Complexity-Weighted Loss and Diverse Reranking for Sentence Simplification: Supplemental Material

Reno Kriz*, **João Sedoc***, **Marianna Apidianaki**△,
Carolina Zheng*, **Gaurav Kumar***, **Eleni Miltsakaki**†,
and Chris Callison-Burch*
* Computer and Information Science Department, University of Pennsylvania
△ LIMSI, CNRS, Université Paris-Saclay, 91403 Orsay
† Choosito, Inc.
{rekriz,joao,gauku,carzheng,ccb}@seas.upenn.edu,
marianna@limsi.fr, eleni@choosito.com

1 Complexity Prediction Models

For both our word complexity and sentence complexity prediction models, beyond calculating the overall mean squared error (MSE), we also calculated MSE by complexity level. In other words, we determine how well our models predict the complexity level for all words/sentences labeled as level i, where $0 \leq i \leq 4$. Note that 4 represents the most complex level, while 0 represents the simplest level. The results are reported in Table 2. This shows that our models also achieve a more balanced performance across levels

2 Training Details

In this section, we show a comprehensive list of all hyperparameters we used when training our default Seq2Seq model, to allow our code to be reproducible by others.[1] This list includes learning rate (LR), learning rate reduction rate (LR reduce), size of embeddings (Embeddings), loss function (loss, we use CE to represent Cross Entropy) among others. These parameters are found in Table 1.

For our extensions to the standard Seq2Seq framework, we use nearly all the same parameters, the only exception being that we change the loss function from cross entropy to our custom complexity-weighted loss function. With this loss, we use $\alpha = 2$ during training. At inference time, we set the beam size $b = 100$, and the similarity penalty $d = 1.0$. After inference, we set the number of clusters to 20, and we compare two separate reranking weightings: one which uses fluency, adequacy, and simplicity (FAS), where $\beta_f = \beta_a = \beta_s = \frac{1}{3}$; and one which only uses fluency and adequacy (FA), where $\beta_f = \beta_a = \frac{1}{2}$ and β_s = 0. Note that our best model uses FA weights.

Parameter	Value	Parameter	Value
Batch size	86	Embeddings	300:300
RNN hidden units	256	LR	0.001
RNN attention	dot	LR reduce	0.7
# of layers	2	Loss	CE
RNN type	LSTM	Min Epochs	1
Dropout inputs	0.2	Max epochs	30
Dropout states	0.2	Max updates	500000
Min vocab freq	3	# Last params	5
Max length	85	Optimizer	Adam
Label smoothing	0	Seed	13

Table 1: Training Hyperparameters for the baseline Seq2Seq model and our extended model.

3 Human Evaluation

In this paper, we ran two different human evaluations to accurately compare our model with other state-of-the-art systems. In our first task, we ask native English speakers on Amazon Mechanical Turk to evaluate the fluency, adequacy, and simplicity of sentences generated by our systems and the baselines; for this task, we model our instructions after that of Zhang and Lapata (2017). Our full instructions are found in Figure 2. We also provide the results of this experiment with additional confidence intervals in Table 3.

For our second task, we run several direct pairwise evaluations. This is inspired by ChatEval, a standardized human evaluation system for pairwise comparisons of chatbots (Sedoc et al., 2018). In this task, we provide the original complex sentence and two simplifications, and ask annotators which sentence is the better simplification. Our instructions for this task are found in Figure 1.[2]

4 Error Analysis

In our paper, we discuss six categories of errors. We now present examples for each category, to

[1] Note that we will also release our code upon publication.

[2] We will release the HTML templates for our human evaluations upon publication.

Task	Model	Correlation	Mean Squared Error					
			Overall	0	1	2	3	4
Word Complexity	Frequency	-0.031	1.9	2.24	0.38	0.64	3.01	6.99
	Length	0.344	1.51	1.03	**0.32**	0.89	2.95	6.90
	LinReg	**0.659**	**0.92**	**0.92**	0.39	**0.49**	**1.17**	**3.27**
Sentence Complexity	Length	0.503	3.72	**0.24**	**0.25**	1.83	4.77	9.12
	CNN	**0.650**	**1.13**	1.78	0.64	**0.43**	**0.71**	**2.03**

Table 2: Pearson Correlation, Overall Mean Squared Error (MSE), and MSE by complexity level for both our word-level and sentence-level complexity prediction models. We also compare to length-based and frequency-based baselines.

