Sequence Tagging in EFL Email Texts as Feedback for Language Learners

Yuning Ding[1], Ruth Trüb[2], Stefan Keller[4], Johanna Fleckenstein[3,5] and Andrea Horbach[1,5]
[1]CATALPA, FernUniversität in Hagen, Germany,
[2]Pädagogische Hochschule der Fachhochschule Nordwestschweiz FHNW, Switzerland,
[3]Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Germany,
[4]Pädagogische Hochschule Zürich, Switzerland, [5]Universität Hildesheim, Germany

Abstract

When predicting scores for different aspects of a learner text, automated scoring algorithms usually cannot provide information about which part of text a score is referring to. We therefore propose a method to automatically segment learner texts as a way towards providing visual feedback. We train a neural sequence tagging model and use it to segment EFL email texts into functional segments. Our algorithm reaches a token-based accuracy of 90% when trained per prompt and between 83 and 87% in a cross-prompt scenario.

1 Introduction

Writing formal emails in English is part of many English as a Foreign Language (EFL) curricula due to its high practical relevance in academic and professional life. However, manual scoring of such writing tasks and the provision of feedback to students are time-consuming tasks for teachers, especially when feedback does not solely consist of a single holistic score per text, but instead consists of more fine-grained feedback such as highlighting certain elements in a learner text and providing feedback for each element.

In this paper, we investigate the task of segmenting EFL learner emails into functional elements relating to their main communicative function (Hyland, 2019). Examples would be the salutation, closing or matter of concern (see Figure 1 for an annotated sample email). We perform the automated segmentation task on the basis of the eRubrix corpus (Keller et al., 2023) consisting of 1,102 semi-formal emails written by Swiss EFL learners at lower secondary level (8th and 9th year of schooling). In these emails, seven different core elements of an email were annotated by trained human raters. We use a neural sequence tagging architecture to automatize the segmentation task and compare it against a simple sentence-based baseline.

Overall, the paper makes the following contributions:

- We present segment annotations on the eRubrix dataset. On the basis of aspects of text quality developed by Keller et al. (2023), we show how the human annotations presented in their study can be transferred to automated span annotations.
- We apply a sequence-tagging architecture that is able to assign the right segment category for 90% of all tokens.
- We show that the automatic segmentation can be applied to new writing prompts almost without performance loss.
- We provide learning curve experiments showing that as little as 50 to 100 emails are enough to train a model that is close to the final performance on the whole dataset.
- We analyze the impact of positional information in the training data, showing that positional information is - unsurprisingly - important in this automatic segmentation task, especially on certain labels like subject line, salutation and closing.
- We discuss how the algorithm can be used as a basis for feedback to language learners and for developing language learning activities in EFL classrooms.

2 Related Work

The interdisciplinary research presented in this paper combines second language writing studies with educational science and natural language processing. In the following section, we therefore discuss related work from these three disciplines.

Yuning Ding, Ruth Trüb, Johanna Fleckenstein, Stefan Keller and Andrea Horbach. Sequence Tagging in EFL Email Texts as Feedback for Language Learners. *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*. Linköping Electronic Conference Proceedings 197: 53–62.

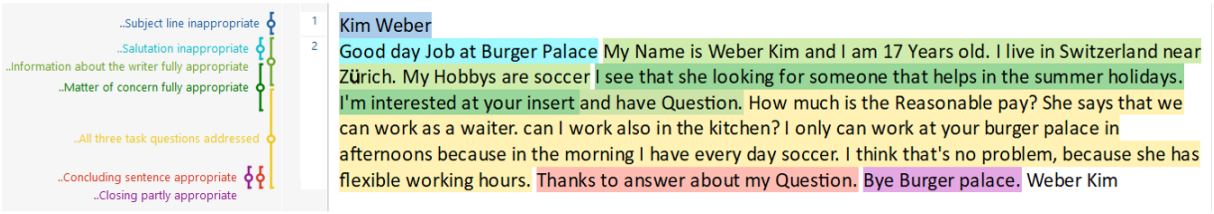

Figure 1: Sample annotation in MAXQDA (Version 22.0.1) for a learner email from the *Burger Palace* task.

