Text Simplification for Leadless Pacemaker Failure Reports in MAUDE

Risako Owan
owan0002@umn.edu

Ariel Roane
roane010@umn.edu

Hao Zou
zou0080@umn.edu

1 Introduction

Text Simplification (TS) is an area of Natural Language Processing (NLP) that relies on both Lexical Simplification (e.g., "an octogenarian" - "a 80-89 year old person") and Syntactic/Structural Simplification (i.e., breaking sentences into simpler constituent parts) to simplify complex documents. TS may rely on other related tasks as well, such as Text Summarization and Concept Simplification, to remove information that is deemed unnecessary and make the semantic content of documents more understandable to a general audience. The simplified sentences outputted by TS methods or models can be used to communicate complex, domain-specific language to a wider general language audience. Within the clinical or biomedical domains, this task is complicated by various phrases and terminology not common in general-domain language. For this reason, neural NLP models trained on general-domain text cannot achieve high accuracy when run on biomedical text without any fine-tuning.

Device-related adverse effects are an area of scrutiny for regulatory bodies across the global market. In order to increase transparency to potential product issues and facilitate product surveillance, the Food and Drug Administration (FDA) in the United States hosts the Manufacturer and User Facility Device Experience (MAUDE) database to store information regarding both mandatory and optional reports on malfunctions, serious injuries, and deaths resulting from the use of a medical device. Patients are able to freely access this database, but may be deterred by the complex lexical and syntactic elements from the medical domain present in the database, along with textual artifacts inherent in the unstructured data stored in MAUDE. However, to date, no TS methods have been applied to the MAUDE dataset to increase accessibility for patients or other non-specialized interested parties.

In order to create simplified records for layperson consumption, we explored various strategies that may be used in TS on MAUDE records. We focused on adverse effect reports for a relatively new medical technology- the leadless pacemaker - which is a hermetically sealed electrical pulse generator that is used to pace the heart under anti-bradycardia medical indications, such as in cases of Sick Sinus Syndrome and Atrioventricular Block/Conduction Disease. Several neural models were trained, including the previous state of the art sequence-to-sequence Neural Text Simplification (NTS) model [7]. Our training data only contained reports for leadless pacemakers so that we did not need to consider report format differences between various devices, but in principle our methods could also be applied to other devices represented in the MAUDE database. To our knowledge this is the first attempt at TS for unstructured malfunction, serious injury, or death reports in MAUDE using neural models trained on complex and simple medical English corpora.

2 Existing Work

Work published on TS applications has greatly increased over the last decade with a focus on neural methods and the increasing availability of large General English corpora. The field has undergone changes throughout its developmental history from rule-based approaches to automatic simplification methods using deep learning models and hybrid approaches. [1] Biomedical TS was introduced as an area of research in 1989 by Rada et al. [12] and it mirrored these trends in general English TS.

However, applications in the clinical or biomedical domains are still limited by the lack of appropriate training data. Several approaches have developed to overcome this limitation, including automatic alignment of complex-simple medical language [13] and leveraging of documents from social media [11]. However, a common problem with these approaches is their limited ability to train models that can transfer to other clinical or biomedical domains that may be more specialized than the original training dataset. Thus the gold standard training dataset for TS in the clinical or biomedical domain continues to be manually simplified text by experts in those domains. This can be prohibitively expensive in some cases, which may result in models trained on General English but tested on medical language. Performance may suffer with this type of transfer learning, depending on how specialized the medical language in the test set may be.

Regarding our current task, related work on TS include neural methods that use recurrent neural network (RNN) architecture or attention mechanisms (Transformers) for sequence-to-sequence (seq2seq) complex to simple translation. Primary related work includes the Open Neural Machine Translation (OpenNMT) framework-implemented Neural Text Simplification (NTS) model [7] trained on English Wikipedia (EW) / Simple English Wikipedia (SEW) and a Fairseq framework implementation [10].

