

A Systems tested

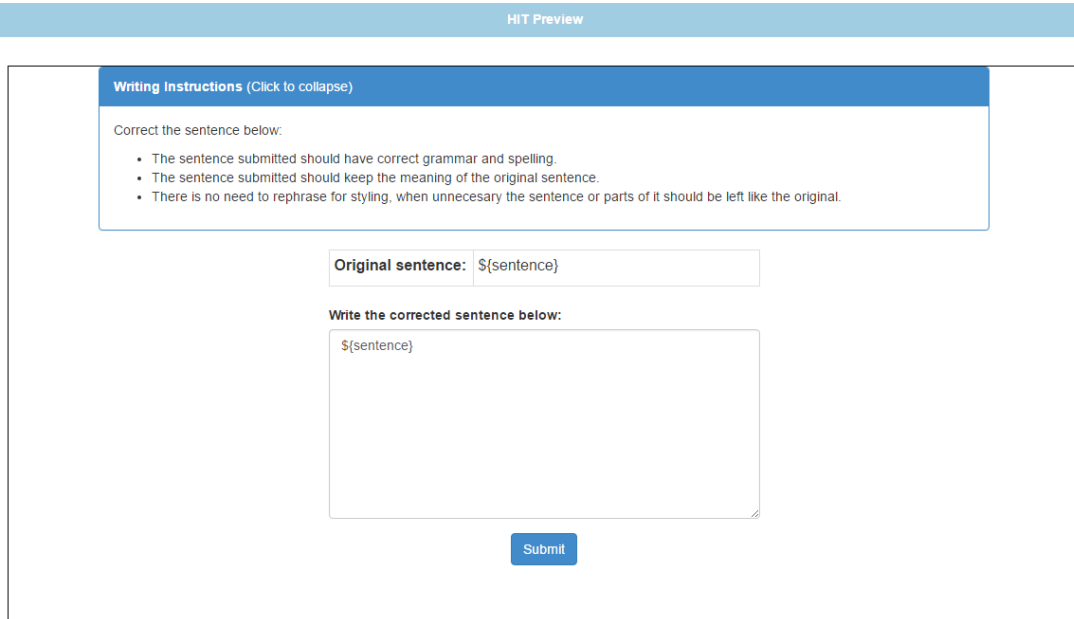
Adam Mickiewicz University (AMU), University of Cambridge (CAMB), Columbia University and the University of Illinois at Urbana-Champaign (CUUI), Indian Institute of Technology, Bombay (IITB), Instituto Politecnico Nacional (IPN), National Tsing Hua University (NTHU), Peking University (PKU), Pohang University of Science and Technology (POST), Research Institute for Artificial Intelligence, Romanian Academy (RAC), Shanghai Jiao Tong University (SJTU), University of Franche Comté (UFC), University of Macau (UMC), Rozovskaya and Roth (2016, RoRo), Junczys-Dowmunt and Grundkiewicz (2016, JMGR) Xie et al. (2016, Char).

Supplementary Material for “Inherent Biases in Reference-based Evaluation for Grammatical Error Correction”

B Collected references

B.1 Grammatical Error Correction

52 sentences with a maximum length of 15 were collected from the NUCLE test corpus (Dahlmeier et al., 2013). For each of the 52 source sentences, we elicited 50 corrections from Amazon Mechanical Turk workers. Workers were native speakers (located in the US) having at least 1000 approved HITs and 98% acceptance rate. Sentences with less than 6 words were discarded, as they were mostly a result of sentence segmentation errors. 4 sentences required no correction according to almost half the workers and were hence discarded. Common methods based on agreement such as Fleiss’s kappa or test questions as the problem we deal with is exactly the low agreement of valid corrections. 4 workers were rejected due to suspicious answers. such were workers that most or all of their work changed nothing except non-alphanumeric characters and were the only ones to keep several source sentences unchanged.



The image shows a screenshot of a HIT Preview interface. At the top, there is a blue header bar with the text "HIT Preview". Below this, there is a white box with a blue header bar that says "Writing Instructions (Click to collapse)". The instructions are as follows:

Correct the sentence below:

- The sentence submitted should have correct grammar and spelling.
- The sentence submitted should keep the meaning of the original sentence.
- There is no need to rephrase for styling, when unnecessary the sentence or parts of it should be left like the original.