2.1 Second Language Writing Studies

A number of theories have been proposed to support students' acquisition of second language writing competences (Matsuda, 2003). Among the most widely used and researched approaches are the genre-based approach and the approach based on text functions (Hyland, 2019, pp. 6-20).

A genre-based approach assumes that all writing is done in a specific social context and that a range of social constraints and choices exist that operate on writers (Hyland, 2019, p. 18). Teaching in this paradigm typically begins with the purposes of communicating before moving on to learning the "stages" of a text which can express these purposes. This often involves the analysis of model texts and typical language structures contained in them.

The approach focusing on text functions is similar in that it relates language structures to meanings. This is achieved by showing students how to compose effective paragraphs for the text functions they want to express, e.g. describing, narrating, or reporting (Hyland, 2019, p. 6). Both the genre-based and the text function-based approach would concur in the view that providing feedback on these core elements of an email can help students to understand the communicative function of an email and to apply them independently in their own writing.

The automated annotation function described in this article can be seen as a technique for enhancing genre-based writing instruction with automated span annotations: it identifies the salient structural elements required in an email to fulfil the communicative function of the text (polite greeting, expression of the writer's purpose, expected response, adequate closing, etc.), highlighting them for learners and laying the basis for feedback relating to specific text functions.

2.2 Multimedia Learning and Feedback Processing

The cognitive theory of multimedia learning (CTML) proposes that people learn more effectively from multimedia sources than from text alone (Mayer, 2001). This assumption is based on the idea that people have limited cognitive processing capacity, and that using a combination of verbal and visual information can help reduce the cognitive load on each channel (Mayer and Moreno, 2003). Research has shown that adhering to certain design principles reduces cognitive load and positively affects learning in multimedia environments (Noetel et al., 2021). The design principles derived from CTML should also pertain to automated writing feedback, but they have seldom been transferred to this context (for an exception see Burkhart et al., 2021). The visualization of different segments of a learner text - as we propose in our study - makes use of the advantages of multimedia learning and should thus support the revision process. The multimedia design principles that are particularly relevant in the context of this study are contiguity, signaling, and segmenting.

Contiguity refers to the relationship between two events or stimuli that are presented close in time or space. In multimedia learning materials, contiguity can be used to help the learner understand the relationship between different pieces of information by presenting them in close proximity to each other. For example, a graphic and a related caption might be presented together to show the relationship between the two. By using spatial contiguity, multimedia learning materials help the learner better understand the relationship between different pieces of information and reduce cognitive load by eliminating the need to search for relevant information (Schroeder and Cenkci, 2018; Burkhart et al., 2021). When transferred to the context of writing and revising, the principle of contiguity can be accomplished by providing

in-text feedback rather than providing feedback in reference to an external rubric or message.

Signaling refers to the use of visual or auditory cues to help the learner understand the material and make connections between different parts of the content. Signaling can be achieved through a variety of means, including visual elements such as arrows, colours, and highlighted text. When used effectively, signaling helps the learner to more easily understand and retain the material presented in the multimedia learning resource (Richter et al., 2016). This principle applies to this study in that a central goal of sequence tagging is to highlight certain parts of the text and to assign different colors to different text elements.

Segmenting means breaking down a large learning sequence into smaller segments. This is often done with audiovisual content, for example, in allowing learners to pause an instructional video between meaningful sequences. According to Clark and Mayer (2011) the rationale for using segmentation is that it allows the learner to take essential processing steps without overloading their cognitive system. Learning has been shown to be more effective when information is presented in segments rather than in one long continuous stream (Rey et al., 2019). Sequence tagging allows us to segment a complex text into smaller parts that are easier to process and therefore more likely to be addressed by the learner.

2.3 Natural Language Processing Perspective

In a study which preceded the one presented here, Horbach et al. (2022) developed an automated scoring model for the emails in the eRubrix dataset. The purpose of that study was to prove that the human scoring of emails presented in Keller et al. (2023) could be generated automatically, and to evaluate the effectiveness of automated feedback based on that algorithm when students revised English emails. In their seminal study, Keller et al. (2023) had shown how a feedback rubric could be developed for English emails based on genre-based principles of writing instruction. They also showed that all aspects of writing quality covered in their rubric could be reliably used by human raters under the time-constraints of a live feedback study, and that the scores provided under such circumstances corresponded to differences in the linguistic quality of the texts, indicating high content validity. Horbach et al. (2022) then demonstrated that the human ratings provided by Keller et al. (2023) could be automatized as a set of binary quality criteria where each score was computed based on the whole text as input. Their study, however, did not automatize the segmentation (Horbach et al., 2022, p. 81). For that reason, it was not possible to draw the learners' attention visually to the specific segments where revisions were necessary. This current study therefore seeks to fill this research gap and provide an automated segmentation model which can be used to provide feedback on learner texts that follows central CTML design principles.

