Small but Mighty: New Benchmarks for Split and Rephrase

Li Zhang[♠*] Huaiyu Zhu[◇] Siddhartha Brahma[♣†] Yunyao Li[◇]
[♠]University of Pennsylvania [◇]IBM Research [♣] Google
zharry@seas.upenn.edu {huaiyu,yunyaoli}@us.ibm.com sidbrahma@google.com

Abstract

Split and Rephrase is a text simplification task of rewriting a complex sentence into simpler ones. As a relatively new task, it is paramount to ensure the soundness of its evaluation benchmark and metric. We find that the widely used benchmark dataset universally contains easily exploitable syntactic cues caused by its automatic generation process. Taking advantage of such cues, we show that even a simple rule-based model can perform on par with the state-of-the-art model. To remedy such limitations, we collect and release two crowdsourced benchmark datasets. We not only make sure that they contain significantly more diverse syntax, but also carefully control for their quality according to a well-defined set of criteria. While no satisfactory automatic metric exists, we apply fine-grained manual evaluation based on these criteria using crowdsourcing, showing that our datasets better represent the task and are significantly more challenging for the models.

1 Introduction

Split and Rephrase is the task of rewriting a presumably long and complex sentence into shorter and simpler sentences, while maintaining the same meaning. For example, one possible way to split the sentence "Voiced by Aoi Koga, Kaguya is the series' titular character, popular among a wide audience." would result in "Kaguya is voiced by Aoi Koga. Kaguya is the series' titular character. Kaguya is popular among a wide audience." While the split sentences have to be coherent, paraphrasing is not enforced. For example, the word "titular" does not have to be replaced. This type of text simplification is challenging as its natural language generation process potentially involves multiple sub-processes such as co-reference resolution, named-entity recognition, semantic role labelling, etc. Split and Rephrase has two main real-world uses: first, to benefit systems whose performance improves with decreasing length of sentences e.g. entity extraction (Zhang et al., 2017) and machine translation (Koehn and Knowles, 2017) by acting as a pre-processing step; second, to benefit human readers, especially those less proficient with the language when reading complex documents such as terms and agreements, in understanding the meaning more easily and accurately (Inui et al., 2003; Siddharthan, 2002).

Datasets of the Split and Rephrase task contain pairs of a *complex sentence* and a presumably meaning-preserving *simplified rewrite* containing multiple simpler sentences. The task was introduced by Narayan et al. (2017), with the release of the WebSplit corpus. Afterwards, Aharoni and Goldberg (2018) proposed the state-of-the-art model to date, a sequence-to-sequence model (Bahdanau et al., 2015) with a copy mechanism (Gu et al., 2016; See et al., 2017) with the observation that most texts are unchanged during a Split and Rephrase operation. Later, Botha et al. (2018) introduced the WikiSplit corpus to be used as large but noisy training data, which the authors reported to be unsuitable as the evaluation data. Also, Sulem et al. (2018) studied the problems of using BLEU as the evaluation metric for this task, while proposing a manually constructed test set called HSplit.

We argue that the widely used benchmark dataset of Split and Rephrase, the WebSplit test set (known as simply WebSplit below), is not suitable for evaluation. Apart from its series of limitations already reported, such as a small vocabulary, unnatural expressions, etc. (Botha et al., 2018), we further show that its complex sentences systematically follow only 3 syntactical patterns marked by lexical cues (§ 2). To demonstrate the implication of such limitations of WebSplit, we show that a simple,

[*] Work done during internship at IBM Research.
[†] Work done during employment at IBM Research.

unsupervised rule-based model with only 3 corresponding operations can perform even slightly better than the state-of-the-art neural model (§ 3).

To remedy the limitations of WebSplit, we crowdsource two new benchmarks with significantly more diverse syntax in the Wikipedia and legal contract domain with hundreds of human-written complex-simple sentence pairs (§ 4). We carefully control for their quality based on 6 well-defined criteria of what constitutes a good Split and Rephrase rewrite. While most related work reports model performance using the widely criticized BLEU score and manual evaluation with no clear rubric, we perform fine-grained model evaluation using these 6 criteria, rated by crowd workers, showing that our benchmarks present models with greater challenges (§ 5).[1]

2 Issues with WebSplit

WebSplit and Wiki-Split are two widely used datasets for the Split and Rephrase task. Because WikiSplit is derived from the edit history of Wikipedia, versions of passages are not necessarily written by Split and Rephrases operations, as the meaning may not be preserved during edits. Hence, WikiSplit is reported by its authors to be noisy and ill-suited for evaluation for this task (Botha et al., 2018).