3 Dataset

Since there existed no readily made regular English to simplified English pairings for the electrophysiology (EP) domain relevant to the leadless pacemaker device, we created our own dataset for training the OpenNMT model and providing a reference simplified record set for model evaluation. We first extracted a subset of the MAUDE dataset for leadless pacemakers (FDA code "PNJ") from 01-JAN-2015 to 31-DEC-2020. Then we collected the following datasets used in other TS model training data:

- ◇ A corpus containing automatically aligned sentences from Wikipedia and the corresponding Simple Wikipedia sentences (585354 rows) [3]
- ◇ A corpus containing automatically aligned medical sentences from Wikipedia and Simple Wikipedia (3390 rows) [13]
- ◇ Medical Paper-Blog Dataset (2973 rows) [11]

To supplement the datasets above and to provide references for the BLEU and SARI evaluation metrics, we manually created a simplified subset of the MAUDE dataset for 67 rows. Fifty-seven (57) rows of this dataset were appended to the training dataset containing the datasets listed above. The remaining 10 rows of the manually simplified MAUDE dataset served as our test data. (See Appendix A.1 for examples.) Manual MAUDE TS was processed by A. Roane who had familiarity with the textual data through experience in post market surveillance for medical devices.

Preprocessing included the removal of abbreviations and acronyms, as the MAUDE records contained the corresponding word before the abbreviation or acronym. In addition, citations, symbols, and report artifacts were removed during the preprocessing step. Report artifacts included tokens that were applied to the original record to mask proprietary or confidential information such as device unique identifiers or patient information. Because the MAUDE dataset adheres to the Freedom of Information Act (5 U.S.C. § 552), these parts of the record were removed before the record was entered into the MAUDE database. Numbers enclosed in parentheses could be ignored since they were usually used in references or to represent numbers, such as in "one (1) person". Finally, all text was converted to lowercase before being input to the models.

Although abbreviations and acronyms were removed from our initial training datasets, we also noted in our initial analysis of the dataset that abbreviation and acronym mappings were not always straightforward due to lexical ambiguity. For example, *PVC* was described in the records as "periventricular" contraction or premature ventricular contraction, and making the acronym plural can change it to *PVCS*. In addition we observed lexical variability as well. For instance, one entry was found listing "premature ventricular contractions" as *PV CS*.

4 Methods

In our project, we used two neural seq2seq models, the NTS model[7] and the Controllable Sentence Simplification model[5]. The first model used the OpenNMT toolkit[4] and the latter used the Fairseq toolkit[10]. Both toolkits are open-source and highly extensible, and both models were chosen because of their functioning Github repositories. These models are commonly used for TS tasks due to their features and are similar in that they both use encoder/decoders, attention mechanisms, and they both employ early stopping. (See Appendix B.2 for details on their differences.)

4.1 Neural Text Simplification model (OpenNMT)

In our OpenNMT model[7], we fine-tuned the model on the combined training dataset described above and tested the model on the remaining 10 rows from the MAUDE dataset, which we had manually simplified entries for. The training was done using the default parameter values in the train.lua file of the Lua Torch library.

4.2 Controllable Sentence Simplification model (Fairseq)

The pretrained FairSeq model was trained on a Transformer with the EW-SEW dataset for Text Simplification purpose. This model aims at simplifying grammar and structure while keeping the underlying information as identical as possible. Unfortunately, computational resource limitations prevented us from being able to fine-tune the model, so we instead did our testing and analysis on the pretrained model. Apart from using the basic Fairseq toolkit, the model adapts a training loss to control the lexical complexity. The training loss was calculated using word frequency and the sentence Flesch-Kincaid Grade Level score. (See Appendix B.2.1 for details on the loss function.)[6]

5 Results

We used BLEU, SARI, and the Flesch-Kincaid Reading Ease score for our evaluations. BLEU generally correlates with meaning and grammar, while SARI correlates with simplicity. The Flesch-Kincaid score is a metric for understanding how difficult some text in English is to understand and the higher the number, the easier to read.

	BLEU	SARI	Change in Avg Flesch-Kincaid Reading Ease
OpenNMT (Epoch 8)	8.6	36.2	163.6
Fairseq	5.4	39.1	88.6

Figure 1: Final Results

6 Discussion

We were surprised to see how much our base models struggled with simplifying the MAUDE records when trained on our training datasets. This may be due to the fact that most of the training data came from Wikipedia, meaning most of the text was not in the biomedical domain. Words can have different meanings and relations with each other based on domains: for example "lead" in "leadless implantable pulse generator" refers to insulated wires used in conventional pacemakers. However, this word is more commonly used as a verb or when referring to pencils or toxic metals in the general domain.