Methodologically, the approach in our study is an instance of a segmenting task where elements in a text are identified based on their function. Such tasks have been used, for example, to identify different parts (like *objective, method, results* and *conclusion*) in scientific abstracts (Hirohata et al., 2008). Mizuta and Collier (2004) identified so-called *rhetorical zones* in biology articles. In the educational domain, our task is related to other NLP tasks with the goal of identifying certain parts within a text either as feedback for learners or teachers, such as argument mining (Wachsmuth et al., 2016; Nguyen and Litman, 2018), where argumentative units are to be marked in essays. We therefore use an architecture that has been previously applied in argument mining tasks (Ding et al., 2022).

3 Data

3.1 eRubrix Dataset

The eRubrix dataset (Keller et al., 2023) contains 1,102 semi-formal emails written by Swiss lower secondary school students in grades 8 and 9. Most of them were in their 6th and 7th year of learning English as a foreign language and between 13 and 16 years old. The learners wrote three emails in randomized order and received feedback and suggestions for improvement in-between from trained human raters (Keller et al., 2023).

3.2 Writing Tasks

The writing tasks in the data-set consisted of three semi-formal emails in which students were asked to make inquiries concerning authentic, real life situations (Keller et al., 2023). In one task, they gathered information about a language school in the UK, in a second task, they inquired about a summer job at a burger restaurant, and in a third

task, they collected information for a holiday at a camping site (Keller et al., 2023). Figure 2 shows the *Burger Palace* task as an example. About 370 emails were written for each task (see Table 1). To avoid the need for anonymization, students were asked to sign their emails using the (gender-neutral) name *Kim Weber*.

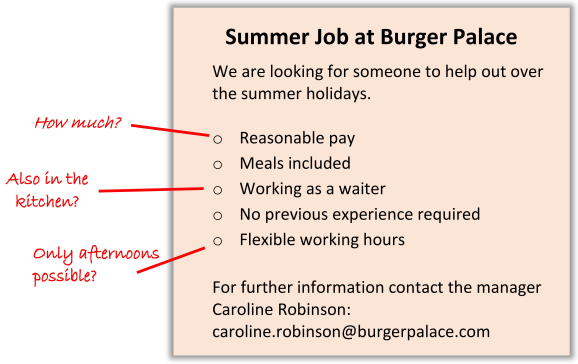

Figure 2: *Burger Palace* task from the eRubrix dataset (Keller et al., 2023, p. 25). The accompanying German instruction translates as follows: *"You want to make some money during your school holidays and are looking for a job. Read the advertisement you found on the internet and look at the notes you took (in red). Write an email to the store manager in which you introduce yourself and say what you are looking for. Inquire about the information in detail by using your notes in red"* (Keller et al., 2023, p. 24).

Prompt	**# emails**	**∅ # tokens (SD)**
Language school	367	97.9 (± 33.0)
Burger restaurant	368	104.1 (± 34.0)
Camping	367	105.0 (± 34.1)

Table 1: Basic dataset statistics.

3.3 Annotation

In Keller et al. (2023), the eRubrix text corpus was first rated on the basis of a rubric specifically developed for providing feedback to the learners. In a second step, the texts were additionally annotated in MAXQDA software by four trained human raters for a more detailed linguistic analysis (Keller et al., 2023). The different text segments were marked according to specific marking guidelines (see Table 2) and coded in terms of text quality for further linguistic analysis. These MAXQDA annotations provided the necessary data to train the automated text segmentation model presented in this paper. 40 texts had been annotated by all four raters (Keller et al., 2023) and were used in this study to calculate the raters' pairwise inter-annotater agreement (IAA) when marking the different segments.