Below the instructions, there is a text input field labeled "Original sentence:" with a placeholder "\${sentence}". Underneath this, there is another text input field labeled "Write the corrected sentence below:" with a placeholder "\${sentence}". At the bottom of the form, there is a blue "Submit" button.

Figure 1: Template for a grammatical error correction annotation task

origin	Other relatives may have the same possibilities to have such kind of disease .
1	It is possible other relatives may have the same kind of disease .
2	It is possible that other relatives may have the same kind of disease .
3	It's also possible for other relatives to have the same kind of disease.
4	Other relatives may also be predisposed to the same kind of diseases.
5	Other relatives may be at risk to have the same disease.
6	Other relatives may have be prone to having such diseases.
7	Other relatives may have similar possibilities to have the same disease.
8	Other relatives may have the possibility of having the same kind of disease.
9	Other relatives may have the possibility to have the same such disease .
10	Other relatives may have the same chance of contracting the same kind of disease.
11	Other relatives may have the same chance of having that kind of disease.
12	Other relatives may have the same chance of suffering such diseases.
13	Other relatives may have the same chances of having the same kind of disease.
14	Other relatives may have the same chance to have the same disease.
15	Other relatives may have the same likelihood of having such a disease.
16	Other relatives may have the same possibilities of having such a disease .
17	Other relatives may have the same possibilities of having such a disease .
18	Other relatives may have the same possibilities of having such a disease.
19	Other relatives may have the same possibilities of having that kind of disease.
20	Other relatives may have the same possibilities of having the same kind of disease.
21	Other relatives may have the same possibilities to develop such a disease.
22	Other relatives may have the same possibilities to have such a disease .
23	Other relatives may have the same possibilities to have such a kind of disease.
24	Other relatives may have the same possibilities to have such a kind of disease.
25	Other relatives may have the same possibilities to have such kinds of diseases.
26	Other relatives may have the same possibilities to have the same kind of disease .
27	Other relatives may have the same possibilities to have this kind of disease .
28	Other relatives may have the same possibilities to have this kind of disease.
29	Other relatives may have the same possibility of developing the same kind of disease.
30	Other relatives may have the same possibility of having said disease.
31	Other relatives may have the same possibility of having the disease.
32	Other relatives may have the same possibility of having this kind of disease .
33	Other relatives may have the same possibility to have such a kind of disease .
34	Other relatives may have the same possibility to have such kind of a disease .
35	Other relatives may have the same possibility to have such kind of disease .
36	Other relatives may have the same possibility to have such kinds of disease .
37	Other relatives may have the same possibility to have such kinds of diseases.
38	Other relatives may have the same possibility to have the same kind of disease.
39	Other relatives may have the same probability of having this disease .
40	Other relatives may have the same probability to have the same kind of disease.
41	Other relatives may have the same risk for developing such a disease.
42	Other relatives may have the same risk of having such a disease.
43	Other relatives may have the same risk of having the disease.
44	Other relatives may have the same strong possibility to have such kinds of disease.
45	Other relatives may possibly also have the same disease .
46	Other relatives may possibly have the same risk of having that kind of disease.
47	Other relatives may possibly have the same such kind of disease .
48	Relatives may also be more prone to similar diseases.
49	Relatives may have the same possibilities to have such kind of disease .
50	Those who are related may have the same chances of acquiring certain diseases.

Table 1: An example of one of the learner language sentences with the highest number of different corrections. The origin sentence is on top.

