Portfolio 8: Summarizing an ACL Paper
Shreya Valaboju, Soham Mukherjee

In modern and traditional academic settings, students have always benefited from receiving helpful feedback on their essays or assignments. Although automatic grading systems already exist, substantial improvements in the quality of those developments are yet to be made. The academic paper, “Your Answer is Incorrect…Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset” focuses on the necessity to improve current answer feedback systems by introducing Short Answer Feedback (SAF) dataset. The SAF dataset can be used to train supervised models that grade answers and provide enhanced feedback to learners. The authors of this paper are Anna Filigera, Tim Steuer, Tobias Meuser, Siddharth Parihar, and Sebastian Ochs, who are all affiliated with Technical University of Darmstadt, Germany - Multimedia Communications Lab. This report will discuss the paper’s research and work, as well as examine the importance of the contributions made in advancing the field of natural language processing (NLP).

The authors’ problem highlights the limitations of current answer feedback systems. One of the main limitations is the lack of elaborate and comprehensive feedback explaining a given score. Specifically, the authors argue that merely providing a score is insufficient for a learner to understand and assess their performance on an assignment or an exam. Learners could benefit from more explanation as to why their chosen answers were incorrect and better utilize their results. The authors explain that advancements in this area are crucial but underwhelming due to the lack of public, content-centered datasets focused on elaborate feedback. Additionally, having teachers manually give elaborate feedback could also be time consuming and cost inefficient. Therefore, the authors hope that their created SAF datasets promote the ability to train understandable feedback models. The recent push to build more explainable and understandable NLP models serves as motivation for the authors in this paper to create SAF datasets.

With that being said, there is a significant amount of prior work that has been done when it comes to short answer feedback, but doesn’t address the limitations that this paper does. The use of error analysis and feedback in language learning has been used extensively in the field of second language acquisition, especially when it comes to identifying similarities in syntax, semantics, etc in between languages. The use of evaluation metrics is often used to quantify this feedback. Examples include precision and recall, but when it comes to making specific quantitative metrics for evaluating the quality of feedback, few are as extensive as using double-language datasets. Previous studies include studies on the effectiveness of feedback in natural language processing tasks, such as named entity recognition and semantic role labeling. Furthermore, there are several natural language explanation datasets, but these datasets only slightly relate to what SAF tries to address. For instance, the authors cite “WorldTree V2,” a dataset that is used to explain the correct answers to multiple choice science questions. However, an issue with this dataset is that the explanations given were simply scientific facts and resources needed to answer the questions correctly, contrary to SAF’s focus on providing feedback. Similarly, the authors reference another dataset that contains solutions to math problems from standardized tests such as the GMAT. Again, this dataset points to reference solutions rather than addressing the specific mistakes or assumptions in an incorrect solution.

While feedback is generated for single-language datasets, there's a lack of attention to the quality of the feedback provided. As discussed previously, there aren't enough metrics that are able to be easily quantified for bilingual datasets such as this one with English and German. As a result, since most data tends to be single language, the creation of the bilingual short answer feedback dataset seeks to address those limitations by accounting for factors in between both languages (second language acquisition). This is a unique contribution that hasn't been done prior when it comes to multi-language datasets.

The Short Answer Feedback dataset (SAF) was created to address the lack of content-focused elaborated feedback datasets in both English and German. Bilingual datasets were used to get a more comprehensive dataset; this set includes short answers, scores, and text explanations of given scores. The purpose is to provide researchers and educators with a resource for developing and evaluating feedback systems that can help learners improve their performance. With its focus on content and bilingual approach, the SAF dataset is a valuable addition to the field of natural language processing.

The authors evaluated their work by assessing the effectiveness of a bilingual SAF dataset in improving the performance of natural language processing models and language learning tools. The performance of two baseline models, a rule based system and a neural network based system was compared using the dataset. The results showed both models achieving higher F1-scores when using the feedback dataset compared to when not using it, demonstrating the dataset's effectiveness in identifying errors and providing feedback. Overall, the authors' evaluation provides evidence that their bilingual short answer feedback dataset is effective in identifying errors and providing feedback, and can be used to improve the performance of natural language processing models and language learning tools.

The work in this paper is foundational to advancing automatic grading systems. The authors provide possible enhancements on top of SAF introduced in this paper. Such future work includes expanding the size of SAF to match other large scale NLP datasets, improving the diversity and complexity of questions by incorporating more domains and languages in addition to German and English, as well as adding neuro-symbolic approaches. It is clear that the contributions made in this paper can serve as a door to more advancements and research on explainable and understandable feedback systems. This is evident given the fact that the paper has already received 7 citations, with author Tobias Meuer holding the most citations. The paper is featured on Long Paper Proceedings of the 2022 ACL (Association of Computational Linguistics) conference.

Link to Paper: ***Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset***
Our Github Links:

https://github.com/Zakenmaru/CS4395_Portfolio
https://github.com/shreyavala/nlp_portfolio