A number of evaluation metrics have been used to calculate the IAA between two annotators in similar span annotation tasks. Ziai and Meurers (2014), for example, evaluated spans in focus annotations by computing agreement on the token level, while Reiter (2015) used boundary edit distance (see Fournier, 2013) on the segmentation of narrative texts. In our evaluation, we used a different span evaluation metric which we also applied in a similar fashion to evaluate human-machine agreement. Spans identified by one annotator were matched against spans found by the second annotator. They were considered true positive if at least 50% of the tokens found by annotator 1 were also identified by annotator 2, and vice versa. Unmatched spans by annotator 1 counted as false negatives, spans by annotator 2 without a counterpart by annotator 1 as false positives. These were combined to compute an overall Kappa score following Brennan and Prediger (1981). With this measure, we reached a pairwise IAA between 0.75 and 1.0. When increasing the required overlap from 50% to 90 %, the IAA was between 0.46 and 1.0 (see Table 3 for the averaged IAA values of all annotator pairs). The average percentage agreement of the four raters, as calculated by the average of their pairwise percentage agreements, ranged between 0.81 and 1.00 for the different criteria. Agreement for *closing* was low mainly because it was unclear to annotators whether the name after the closing should also be marked or not.

Together with the segmentation, annotators also assigned a quality label to each segment, indicating whether the content and form of the segment was appropriate (not used in this study). The annotator for the final gold standard was selected based on a many-facet Rasch analysis (Eckes, 2011) of these quality assessments, i.e. the rater whose ratings were the most balanced in terms of severity and leniency was selected.

Table 3 also shows basic statistics for the dataset. Elements are listed in order of their typical appearance in the text. We see that elements occurring later (*concluding sentence, closing*) have higher chances of being missing as learners often did not finish the email in time. We

Label	Annotation guidelines
Subject line	Code the whole subject line. If missing, code first letter of the email.
Salutation	Code the salutation including name and punctuation.
Information about writer	Code the introductory information about the writer including punctuation. Could be multiple sentences. Code entire extract, even if it contains a different type of information in between (e.g. matter of concern)
Matter of concern	Code the introductory information about the matter of concern including punctuation. Could be multiple sentences. Code entire extract, even if it contains a different type of information in between (e.g. information about the writer)
Task questions addressed	Code entirety of questions, including punctuation. If missing, code punctuation mark of previous sentence (or last letter if no punctuation present), where the questions would usually appear. Could be multiple sentences. Code entire extract even if there is additional information in between.
Concluding sentence	Code entirety of the concluding sentences, including puntuation. Could be multiple sentences, but it should be distinct from the questions.
Closing	Code entire closing, including punctuation, but do not include "Kim Weber". If closing is missing, insert code over last letter/character in the email or if only "Kim Weber" is present code the entire name.

Table 2: Guidelines for marking the segments in the eRubrix dataset

			50% overlap		90% overlap	
Label	**# segments**	**avg. length**	**∅ % agreem.**	κ	**∅ % agreem.**	κ
Subject line	1020	4.1	0.99	0.98	0.99	0.98
Salutation	1090	2.9	1.00	1.00	0.99	0.99
Information about writer	916	9.3	0.84	0.79	0.79	0.72
Matter of concern	1023	22.4	0.91	0.87	0.76	0.68
Questions	1015	45.2	0.96	0.95	0.73	0.64
Concluding sentence	747	10.2	0.93	0.91	0.76	0.69
Closing	697	2.1	0.81	0.75	0.60	0.46

Table 3: Number of segments per label as identified within the entire dataset, average length in tokens, and inter-annotater agreement. Average percentage agreement of all rater pairs, and kappa calculated according to Brennan and Prediger (1981). The segments were counted as agreement if either 50 or 90 percent of a segment matched with that of the second rater.

also see that individual elements have a very different average length with the *question* part by far the largest element on average.

In the original annotation setup, it was possible to annotate overlapping segments. It happened 93 times in the whole dataset, the majority of these cases (81) being overlaps between *matter of concern* and *information about the writer*. As our algorithm cannot work with overlapping segments, we ended a segment as soon as a new overlapping segment started, i.e. in cases of an overlap, the segment starting earlier was cut short.

