# Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification

**Neha Srikanth**
Department of Computer Science
The University of Texas at Austin
nehasrik@utexas.edu

**Junyi Jessy Li**
Department of Linguistics
The University of Texas at Austin
jessy@austin.utexas.edu

## Abstract

Much of modern-day text simplification research focuses on sentence-level simplification, transforming original, more complex sentences into simplified versions. However, *adding* content can often be useful when difficult concepts and reasoning need to be explained. In this work, we present the first data-driven study of content addition in text simplification, which we call *elaborative simplification*. We introduce a new annotated dataset of 1.3K instances of elaborative simplification in the Newsela corpus, and analyze how entities, ideas, and concepts are elaborated through the lens of contextual specificity. We establish baselines for elaboration generation using large-scale pre-trained language models, and demonstrate that considering contextual specificity during generation can improve performance. Our results illustrate the complexities of elaborative simplification, suggesting many interesting directions for future work.

## 1 Introduction

Text simplification aims to help audiences read and understand a piece of text through lexical, syntactic, and discourse modifications, while remaining faithful to its central idea and meaning (Siddharthan, 2014). It remains an important task, improving text accessibility for children (De Belder and Moens, 2010; Kajiwara et al., 2013), language learners (Yano et al., 1994; Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014; Paetzold, 2016), and those with language impairments (Carroll et al., 1998; Rello et al., 2013). Text simplification can also be a useful pre-processing step for other NLP tasks such as machine translation (Chen et al., 2012; Štajner and Popovic, 2016) and summarization (Vanderwende et al., 2007; Silveira and Branco, 2012).

With the introduction of large, parallel corpora (Zhu et al., 2010; Woodsend and Lapata, 2011; Coster and Kauchak, 2011; Xu et al., 2015), text simplification research has rapidly advanced in recent years, especially in sentence simplification (Alva-Manchego et al., 2020). However, document simplification involves rich linguistic phenomena that cannot be easily characterized by sentence-level transformations of text, e.g., the omission and addition of content (Petersen and Ostendorf, 2007; Siddharthan, 2014).

| **Original Text** |
|---|
| Results, she said, "could help the team better understand ancient Egyptian health" and, correspondingly, modern-day health. For instance, some mummies still have arteries in their mummified remains, Miller-Thomas said. And, sometimes, scientists can tell if those arteries had hardened. |
| **Simplified Text** |
| The scans could help the team understand about ancient Egyptians' health. For example, some mummies still have arteries. **An artery is a tube that moves blood through the body. The artery could show if the person had been healthy or not.** |

Figure 1: Elaborative simplification with two elaborations of varying contextual specificity.

This paper presents the first data-driven, dedicated study of *elaborative simplification*, which involves inserting elaborations in the form of definitions, explanations or clarifications to improve readability by providing readers with necessary additional context. Effective elaborations must provide background in a *contextual* manner, adding relevant information to the surrounding text.

Figure 1 shows an example. The original text snippet explains that scientists study mummy arteries to see whether they are hardened. In the corresponding simplified text, we see two elaborations inserted – one, in green, simply defines an artery, and the second, in blue, states the implication of hardened arteries. The content of both elaborations is semantically absent from the original text.



*Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137

Our goal is to provide resources and directions toward understanding and generating naturally occurring elaborations. We present an annotated dataset of 1.3K instances of elaborative simplification in the Newsela corpus (Xu et al., 2015). We automatically identify candidate elaborations from simplified documents, and have human annotators verify candidates. We find that many elaborations require multi-hop reasoning, inference, commonsense reasoning, and relevant information retrieval, making it an interesting testbed for a bevy of related tasks.

The previous example highlights two elaborations on opposite ends of the spectrum – the first requires little context, while the second is highly contextualized, drawing a conclusion from content presented in the original text. To this end, we characterize elaborations by annotating their *contextual specificity*, i.e., the extent to which the added content is specific to the current topic under discussion.

We reveal that our dataset contains a fairly balanced distribution of contextual specificity. Qualitatively, while inserting definitions may help provide background about entities, highly contextualized elaborations interpreting or clarifying content can help readers understand the larger implications or significance of ideas presented in the original text. We propose the primary task of generating elaborations given document context. We present baselines for elaboration generation mainly using GPT-2 (Radford et al., 2019), and discuss some of the challenges, especially with respect to the contextual specificity of added content.

We find that generation quality can be improved by selecting an elaboration with an appropriate predicted contextual specificity level. However, existing methods struggle to effectively incorporate input context to generate elaborations. We hope that this study will motivate advancement in elaborative simplification.

In summary, our main contributions include:

1. Introduction of elaborative simplification, a previously understudied phenomenon in text simplification;

2. A new, annotated dataset of 1.3K naturally occurring elaborations in the Newsela corpus and their contextual specificity;

3. Analysis of the challenges of elaborative simplification for pre-trained language models through performance of our baselines.

We make our annotations and code available at https://github.com/nehasrikn/elaborative-simplification.

## 2 Data and Annotation

Elaborative simplification involves the *insertion* of content to make simplified text easier to understand. We present an annotated dataset of 1.3K elaborations from the Newsela corpus (Xu et al., 2015), which contains English news articles manually simplified by professional editors. We describe the scope of our elaborative simplification study (§2.1), strategies for trusted annotators to extract elaborations (§2.2) and rate contextual specificity (§2.3), and scaling up annotation through crowdsourcing with rigorous quality control (§2.4).

### 2.1 What is an elaboration?

We consider a sentence an elaboration if it contains new content (e.g. statements about entities, actions, or concepts) present in the simplified document, but semantically missing from the original document. Note that while elaborations can contain multiple sentences, we define our label at the sentence level. Past simplification research has focused on operations such as substitution and deletion, but simplifying a piece of text that may contain unknown or difficult concepts could involve inserting simple explanations as well. As we highlight in §6, others have shown that audiences such as new language learners benefit from elaboration or explanation insertion (and conversely, that unfamiliar concepts negatively impact reading comprehension), though computational approaches till date have been largely limited to definition retrieval.

