelled by users or not recognized by the system, and grammar usage is already covered by the previous metric. In a practical setting, focusing on stem words will result in more hits, whereas using declined words will often result in zero hits, which would defeat the purpose.

To focus on newly learned words of a particular course section only, we collected all stem words of all sections. If a word was already part of a previous section, it was removed from the current one. This resulted in a list, for every section of the course, only containing words that the user was confronted with for the first time in the particular section. In contrast to our approach for grammar, where we used receptive texts only, we use all texts of all sections, independently of them being part of a task or a reading/listening text. It could be observed that people mainly focus on vocabulary learning and submissions with a good rating use learned words of the section, which the learner had not necessarily encountered previously.

Subsequently, we apply different methods to predict the tutor's rating of users' submissions. We searched for correlations and an optimal regression, we use the thresholds as defined above and observe a neural network to compare the accuracy.

4 Results & Evaluation

An experienced tutor rated 400 users' submissions of a specific section of the online course "Deutsch Online"[5]. Texts could be labeled with the following ratings: 1 = very good, 2 = good/acceptable, 3 = not acceptable, 4 = too bad/out of scope/fake.

Figure 2 shows all ratings concerning the features we observed. Each dot represents a single user text according to the observed feature. The dataset is as expected. The more POS trigrams that exist in the submission that also occur in the course, the better the rating (Fig. 2 a). Not acceptable submissions (rating 3) use fewer POS trigrams that exist in the course, but submissions that are out of scope contain more trigrams that are not part of the course trigrams (Fig. 2 b). Submissions get bad ratings the more mistakes that are found (Fig. 2 c). Texts that are out of scope (e.g. because of being a fake submission) have fewer mistakes on average, as expected.

Fig. 2: a) POS trigrams intersection of users' submissions and course content. b) The number of POS trigrams that occur in the users' submissions, but which are not part of the course content. c) The number of found mistakes in the users' submissions. d) The intersection of stem words in the users' submissions, and course content.

The more stem words, which are included in the online course and used by the learner, the better the rating (Fig. 2 d). This general observation shows that there are some remarkable patterns concerning our metrics and ratings. To understand, whether a linear relationship between our meta-data and the rating exists, we used the Pearson Coefficient:

[5] https://www.goethe.de

$$r([3POSInter;\ WORDS;\ 3POSSuper;\ MISTAKES]) = [-0.26;\ 0.45;\ -0.03;\ 0.05]$$

This rough analysis shows that our observed superset has almost no linear correlation. Mistakes in submissions have a low impact on the rating. It makes sense that the superset alone is no linear indicator for a rating and it should only be used to determine whether a submission is out of scope. A detailed rating solely using this metric is impossible. The feature "$WORDS$" has a moderate correlation only and the correlation of $3POSInter$ is weak. Mistakes also have a very small linear influence on the rating. To see whether other non-linear dependencies are existing, we used a regression approach to predict all ratings. The resulting RMSE shows that this methodology cannot be used for good predictions (standard deviation = 1.003):

$$RMSE([3POSInter;\ 3POSSuper;\ MISTAKES;\ WORDS]) = [1.04; 1.00; 1.01; 1.09]$$

The absolute error is in the range of 0.87 and 0.92, which shows that on average the predicted rating lies in the right direction. Furthermore, combinations of all our features for a multi regression were tested but resulted in $RMSE = 1.14$, which is an even higher error rate than for single features.

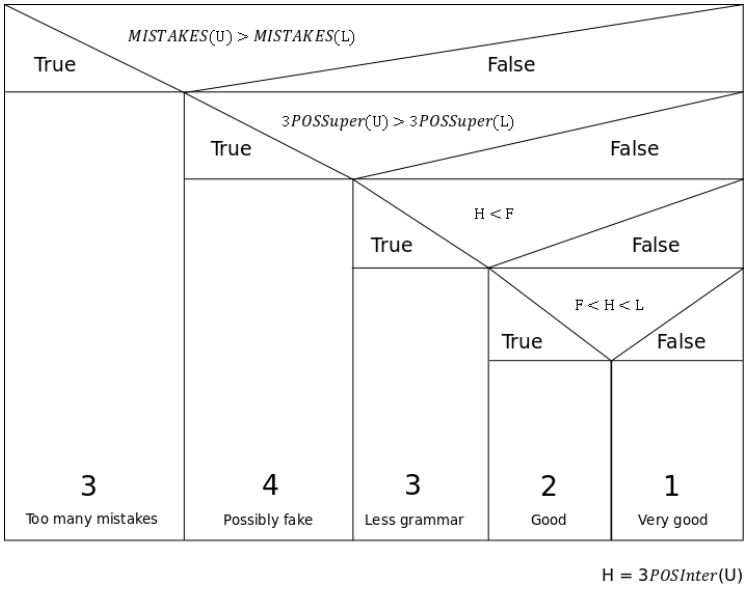

Fig. 3: Hierarchical decision tree for an algorithmic rating of users' submissions using different thresholds.