WebSplit is used in multiple previous works as the evaluation benchmark. It was created by automatically matching sentences in the WebNLG corpus (Gardent et al., 2017) according to partitions of their meaning representations. The dataset has been shown to have various limitations, such as unnatural expressions, repetition of phrases (Botha et al., 2018), etc.

Furthermore, our preliminary study shows that WebSplit contains several recurring syntactic patterns marked with lexical cues. To demonstrate this, we randomly sample 100 complex sentence from the test set, and are able to categorize them with only 3 syntactical patterns marked by lexical cues (underlined), at which some almost trivial Split and Rephrase operations can take place:

relative clause (rc) (48 out of 100): Scott Adsit voiced Baymax which was created by Duncan Rouleau.

conjunction (conj) (46 out of 100): Above the Veil is from Australia and was preceded by Aenir and Castle.

participle (part) (13 out of 100): Serving the city of Alderney, the 1st runway is made from Poaceae.

It can be further noticed that most complex sentences in WebSplit are short and require only one Split and Rephrase operation. We next show that a rule-based model which only exploits these patterns can perform on par with the state-of-the-art neural model on WebSplit.

3 Rule-Based Model

We design a simple rule-based model to exploit the syntactic cues widely present in WebSplit.

3.1 Algorithm

The rule-based model requires no training data and only uses semantic role labeling (He et al., 2017) and dependency parsing (Dozat and Manning, 2016), running on AllenNLP (Gardent et al., 2017). Given a complex sentence, the model makes 3 splits when applicable. First, using semantic role labeling, the model identifies a Relational Argument and makes a split with the Relational Argument replaced by the Subject Argument. Second, The model looks for the word "and", making a split accordingly. Third, using dependency parsing, the model looks for a node which is joined by the clause, which is extracted, prepended with the subject, and split as a new simple sentence, while the rest of the original complex sentence is split as another new simple sentence.[2]

3.2 Performance

The rule-based model and the state-of-the-art seq2seq model trained on WikiSplit (Aharoni and Goldberg, 2018; Botha et al., 2018) are evaluated using BLEU (Papineni et al., 2002) on WebSplit. The rule-based model achieves a BLEU of 61.3, outperforming the neural model which achieves a BLEU of 56.0. The two models are also evaluated manually on 100 randomly sampled examples, with an identical accuracy of 64% (the criteria of correctness is described in § 4.2.2). While the rule-based model is imperfect and can likely improve with more and better defined rules, it serves as a strong baseline that exploits the syntactical cues in WebSplit and potentially other benchmarks generated in a similar fashion. The strong performance of such a simplistic model highlights the need of more difficult and diverse benchmark data to better capture the complexity of the Split and Rephrase task.

[1] We will release the datasets, code to process them, and crowdsourcing designs soon.

[2] The detailed algorithm is shown in the Appendix A.

4 New Benchmark Datasets

Considering the limitations of WebSplit, an ideal benchmark must not only be challenging with diverse patterns, but also ensure that the rewrites are strictly meaning-preserving Split and Rephrase. With these two goals, we collect two benchmark datasets, Wiki Benchmark (Wiki-BM) from Wikipedia and Contracts Benchmark (Cont-BM) from the legal documents. These two datasets are to be used as gold standard for the evaluation of Split and Rephrase. To systematically control for the quality, we define 6 criteria of what constitutes a good Split and Rephrase, and validate the collected rewrites based on these criteria.

4.1 Collecting Complex Sentences

First, we gather complex sentences as the input for the Split and Rephrase operation.

4.1.1 Wiki Benchmark (Wiki-BM)

While the simplified rewrites in the WikiSplit dataset are not guaranteed to be meaning preserving and cannot be used in a benchmark, the original complex sentences are semantically and syntactically diverse, with adequate complexity. From the 5000 complex sentences from the WikiSplit test set, we randomly select 500 for budget reasons with only alphanumerical characters, whitesspaces, commas and periods, and manually inspect them to ensure that they are well-formed.