**Scope.** We intentionally choose to study *how* concepts are elaborated, posing a scenario where an author has the freedom to specify where to elaborate, and our system generates an appropriate elaboration. We do this for two main reasons: first, understanding how to elaborate can be utilized in a system where users specify what to elaborate on, in the spirit of personalized simplification (Paetzold and Specia, 2016; Bingel et al., 2018). Second, determining *when* to elaborate is arguably pragmatically more complex, in that the need for elaboration often relies on the writer's belief about their readers' background, knowledge, and reading ability, as well as their own judgments on how often to elaborate. For example, in the extreme case, inserting an elaboration after every sentence could prove

| Original Text | Simplified Text | Specificity |
|---|---|---|
| A new standard would put more areas of the country in violation of air quality standards and place parts of the West in a tough spot between a rising baseline of ozone and stricter federal limits. Limiting pollution flowing in from Asia would require an internal treaty, said Owen Cooper, an atmospheric scientist at the Cooperative Institute for Research in Environmental Sciences in Boulder, Colorado. | Smog has become much worse in California's cities like Bakersfield, Fresno, and Los Angeles. It will not be easy to stop pollution from China, said scientist Owen Cooper. The U.S and China would have to work out a deal. **A deal between two countries is called a treaty.** Such an agreement is not likely, said Cooper. | Low |
| "It was something kind of fun for the country." The artwork, which will be open to the public from Saturday until Oct. 31, adds a new element to the Mall, the stretch of green space, museums and memorials from the Lincoln Memorial to the U.S Capitol, known as "the nation's front lawn." "This is the perfect environment of science and art coming together," she said. | The National Mall is the stretch of parks and museums in Washington D.C. Its nickname is the "nation's front lawn." It is called this because it is a green, open space and is next to some of the country's most important government buildings. **Every year millions of tourists visit it.** This October, there is an extra reason to go. | Medium |
| Claudia gets straight A's at one school, somewhat lower grades at her other. But as years pass and coursework gets more complex, the odds rise against her. Eventually, about 90 percent of kids living in seasonal worker housing drop out of school, according to the San Jose-based nonprofit human rights organization Human Agenda. | Claudia goes to two different schools each year. She gets straight A's at one. Her grades are lower at the other. **Switching between schools makes it more difficult to learn.** It's easy for kids like Claudia to fall behind. Nine out of every 10 children living in farmworker camps drop out of school, says Human Agenda. | High |
| We really wanted to go the next mile to nail down an Earth and to tell if there are moons," he said. "The extended mission, with extra transits, would have told us that." Added UC Berkeley's Gould: "The longer you go ... the more certain you are that it is a planet." Because Kepler's data flow has stopped, it is even more important to understand the existing data and look more closely for subtle patterns that might suggest an Earth-like planet. | We really wanted to go the next mile to nail down an Earth and to tell if there are moons," he added. "The extended mission, with extra transits, would have told us that." **Kepler is not providing any more information.** So it is even more important to understand what it has already found and look more closely for patterns that might suggest a planet like Earth. Computer experts like Erik Petigura will have to crunch the numbers to make sense of it all. | N/A *(Not an Elaboration)* |

Figure 2: Example candidate elaborations. Rows 1–3 contain verified elaborations. Row 4 contains a rejected candidate. We include the original and simplified text regions, highlighting the candidate elaboration, and its corresponding level of contextual specificity in Column 3.

useful for children or readers with no background knowledge about the document content, but may be unnecessary for adults or those with sufficient knowledge.

**Task.** We introduce the primary task of elaboration generation: given some document context $C$ consisting of text from the original and/or simplified documents, generate an elaboration $E$.

## 2.2 Extracting Elaborations

Detecting elaborative simplification requires crafting a way to reliably extract sentences containing new content in simplified documents. Asking humans to read and annotate every sentence in each document is prohibitively costly. To streamline this process, we first obtain candidate elaboration sentences with automatic sentence alignment, then use human annotation to extract true elaborations.

**Candidate extraction.** Each set of articles in the Newsela corpus consists of multiple simplified articles ranging from grades 3–12. We choose the article written for the lowest grade level as our simplified document (we leave investigating simplified documents across higher grade levels as future work). Using the approach from Zhong et al. (2020), we then align sentences from the original and simplified documents by thresholding the cosine similarity of sentence vector representations using Sent2Vec (Pagliardini et al., 2018). We then consider sentences in the simplified document that are not aligned with any sentence in the original document as *candidate elaborations*. Of the 54,892 sentences across the 1,042 simplified documents (on average, 52 sentences per document), 6,207 were extracted as candidate elaborations.

**Human verification.** Before crowdsourcing, we conducted a pilot study of elaboration verification with two sets of annotators: (1) Expert annotators (one graduate student, one undergraduate, both native speakers of English) who studied the data extensively; (2) 13 trusted undergraduate volunteer annotators at our university, also native English speakers. They received detailed instructions, but no face-to-face training. This allowed us to gauge task scalability and to gather feedback to design our crowdsourcing protocol. The 13 annotators each annotated a subset of 50 randomly selected documents (a total of 301 candidate elaborations) from our corpus. Each candidate elaboration was annotated by 2 to 4 annotators.

For each original-simplified document pair, we provided annotators with the entirety of both documents. We asked them to identify whether each candidate elaboration truly contained semantically new content, and to provide a rationale for their annotation. We *aggregated* the annotations for each

candidate elaboration by taking the mode of all responses. The expert annotation consisted of 150 of these candidate elaborations under the same setup. Figure 2 shows some examples of verified and rejected candidate elaborations.

**Agreement.** Cohen's Kappa among the two expert annotators is 0.75, indicating substantial agreement (Artstein and Poesio, 2008). Cohen's Kappa between expert annotations and aggregated student annotations is also substantial, at 0.67. Krippendorff's alpha among the 13 student annotators is 0.37. As in complex NLP annotations (Nye et al., 2018), although there is subjectivity among individual annotators due to the complicated nature of the task, their aggregated judgment can be of as high quality as trained expert annotators.