We attempted to overcome this problem of training data sets through manual conversion of a subset of 67 MAUDE records to simplified records. However, no major gains in performance were observed after fine-tuning the OpenNMT NTS model using these simplified records. Lack of suitable training data sets in the clinical or biomedical domains is a well-known problem in the use of these systems for NLP tasks. In a sense, this projects represents an application of Transfer Learning in regard to using models trained primarily on EW-SEW complex/simple sentence pairs. The ways in which words can relate to each other within a sentence can differ greatly across domains, so that even if

there are many shared common words, that doesn't necessarily mean that the information gained from the training data will be transferable to the test set.

Text simplification is a form of paraphrasing and this task can sometimes bring in unverified information and/or biases. For example, our FairSeq base model changed the phrase, "It was reported...", to "He said that...". Not only does the text itself give no reference about the gender of the doctor treating the patient, but it also brings in widespread existing gender biases[2]. FairSeq's method of simplification raises the question of what kind of simplification processes are necessary in order to make medical text more public friendly, including reflecting the information in a way that is most likely to communicate the desired meaning of the text.

6.1 Simplified Sentences in the Medical Domain

One interesting thing to note is that, while general English text simplification often involves shortening sentences, those of the medical domain often require longer sentences. In regard to TS in the clinical or biomedical domain, longer simplified sentences are expected due to lexical simplification of highly specialized medical jargon.* For instance, the adverse effect of "angina" may be translated to "pain in the chest", which represents a shift from one (1) to four words (4), increasing the length of the original sentence. However, the sentence outputs of our models were significantly shortened. This is likely due to the training data set, which was mainly representative of General English Language, and not a specific clinical domain, such as electrophysiology (EP). Thus words encountered outside of the vocabulary of the models was either not substituted in the simplified sentence (e.g., "angina" -> "angina"), or removed from the output sentence. This could be overcome with a more appropriate training data set specific to the EP domain. This would also potentially be useful in simplifying "medical metaphors", such as "the delivery system showed a goose-neck", to more simplified phrases, such as "the delivery system had a very curved shape/u-shape".

6.2 Appropriateness of Evaluation Metrics

Both BLEU and SARI utilize a reference set of sentences to evaluate the quality of a simplified translation. Translations that contain greater frequency of n-grams from the reference sentences receive higher scores. SARI also takes into account the original sentence overap with the simplified sentence. Our results showed low BLEU scores signifying that the n-grams generated by the manually simplified sentences in the reference set did not share a significant amount of n-grams in common with the automatically simplified sentences. Flesch-Kincaid Reading Ease (FKRE) is the linear combination of the number of words, sentences, and syllables in a record and can range from 121.22 to -∞, with maximum readability at 121.22. Thus FKRE is biased to several changes in word, sentence, and syllable counts in the target record that may not be representative of a true simplification.

7 Limitation and Challenges

7.1 Reproducibility

Reproducibility issues were encountered in implementing the OpenNMT NTS model. These issues centered around running the original Torch model, written in Lua, compared to a python implementation (Pytorch). As the Nisioi et al. (2017) model utilized Lua Torch, version compatibility issues were encountered between the latest version of CUDA and Torch, as well as the lack of online support related to the older versions of CUDA and Torch. Because Torch is no longer updated compared to Pytorch, much of the troubleshooting information was related to Pytorch only, resulting in less available support. In addition, because the model needed significant computing resources, a University of Minnesota lab computer was used. This prevented freely changing versions of the required packages. The solution was to update the CUDA package on the shared resource after checking with other users of that resource, but this is a limitation of the current work that should be addressed in the future.

*We would like to extend a thanks to Professor Pakomov for pointing this out to us.

7.2 Challenge of Finding and Creating Appropriate Datasets

Creation of an appropriate dataset for training the OpenNMT and Fairseq models was difficult. Various methods have been developed to automatically align candidate complex-simple datasets to create training data in the clinical or biomedical domains. Opportunities for dataset creation have previously leveraged Wikipedia and social media accounts of medical content as previously mentioned. Although there are layperson readable documents created for patient consumption- such as Informed Consent Forms used to communicate risk in clinical trials, available patient-language glossaries, or publicly available website on medical phenomenon- there are no available datasets that link these documents to used complex language. Creating automatic methods to link these available General English documents with corresponding language used in a specific clinical or biomedical domain may help to address the training dataset challenge.