4.1.2 Contracts Benchmark (Cont-BM)

We collect sentences from publicly available legal procurement contracts online, and contract templates within IBM with no confidential information. We randomly sample and inspect 500 sentences in the same manners as above.

4.1.3 Syntactical Diversity

To demonstrate that our complex sentences are syntactically diverse and are not plagued by patterns analyzed before, we randomly sample 100 complex sentences from each benchmark to annotate them by syntactical patterns. In addition to the 3 patterns outlined before, we define the following new patterns (the examples are truncated to save space):

prepositional phrase (prep): The mausoleum was built in 1894 along the lines specified by Frazer.

adverbial phrase (adv): Except as may be otherwise specified, Supplier shall invoice Buyer.

apposition clause (appos): Leila married the movie director Ruy Guerra, father of her only daughter.

Patterns	WebSplit	Wiki-BM	Cont-BM
rc	48	34	29
conj	46	71	66
part	13	34	28
prep	5	12	66
adv	0	19	38
appos	2	10	0
inf	0	5	10
patterns/sent	1.22	1.78	2.37

Table 1: Counts of syntactic patterns for splitting in 100 random examples from each of WebSplit, Wiki Benchmark, and Contracts Benchmark. Note that each complex sentence may have more than one pattern.

Entry	WebSplit	Wiki-BM	Cont-BM
Rewritten by human	No	Yes	Yes
# complex	930	403	406
# simple	43958	720	659
# toks/complex	20.6	29.6	41.5
# sents/simple	3.7	3.0	3.0

Table 2: Comparison of statistics among WebSplit, Wiki Benchmark, and Contracts Benchmark.

infinitive clause (inf): Nimfa was forced to take part of a devilish plan to fool the Saavedra family.

The counts from the manual annotation are shown in Table 1. Wiki-BM has more diverse patterns and number of patterns per complex sentence than WebSplit, while Cont-BM has the most. The difference of complexity in the 3 benchmarks would be beneficial for evaluation.

4.2 Collecting Simplified Rewrites

We ask a set of crowd workers to Split and Rephrase the gathered complex sentences on Amazon Mechanical Turk, and another set to ensure their quality[3]. We divide the crowdsourcing workflow into two phases.

4.2.1 Phase 1: Rewrite

For each complex sentence, we ask 3 crowd workers to rewrite it by splitting and rephrasing, with the option to flag the complex sentence as too simple or too problematic to split, which we later discard. We require Master Qualification, and pay $0.2 per HIT for the complex sentences from Wiki-BM and $0.4 per HIT for the more challenging Cont-BM. This Phase costs $1,125 in total.

4.2.2 Phase 2: Rate

For each crowdsourced rewrite submitted in Phase 1, we ask 2 different crowd workers to evaluate its

[3] Detailed guidelines are shown in the Appendix B.

WebSplit	sensical	grammatical	no miss fact	no new fact	correct split	enough split	*correct*	BLEU
seq2seq	71.6%/4.55	64.0%/4.40	94.30%	94.30%	87.30%	79.00%	50.1%	62.6%
rule	72.2%/4.35	58.2%/4.01	92.70%	95.70%	83.30%	82.90%	51.7%	65.9%
Wiki-BM	sensical	grammatical	no miss fact	no new fact	correct split	enough split	*correct*	BLEU
seq2seq	55.4%/4.22	47.8%/3.93	98.30%	99.70%	80.30%	76.30%	37.3%	87.0%
rule	59.8%/4.06	54.2%/3.85	94.30%	94.30%	78.70%	47.70%	28.5%	77.2%
human	84.9%/4.76	76.8%/4.61	95.00%	93.70%	88.30%	88.00%	68.4%	77.8%
Cont-BM	sensical	grammatical	no miss fact	no new fact	correct split	enough split	*correct*	BLEU
seq2seq	29.4%/3.45	25.0%/3.04	92.70%	99.00%	52.30%	63.00%	16.7%	78.6
rule	57.9%/3.97	54.8%/3.89	97.70%	96.70%	83.30%	44.70%	25.0%	79.2
human	78.2%/4.53	72.6%/4.43	95.30%	96.30%	93.70%	85.00%	63.3%	73.0%

Table 3: Average crowd ratings by criteria, model and benchmark. For the first two criteria which are on the scope of *0–5*, we report the percentage of *5* and the average. For the rest which are *yes–no* questions, we report the percentage of *yes*.

quality, based on the following fine-grained criteria:

1. Is it sensical (scale of 0-5)?
2. Is it grammatical (scale of 0-5)?
3. Does it miss any existing facts (yes/no)?
4. Does it introduce new facts (yes/no)?
5. Does it have splits at the wrong place (yes/no)?
6. Should some of its sentences be further split (yes/no)?

