

Evaluating Factual Consistency of Texts with Semantic Role Labeling

Jing Fan* and Dennis Aumiller* and Michael Gertz

Institute of Computer Science, Heidelberg University

j.fan@stud.uni-heidelberg.de

{aumiller, gertz}@informatik.uni-heidelberg.de

Abstract

Automated evaluation of text generation systems has recently seen increasing attention, particularly checking whether generated text stays truthful to input sources. Existing methods frequently rely on an evaluation using task-specific language models, which in turn allows for little interpretability of generated scores. We introduce **SRLScore**, a reference-free evaluation metric designed with text summarization in mind. Our approach generates fact tuples constructed from Semantic Role Labels, applied to both input and summary texts. A final factuality score is computed by an adjustable scoring mechanism, which allows for easy adaption of the method across domains. Correlation with human judgments on English summarization datasets shows that **SRLScore** is competitive with state-of-the-art methods and exhibits stable generalization across datasets without requiring further training or hyperparameter tuning. We experiment with an optional co-reference resolution step, but find that the performance boost is mostly outweighed by the additional compute required. Our metric is available online at: <https://github.com/heyjing/SRLScore>

1 Introduction

One of the remaining issues that prevents productive deployments of neural text summarization systems is the low correlation of system outputs with human preferences. Among those, *factuality*, i.e., the agreement of facts in the generated summaries with those present in the input text, is not part of the general training objectives of models, which frequently leads to hallucinated facts that are detrimental to perceived system performance (ter Hoeve et al., 2020; Fabbri et al., 2021). Prior work has therefore introduced metrics for automated testing of factuality in generated text (Goodrich et al., 2019; Kryscinski et al., 2020; Yuan et al., 2021),

which allows for a more nuanced verification of model capabilities. In particular, one of the first relevant works by Goodrich et al. (2019) introduces the idea of representing text as a series of "fact tuples", in their case as (subject, predicate, object) triplets. Their method exhibits some assumptions about the underlying data, which hampers correlation with human ratings. For example, subject or object may vary for the same sentence meaning expressed using different syntactic structures, e.g., active and passive forms. Semantic Role Labeling (SRL), however, allows for a syntactically independent meaning representation. Our metric, **SRLScore**, improves factuality evaluation, building on fact tuples similar to Goodrich et al. It distinguishes itself in several ways from existing approaches, though:

1. To account for a more nuanced fact representation, we employ SRL to produce abstract representations of sentences that are *independent of their syntactic formulations*.
2. Fact tuples in **SRLScore** are generated on the *input text* instead of gold summaries; as a consequence, our method is reference-free, and may be applied for evaluation irrespective of the availability of labeled datasets.
3. We introduce a novel weighting scheme for fact tuple comparison, where adjustable weights allow for user optimization.
4. Finally, we experiment with extensions along different parts of the pipeline, including an optional co-reference resolution step and alternative similarity scoring functions.

Notably, **SRLScore** entirely relies on publicly available software components and may be used without any further domain adaption required. While our experiments are performed on English, we argue that the transfer of our approach to other languages is possible given only the existence of a language-specific tokenizer and a sufficiently good SRL tagger. Furthermore, **SRLScore** offers the

*Both authors contributed equally to this work.