### 2.3 Contextual Specificity

At first glance, it seemed that elaborative simplification might simply involve retrieving definitions (Paetzold and Specia, 2016) or crafting informative post modifiers (Kang et al., 2019). However, while annotating candidate elaborations, we noticed that elaborations in our corpus took a variety of forms.

To better understand content addition, we conducted an extensive study of elaborations and found that often times, clarification or analysis sentences specific to document context are inserted to aid comprehension or facilitate connections between content in the original text. Notably, elaborations vary in their *contextual specificity*, i.e., the degree to which an elaboration is specific to the context.[1] For example, while simple definitions can be inserted into several different documents mentioning the same entity (low contextual specificity), some elaborations containing clarifications, commonsense reasoning applied to document content, or explicit inference are more contextually specific, as illustrated in Figure 2.

This formulation is inspired by prior work in text specificity (Li et al., 2016; Ko et al., 2019) which is related to how a sentence "stands on its own" or sentence "decontextualization" as in Parikh et al. (2020). As we discuss in §2.4, contextually specific elaborations tend to have slightly lower sentence specificity, thus depending on the surrounding context to enhance understanding.

We ask the pair of experts from the previous pilot to annotate 116 randomly chosen verified elaborations for contextual specificity. Each expert was again given the entirety of the original and simplified documents with the highlighted elaboration, and asked to label its contextual specificity on a scale of 1–3 (low/medium/high). Their Fleiss' Kappa showed moderate agreement (Landis and Koch, 1977) with $\kappa = 0.57$. Spearman's correlation between the two annotators is 0.72. To enable collection, study, and modeling of this linguistic knowledge at scale, we gather contextual specificity ratings during crowdsourcing.

### 2.4 Crowdsourcing

Annotating elaboration verification and contextual specificity requires careful reading and thoughtful reasoning over text. For the pilot described in §2.2, we provided thorough instructions and example documents and annotations. While these trusted annotators delivered high quality, reliable annotations, they ultimately cannot annotate a dataset of the scale supervised systems require. To remedy this, we use Amazon Mechanical Turk to collect labels at scale, albeit with slightly more noise. Our rationale is that models can tolerate this during training, and we ensure cleaner validation and test sets through expert annotations.

**Task setup.** We ask workers to annotate elaboration verification and contextual specificity in a single task (HIT). For each candidate elaboration, we provide crowdworkers with the text region from the simplified document containing the elaboration, and the aligned text region from the original document. We ask crowdworkers to categorize each candidate as a true elaboration, not an elaboration, or indicate that the snippets were unrelated. If true elaboration is selected for a candidate, we asked them to rate its contextual specificity[2]. From feedback during our expert pilots, we determined that providing entire documents was often distracting, proving necessary only in rare cases where content was drastically rearranged. Instead, we display text regions of 5–7 sentences from both the simplified and original documents. The simplified text region contains the candidate elaboration and surrounding sentences, and the original text region contains sentences that are aligned with neighboring sentences

[1] We draw a distinction between contextual specificity and contextual relevance (as in Kang et al. (2019)).

[2] During crowdsourcing we utilized a 5-point scale, but aggregated the labels to a 3-point scale because the two scores on either end of the scale are not distinctive (i.e., are subjective).

of the elaboration in the simplified text region. We compose HITs that consist of ∼4 candidates from the same article.

**Quality control.** To ensure high quality annotations, we ask crowd workers to provide a rationale for each rating decision, as in §2.2. These rationales provide insight into worker interpretations of our task, allowing us to actively curate annotations to only include reliable annotations in our dataset. For example, using this method, we were able to remove annotations where crowd workers inflated specificity ratings due to coreferent entity mentions (i.e *"It is a tube that moves blood"* as opposed to *"An artery is a tube that moves blood"*).

In addition, we require all crowd workers to reside in the US, UK, Canada, Australia, or New Zealand, and to have completed ≥ 100 HITs with an acceptance rate of 95%. Each elaboration is annotated by 5 different crowdworkers. Through active monitoring and small batches of HIT releases, we identified a set of workers that we trust and invite back to the task. Initially, we pay $0.15 – $0.23/HIT, and retroactively pay trusted workers at the rate of $8/hr after work time information is obtained.

**Agreement between trained and crowdsourced annotators.** For both tasks, we aggregate crowdsourced labels by taking the mode of all responses[3]. Cohen's Kappa of elaboration verification between crowdworkers and experts is 0.37 (fair). To measure contextual specificity agreement between crowdworkers and experts, we use Krippendorff's alpha with an ordinal distance metric, aggregating Turker and expert responses using the mode to obtain an agreement value of $\alpha = 0.47$, indicating moderate agreement (Artstein and Poesio, 2008). We attribute the disparity between inter-expert agreement and expert-crowdworker agreement to the challenge and subjectivity of this task, especially amongst untrained crowd workers. Though crowdsourcing our data does result in a slightly noisier training set, we are able to collect data for supervised learning and analysis at scale.

**Dataset analysis.** Using Mechanical Turk, we annotated 4,178 out of the 6,207 candidate elaborations from 1,042 documents. We obtained 1,299 verified elaborations, establishing an approximate 32% conversion rate from candidate to verified elaborations. Note that since candidate elaborations are obtained automatically, this does not accurately reflect the true elaboration rate per document, but rather a lower bound. On average, the elaborations in are corpus are 7–13 tokens long.

[3]Using the mean as an aggregation function resulted in noisier labels.

| | Low | Medium | High | **Total** |
|---|---|---|---|---|
| Train | 303 | 349 | 397 | **1049** |
| Valid | 71 | 39 | 24 | **134** |
| Test | 42 | 34 | 40 | **116** |
| Total | 406 | 423 | 470 | **1299** |

Table 1: Dataset distribution by contextual specificity.