We require Master Qualification, and pay $0.07 per HIT[4]. This Phase costs $508.

In each benchmark, we now have 500 complex sentences, each with 3 rewrites, each with 2 ratings. For each rating, if the worker answers 5 for the first two criteria, and chooses "no" for last four criteria, we denote this rating as *correct*. For each rewrite, if both of its ratings are *correct*, we denote this rewrite as *perfect*. To ensure high quality of the gold standard, we only keep the rewrites that are *perfect* as gold standard corresponding to their complex sentences in our benchmarks.

4.3 Descriptive Statistics

Some descriptive statistics and the comparison with WebSplit are shown in Table 2. While our new benchmarks are smaller than WebSplit, we argue that a small number of human-written, high quality ground-truth simple rewrites are better suited for evaluation than a larger number of automatically generated, noisy ones.

While similar to HSplit (Sulem et al., 2018), our benchmarks include several additional features, such has much more complex sentences from the legal domain, a clear set of rubrics for evaluation, and crowdsourced human judgements to scale.

5 Model Performance

Previous work reports the model performance on this task using two metrics: BLEU on the entire benchmark and manual ratings on a small subset. However, BLEU has long been shown to have little correlation with human judgements in text simplification[5] (Sulem et al., 2018). While other alternatives exist, the focus of our work is not the metrics, but rather the quality and difficulty of benchmarks, which can be illustrated no better than by human evaluation. Previously, manual evaluation has been done without a well-established rubric on what makes a Split and Rephrase rewrite correct. To address these problems, we use crowdsourcing following the process of Phase 2, by asking 3 crowd workers to rate model outputs based on the 6 fine-grained criteria described above.

Table 3 shows the average crowd ratings and BLEU score for each combination of a model and a benchmark. We consider the state-of-the-art seq2seq model trained on WikiSplit (Botha et al., 2018) and our rule-based model. We use all rewrites from Phase 1 including those not included in our benchmarks to measure human performance.

Both the rule-based and seq2seq model have large rooms for improvement, as they significantly underperform crowd workers in almost all criteria, with significantly lower performance in our proposed benchmarks than in WebSplit. Even for

[4] The pay exceeds the prorated US minimum wage.

[5] We reinforce this claim in Appendix C.

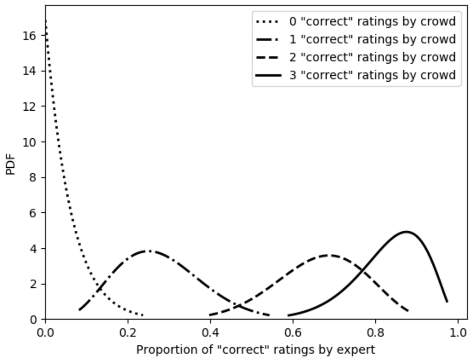

Figure 1: Beta distributions with Laplace smoothing of the proportion of *correct* ratings by expert.

crowd workers, the percentage of overall *correct* is less than 70% in our new benchmarks, whose complex sentences are much more challenging to Split and Rephrase.

6 Reliability of the Crowd

Can we use crowdsourcing to evaluate models no less reliably as experts or authors, as done in previous work? As experts of this task, we manually rate a subset of model-output rewrites as the ground truth for rating, and compare it against the crowd's rating. Since there are 3 benchmarks and 3 models (including human, whose outputs are crowd rewrites we have collected in Wiki-BM and Cont-BM, **but not** WebSplit), there are 8 combinations in total. From the crowd ratings of these combinations, we assign each complex–output pair into one of 4 buckets, determined by the number of *correct* ratings out of 3 crowd ratings. For each bucket, we sample 2 complex–output pairs. In total, $8 \times 4 \times 2 = 64$ complex–output pairs are sampled. The expert rates them independently following the same 6 criteria as the crowd workers. This gives the proportion of expert's *correct* ratings among each bucket.