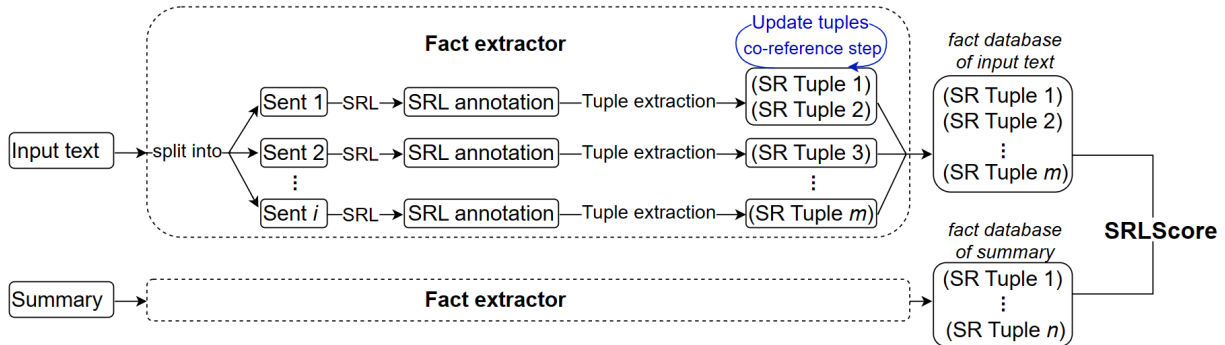


Figure 1: Visual explanation of **SRLScore**. An input text and its associated summary are transformed into a series of fact tuples (*SR Tuple*) through extraction from SRL (and optional co-reference) annotations. The final factuality score is computed based on the similarity of the summary facts with fact tuples generated from the input text.

additional benefit of being an *interpretable* metric, due to its composition on top of fact tuples. In comparison, metrics used for factuality evaluation that are based on the intermediate presentations of language models, e.g., *generation perplexity* (Zhang et al., 2020; Thompson and Post, 2020; Yuan et al., 2021), cannot present insightful reasons *why* a particular score was achieved. Furthermore, it has been empirically demonstrated that generation-based evaluators exhibit a *self-preference* of outputs generated by models similar to the factuality evaluator (Fabbri et al., 2021; Liu et al., 2023). This makes them a questionable choice over interpretable metrics. We empirically show that the correlation of **SRLScore** with human ratings is on par with existing methods, and perform several ablations to study the impact of algorithmic choices within our pipeline.

2 Related Work

Automated analysis of (abstractive) summaries became more relevant in recent years, with the influx of generic summarization systems becoming available (Nallapati et al., 2016; See et al., 2017; Lewis et al., 2020). In particular, Goodrich et al. (2019) were the first to propose a reference-based estimator for factuality of generated summaries. As mentioned, their approach is based on a tuple representation of "facts" in the generated and gold summary. Fact tuples are extracted based on a weakly supervised end-to-end tagger and subsequently compared on the basis of matching arguments. Notably, no readily available implementation of their method currently exists.

Later work has proposed alternative metrics based on textual entailment (Falke et al., 2019; Mishra

et al., 2021) and Question Answering (QA) (Wang et al., 2020; Durmus et al., 2020), where agreement of answers to questions on the reference and summary are used for estimating factuality. However, QA-based metrics require additional task-specific fine-tuning on generic datasets, which makes the adoption to new domains fairly expensive.

The only other work that to our knowledge utilizes some form of SRL-based factuality estimation is presented by Fischer et al. (2022). In comparison to **SRLScore**, their method aggregates "role buckets" at the document level, instead of creating sentence-specific fact tuples. Empirically, their implementation has lower correlation with human ratings than compared approaches, which is contrary to our own findings.

Li et al. (2022) frame factuality estimation as an in-filling task, where fact statements are withheld as masked tokens in a generated summary, and a separate model is trained to predict missing facts. Notably, this relies on the assumption that the majority of factual mistakes stems from noun phrases and entity mentions (Pagnoni et al., 2021).

An alternative body of literature has explored the possibility to exploit Language Models (LMs) directly for estimating factual consistency: Some works, such as BertScore (Zhang et al., 2020), use LM-generated representations to generate alignments for scoring. In comparison, PRISM (Thompson and Post, 2020) or BARTScore (Yuan et al., 2021) directly use model perplexity as a factuality estimate. Xie et al. (2021) explore masking approaches, which fall somewhere between the works of Li et al. (2022) and BARTScore; their framing of counterfactual estimation still relies on model-based likelihood scores for computation.