C Validity

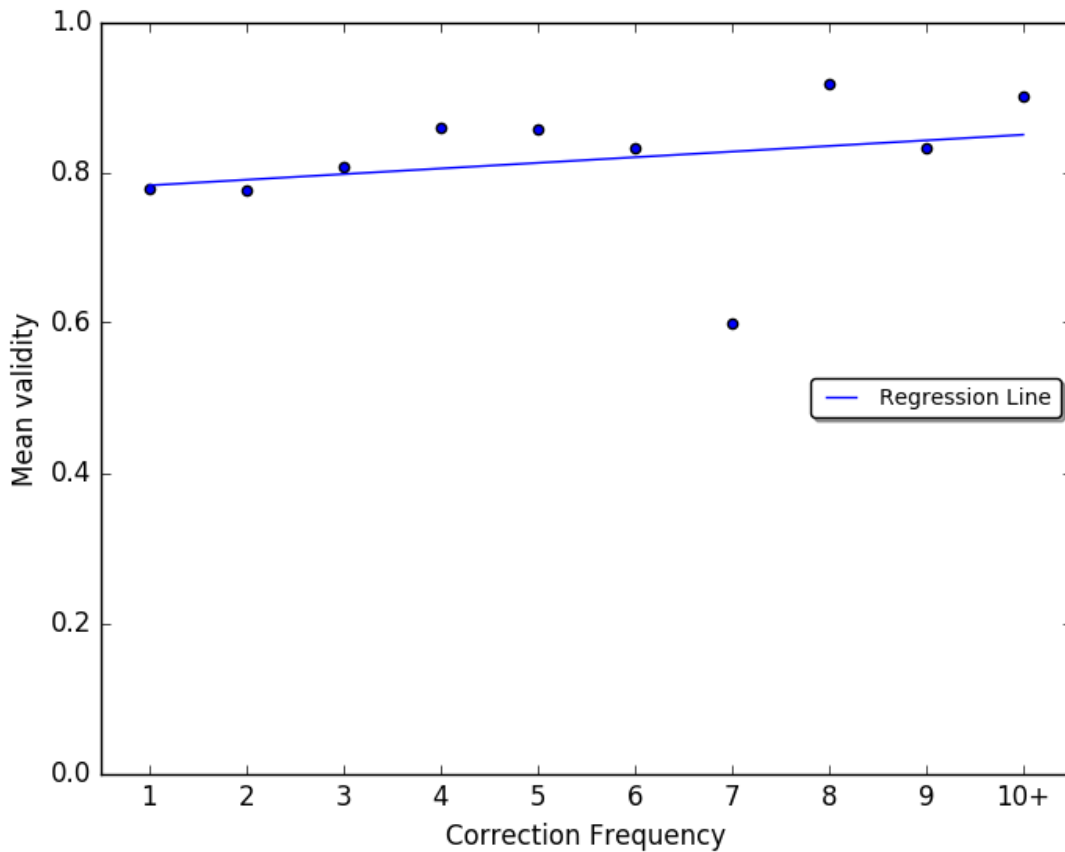


Figure 2: The mean frequency (y -axis) in which a correction that was produced a given number of times (x -axis), was judged to be valid.

D Accuracy - Poisson Binomial Distribution

The analytic tools we have developed support the computation of the entire distribution of the accuracy, and not only its expected values. From the Equation

$$Acc(C; X, Y) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{C(x_i) \in Y_i}. \quad (1)$$

we see that Accuracy has a Poisson Binomial distribution (i.e., it is a sum of independent Bernoulli variables with different success probabilities), whose success probabilities are $P_{y, Y \sim \mathcal{D}_i}(y \in Y)$, which can be computed, as before, using our estimate for \mathcal{D}_i . Estimating the density function allows for the straightforward definition of significance tests for the measure, and can be performed efficiently (Hong, 2013). An implementation of this and other methods for efficiently computing and approximating Poisson Binomial Distributions and the estimated density functions can be found in <https://github.com/borgr/PoissonBinomial>.

E Type conservatism and prevalence

TYPE	CHANGE
CONTR	0.971
NOUN:NUM	0.955
ORTH	0.893
ADJ:FORM	0.810
NOUN:INFL	0.713
DET	0.672
VERB:SVA	0.643
MORPH	0.615
VERB:FORM	0.538
SPELL	0.537
VERB:INFL	0.500
ADJ	0.460
CONJ	0.430
ADV	0.416
PREP	0.334
WO	0.315
NOUN:POSS	0.290
VERB:TENSE	0.289
NOUN	0.282
PART	0.268
PRON	0.239
PUNCT	0.201
OTHER	0.174
VERB	0.151

Table 2: Number of mean corrections of systems divided by the number of corrections by references on NUCLE dataset(Dahlmeier and Ng, 2012). Data is based on Bryant et al. (2017).