4 Experimental Study

4.1 Experimental Setup

We use a sequence tagging architecture which has been successfully applied for structure-related tasks such as argument mining (Ding et al., 2022), as shown in Figure 3. In this architecture, tokens with a Inside-Outside-Beginning (IOB) tag representation of the gold-standard annotations are used as the input to a pretrained language model for token classification. We considered different pretained models and decided for RoBERTa (Liu et al., 2019) based on the Huggingface implementation [1] as it provided the best performance. We train the model for 10 epochs with a batch size of 16, CrossEntropyLoss as loss function, a learning rate at 1e-5 and an Adam optimizer.

We compare this model against several baselines: In the **random sentence baseline**, we split the data into individual sentences using the NLTK tokenizer[2] and assign each sentence a random label. In the **sentence order baseline**, we tag the first four sentences as *subject line*, *salutation*, *information about the writer* and *matter of concern* respectively, the last two sentences as *concluding sentence* and *closing*, and anything in-between as *questions*.

To examine the influence of the writing prompt, we train and test our model under several conditions: In the **all condition**, we employ 10-fold cross-validation on the complete dataset across all 3 prompts. In a **per-prompt condition**, we cross-validate on the *Language School*, *Burger Restaurant* and *Camping* prompt individually. Differences in the performance between **all** and the three **per-prompt conditions** (or rather a lack thereof) might be due to more training data available in the **all condition**. Therefore, we also introduce an **all-reduced condition** where we use only one third of the **all condition** to make the dataset size comparable to the per-prompt training sets. In a **cross-prompt condition**, we train on one prompt and test on one of the other two prompts. For each fold, we use the run with the best performance on the validation dataset.

Evaluation We follow a span evaluation F1 metric used also in similar tasks[3]. For this score, identified spans are matched against gold spans and considered a true positive if at least 50 percent of the gold span tokens are covered by the identified spans, and vice versa as described in Section 3. Unmatched gold spans count as false negatives, spans in the results without a gold counterpart as false positives. These are combined to compute an overall F-score. This score gives a good overall impression but does not account for exact matches at the segment boundaries. Therefore, we also evaluate accuracy on the token level.

4.2 Experiment 1: Prompt-Specific vs Generic Annotation

Table 4 shows the segmentation results for the two baselines, followed by the **all**, **all-reduced** and **prompt-wise** conditions.

Unsurprisingly, the **random sentence baseline** does not perform well. That also the **sentence order baselines** shows mediocre results can be taken as an indicator that the segmentation task is non-trivial.

The machine learning results show a high performance overall with token-wise accuracy between .88 and .91 and F1 scores between .84 and .89. The difference between the **all condition** and the other conditions is minimal, both for prompt-specific models and the **all-reduced** condition, indicating that the smaller models have already been provided with enough data to perform well.

4.3 Experiment 2: Cross-Prompt Segmentation

Experiment 2 investigates the model transfer potential from one email writing task to another. The lower half of Table 4 presents the results when a model trained on one prompt is applied to the other two prompts individually. Performance is slightly