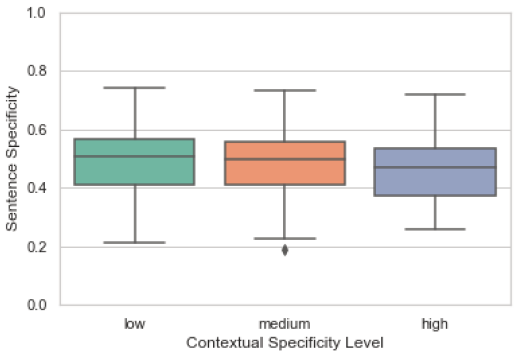


Figure 3: Sentence specificity distribution of elaborations across contextual specificity levels.

To ensure finetuning and evaluation quality, we use the expert-annotated subset of our data for the test set, and sought additional expert annotations for the validation set as well. Table 1 shows our dataset size across splits, stratified by contextual specificity. Our dataset contains a relatively uniform distribution of specificity levels, confirming our qualitative analysis that the contextual specificity of added content is diverse.

**Sentence Specificity.** As mentioned in §2.3, we explore the nature of sentence specificity of elaborations by running the sentence specificity predictor from Ko et al. (2019) on all standalone elaborations across all splits in our dataset. Sentence specificity predictions range on a continuous scale from 0 (very general) to 1 (highly detailed). Figure 3 shows the sentence specificity distribution across contextual specificity levels. The correlation between contextual and sentence specificity is $\tau = -0.11$, and is statistically significant. This negative correlation illustrates some of the intuition behind contextual specificity – only when highly contextualized elaborations are inserted into documents do they facilitate document understanding.

## 3 Elaboration Generation

We frame elaborative simplification as a natural language generation task, and describe a process mimicking editors elaborating as they compose a simplified document from the beginning (i.e. elaborations may be generated based only on the preceding simplified/original context) [4]. Elaboration generation is a challenging test for a model's ability to produce relevant and effective elaborations ranging in contextual specificity given snippets of context from documents in our corpus. We investigate the abilities of pre-trained language models to generate elaborations, establishing baselines in §3.1 and incorporating contextual specificity in §3.2. We find that selecting elaborations of appropriate levels of predicted contextual specificity can help improve elaboration generation results.

### 3.1 Baseline Elaboration Generation

We generate elaborations using GPT-2 (Radford et al., 2019), a large-scale pre-trained language model which has been shown to be effective in a range of generation tasks, including in recent efforts to elicit world and commonsense knowledge (Zhou et al., 2020; Shwartz et al., 2020).

Formally, we generate elaborations by conditioning on some document context, $C$. In this baseline setting, we generate sequences via greedy decoding. We utilize context from the original document ($C_o$) and from the simplified text ($C_s$). To understand the role that context plays in elaboration generation, we elicit elaborations from the language model by providing it one of the following: **(1)** 2 sentences prior to the gold elaboration in the simplified document ($C_{2s}$), **(2)** a concatenation of 2 sentences prior to the gold elaboration from the simplified document and the corresponding aligned region in the original document ($C_{2s} + C_o$), **(3)** 4 sentences prior to the gold elaboration in the simplified document ($C_{4s}$).

**Finetuning.** We finetune GPT-2 on the set of simplified documents written for the lowest grade level in the Newsela corpus, as well as on our dataset of verified elaborations excluding the test set. We found that such fine-tuning substantially improves generation quality (c.f. Appendix B.1).

[4] We explored elaboration generation as a post-processing task after document simplification (Appendix B.2). From preliminary results, we find it to be a more nuanced task which we leave for future work.

### 3.2 Specificity-guided Generation

As discussed in §2.3, elaborations in our corpus are notably diverse in terms of their contextual specificity. Producing elaborations of appropriate contextual specificity is important, e.g., inserting an unnecessary definition instead of explaining a central concept can be ineffective or detrimental to readers' understanding. Rows 1-2 in Figure 4 show examples where the elaboration generated by the model in §3.1 does not match the level of contextual specificity of the gold elaboration, motivating our exploration of including contextual specificity and its prediction to aid elaboration generation.

**Contextual specificity prediction.** We build a model to classify the level of contextual specificity of an elaboration as low, medium, or high to incorporate downstream during generation. We leverage BERT (Devlin et al., 2019) for this task. Appendix A explores this auxiliary task further to understand modern NLP models' ability to capture this linguistic information.

We train the model on $(E, s)$ pairs, where $E$ is an elaboration, and $s$ is its labeled contextual specificity. We feed $E$ as input to BERT, and then feed the `[CLS]` token embedding into an output layer for classification. We freeze the BERT parameters since fine-tuning yielded unstable results. We utilize `bert-base` from the HuggingFace library (Wolf et al., 2019). After tuning on the validation set, we train for 5 epochs, using a batch size of 32 and a learning rate of 2e-3. We use the default dropout rate of 0.1 for self-attention layers, but refrain from adding dropout on our linear layer.

This contextual specificity model achieved an accuracy of $56.8 \pm 1.5$, a macro-averaged F1 score of $55.3 \pm 1.6$, a Spearman correlation of $47.5 \pm 2.6$, and a mean absolute error of $0.552 \pm 0.01$, averaged across 15 randomly initialized runs. This performance is better or on par with other models that incorporate document context in different ways (Appendix A). We find contextual specificity prediction to be a challenging task for BERT. Prediction of *expected* contextual specificity (i.e prediction from context alone, without the elaboration) was particularly difficult, and we leave building stronger models in this setting to future work.

**Generation.** We investigate the importance of contextual specificity in generating effective elaborations by comparing sequences generated in 3 ways:

1. **Greedy:** Generate elaborations via greedy decoding. This setting was discussed in §3.1.
2. **Top-k:** Sample a sequence from the language model using top-k sampling (Fan et al., 2018), without considering contextual specificity.
3. **Contextual specificity-informed sampling**, shorthand **Contextual:** Sample sequences using top-k sampling until we have 3 elaborations of low, medium, and high contextual specificity, as predicted by the contextual specificity model, and select the sequence with predicted contextual specificity matching the gold specificity level.