Model	Fluency	Adequacy	Simplicity	All
Hybrid	2.79** (± 0.08)	2.76 (± 0.08)	2.88** (± 0.08)	2.81 (± 0.07)
DRESS	**3.50** (± 0.08)	**3.11**** (± 0.08)	3.03 (± 0.08)	**3.21** (± 0.07)
DMASS	2.59** (± 0.08)	2.15** (± 0.08)	2.50** (± 0.08)	2.41 (± 0.07)
S2S-All-FAS	3.35 (± 0.08)	2.50** (± 0.08)	**3.11** (± 0.08)	2.99 (± 0.07)
S2S-All-FA	3.38 (± 0.08)	2.66 (± 0.08)	**3.08** (± 0.08)	3.04 (± 0.07)
Reference	3.82** (± 0.06)	3.23** (± 0.06)	3.29** (± 0.06)	3.45** (± 0.05)

Table 3: Average ratings of crowdsourced human judgments on fluency, adequacy and complexity. Ratings significantly different from S2S-All-FA are marked with * ($p < 0.05$) or ** ($p < 0.01$); statistical significance tests were calculated using a student t-test. We also provide 95% confidence intervals for each rating.

better understand what needs to be improved in future work. These examples will each show the original complex sentence (Complex), the reference simple sentence (Simple), and our generated simple sentence (Best).

1. Long and complex sentences with multiple clauses

   (a) *Complex*: Turkey has long enshrined the secular ideals of founding father Mustafa Kemal Ataturk, particularly in an education system that until recently banned Islamic headscarves in schools and made schoolchildren begin the day reciting an oath of allegiance to Ataturk's legacy.
   Reference: Schools in Turkey had banned headscarves.
   Simple: They made schoolchildren to Ataturk's history.

   (b) *Complex*: And Wal-Mart, which imports more fruits and vegetables from Mexico than any other U.S. company, announced its effort to force improvements up and down its supply chain.
   Reference: Experts said Wal-Mart is an important company.
   Simple: Wal-Mart used more fruits and vegetables from the company.

2. Need for anaphora resolution

   (a) *Complex*: He is the creative director of Rethink Leisure & Entertainment , which is working on several projects in China and elsewhere in Asia .
   Reference: He is with Rethink Leisure & Entertainment.
   Simple: He is working on several projects in China.

   (b) *Complex*: Teachers there say Richie reads like a high school student.
   Reference: He reads like a high school student.
   Simple: Richie says he is a high school student.

3. Simplifying the wrong part of the sentence

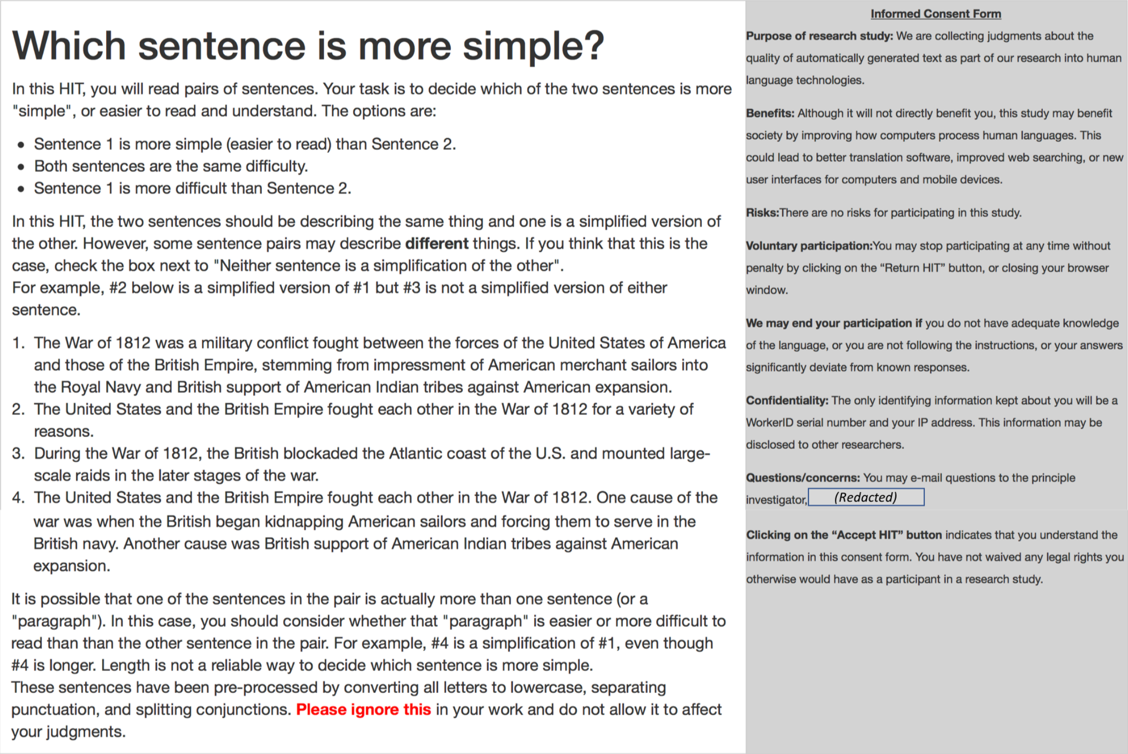
   (a) *Complex*: Parks deliberately maintained her image as shy and proper, said Adrienne Cannon, an expert on African-American history.
   Reference: Adrienne Cannon studies African-American history.
   Simple: She is an expert on African-American history.