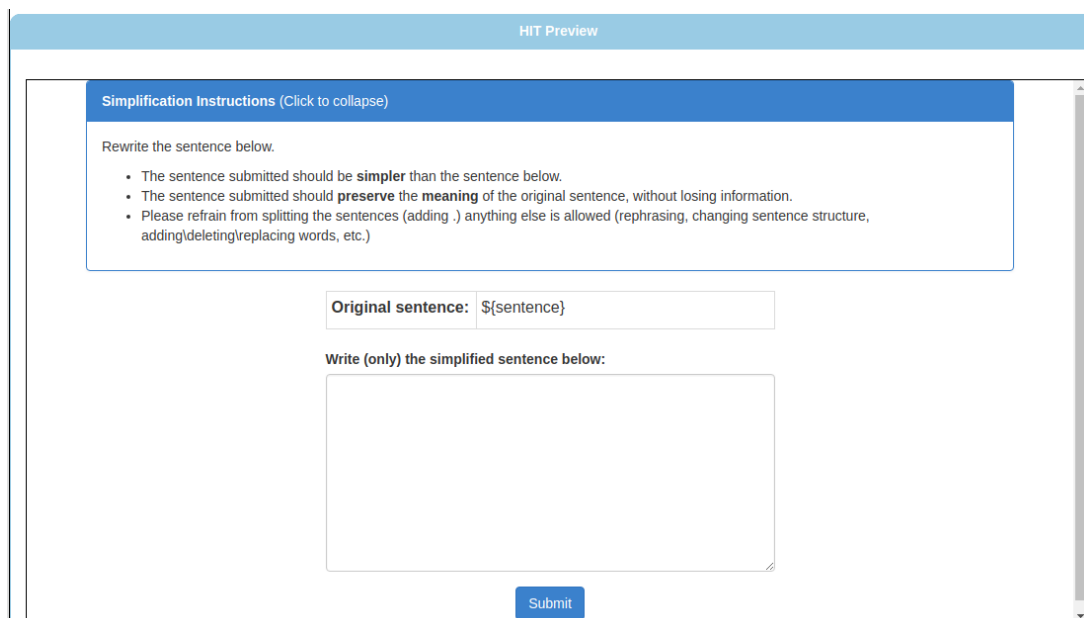
F SARI with Multiple references

Most RBMs define their multi-reference score to be the maximum score the output attains against any single one reference. SARI takes a different approach and combines multiple references to yield its score, possibly in order to compensate for the necessarily limited coverage of the available references. This yields non-monotonic behavior of SARI with respect to increasing the number of references, which makes the biases incurred by low coverage less predictable, but not less significant. For instance, a set of diverse set of references would span a large space of possible combinations, making SARI more permissive. A set of references of the same size, which only differ little from one another would yield a more conservative SARI score.

SARI is defined as the average of 3 scores, based on how well the system output kept the words it should keep, deleted the words it should delete, and added the words it should add (all with respect to the reference). Each of these scores behaves differently when increasing the number of references. For a perfect recall for addition can only be obtained if all additions suggested by any reference are added. For a perfect precision for addition all additions should be found in at least one reference, acting similar to max operation. For a perfect precision on keeping all the references should keep all agree on keeping everything the system kept, and as we showed reference don't tend to agree on it. For a perfect recall the system should everything that at least one of the references chose to keep. Note, that the two terms together mean that for a perfect keep score a necessary condition is that all **references** should agree on what to keep. A perfect precision of deletion acts similarly to the one over keeping, but being precision oriented SARI ignores the deletion recall. Thus, a perfect deletion would be one to delete only words, and at least one, that all references agree on.

G Reference Crowdsourcing for Simplification

In order to perform our simplification experiments, we collected additional references for sentences from the corpus presented in Xu et al. (2016), using Amazon Mechanical Turk with similar annotation guidelines to those used by Xu et al. Overall we collect 50 references for 45, and 100 and 150 references for two more sentences. The latter two have 70 and 126 references that occur only once and none reoccurring more than 6 times, supporting the claim that the size of the space of valid references for TS is huge. Standard worker rejection techniques were used, but no worker had to be rejected.



HIT Preview

Simplification Instructions (Click to collapse)

Rewrite the sentence below.