7.3 Metaphors in Medical Text

Another issue encountered in this work was the use of domain-specific metaphors represented in the MAUDE record. For instance, the following sentence was used to describe the shape of a device delivery catheter: *The Delivery System showed a 'Gooseneck' and the physician commented the top felt softer than before.* Within the context of a MAUDE record, this description may be important as it could describe a fault-state with the device that could lead to patient harm. However, a layperson may have the same difficulty understanding the metaphor, as well as its importance, as they do understanding a complex medical term. TS models that attempt to simplify this language need target their audience carefully when producing the target output. If the audience is a layperson, it may be better to describe the importance or implications of what the shape may mean rather than just the shape itself. If the audience of the TS output is the device manufacturer, it may be better to describe the shape so that the complaint could be factored into risk analysis of what the shape may mean for potential device faults/hazards or fault conditions. Word sense disambiguation could potentially be applied in this scenario as well to clarify the sense of the metaphor within the record.

8 Future Work

8.1 Grammar Checking

Once we are able to improve our text simplification process, the next step would be to make sure it is grammatically correct. Deep learning models are fundamentally statistical and we cannot use the output as is in professional settings without some additional checks and edits. Gector is a grammar error model developed by Grammarly; it uses a sequence tagging approach with a pretrained Transformer encoder. We believe that this grammar error correction model should be paired with our text simplification model in future stages. [9]

8.2 Lexical Simplification

Lexical Simplification for complex medical terms is another strategy that may improve the performance of our models. After we get the NTS model outputs, we can further simplify these documents through clinical or biomedical Complex Word Identification (CWI), then select the best candidates to replace them. Utilizing CWI we can both increase the readability for laypeople and decrease the complex level for individual words.

9 Conclusion

The purpose of our project was to simplify adverse event records stored in MAUDE for better lay-patient understanding of risks that are present during medical device procedures. We explored the OpenNMT toolkit-implemented NTS model and the Fairseq-implemented Controllable Sentence Simplification model to simplify textual records in the MAUDE database. In addition, we created a training dataset by combining the EW-SEW dataset, the Medical Paper-Blog dataset, and a manually simplified MAUDE dataset. A more specialized dataset was likely required for better performance in our task and the appropriateness of our evaluation methods should be revisited. In conclusion, this project highlighted the unique difficulties associated with running TS tasks on biomedical datasets.

References

[AS14] M.A. Angrosh and A. Siddharthan. "Text simplification using synchronous dependency grammars: Generalising automatically harvested rules". In: 2014, pp. 16–25.

[Bel+21] Deborah Belle et al. ""I Can't Operate, that Boy Is my Son!": Gender Schemas and a Classic Riddle". In: Sex Roles, 2021. URL: https://doi.org/10.1007/s11199-020-01211-4.

[Hwa+15] William Hwang et al. "Aligning sentences from standard wikipedia to simple wikipedia". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 211–217.

[Kle+17] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 67–72. URL: https://www.aclweb.org/anthology/P17-4012.

[Mar+20a] Louis Martin et al. "Controllable Sentence Simplification". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille: European Language Resources Association (ELRA), May 2020, pp. 4689–4698. URL: https://arxiv.org/pdf/1910.02677.pdf.

[Mar+20b] Louis Martin et al. *Controllable Sentence Simplification*. 2020. arXiv: 1910.02677 [cs.CL].

[Nis+17] Sergiu Nisioi et al. "Exploring Neural Text Simplification Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 85–91. DOI: 10.18653/v1/P17-2014. URL: https://www.aclweb.org/anthology/P17-2014.

[NKA19] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. "Controllable Text Simplification with Lexical Constraint Loss". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 260–266. DOI: 10.18653/v1/P19-2036. URL: https://www.aclweb.org/anthology/P19-2036.

[Ome+20] Kostiantyn Omelianchuk et al. "GECToR – Grammatical Error Correction: Tag, Not Rewrite". In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA â†' Online: Association for Computational Linguistics, July 2020, pp. 163–170. URL: https://www.aclweb.org/anthology/2020.bea-1.16.

[Ott+19] Myle Ott et al. *fairseq: A Fast, Extensible Toolkit for Sequence Modeling*. 2019. arXiv: 1904.01038 [cs.CL].

[Pat+20] Nikhil Pattisapu et al. "Leveraging Social Media for Medical Text Simplification". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1141–1144.

[Rad+89] Roy Rada et al. "Development and application of a metric on semantic nets". In: *IEEE transactions on systems, man, and cybernetics* 19 (1989), pp. 17–30.