[1] https://huggingface.co/roberta-base

[2] https://www.nltk.org/api/nltk.tokenize.html

[3] https://www.kaggle.com/competitions/feedback-prize-2021/overview/evaluation

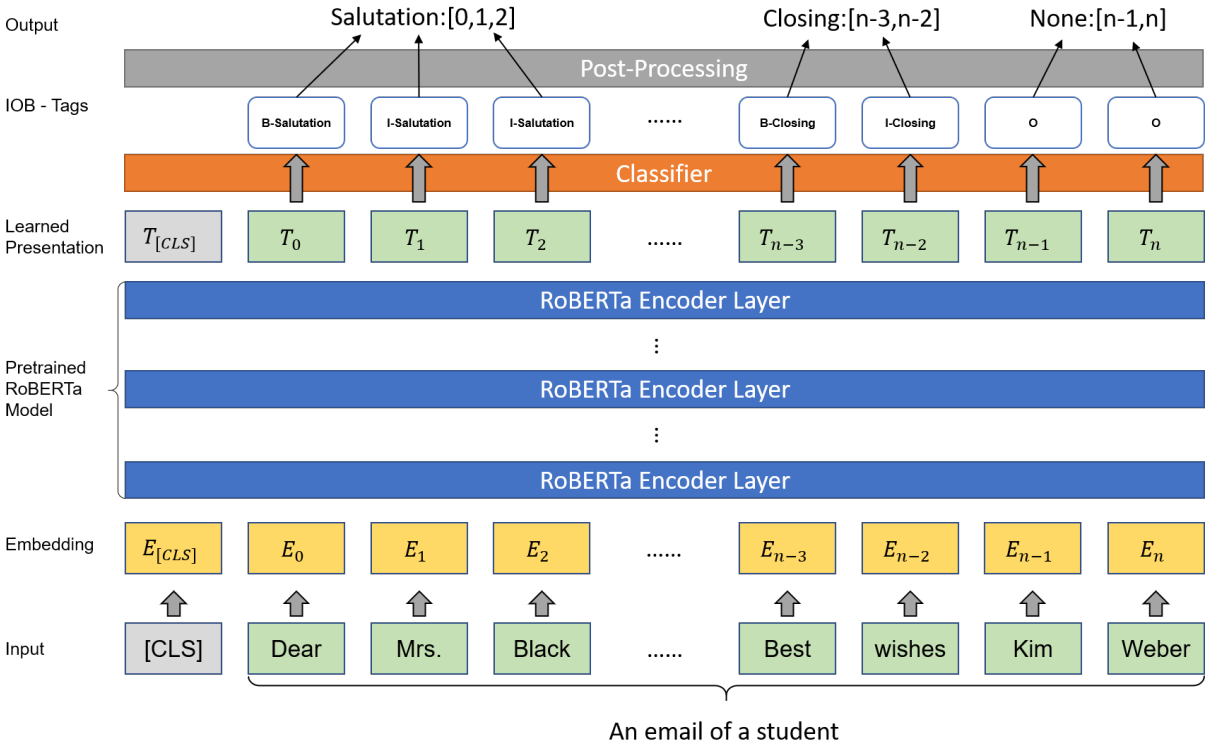

Figure 3: Adapted sequence labeling architecture from Ding et al. (2022).

Train	Test	F1	Acc.
Random Sentence Baseline		.06	.12
Sentence Order Baseline		.30	.42
All (CV)		**.89**	**.90**
All-reduced (CV)		.87	.89
Language school (CV)		.85	.88
Burger restaurant (CV)		.84	.88
Camping (CV)		**.88**	**.91**
Language school	Burger restaurant	.84	.87
Language school	Camping	.85	.87
Burger restaurant	Language school	.81	.83
Burger restaurant	Camping	.86	.87
Camping	Language school	.83	.84
Camping	Burger restaurant	.84	.84

Table 4: Segmentation results for two baselines and when training a generic or a prompt-based classifier (upper half) and for cross-prompt transfer (lower half).

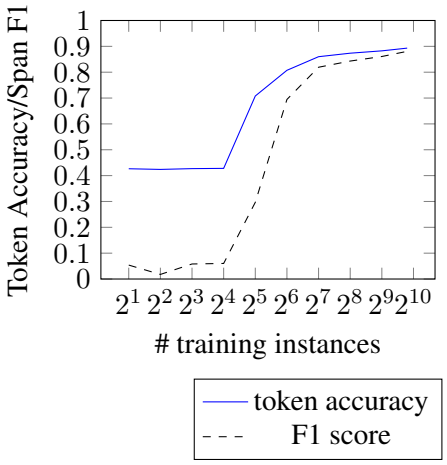

Figure 4: Learning curve experiment

lower than for the prompt-specific models, indicating that prompt-specific lexical material is certainly important. The criterion *salutation* can be best predicted in the cross-prompt segmentation, since it has a fixed form like "Dear xxx". *Subject line* can also be well predicted without context because it always spans over the first line of the email.

4.4 Experiment 3: The Influence of Training Data Sizes

In a practical application scenario when a teacher wants to train a model for a new prompt, it is important to know how much labeled data is required, since human annotation effort is often a crucial factor for creating a machine learning model.

Therefore, we perform learning curve experiments, in which we systematically vary the amount of training data. We use the **all condition** and 90% of the data for the training, while saving 10 % for testing.