These statistics allow us to fit a beta distribution for expert rating conditional on each crowd rating bucket, using Laplace prior smoothing. The results are shown in Figure 1. Each distribution corresponds to a bucket with 0, 1, 2, or 3 out of 3 *correct* crowd ratings. For example, the right-most curve represents the probability density function where both the expert and the 3 crowd raters agree on a *correct* rating. According to the figure with a 90% one-sided confidence, when all 3 crowd raters rate a rewrite as *correct*, the expert also rates *correct* in more than around 80% of the samples; when none of the 3 crowd raters rate a rewrite as *correct*, the expert rates *correct* for less than around 10% of the samples.

This shows that crowdsourcing can be a reliable way to evaluate models for this task, with variable reliability depending on the number of raters per sample and their agreement.

7 Conclusion

After showing the flaws of the current benchmark in Split and Rephrase, we release two crowdsourced benchmarks containing significantly more diverse syntax. Using fine-grained crowdsourcing evaluation on 6 well-defined criteria, we show that they provide a greater challenge to models.

Acknowledgments

We thank Qing Lyu, Ranit Aharonov, Yannis Katsis, Lucian Popa, Nancy Xinru Wang and the anonymous reviewers for their valuable feedback. We also thank Shashi Narayan for providing helpful references on text simplification at the early stage of this work.

References

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke S. Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *ACL*.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *IWP@ACL*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Advaith Siddharthan. 2002. An architecture for a text simplification system. *Language Engineering Conference, 2002. Proceedings*, pages 64–71.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A Algorithm of the Rule-Based Model

Given a complex sentence, the model runs the following processes once each.

Wh Handling Using semantic role labeling, the model looks for a Relational Argument (R-ARG), and the Subject Argument (asserted to be the ARG preceding the R-ARG). Then, a split is made with the Relational Argument replaced by the Subject Argument.

Conjunction Handling The model looks for the word "and". Using semantic role labeling, if the word following "and" is an argument (ARG), assert that "and" is followed by a sentence, and a split is made. Or, if the word following "and" is a verb (V), the model asserts the Subject Argument to be the ARG preceding the V; a split is made with "and" replaced by the Subject Argument.

Insertion Handling Using dependency parsing, the model looks for a node with type *participle modifier, relative clause modifier, prepositional modifier, adjective modifier,* or *appositional modifier*. The clause with the node as the root is extracted, prepended with the subject, and split as a new simple sentence. The rest of the original complex sentence is split as another new simple sentence.

B Crowdsourcing Guidelines

B.1 Guidelines of Phase 1: Rewrite

Instructions: A long, complex sentence is hard to understand for many people. Please try to rewrite such a sentence by splitting and rephrasing it as several shorter and simpler sentences. A good example:

- **Original**: Jonathan Thirkield, currently living in New York City, is an American poet who is known to be prolific.
- **Rewritten** (good): Jonathan Thirkield is an American poet. Jonathan Thirkield is known to be prolific. Jonathan Thirkield is currently living in New York City.
- **Rewritten** (good): Jonathan Thirkield is an American poet. He is currently living in New York City. He is known to be prolific.

Your rewrite must satisfy the following requirements:

1. Grammatical
   Rewritten (bad: ungrammatical): Jonathan Thirkield currently living in New York City. Jonathan Thirkield is an American poet. He

Benchmarks+Models	sensical	grammatical	no miss fact	no new fact	correct split	enough split	*correct*
WebSplit+s2s	.367	.273	-.037†	-.001†	.184	.046†	.303
WebSplit+rule	.491	.456	.118†	.048†	.425	.276	.480
Wiki-BM+s2s	.231	.319	.190	-.005†	.167†	.256	.412
Wiki-BM+rule	.438	.561	.083†	.075†	.512	-.035†	.232
Cont-BM+s2s	.345	.329	.402	-.062†	.191	.215	.255
Cont-BM+rule	.277	.190	-.007†	.064†	.148†	.115†	.098†
WebSplit+all models	.433	.348	.029†	.023†	.289	.161	.326
Wiki-BM+all models	.340	.425	.179	.122†	.328	.243	.217
Cont-BM+all models	.313	.271	.228	.01†	.199	.142	.165
all benchmarks+s2s	.237	.233	.208	.064†	.167	.146	.141
all benchmarks+rule	.388	.400	.081†	.063†	.347	.089†	.230
all benchmarks+all models	.362	.393	.172	.068†	.315	.168	.251