The majority of prior work expresses metric perfor-

Sentence 1								
Mueller	gave	a book	to	Mary	yesterday	in Berlin	secretly	
Agent	Verb	Patient		Recipient	Time	Location	Manner	

Sentence 2								
A book	was	given	to	Mary	by Mueller	yesterday	in Berlin	secretly
Patient		Verb		Recipient	Agent	Time	Location	Manner

Sentence 3								
Mueller	met	with	senators	in a private room	to provide more details			
Agent	Verb		Patient	Location	Purpose			

Mueller	met	with	senators	in a private room	to	provide	more details
Agent						Verb	Patient

Figure 2: Examples of semantic role label annotations. Labels may remain consistent across different syntactic forms (Sentence 1 & 2). A single sentence can also include several relations at the same time (Sentence 3).

mance in terms of correlation with human factuality ratings. Notably, annotations exist for subsets of the popular CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2017) and XSUM summarization corpora (Narayan et al., 2018). Where Wang et al. (2020) collect user annotations from crowd workers, Fabbri et al. (2021) additionally sample expert judgments, and find that expert ratings tend to be more representative. Maynez et al. (2020) study several aspects of summarization evaluation beyond just factuality, but do not disclose the background of annotators for evaluation.

Generally, reliably evaluating correlation of summarization metrics with human preferences is no easy task, either: Deutsch et al. (2022) show that system-level evaluation metrics for text summarization rarely outperform simplistic metrics, such as ROUGE (Lin, 2004), to a statistically significant degree. Partially, this can be attributed to the small number of human-annotated samples available, generally less than 1000 different instances.

3 SRLScore

Our factual consistency metric, called **SRLScore**, is implemented as a two-stage process: first, extracting fact tuples using Semantic Role Labeling (SRL) on both the source texts and the summary texts, and then determining a factuality score based on tuple comparison. The measure outputs human-interpretable scores between 0 and 1, where a higher score indicates greater factual consistency of a summary text. In this section, we detail the algorithmic choices and present an adaptive weighting scheme for computing the final factuality scores.

3.1 Generating Fact Tuples with Semantic Role Labeling

As Figure 1 shows, we operate on the sentence level, primarily because existing SRL tools work well on this level of granularity (Shi and Lin, 2019; Xu et al., 2021). The goal of our fact extractor is to produce a *fact database* comprised of semantic role tuples for each input text.

The primary task of SRL is to find all role-bearing constituents in a sentence and label them with their respective roles (Márquez et al., 2008). Typical semantic roles include *agent*, *patient/theme*, *recipient*, *goal*, *instrument*, *manner*, *time*, *location* and so on. From the many semantic labels available, we include seven roles based on availability in tagging schemes to construct a fact tuple: *agent*, *negation*, *relation*, *patient*, *recipient*, *time*, and *location*. We further note that not every sentence needs to contain *all* of these roles; absent labels are represented by *None* in this work. Importantly, roles reveal the semantic relations between a predicate (verb) and its arguments, which implies that one can generate several fact tuples from a single sentence, depending on the number of verbs in it. To illustrate an exemplary fact tuple, the extracted semantic tuple from sentence 1 in Figure 2 is (Mueller, None, gave, a book, Mary, yesterday, in Berlin).

3.2 Scoring Texts by Comparing Fact Tuples

Once fact tuples for both the input and summary texts are generated, the second step in our pipeline is to compute a factual accuracy score. We implement a dynamic weighting system, which crucially improves over a naive comparison, as we empirically show in Section 4.6. Furthermore, we describe the drop-in replacements for exact matching during similarity computation.

Scoring Algorithm. Given an input text R and summary text S , let F_R and F_S be *fact databases*, representing the semantic information contained in R and S , respectively. Individual fact tuples are represented as an ordered list of fact arguments, e.g., $f = (\text{agent}, \text{negation}, \text{relation}, \text{patient}, \text{recipient}, \text{time}, \text{location}) \in F$. Particular arguments in a fact tuple are referred to by their index position, meaning $\text{agent} = f^0$, $\text{negation} = f^1$, and so on. We further assume that there exists a scoring function that expresses the *factual support of summary tuple* f_s , given an input tuple f_r , denoted as $S(f_s|f_r)$. To obtain a factuality score, we attempt to extract the best match $\hat{f}_r \in F_R$ for each sum-

mary fact $f_s \in F_s$ where \hat{f}_r maximizes the support score $S(f_s|\hat{f}_r)$. Importantly, we differ from, e.g., Goodrich et al. (2019), by considering the entirety of F_R , instead of subsets that match both the agent and relation of the fact tuple. The factual accuracy is then the average across all maximized tuple scores in F_S . With that, **SRLScore** is defined as:

$$\text{SRLScore}(R, S) := \frac{1}{|F_S|} \sum_{f_s \in F_s} \max_{f_r \in F_R} S(f_s|f_r) \quad (1)$$