   (b) *Complex*: His father owned the home when the lava flowed slowly to the coast.
   Reference: His father still owned the home.
   Simple: The river cut slowly to the coast.

4. Poor substitution due to word embeddings proximity

   (a) *Complex*: Just until recently, the education system had banned Islamic **headscarves** in schools and made schoolchildren begin the day reciting a pledge of allegiance to Ataturk's legacy.
   Reference: Schools in Turkey had banned headscarves.
   Simple: Just until recently, the education officials had banned Islamic **flags**.

   (b) *Complex*: In Beijing kite circles, Fei is widely known as the elder **statesman**.
   Reference: In Beijing, Fei is widely known as an



Which sentence is more simple?

In this HIT, you will read pairs of sentences. Your task is to decide which of the two sentences is more "simple", or easier to read and understand. The options are:

- Sentence 1 is more simple (easier to read) than Sentence 2.
- Both sentences are the same difficulty.
- Sentence 1 is more difficult than Sentence 2.

In this HIT, the two sentences should be describing the same thing and one is a simplified version of the other. However, some sentence pairs may describe **different** things. If you think that this is the case, check the box next to "Neither sentence is a simplification of the other".
For example, #2 below is a simplified version of #1 but #3 is not a simplified version of either sentence.

1. The War of 1812 was a military conflict fought between the forces of the United States of America and those of the British Empire, stemming from impressment of American merchant sailors into the Royal Navy and British support of American Indian tribes against American expansion.
2. The United States and the British Empire fought each other in the War of 1812 for a variety of reasons.
3. During the War of 1812, the British blockaded the Atlantic coast of the U.S. and mounted large-scale raids in the later stages of the war.
4. The United States and the British Empire fought each other in the War of 1812. One cause of the war was when the British began kidnapping American sailors and forcing them to serve in the British navy. Another cause was British support of American Indian tribes against American expansion.

It is possible that one of the sentences in the pair is actually more than one sentence (or a "paragraph"). In this case, you should consider whether that "paragraph" is easier or more difficult to read than than the other sentence in the pair. For example, #4 is a simplification of #1, even though #4 is longer. Length is not a reliable way to decide which sentence is more simple.
These sentences have been pre-processed by converting all letters to lowercase, separating punctuation, and splitting conjunctions. **Please ignore this** in your work and do not allow it to affect your judgments.

Informed Consent Form

Purpose of research study: We are collecting judgments about the quality of automatically generated text as part of our research into human language technologies.

Benefits: Although it will not directly benefit you, this study may benefit society by improving how computers process human languages. This could lead to better translation software, improved web searching, or new user interfaces for computers and mobile devices.

Risks: There are no risks for participating in this study.

Voluntary participation: You may stop participating at any time without penalty by clicking on the "Return HIT" button, or closing your browser window.

We may end your participation if you do not have adequate knowledge of the language, or you are not following the instructions, or your answers significantly deviate from known responses.

Confidentiality: The only identifying information kept about you will be a WorkerID serial number and your IP address. This information may be disclosed to other researchers.

Questions/concerns: You may e-mail questions to the principle investigator, *(Redacted)*

Clicking on the "Accept HIT" button indicates that you understand the information in this consent form. You have not waived any legal rights you otherwise would have as a participant in a research study.

Figure 1: Instructions for our pairwise human evaluation.

expert.
Simple: In Beijing, Fei is considered a **doctor**.

5. Training confusion due to misalignments in training corpus

   (a) *Complex*: He heard the applause.
   Reference: The winning word was "magician."
   Simple: But he heard the song.

   (b) *Complex*: More than 300,000 students across the country signed up to participate.
   Reference: Rogers Middle School took part in the program.
   Simple: There were more than 300,000 students.

6. Acceptable simplifications that were not available in the training

   (a) *Complex*: He later tried to buy his freedom.
   Reference: Henson tried to buy his freedom later.
   Simple: The man tried to buy his freedom.