In practice, one would ideally use a contextual specificity model trained *without* the elaboration itself (i.e., *Context-Only* models in Appendix A) to predict the appropriate level of contextual specificity of a generated elaboration. However, since we leave to future work to build a strong model presented with this setup, we instead utilize the gold specificity label and explore the upper bound with our generation experiments.

We use sampling-based decoding strategies to achieve contextual specificity diversity because we find that while beam-based decoding methods may result in sequences with diverse *content*, they do not necessarily result in sequences with diverse *contextual specificity.*

### 3.3 Experimental Settings

We use GPT-2 medium from the HuggingFace library (Wolf et al., 2019) to finetune and generate elaborations. We finetune GPT-2 on documents simplified for the lowest-grade level in the Newsela corpus for 3 epochs with a learning rate of 1e-5 and a batch size of 32. For sampled sequences, we use top-k sampling with $k = 40$, and a temperature of $t = 0.45$, tuned on validation data.

## 4 Generation Evaluation

As elaboration generation is a new task, we include BLEU scores for completeness and emphasize human evaluation, which provides important insight early on in the study of a new phenomenon.

### 4.1 Automatic Evaluation

We report BLEU (Papineni et al., 2002), a standard metric in generation tasks. Table 2 shows corpus BLEU-1 and BLEU-2 scores on our test set. As illustrated in Table 2, the best models, as reflected by BLEU, are those finetuned on the Newsela simplified corpus, with four sentences from the simplified document before the gold elaboration as context.

| | Greedy | | Top-k | | Contextual | |
|---|---|---|---|---|---|---|
| Context | B-1 | B-2 | B-1 | B-2 | B-1 | B-2 |
| $C_{2s}$ | 20.8 | 6.77 | 20.4 | 6.12 | 21.4 | 7.26 |
| $C_{2s} + C_o$ | 18.7 | 5.66 | 17.2 | 4.32 | 19.0 | 5.31 |
| $C_{4s}$ | 20.8 | 5.54 | 19.7 | 6.06 | **22.4** | **7.56** |

Table 2: BLEU-1 and BLEU-2 scores for elaborations generated by GPT-2, finetuned on the Newsela simplified document corpus. Results for our best model, which we conduct human evaluation on, are in bold.

| System | Greedy | Top-k | Contextual |
|---|---|---|---|
| % selected | 53.2 | 44.9 | 58.0 |

Table 3: Percentage of annotations for which users selected elaborations generated by each model.

While BLEU captures lexical overlap between generated and gold elaborations, it is also criticized due to poor correlation with human judgments (Liu et al., 2016; Novikova et al., 2017; Chaganty et al., 2018), as it fails to capture semantic similarity or reward multiple plausible hypotheses. During manual inspection of these sequences, we find that elaborations produced after finetuning GPT-2 can be semantically plausible, coherent, and elaboration-like. Content that is pertinent and new, but that does not overlap with the content in the gold elaboration is not rewarded. In some cases, staying true to the content of the gold elaboration is likely unnecessary, as long as the contextual specificity is comparable (see row 4 in Figure 4). To that end, we also perform a human evaluation study of generated elaborations, given that the purpose of elaborations is largely to make simplified text easier to understand for readers.

### 4.2 Human Evaluation

We set up our human evaluation similar to Panthaplackel et al. (2020), providing a pair of expert evaluators elaborations generated by our $C_{4s}$ model (see Table 2) in each of the three setups (greedy, top-k, contextual), and ask them to select the sequence they thought was most coherent, topical, semantically plausible, and elaboration-like. We allow selection of multiple sequences if they are equally good, and no selection if all sequences are poor. We report human evaluation results as the percentage for which evaluators chose the sequence as higher

| Simplified Text | Generated Elaborations |
|---|---|
| It also has a high amount of smoking for a big city, according to a health department report. In Philadelphia, smoking is more common among the poor than people who aren't poor, said Giridhar Mallya. He works for the Philadelphia Department of Public Health. Being poor can be very hard on a person's mind. **It is not easy to relax, or to feel positive**.<br>***Contextual Specificity:*** *High* | **Greedy:** "It's not just the money," he said.<br>**Top-k:** Mallya said the study showed that the poorest people in the city smoke a lot more than the richest people.<br>**Contextual:** It can make them feel guilty. |
| Workers at the Danish zoo said they had no choice but to kill the giraffe. The reason why is complicated. They said its genes were too much like the genes of other giraffes. All plants and animals have genes. They play a big part in what animals and plants look and act like. Genes are passed down from parents. **They determine eye color, hair color, athletic ability, health and more.**<br>***Contextual Specificity:*** *Low* | **Greedy:** The scientists found that the giraffe's genes were very similar to those of other giraffes.<br>**Top-k:** Scientists have known for years that the giraffe's genes were more like those of other giraffes.<br>**Contextual:** The genes are passed down from generations. |
| Because they look like a poisonous butterfly, they are left alone by predators. Looking like another species works just fine for them. While they avoid getting eaten, the male and the masculine-colored female are still at risk. For a long time, biologists have studied how animals try to look like other animals. It teaches them a lot about natural selection. Natural selection is sometimes called "survival of the fittest." **It means that animals that can change will survive.**<br>***Contextual Specificity:*** *Low* | **Greedy:** It is a way of thinking about how animals live.<br>**Top-k:** It means that the more an animal looks like another animal, the more likely it is to survive.<br>**Contextual:** It's a way of thinking about how organisms adapt to change. |
| The newly found cat is probably part of a separate branch of the cat family, said Tseng. The modern snow leopard is most likely part of the same branch. Both are able to live high in the Tibetan mountains. Big cats present serious problems for paleontologists. **They have existed millions of years longer than people have.**<br>***Contextual Specificity:*** *High* | **Greedy:** They are slow and slow to change.<br>**Top-k**: Paleontologists study ancient animals.<br>**Contextual**: They are difficult to study. |
| Reporters and photographers watched. The big cat crept out of the cage. He looked around, and then ran off down a dirt road. The panther disappeared into the forest. **Wildlife workers were setting him free**.<br>***Contextual Specificity:*** *High* | **Greedy:** The news was not good.<br>**Top-k:** Then, it disappeared into the darkness.<br>**Contextual:** The news of the escape made people angry. |