Figure 4 plots labeled data on the x-axis vs segmentation performance (accuracy and F1) on the y-axis, showing that the algorithm is able to learn most of its performance from very few training instances. The curve flattens out in the end indicating that adding more training data will most likely not substantially improve performance any further.

4.5 Experiment 4: The Influence of Positional Information

Positional information is obviously important for the task as most elements typically appear at a certain position within the email. When students make errors in organizing their emails, i.e. when email elements do not appear in the expected location, one would expect a feedback that addresses this misplacement. It is thus important to correctly identify misplaced segments. As a worst-case scenario for emails in the wrong order, we therefore shuffle segments in emails randomly, i.e. we use gold standard information about email boundaries but randomly vary the order in which the elements appear. We use these scrambled emails in several ways. To assess the contribution of positional information in our original tagging models, we use scrambled test data (keeping the training data as is). To check how to make models more robust against misplacements, we train a model on scrambled training data, testing on both unchanged and scrambled test data.

Table 5 shows the results. We can observe a performance loss when using our normally trained model on scrambled test data (**scramble test**), indicating that the model indeed learns in part to rely on positional information and performs worse on test data that does not follow this convention. When also scrambling the training data, i.e. forcing the model to ignore positional information, scrambled test data can be handled with a similar performance to the baseline (compare **unscrambled** with **scramble both**), indicating that the data is somewhat redundant and that the same information can be learned without the positional information.

Setup	F1	Acc.
All (CV) - unscrambled	.89	.90
All (CV) - scramble test	.60	.78
All (CV) - scramble train	.85	.91
All (CV) - scramble both	.89	.92

Table 5: Segmentation results when training and/or testing on scramled data.

When comparing the performance on individual labels, we find that some labels, such as *subject line*, *salutation* and *closing* benefit more from positional information than others, i.e. for these labels there is a larger performance drop if positional information is missing.

4.6 Error Analysis

A confusion matrix between individual labels in the **all** condition (see Table 6) provides further information about the behavior of the algorithm. As can be seen in Table 6, most confusions occur between labeled segments and text segments without any label rather than between two labeled segments. This shows that assigning correct segment boundaries is sometimes difficult, resulting in segments without a counterpart with sufficient overlap. A comparison of the number of unmatched gold standard labels (1062) and unmatched predicted labels (277) shows that the algorithm tends to not assign a label rather than assign one.

When looking at the (substantially fewer) cases of confusion between two labels, most confusions unsurprisingly concern labels one would expect to be adjacent in an email, such as *matter of concern* and *information about the writer*. This corresponds to human annotation, as most overlapping annotations were found between these two labels. It often happens when the *information about the writer* is surrounded by *matter of concern* segments. Take the following sentences as an example: *I am interested to help you out over the summer holidays. I am 14 years old and my name is Kim Weber. I would like to earn some money in the summer holiday and i thought this is the right place to work in the summer holiday.* The first and

	Subject line	Salutation	Info. about writer	Matter of concern	Questions	Conclud. sent.	Closing	None
Subject line	917	1	0	0	0	0	0	3
Salutation	5	976	0	0	0	0	0	4
Info. about writer	0	2	751	15	1	0	0	63
Matter of concern	0	0	10	841	2	0	0	68
Questions	0	0	0	1	893	0	0	21
Conclud. sent.	0	0	0	0	0	640	2	43
Closing	0	0	0	0	0	11	537	75
None	5	10	245	297	185	162	108	N.A.

Table 6: Confusion matrix between gold standard (columns) and results in the *all* setting (rows)

the last sentence illustrate the *matter of concern*, whereas the sentence in-between was double annotated with both *matter of concern* and *information about the writer*.

5 Discussion & Practical Applications

With the developed technology, we envision two application scenarios. First, automatic segmentation could be used to provide formative feedback to students by showing them not only how their text was scored automatically, but also where the algorithm thought it had found the respective passages, pointing at the location where a revision could take place. According to CTML principles, this should reduce cognitive load and thus positively affect learning. Contiguity can be achieved by presenting feedback within the text rather than in the margins. By being able to highlight and assign colours to certain parts of the text, signaling can support the learners' understanding. Most importantly, the segmentation of the text can break a complex task down into smaller parts. Students can revise their text step-by-step rather than being faced with a lot of information at once. Especially when combined with evaluative feedback (automatic quality assessment) on the segment level, the reduction of cognitive load in the revision process may lead to higher feedback uptake and better learning outcomes. In addition, such formative feedback could also be enriched with automatic quality assessment similar to the study by Horbach et al. (2022). From an NLP perspective, the quality of automatic scoring, in turn, might also benefit from segmentation in that only relevant parts of the email would be fed into the scoring algorithm.