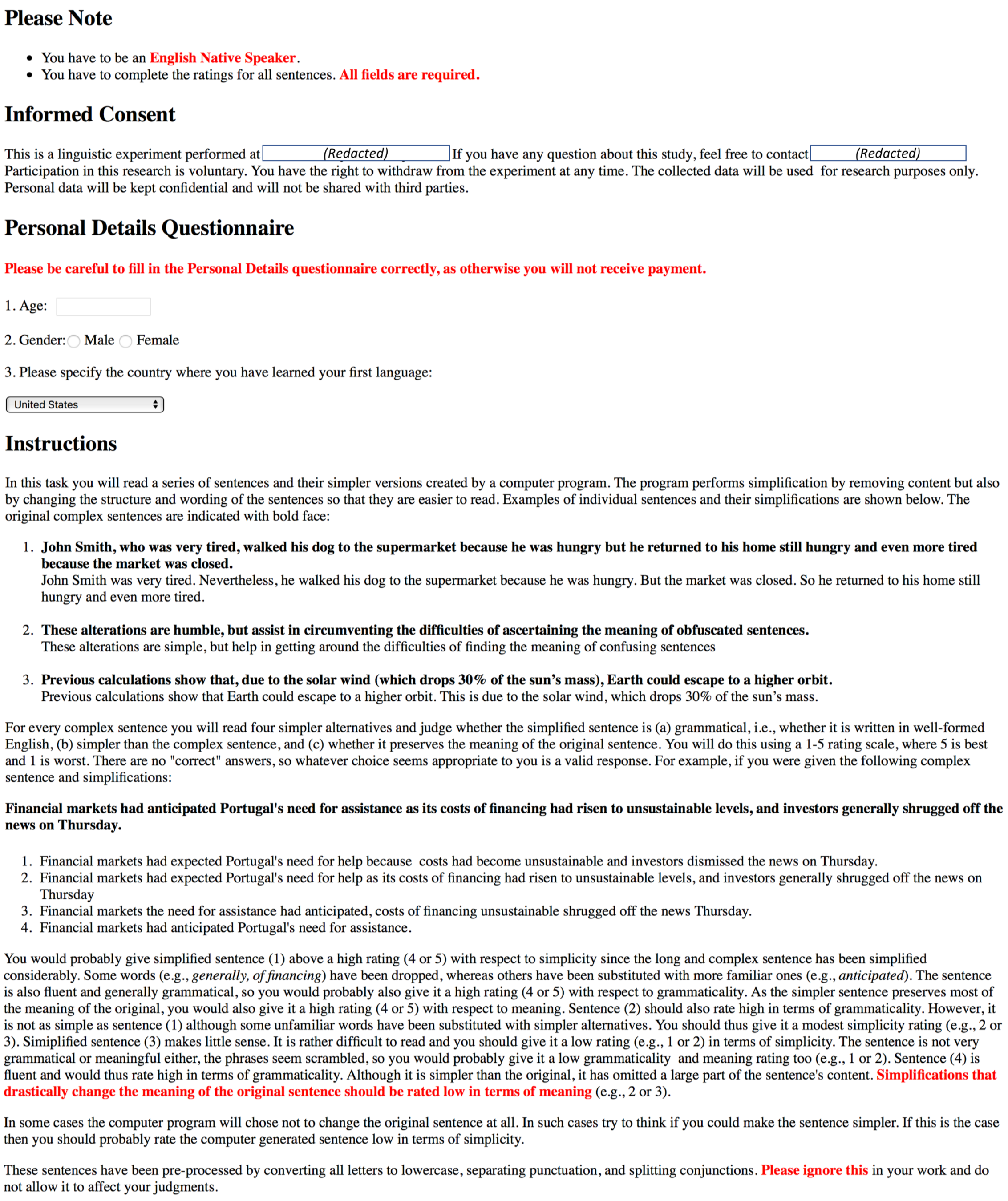
   (b) *Complex*: Middle school was a rough couple of years for Talmus' daughter, Lili Smith.
   Reference: She had a hard time in middle school.
   Simple: School was a rough couple of years.

6a and 6b are examples of good simplifications that would however be penalized by our metrics. 5b is a poor simplification because the model has not captured that the original sentence conveyed important information about the subject and made the poor decision to omit. 4a and 4b are examples displaying the limitation of the deep learning algorithm. It correctly captured the semantic proximity of the substitutions but it failed to recognize that they are not interchangeable in all contexts. 3a is an example of making a poor choice of which part of the sentence is important to keep and which could be omitted. 2a and 2b are examples of poor simplification because the model picked the wrong pronoun or reference. 5a shows that the model was trained on bad pairs of sentences. Clearly, the gold sentence is not the simplification of the target. A more complex sentence was split into more than one sentences and the alignment only picked one. 1a and 1b are examples showing the challenge of trying to simplify long and complex sentences before splitting them.