Figure 4: Examples of generated elaborations with the different decoding strategies described in §3. Exs. 1–3 are cases where selecting a contextually-appropriate generated elaboration was effective. Ex. 4 is a relevant, sound elaboration with no content overlap with the gold elaboration, hence not rewarded by automatic metrics. Ex. 5 is a difficult case where context is essential – the generated elaboration is not pertinent to document context.

quality. Two annotators each annotated all 116 examples in our test set, resulting in 232 evaluations total. Table 3 shows these results. We calculate human agreement via Cohen's kappa with MASI distance (Passonneau, 2006), obtaining $\kappa = 0.51$, indicating moderate agreement (Artstein and Poesio, 2008). This round of evaluation confirmed that incorporating contextual specificity is helpful, consistent with our findings with BLEU.

## 5 Analysis and Discussion

We observe that GPT-2, finetuned on simplified text from the Newsela corpus, is able to adopt elaborative style (i.e short sentences of 7–13 tokens with limited vocabulary), see Figure 4. We find that the model can be effective at generating simple definitions and reasoning. However, the content contained in the elaborations is often not anchored in the document itself – generated sequences seem relevant to the snippet of context provided, but less so when placed in the larger document (see row 5 of Figure 4).

**Original Text.** We observe that our best model involves context only from the simplified document. We attribute the drop in performance of models with $C_o$ as a part of input largely to the crude incorporation of content from the original document, which is stylistically starkly different from simplified text, most notably in terms of length and vocabulary complexity. Since one of the main sources of relevant content during simplification is the original document, better methods to incorporate text or information from the original document is an important direction for future work.

| Low | Medium | High |
|---|---|---|
| Cushing died in a battle in the War of 1812. | He was captured and taken to a prison. | Cushing was a hero, his supporters said. |
| A government shutdown is when there are no government services. | A large minority of Germans thought American lawmakers behaved badly, said a poll released Tuesday. | Many lawmakers and their supporters blame the news coverage of their actions. |
| Football is the national and most popular sport in the United States. | In 2010, just 1 percent of its subscribers played fantasy sports. | The league is also popular with high school and college students looking to build a fan following. |

Figure 5: Example generated elaborations of varying contextual specificity.

**Effectiveness of contextual specificity.** Decoding with top-k sampling allowed GPT-2 to generate low, medium, and high contextualization sequences. A few examples of generated elaborations with varying contextual specificity that were conditioned on the same context are shown in Figure 5.

For most of our models, we do see an improvement when appropriately contextually specific sequences are chosen (rows 1–3 in Figure 4), suggesting the importance and need for further improvement of contextual specificity models.

While our methods take contextual specificity into account, they do not consider factuality or larger document relevance. An improved decoding scheme considering these could promote sequences that better align with larger document context.

**Retrieval.** Elaborations of medium to high contextual specificity often involve external knowledge not readily available from the simplified or original text. For example, generating factually correct details about a certain event or entity with little to no background on the event the document is referring to can prove challenging for pre-trained language models. To that end, generating truly effective elaborations of medium to high contextual specificity may require some type of retrieval module.

## 6 Related Work

Text simplification has been studied extensively (Siddharthan, 2014), especially at the sentence level. Recent progress has largely been driven by adapting monolingual translation for sentence simplification (Wubben et al., 2012; Wang et al., 2016; Xu et al., 2016; Zhang and Lapata, 2017; Dong et al., 2019; Kriz et al., 2019). This paradigm, while effective at transforming text, does not suffice when *new* content needs to be generated. A recent survey (Alva-Manchego et al., 2020) identifies explanation generation in simplification as an understudied area in dire need of new resources and methods. We tackle content addition, framed as explanation generation during simplification, and name it broadly as *elaborative simplification*.

The need for elaborative simplification is highlighted in prior hand-coded analysis (Yano et al., 1994), which showed that language learners and other audiences benefit from insertion of relevant elaborations and explanations, and that new or unfamiliar concepts negatively impact reading comprehension (Kintsch and Vipond, 1985). However, existing computational approaches are limited to the retrieval of definitions (Damay et al., 2006; Kandula et al., 2010; Eom et al., 2012; Paetzold and Specia, 2016), or constrained tasks such as post-modifier generation (Kang et al., 2019).

## 7 Conclusion

We presented the first data-driven study of elaborative simplification, i.e., content insertion during text simplification. We constructed a new corpus of 1.3K verified elaborations, observing a spectrum of contextual specificity and rich types of added content. We developed baselines for elaboration generation using pre-trained language models and found that considering contextual specificity could improve generation quality. We discussed some of the challenges of generating elaborations, and call for techniques to address elaborative simplification.

## Acknowledgments

We thank Greg Durrett for reviewing an early draft of this paper, and Joel Tetreault and the UT Austin Computational Linguistics group for valuable feedback and discussions. This work was partially supported by NSF Grant IIS-1850153. We also acknowledge the Texas Advanced Computing Center (TACC) at UT Austin for providing the computational resources for many of the results within this paper.

## References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of COLING 2012*, pages 545–560, Mumbai, India. The COLING 2012 Organizing Committee.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Jerwin Jan S Damay, Gerard Jaime D Lojico, Kimberly Amanda L Lu, D Tarantan, and E Ong. 2006. SIMTEXT: Text simplification of medical literature. In *Proceedings of the 3rd National Natural Language Processing Symposium-Building Language Tools and Resources*, pages 34–38.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Soojeong Eom, Markus Dickinson, and Rebecca Sachs. 2012. Sense-specific lexical information for reading assistance. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 316–325, Montréal, Canada. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, pages 366–370.