Second, segmentation could be the basis for the generation of various activity types useful for teaching students how to write an email. In particular, such activities could be set up with the texts written by the learners themselves. These could be identification tasks (*Please indicate where the* Matter of Concern *is in this email.*), reordering tasks (*Please bring these email segments into the right order.*), gap-filling tasks (*Which part is missing here?*) and many more. When combined with an automated model for judging the quality of the segments, further activity types may become possible such as judgment tasks (*Which texts have a suitable concluding sentence?*) or comparison tasks (*Which salutation is more appropriate in terms of register?*). A crucial advantage of generating such activities from automatically segmented texts is that arbitrary emails could be integrated into language-learning tasks, including emails the learners themselves have written.

6 Conclusion

We showed in this study that the individual segments of a formal email can be predicted with high accuracy, making segmentation a suitable instrument to give feedback in an EFL context. We have outlined ways how segmentation could be used to generate language learning tasks and - together with automatic scoring - could be used to generate formative feedback for language learners. We will explore these directions further in future work.

7 Acknowledgements

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany, and partially within the KI-Starter project "Explaining AI Predictions of Semantic Relationships" funded by the Ministry of Culture and Science, Nordrhein-Westfalen, Germany.

References

Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Christian Burkhart, Andreas Lachner, and Matthias Nückles. 2021. Using spatial contiguity and signaling to optimize visual feedback on students' written explanations. *Journal of Educational Psychology*, 113(5):998.

RC Clark and RE Mayer. 2011. Applying the segmenting and pretraining principles: Managing complexity by breaking a lesson into parts e-learning and the science of instruction.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2022. Don't drop the topic - the role of the prompt in argument identification in student writing. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 124–133, Seattle, Washington. Association for Computational Linguistics.

Thomas Eckes. 2011. Introduction to many-facet rasch measurement. *Franfurt am Main: Peter Lang*.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Andrea Horbach, Ronja Laarmann-Quante, Lucas Liebenow, Thorben Jansen, Stefan Keller, Jennifer Meyer, Torsten Zesch, and Johanna Fleckenstein. 2022. Bringing automatic scoring into the classroom–measuring the impact of automated analytic feedback on student writing performance. In *Swedish Language Technology Conference and NLP4CALL*, pages 72–83.

Ken Hyland. 2019. *Second language writing*. Cambridge university press.

Stefan D Keller, Ruth Trüb, Emily Raubach, Jennifer Meyer, Thorben Jansen, and Johanna Fleckenstein. 2023. Designing and validating an assessment rubric for writing emails in english as a foreign language. *Research in Subject-matter Teaching and Learning (RISTAL)*, 6(1):16–48.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Paul Kei Matsuda. 2003. Process and post-process: A discursive history. *Journal of second language writing*, 12(1):65–83.

Richard E Mayer. 2001. *Multimedia learning*. Cambridge University Press.

Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1):43–52.

Yoko Mizuta and Nigel Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 29–35.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Michael Noetel, Shantell Griffith, Oscar Delaney, Taren Sanders, Philip Parker, Borja del Pozo Cruz, and Chris Lonsdale. 2021. Video improves learning in higher education: A systematic review. *Review of educational research*, 91(2):204–236.

Nils Reiter. 2015. Towards annotating narrative segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 34–38.

Günter Daniel Rey, Maik Beege, Steve Nebel, Maria Wirzberger, Tobias H Schmitt, and Sascha Schneider. 2019. A meta-analysis of the segmenting effect. *Educational Psychology Review*, 31:389–419.

Juliane Richter, Katharina Scheiter, and Alexander Eitel. 2016. Signaling text-picture relations in multimedia learning: A comprehensive meta-analysis. *Educational Research Review*, 17:19–36.

Noah L Schroeder and Ada T Cenkci. 2018. Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review*, 30:679–701.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pages 1680–1691.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 159–168.