Table 4: Spearman's correlation between sentence-level BLEU and human judgement on 6 criteria by combinations of benchmarks and models. †: the correlation coefficient is not statistically significant with $\alpha = .05$.

is known to be prolific.

2. Sensical and understandable
   Rewritten (bad: non-sensical): Jonathan Thirkield lives in prolific New York City. He is an American poet.
3. Has the same meaning as the original complex sentence, with no new facts and no missing facts (show/hide examples)
   Rewritten (bad: new facts): Jonathan Thirkield is a best- selling American poet. He is currently living in New York City. He is known to be prolific.
   Rewritten (bad: missing fact): Jonathan Thirkield is an American poet. He is currently living in New York City. (does not mention prolific)
4. Split into appropriate number of short sentences (at least two), not too few or too many. If the sentence is too simple to be split, write SIMPLE as your response.
   Rewritten (bad: too few splits): Jonathan Thirkield, currently living in New York City, is an American poet. He is known to be prolific.
   Rewritten (bad: too many splits): Jonathan Thirkield is a poet. He is American. He is currently living somewhere. That somewhere is New York City. He is prolific. Such is known. (too many unnecessary splits)
5. Do NOT use pronouns (it, she, he, they, this, that) if they are ambiguous
   Rewritten (bad: ambiguous pronoun): Walt Whitman is an American poet. Jonathan Thirkield is also an American poet. He is living in New York City.

Your rewrite will be validated by others. You might not receive payment if your rewrite does not satisfy the requirements. You may skip this HIT if you find splitting the given sentence too hard. However, if you manage to appropriately split a sentence which many other workers have skipped, you will receive a bonus.

B.2 Guidelines of Phase 2: Rate

Instructions: Read the two pieces of text below. The second text is an attempt to rewrite the first text, by splitting and rephrasing it into several shorter sentences to be understood more easily. Your job is to judge if this rewrite is good.

1. The Rewritten text **makes sense**
2. The Rewritten text **is grammatical**
3. Does the Rewritten text miss some facts that are present in the Original text?
4. Does the Rewritten text have new facts that are not present in the Original text?
5. Does the Rewritten text split the Original text at the wrong place or unnecessarily?
6. Does the Rewritten text have one or more sentences that should be further split?

Each question is accompanied by a positive and negative example, the same as in the previous section. The crowd workers answer the first two questions by dragging a draw bar between "Strongly Disagree" and "Strongly Agree", and the last four questions by choosing "yes/no" radio boxes.

C Correlation Between BLEU and Crowd Workers

Does BLEU correlate with human judgement on a large scale? To answer this, we collect crowd-sourced ratings of model outputs. With 3 benchmark datasets (WebSplit, Wiki-BM and Cont-BM) and two models (seq2seq and rule-based), we sample 100 complex sentence and output rewrite pairs from each combination, resulting in 600 in total.[6] Then, we run the same crowdsourcing project as Phase 2 (Sec. 5.2.2) with these 600 pairs, for each of which we collect ratings from 3 crowd raters. The crowd raters are asked to rate based on the same 6 criteria as before (Sec. 3.1). As defined before, if a rating includes 5 for the first two criteria and "no" for the other four, it is considered *correct*.

The Spearman's correlation coefficients between sentence-level BLEU and crowd ratings in each 6 criteria are shown in Table 4. While BLEU has higher correlation with crowd raters on whether the rewrite is sensical or grammatical, most correlation coefficients are less than .5, and many do not imply a positive correlation at all.

This reinforces the claim that BLEU is not a suitable evaluation metric for the Split and Rephrase task, because it has little correlation with human (crowd) judgement.

[6] Additionally, we sample 100 pairs each directly from Wiki-BM and Cont-BM with 3 crowd rewrites. These 600 pairs are used to measure human performance, but are not used in this section because they themselves are ground truth.