Jun Seok Kang, Robert Logan, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian. 2019. PoMo: Generating entity-specific post-modifiers in context. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 826–838, Minneapolis, Minnesota. Association for Computational Linguistics.

Walter Kintsch and Douglas Vipond. 1985. Reading comprehension and readability in educational practice and psychological theory. *Perspectives on learning and memory*, pages 329–365.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of AAAI*, pages 6610–6617.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3921–3927, Portorož, Slovenia. European Language Resources Association (ELRA).

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics

for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 79–83, Osaka, Japan. The COLING 2016 Organizing Committee.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Sheena Panthaplackel, Pengyu Nie, Milos Gligoric, Junyi Jessy Li, and Raymond Mooney. 2020. Learning to update natural language comments based on code changes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1853–1868, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

David Pellow and Maxine Eskenazi. 2014. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93, Gothenburg, Sweden. Association for Computational Linguistics.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Sara Botelho Silveira and António Branco. 2012. Enhancing multi-document summaries with sentence simplification. In *Proceedings of IJCAI*.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of AAAI*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yasukata Yano, Michael H Long, and Steven Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning*, 44(2):189–219.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

## A Contextual Specificity Prediction

We further explore the auxiliary task of contextual specificity prediction introduced in §3.2, prompted by the observation of diverse elaborations in our corpus. Formally, the task involves predicting the contextual specificity $s$ of an elaboration $E$ as low, medium, or high, given some document context $C$.

### A.1 Methods

As described in §3.2, we use BERT (Devlin et al., 2019) for this classification task. We do so in two settings based on surrounding text and/or the actual elaboration. Settings which include the elaboration can aid generation models by utilizing generated hypothesis elaborations and surrounding text to select sequences that are appropriately contextually specific. Settings that operate off context alone capture the *expected* level of specificity. In addition to the $E$-only model presented in §3.2, we explore combinations of $E$, $C_o$ (original document context) and $C_{4s}$ (4 sentences prior to the gold elaboration from the simplified document).

**With elaboration.** We feed the input sequence into BERT and use `[CLS]` token representation of the sequence, projecting it using a weight matrix $W \in \mathbb{R}^{dx3}$. Input sequences with the elaboration consist of `[CLS]` $C$ `[SEP]` $E$, where $C$ is either $C_o$ or $C_{4s}$, or both. When both types of context are used, we learn a representation for a separation token `[CONTEXT_SEP]` to distinguish between the two, and use $C = C_o$`[CONTEXT_SEP]`$C_{4s}$.

**Context only.** While contextual specificity clearly involves the elaboration itself, context-only models help us understand whether it is predictable from context alone, and simulate a realistic setting during simplification, when these models may be incorporated before the actual elaborative text is generated. Input to these models is crafted similarly, but excluding $E$ from the sequence.

### A.2 Experiments and Analysis

We train on $(E, s)$ pairs, and utilize `bert-base` from the HuggingFace Transformers library. We feed the sequence representation from the `[CLS]` token embedding into an output layer for classification [5]. For each setting, we train for 5 epochs, using a batch size of 32, and a learning rate of 2e-3. We use the default dropout rate of 0.1 for self-attention layers, but refrain from adding dropout on our linear layer.

**Results.** We use the same four metrics to evaluate our results – two classification metrics (accuracy, macro-averaged F1), and two regression metrics (Spearman's correlation and mean absolute error), and we again report mean performance over 15 different, randomly initialized runs. Results are shown in Table 4, and suggest that this is a challenging task, even for powerful pre-trained language models. The best predictor of contextual specificity, in terms of correlation and MAE, is context in the form of 4 sentences before the elaboration combined with the elaboration itself. However, the elaboration-only model performs the best in terms of accuracy and F1.

**Original Text Presence.** In all settings in which the aligned snippet of text from the original document was fed in as partial or complete input to the model, we see a reduction in performance. Compared to text from the simplified document, text from the original document is stylistically distinct. Consequently, when jointly fed in as context with simplified text, the input is largely incoherent, potentially impacting the model. We leave studying more effective ways of incorporating context from the original document to future work.

**Qualitative Analysis.** In cases where linguistic cues explicitly indicate the level of contextual specificity, our model performs well—i.e when definitions are inserted as *"A is B"* or reasoning is inserted as *"A but B"* or *"The reason for A is B"*. However, predicting the contextual specificity of more nuanced sentences may require an improved method of modeling surrounding context. For example, when the elaboration contains a definition of a term from a different sentence using coreferent mentions, our model predicts a higher level of contextual specificity. In general, our model over-predicts highly contextualized elaborations, and under-predicts lower levels of contextual specificity. Medium contextual specificity was hardest for our models to predict accurately.

**Amount of context.** To understand the impact of the amount of context on performance, we vary the number of sentences ($\{2, 4, 6\}$) before the elaboration to feed into our best performing model involving context ($C_s + E$). Table 5 shows these results. We see that merely increasing the amount of con-

[5] We tried finetuning our contextual specificity prediction models on our elaboration dataset, but found that our dataset was too small to yield stable results.

| | Context | Acc. | F1 | Correlation | MAE |
|---|---|---|---|---|---|
| *Context Only* | $C_o + C_{4s}$ | $45.2 \pm 3.0$ | $43.1 \pm 2.8$ | $27.8 \pm 4.9$ | $0.729 \pm 0.05$ |
| | $C_{4s}$ | $46.4 \pm 2.9$ | $44.9 \pm 3.0$ | $32.4 \pm 4.4$ | $0.679 \pm 0.04$ |
| | $C_o$ | $37.9 \pm 4.6$ | $36.1 \pm 5.4$ | $20.2 \pm 1.0$ | $0.813 \pm 0.07$ |
| *With Elaboration* | $E$ | $\mathbf{56.8 \pm 1.5}$ | $\mathbf{55.3 \pm 1.6}$ | $47.5 \pm 2.6$ | $0.552 \pm 0.01$ |
| | $C_o + C_{4s} + E$ | $50.5 \pm 3.8$ | $48.3 \pm 4.0$ | $40.4 \pm 5.8$ | $0.628 \pm 0.05$ |
| | $C_{4s} + E$ | $55.3 \pm 3.3$ | $54.0 \pm 2.5$ | $\mathbf{50.8 \pm 4.1}$ | $\mathbf{0.545 \pm 0.03}$ |
| | $C_o + E$ | $43.7 \pm 1.8$ | $41.7 \pm 2.0$ | $26.7 \pm 3.6$ | $0.749 \pm 0.03$ |