- The sentence submitted should be **simpler** than the sentence below.
- The sentence submitted should **preserve** the **meaning** of the original sentence, without losing information.
- Please refrain from splitting the sentences (adding .) anything else is allowed (rephrasing, changing sentence structure, adding/deleting/replacing words, etc.)

Original sentence: \${sentence}

Write (only) the simplified sentence below:

Submit

Figure 3: Template for a simplification annotation task

G.1 Simplification reranking

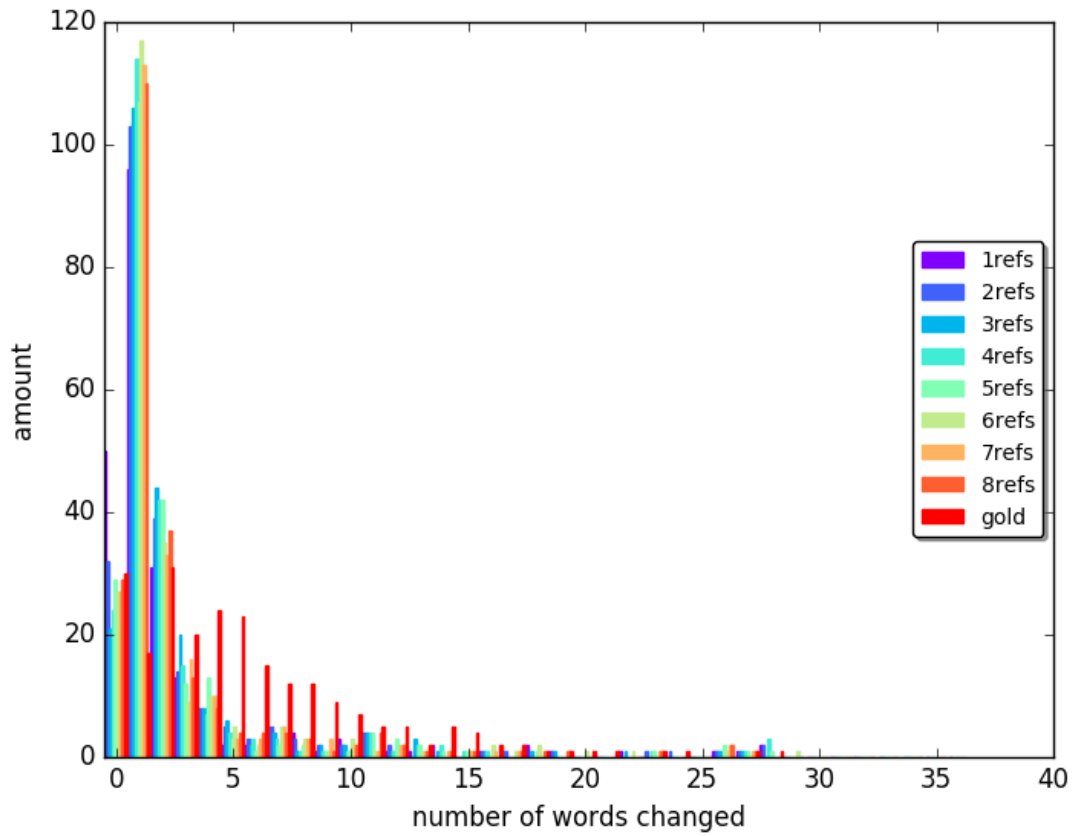


Figure 4: Reranking results with Nisioi et al. (2017) and MAX-SARI

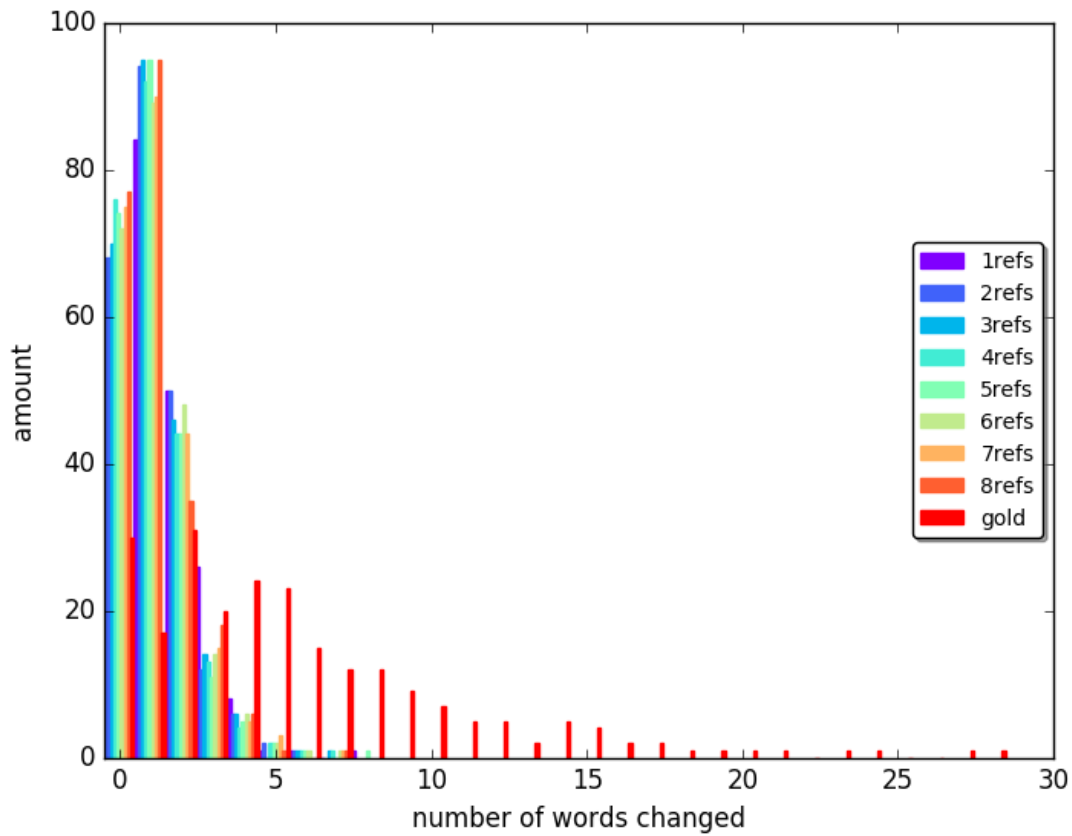


Figure 5: Reranking results with Moses and SARI

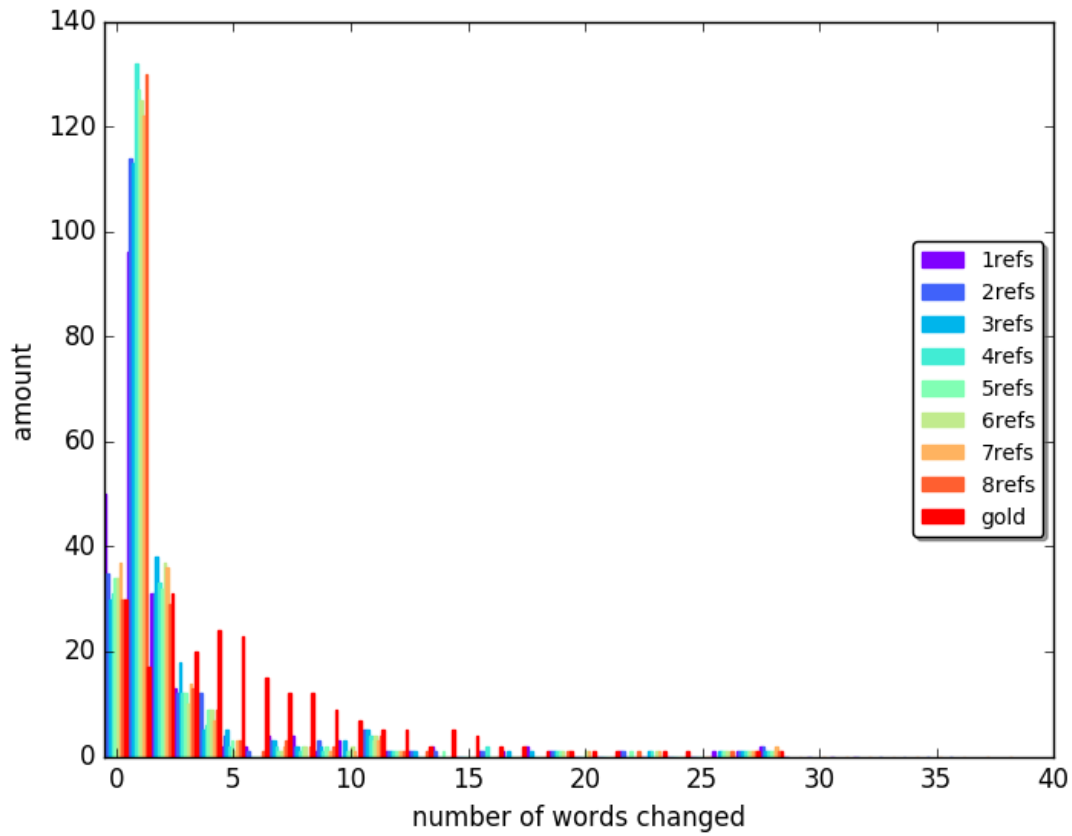


Figure 6: Reranking results with Nisioi et al. (2017) and SARI

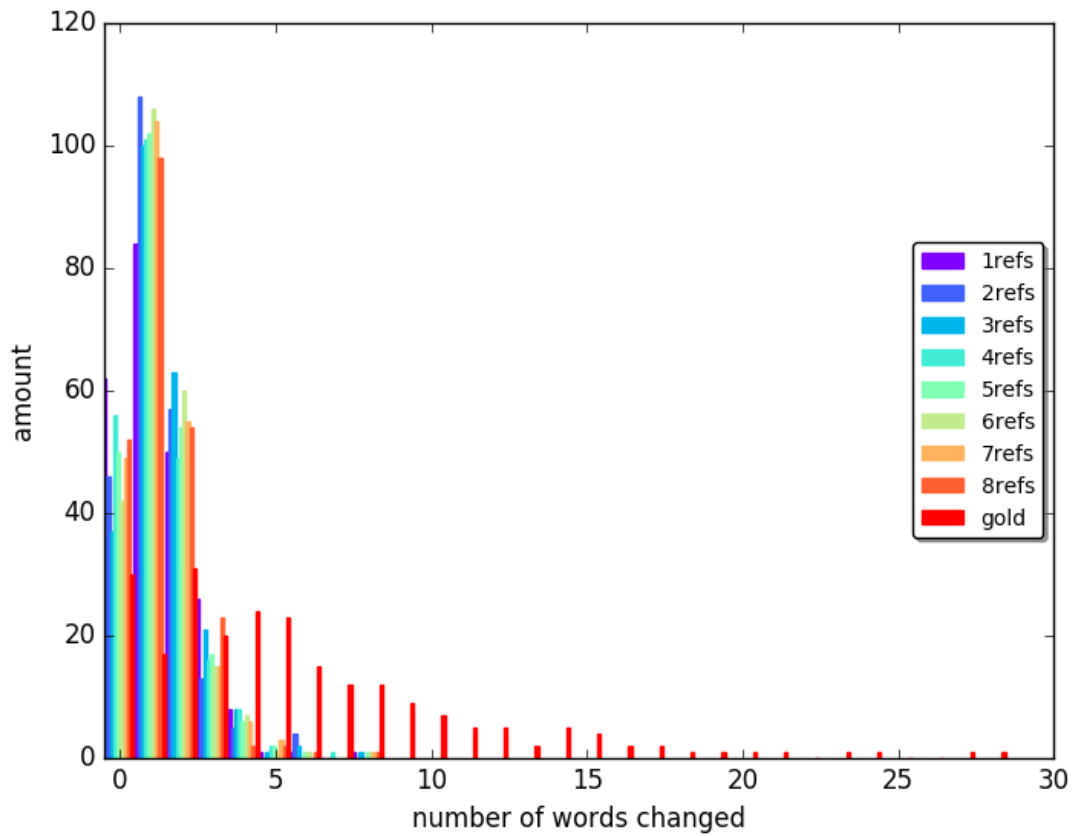


Figure 7: Reranking results with Moses and MAX-SARI

References

- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572. Association for Computational Linguistics.
- Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.
- Yili Hong. 2013. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51.
- Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Austin, Texas. Association for Computational Linguistics.
- Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 85–91.
- Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proc. of ACL*, pages 2205–2215.
- Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.
- 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.