References

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2018. Chateval: A tool for the systematic evaluation of chatbots. In *Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*.

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. Association for Computational Linguistics.



Please Note

- You have to be an **English Native Speaker**.
- You have to complete the ratings for all sentences. **All fields are required.**

Informed Consent

This is a linguistic experiment performed at *(Redacted)* If you have any question about this study, feel free to contact *(Redacted)* Participation in this research is voluntary. You have the right to withdraw from the experiment at any time. The collected data will be used for research purposes only. Personal data will be kept confidential and will not be shared with third parties.

Personal Details Questionnaire

Please be careful to fill in the Personal Details questionnaire correctly, as otherwise you will not receive payment.

1. Age:

2. Gender: ○ Male ○ Female

3. Please specify the country where you have learned your first language:

United States

Instructions

In this task you will read a series of sentences and their simpler versions created by a computer program. The program performs simplification by removing content but also by changing the structure and wording of the sentences so that they are easier to read. Examples of individual sentences and their simplifications are shown below. The original complex sentences are indicated with bold face:

1. **John Smith, who was very tired, walked his dog to the supermarket because he was hungry but he returned to his home still hungry and even more tired because the market was closed.**
   John Smith was very tired. Nevertheless, he walked his dog to the supermarket because he was hungry. But the market was closed. So he returned to his home still hungry and even more tired.

2. **These alterations are humble, but assist in circumventing the difficulties of ascertaining the meaning of obfuscated sentences.**
   These alterations are simple, but help in getting around the difficulties of finding the meaning of confusing sentences

3. **Previous calculations show that, due to the solar wind (which drops 30% of the sun's mass), Earth could escape to a higher orbit.**
   Previous calculations show that Earth could escape to a higher orbit. This is due to the solar wind, which drops 30% of the sun's mass.

For every complex sentence you will read four simpler alternatives and judge whether the simplified sentence is (a) grammatical, i.e., whether it is written in well-formed English, (b) simpler than the complex sentence, and (c) whether it preserves the meaning of the original sentence. You will do this using a 1-5 rating scale, where 5 is best and 1 is worst. There are no "correct" answers, so whatever choice seems appropriate to you is a valid response. For example, if you were given the following complex sentence and simplifications:

Financial markets had anticipated Portugal's need for assistance as its costs of financing had risen to unsustainable levels, and investors generally shrugged off the news on Thursday.

1. Financial markets had expected Portugal's need for help because costs had become unsustainable and investors dismissed the news on Thursday.
2. Financial markets had expected Portugal's need for help as its costs of financing had risen to unsustainable levels, and investors generally shrugged off the news on Thursday
3. Financial markets the need for assistance had anticipated, costs of financing unsustainable shrugged off the news Thursday.
4. Financial markets had anticipated Portugal's need for assistance.

You would probably give simplified sentence (1) above a high rating (4 or 5) with respect to simplicity since the long and complex sentence has been simplified considerably. Some words (e.g., *generally, of financing*) have been dropped, whereas others have been substituted with more familiar ones (e.g., *anticipated*). The sentence is also fluent and generally grammatical, so you would probably also give it a high rating (4 or 5) with respect to grammaticality. As the simpler sentence preserves most of the meaning of the original, you would also give it a high rating (4 or 5) with respect to meaning. Sentence (2) should also rate high in terms of grammaticality. However, it is not as simple as sentence (1) although some unfamiliar words have been substituted with simpler alternatives. You should thus give it a modest simplicity rating (e.g., 2 or 3). Simiplified sentence (3) makes little sense. It is rather difficult to read and you should give it a low rating (e.g., 1 or 2) in terms of simplicity. The sentence is not very grammatical or meaningful either, the phrases seem scrambled, so you would probably give it a low grammaticality and meaning rating too (e.g., 1 or 2). Sentence (4) is fluent and would thus rate high in terms of grammaticality. Although it is simpler than the original, it has omitted a large part of the sentence's content. **Simplifications that drastically change the meaning of the original sentence should be rated low in terms of meaning** (e.g., 2 or 3).

In some cases the computer program will chose not to change the original sentence at all. In such cases try to think if you could make the sentence simpler. If this is the case then you should probably rate the computer generated sentence low in terms of simplicity.

These sentences have been pre-processed by converting all letters to lowercase, separating punctuation, and splitting conjunctions. **Please ignore this** in your work and do not allow it to affect your judgments.

Figure 2: Instructions for our human evaluation on judging fluency, adequacy, and simplicity.