[VSL19] L. Van Den Bercken, R.-J. Sips, and C. Lofi. "Evaluating neural text simplification in the medical domain". In: 2019, pp. 3286–3292. DOI: 10.1145/3308558.3313630.

Appendix

A Dataset examples

A.1 Training dataset

Medical EW-SEW	Original	lymph node is small ball shaped organ of the immune system distributed widely throughout the body including the armpit and stomach gut and linked by lymphatic vessels lymph nodes are garrisons of and other immune cells
Medical EW-SEW	Simplified	lymph node is an organ consisting of many types of cells and is part of the lymphatic system
Medical Paper-Blog	Original	the human body can adapt to high altitude by breathing faster , having a higher heart rate , and adjusting its blood chemistry .
Medical Paper-Blog	Simplified	the human body can deal with high altitude by breathing faster , having a higher heart rate , and changing the blood itself to have more red blood cells that can carry oxygen .
Manual MAUDE	Original	it was reported that the patient experienced extra-cardiac stimulation. the leadless implantable pulse generator was inactivated and a new ipg implanted. no further patient complications have been reported as a result of this event. additional manufacturer narrative if information is provided in the future, a supplemental report will be issued.
Manual MAUDE	Simplified	It was reported that the patient experienced stimulation outside of the heart. The device was turned off and a new device implanted. No additional complications have been reported. If more information is given in the future, an add-on report will be made.

Table 1: Training Dataset Examples

B Model Architecture and Details

B.1 OpenNMT model

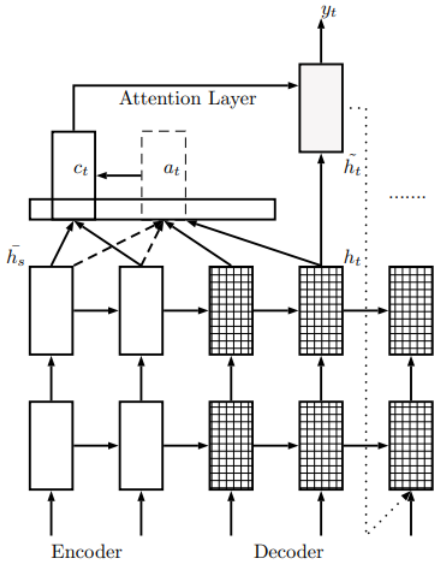

Figure 2: OpenNMT model: Architecture of the neural simplification model with global attention and input feeding. [7]

B.2 Fairseq model

B.2.1 Loss Function with Word Level

To control the lexical complexity, this model weighed a training loss of a text simplification model considering words that frequently appear in the sentences of a specific grade level. Different sentences have different word grade level, Figure 2 shows examples for that. [8]

Grade	Examples
12	According to the Pentagon , 152 female troops have been killed while serving in Iraq and Afghanistan .
7	The Pentagon says 152 female troops have been killed while serving in Iraq and Afghanistan .
5	The military says 152 female have died .

Table 2:

The weight $f(w, l)$ was used to represent the relevance of the word w at grade level l. Cross-entry loss was used for this sequence-to-sequence model, when the model outputs $o = [o_1, ..., o_N]$ (N is the size of the vocabulary) at a certain time step, the loss is as follows:

$$L(o, y) = -ylogo^T = -logo_c[6]$$

where $y = [y_1, ..., y_N]$ is a one-hot vector. The model adds weights to the loss function based on the level of words. As for $f(., .)$, TFIDF and PPMI were used and held the assumption that words frequently appeared in sentences of level l have the same level. TFIDF was computed regarding sentences of the same level as a document. Pointwise mutual information (PMI) estimated the strength of a cooccurrence between word and grade level. However, words with negative PMI scores have a negative correlation against grade level, hence, the words with a negative PMI were ignored and positive-PMI (PPMI) function was used.