We can find thresholds according to our metrics and the existing user submissions. According to our findings, all thresholds combined result in a decision tree to rate grammar. It can be found in Fig. 3, with U as the user's submission, L as the previously learned

parameters and H as the intersection of the user submissions' POS tags and the POS tags of the lecture.

Testing this approach in cross-validation shows that 70% of all ratings can be predicted correctly, by using the hierarchical decision tree. In this paper, we do not differentiate between the granularity of good ratings, as the main concern is to find a way to detect acceptable and not acceptable submissions. More granular observations are part of further studies. Using the mean value of existing stem words in the existing user submissions related to each rating also results in fixed thresholds. It cannot be varied much as is visible in Fig. 2 d): user submissions contain up to 6 words that they have previously learned. Thus, more granularity is not possible. Merging rating 1 and 2 to be an acceptable solution and 3 and 4 to be an unacceptable solution results in a more robust rating. In our case:

$$R_{1,2} = \emptyset(Rating1 \cup Rating2) = 4.2;\ R_{3,4} = \emptyset(Rating3 \cup Rating4) = 3.2$$

Using $R_{1,2}$ as a threshold, 63.2% of the ratings could be predicted. Using $R_{3,4}$ as threshold results in a prediction accuracy of 63.9%. Each value lower or higher of $R_{3,4}$ decreases the accuracy. Thus it is recommended to use $R_{3,4}$ as a threshold.

Having fixed thresholds by looking at our features separately, shows that we reached a limit in prediction concerning our observed features using threshold-based methodologies. To overcome this problem, we use a neural network that gathers all features together to find any possible existing pattern amongst them. As this is a black-box approach, we can only see how good the prediction works, without receiving any insights into the internal patterns. Our neural network has a two-layer architecture with "hard sigmoid" as activation functions and "Nadam" as the optimizer. All available features are used for training ($3POSInter, 3POSSuper, WORDS, MISTAKES$). Our data was balanced to avoid overfitting problems. We used Gridsearch [Sc17] to find the optimal hyper-parameters combination. With 1500 epochs and a batch size of 50 we achieved an accuracy of 69.92% in 10-fold cross-validation. Limiting our training data to the feature "$WORDS$" only, the neural network achieves an accuracy of 63.85% in 10-fold cross-validation. Using $3POSInter, 3POSSuper, MISTAKES$ as done within the previous approach, we achieve 63.46%. If we use these features separately for training, the achieved accuracy is lower ($3POSInter$: 61.54%, $3POSSuper$: 55.38%, $MISTAKES$: 46.15%). By combining these features we achieve the best accuracy; this is comparable to the result when using our trained thresholds. To conclude our results, we see that the best achievable accuracy is 70% with both of our used methodologies. The accuracy of using words only is nearly the same for both methods (63.85% and 63.9%); the difference is just a matter of fine-tuning.

5 Discussion

Neumeyer et al. [Ne00] measured the human-to-human correlation in rating spoken sentences by teachers. Raters had an inter-rater reliability among each other between 0.61 and

0.72 at the sentence level. Compared to our result, machine-to-human reliability of the same ratings achieved 0.7, which is acceptable for practical applications.

Both of our methodologies show very similar results when predicting tutor ratings for open writing tasks. The threshold approach has the advantage that explainable ratings can be created, which is impossible when using the neural network as a black-box. As the results are very similar, it is recommended to use the threshold approach, allowing for explainable ratings, for further applications. The application of our meta-data extraction requires a separation of the course's texts into receptive parts and new content. This is an important step to achieve meta-data for more granularity but it requires manpower.

Alternatively, to find possible mistakes due to the non-existence of POS trigrams in the overall course, a third-party tool can be used to find mistakes, the quantities found can be used as a feature. This may be necessary if the whole course does not teach and use all existing grammatical constructs of a language. Furthermore, such a tool could also find spelling mistakes that could help in the prediction of ratings.

Even though teacher training and examiner certification have the objective to give guidelines on the correction of written texts, the current definitions, to a certain extent, leave room for interpretation for the correcting teacher. There is no hard limit that can be defined between the different grades, as it is up to the teacher to decide whether a text is in general understandable or whether only parts of it are. Our approach of automatic feedback on written tasks can support teachers if they are undecided as to which of two grades is the best one.

Improvements can be understood in two directions. We expect that teachers will be able to provide more standardized feedback based on the automatic evaluation of learners' texts and they will also be able to focus more on the content itself and less on the correction of formal problems and errors to support the language learning process. In addition, the time spent on the correction of learner texts will be reduced. Having tutoring as the main factor on the cost side of online courses, this could make online courses more financially accessible to learners. Currently, these courses can be found in the premium sector of the language learning market. It is necessary to provide additional information as part of the given feedback to make the understanding of the prediction easier. As in many parts of this proposal, it is important to state that technical analyses and didactic approaches need to be linked to each other.