Table 4: Contextual Specificity Prediction results, including accuracy, macro-averaged F1, Spearman's correlation, and Mean Absolute Error, reported across 15 runs. We bold our best results. The performance differences between (1) $C_{4s} + E$ vs $E$, (2) $C_o + C_{4s}$ vs $C_{4s}$, and (3) $C_o + C_{4s} + E$ vs. $C_{4s} + E$ are not statistically significant.

| | Acc. | F1 | Corr | MAE |
|---|---|---|---|---|
| $C_{2s}$ | $53.6 \pm 1.8$ | $51.2 \pm 3.1$ | $\mathbf{51.7 \pm 6.8}$ | $0.566 \pm 0.05$ |
| $C_{4s}$ | $\mathbf{55.3 \pm 3.3}$ | $\mathbf{54.0 \pm 2.5}$ | $50.8 \pm 4.1$ | $\mathbf{0.545 \pm 0.03}$ |
| $C_{6s}$ | $53.9 \pm 4.0$ | $52.0 \pm 3.6$ | $44.3 \pm 5.5$ | $0.591 \pm 0.04$ |

Table 5: Mean performance of $C_s + E$ model over 15 runs with varying amounts of pre-elaboration context.

text fed to the model does not translate to stronger results – considering overall performance, 4 sentences before the elaboration from the simplified document performed best.

## B Elaboration Generation

### B.1 GPT-2 Finetuning

We explore generation with GPT-2 across varying finetuning settings – (1) zero shot (no finetuning, only relying on GPT-2's pre-training), (2) finetuning on the set of simplified documents in the Newsela corpus (excluding documents from the test set), and (3) finally on our elaboration corpus. We utilize the same 3 decoding schemes described in § 3.2 across these different finetuning settings. We used a temperature of $t = 0.7$ for the zero shot setting, and $t = 0.45$ for finetuned settings. For finetuning on our elaboration corpus, we trained for 3 epochs with a batch size of 8 and a learning rate of 1e-3. We report BLEU-1 and BLEU-2 as described in § 4.1. As BLEU metrics for setting 2 are already included in Table 2, we report metrics for zero-shot generation (Table 6), and for generation after finetuning on our elaboration corpus (Table 7). Comparatively, finetuning GPT-2 on the set of simplified Newsela documents yielded the best performance, and we attribute this to there being strictly more data in that setting as opposed to our corpus of verified elaborations.

| **Pre-trained GPT-2** | Greedy | | Top-k | | Contextual | |
|---|---|---|---|---|---|---|
| Context | B-1 | B-2 | B-1 | B-2 | B-1 | B-2 |
| $C_{2s}$ | 12.48 | 2.71 | 9.82 | 2.04 | 11.93 | 2.66 |
| $C_{2s} + C_o$ | 12.21 | 2.58 | 9.80 | 2.08 | 10.86 | 2.82 |
| $C_{4s}$ | 13.46 | 3.35 | 11.78 | 2.43 | 13.80 | 3.89 |

Table 6: BLEU-1 and BLEU-2 for generation after finetuning on our elaboration corpus.

| **Fine-tuned GPT-2: Elaboration Corpus** | Greedy | | Top-k | | Contextual | |
|---|---|---|---|---|---|---|
| Context | B-1 | B-2 | B-1 | B-2 | B-1 | B-2 |
| $C_{2s}$ | 20.9 | 6.82 | 19.11 | 5.32 | 19.38 | 5.47 |
| $C_{2s} + C_o$ | 11.89 | 2.78 | 12.72 | 2.77 | 14.2 | 3.05 |
| $C_{4s}$ | 20.17 | 5.87 | 16.89 | 4.09 | 18.97 | 5.16 |

Table 7: BLEU-1 and BLEU-2 for the zero-shot generation setting.

### B.2 Generation with BART

In addition to GPT-2, we experimented with BART (Lewis et al., 2020), a pre-trained sequence to sequence model. The encoder-decoder nature of BART allows us to explore elaborative simplification as a post-processing/post-editing scenario, where the model can receive context both preceding *and* following the elaboration in the simplified text.

We finetune `bart-base` available via the HuggingFace Transformers library, and feed in four different types of context (1) $C_{2s}$, (2) $C_{4s}$, (3) $C_{2s+}$, (4) $C_{4s+}$. The latter two context settings utilize two and four sentences before and after the elaboration (without the elaboration itself). In all settings, the gold elaboration was the target. We finetune for 3 epochs, with a batch size of 2, and a learning rate of 1e-4, and generate elaborations via greedy decoding. Results are shown in Table 8.

We find that BART is able to adopt elaborative

| | $C_{2s}$ | $C_{2s+}$ | $C_{4s}$ | $C_{4s+}$ |
|---|---|---|---|---|
| B-1 | 18.9 | 21.5 | 20.2 | 20.1 |
| B-2 | 5.05 | 6.68 | 6.02 | 6.18 |

Table 8: BLEU-1 and BLEU-2 for greedy generation with BART.

style, generating short sequences with limited vocabulary, however we observe that the smaller size of our corpus affected BART's ability to generate coherent, diverse elaborations. In addition, we note that framing elaborative simplification as a post-processing task is a more difficult, nuanced setting – the generated elaboration to be inserted must maintain the flow of the text and blend with the content present subsequent sentences. Elaborative simplification in this setting is another interesting, rich direction for future work.