TFIDF:

$$TFIDF(w, l) = P(w|l) * log\frac{D}{DF(w)}[6]$$

where $P(w|l)$ is the probability that word w appears in the sentences of grad level l. D represents the number of grade levels. $DF(w)$ means the number of grade levels that the word w appears. [6]

PPMI:

$$PMI(w, l) = log\frac{P(w|l)}{P(w)}$$

$$PPMI(w, l) = MAX(PMI(w, l), 0)[6]$$

where $P(w)$ is the probability of word w appearing in the whole training sentences. $P(w|l)$ is the probability that word w appears in the sentences of grad level l. [6]

B.3 Comparision of Model Architecture

	OpenNMT	Fairseq
Torch	Lua	Pytorch
Architecture	2 LSTM layers with input feeding	Base Transformer Architecture
Attention	Global	Self
Embedding Dimension	500	512
Vocab size	50K	10K
Vocab size	50K	10K
Beam Size	5-12	8
Dropout	0.3	0.2
Optimizer	SGD	Adam

Table 3: Differences between our OpenNMT model and Fairseq model

C Model Results

C.1 Text Simplification Examples

Original:
it was reported that the patient experienced extra-cardiac stimulation. the leadless implantable pulse generator was inactivated and a new ipg implanted. no further patient complications have been reported as a result of this event. additional manufacturer narrative if information is provided in the future, a supplemental report will be issued

Source	Simplified
OpenNMT	it was reported that the patient experienced extra-cardiac the leadless implantable pulse and a new ipg implanted. no further patient complications have been reported as a result of this event. additional manufacturer narrative
Fairseq	He said that the patient experienced extra-cardiac stimulation , the leadless pulse generator was inactivated and a new ipg implanted , but no more patient complication complications have been reported as a result of this event , and more manufacturer narrative if information is provided in the future
Manual Simplification	It was reported that the patient experienced stimulation outside of the heart. The device was turned off and a new device implanted. No additional complications have been reported. If more information is given in the future, an add-on report will be made.

Figure 3:

C.2 OpenNMT Evaluation Scores During Training

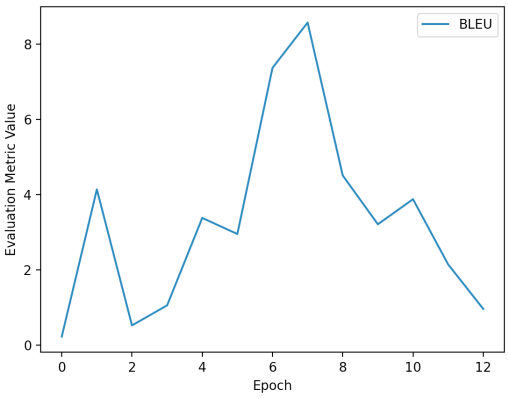

Figure 4: Bleu Scores

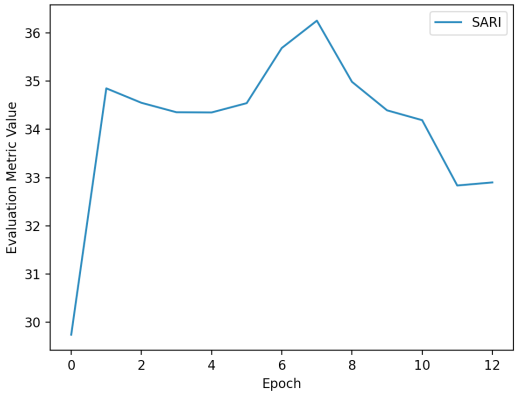

Figure 5: SARI Scores

D Github link

https://github.umn.edu/OWAN0002/HINF5610

E Member contributions (names in alphabetical order)

Everyone

- ◇ Conducted literature reviews
- ◇ Participated in once/twice-a-week meetings
- ◇ Explored MAUDE data
- ◇ Brainstormed project plan
- ◇ Found and added existing training datasets (Simplified Wikipedia (General and Medical) and the Medical Blog dataset)
- ◇ Ran the model on test data for each epoch
- ◇ Worked on the proposal, paper, and presentation
- ◇ Observed and analyzed results

Risako Owan

- ◇ Explored, made adjustments, and retrained Neural Text Simplification (OpenNMT) Lua model
- ◇ (Unused) Explored Gector (A grammatical error correction model by Grammarly) - could not be run within the 24-hour limit on the v100 queue in MSI.
- ◇ Participated in Q&A

Ariel Roane

- ◇ Organized literature review content in Excel sheet
- ◇ Provided insight on MAUDE dataset uses and contents
- ◇ Manually simplified MAUDE dataset subset
- ◇ Created script to pull data from MAUDE
- ◇ Created Flesch-Kincaid script
- ◇ Participated in Q&A

Hao Zou

- ◇ Explored Controllable Sentence Simplification (Fairseq) model
- ◇ Ran the model on test data
- ◇ (Unused) Explored Neural Text Simplification Python model