6 Future Work

We have seen that we are unable to achieve more than 70% accuracy with our used meta-data. Further research could focus on collecting more meta-data to be used as features, but the result may still vary depending on the tutor. As the overall aim is to improve the accuracy of predicted ratings, we would like to explore more features that can be derived from

user submissions. One idea is to get insights into semantics. Our proposed solution only takes grammar and used words into account. Words themselves determine the context, but whether a written text is appropriate with regard to the context has not yet been examined. Taking an algorithmic approach to understand in which relation words are used could be helpful to distinguish between good and unacceptable submissions. Another extension is the combination of words in general, not just searching for new words that only co-occur in the online course. Looking at bi- or trigrams of words gives a more in-depth overview of the context.

Future scenarios beyond the practical use for the correction of learner texts in language courses could involve placement tests. They detect the current language level of the learner and are normally used at the beginning of a language course or to certify a level for job applications. Available automatic tests for German as a foreign language are limited to the receptive skills of reading and writing. The productive skills of writing and speaking are either assessed by a tutor, incurring additional costs, or are simply left out. This can lead to an inaccurate evaluation of the language level which could result in the need for a class change at the beginning of a new course. More precise placement including an automatic evaluation of the learner's level would improve the learning experience, reduce frustration due to being placed at the wrong level and would give more precise information on language qualification for personalized training programs.

The application could cluster typical errors and problem areas, conduct the learner to adaptive training of these errors and generally guide the learning process according to the student's need. Often automatic translations are used and in many cases the learner is unable to understand or evaluate whether the translations are meaningful in the larger context. Moreover, translations may only be helpful to finish off a task but they do not normally support the real learning process. At a more advanced stage, after having proven high accuracy in placement tests, automatic correction could also support official language certification. Exemplarily, the writing section is corrected by two grading persons and in the case of a large discrepancy, a third corrector will review and take the final decision. A trained automatic correction tool could be used to replace the first correction stage and could subsequently be compared to a manual correction, allowing students to receive their final exam marks within a shorter time period. Besides some investigation is required to understand how tutors use automatic rating, whether they guide or manipulate them.

7 Conclusion

In this paper, we explored the prediction of ratings for open writing tasks to support tutors of online language learning courses. The novelty of our approach is the combination of online course contents and existing submissions to derive new features. Our different experiments have shown that we can predict tutor ratings with an accuracy of 70%. The results both using fixed thresholds and a neural network are similar. Thus, for this area of

learning, the more explainable method, i.e. fixed thresholds, should be used. As the predicted result is explainable, generated feedback can be more granular to give recommendations, e.g. to use more newly learned words or to create more complex sentences. Our approach can help tutors in online courses, as the prediction gives tutors advice to enable more consistent ratings among different tutors. Our practical results are promising and observing more features could help to optimize the accuracy in further investigations.

Acknowledgments: This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII116 (Weizenbaum-Institute) and the Goethe Institute e.V. The responsibility for the content of this publication remains with the authors.

Bibliography

[AYR16] Alikaniotis , D.; Yannakoudakis, H.; Rei, M.: Automatic Text Scoring Using Neural Networks, in ACL, 2016.

[BJK08] Beuningen, C. v.; Jong, N. d.; Kuiken, F.: The Effect of Direct and Indirect Corrective Feedback on L2 Learners' Written Accuracy, in ITL - International Journal of Applied Linguistics, Volume 156, Issue 1, 2008.

[Co01] Council of Europe, A Common European Framework of Reference for Languages: Learning, Teaching, Assessment, in Council for Cultural Co-operation, Strasbourg, Cambridge University Press, 2001.

[Co17] Council of Europe, The CEFR Levels, 2017, Available: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions [Accessed 31 03 2020]

[He99] Heeman, P. A.: POS Tags and Decision Trees for Language Modeling, 129-137, 1999.

[KF13] Klebanov, B. B.; Flor, M.: Word Association Profiles and their Use for Automated Scoring of Essays, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, p. 1148–1158, 2013.

[Kl84] Klein-Braley, C.: Advance Prediction of Difficulty with C-Tests. 1984

[LK10] Lavoie, A.; Krishnamoorthy, M.: Algorithmic Detection of Computer Generated Text, 2010.

[Mc15] McNamara, D. et al.: A hierarchical classification approach to automated essay scoring, in Assessing Writing Volume 23, Elsevier Ltd., 2015, pp. 35-59.

[Ne00] Neumeyer, L. et al.: Automatic scoring of pronunciation quality, in Speech Communication, Volume 30, Issues 2–3, Elsevier Science B.V., 2000, pp. 83-93.

[Sc17] Scikit-learn developers, Tuning the hyper-parameters of an estimator, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/grid_search.html. [Accessed 30 07 2018].

[Tr96] Truscott, J.: The Case Against Grammar Correction in L2 Writing Classes, in Language Learning: A Journal of Research in Language Studies, Volume46, Issue2, 1996, pp. 327-369.