The final part of this scoring system is the computation of factual support $S(f_s|f_r)$. Tuples are scored by comparing the corresponding attributes of each tuple, formally:

$$S(f_s|f_r) := \sum_i \mathbb{1}_{f_s^i \neq \text{None}} \cdot \text{sim}(f_s^i, f_r^i) \cdot w_i, \quad (2)$$

where the summation over i addresses all attributes of the fact tuples, $\mathbb{1}_{f_s^i \neq \text{None}}$ represents an indicator function considering only non-empty arguments f_s^i (zero otherwise), and w_i assigns static weights to arguments in position i . Generally, it should be assumed that the weights allow for a maximum factuality score of 1, i.e., $\sum_i w_i = 1$. Finally, $\text{sim}(f_s^i, f_r^i)$ is the pairwise argument similarity of f_s^i and f_r^i . We consider different similarity metrics, as described in the following paragraphs.

Dynamic Weighting System. The generic weighting in Equation (2) does not necessarily apply to the particular case of evaluating factual consistency in summarization, since a summary is still factually correct even if it leaves out particular aspects (e.g., dropping the date of an event), which were present in the input text. With static weights, however, absent arguments are still contributing to the scoring of the tuple f_s , which means that leaving arguments out might potentially be considered as a penalization of factuality. To address this issue, we introduce a weight re-normalization factor, W_{norm} , that distributes the static weights w_i across only those attributes that are present in the current summary fact. In particular, this also increases penalties for actual mistakes over simple fact omission. The weight normalization is defined as follows:

$$W_{norm} := \frac{1}{\sum_i \mathbb{1}_{f_s^i \neq \text{None}} \cdot w_i} \quad (3)$$

With re-normalization enabled, we replace the existing computation of $S(f_s|f_r)$ by the product $W_{norm} \cdot S(f_s|f_r)$.

String Similarity Methods. We experiment with different methods to calculate the pairwise similarity $\text{sim}(f_s^i, f_r^i)$: exact matching (in line with prior work), but also approximate matching functions, such as word vector similarity¹ and ROUGE-1 precision (Lin, 2004). Computation of similarity with vectors and ROUGE each have their own respective strengths. Word vectors offer the highest flexibility in terms of recognizing argument similarity, enabling semantic comparison instead of purely syntactic equivalence. ROUGE-1 similarity does not offer the same level of flexibility in terms of matching, but shines with its comparatively faster computation, while still recognizing partial matches.

3.3 Improved Surface Form Invariance with Co-reference Resolution

In light of the fact that sentence-level SRL extraction misses co-references of the same entity across the texts, we integrate an optional component that takes co-reference resolution into account during the tuple generation. Concretely, we employ an off-the-shelf co-reference resolution tool (Lee et al., 2017) to identify and store all reference clusters in an external *entity dictionary*. There, all linguistic expressions that refer to the same entity will be grouped together, which allows for later disambiguation. As shown in Figure 3, if an extracted semantic role tuple contains co-references, a single fact tuple will be *expanded* into multiple tuples, representing the Cartesian product over all synonymous entity surface forms.

The key idea here is to enable a better matching of potential facts across input texts and summaries, effectively increasing the recall of matches. The disadvantage is that this directly affects the runtime of our method by a strong factor, since the additional tuples in F_S and F_R will undoubtedly increase the number of comparisons.

4 Experiments

We empirically demonstrate the performance of our method through a number of experiments on two popular datasets for factual consistency evaluation, which are covered in this section. We further share implementation details and the choices for extracting SRL tuples and extracting co-reference clusters.

¹We use spaCy’s vector similarity, see <https://SpaCy.io/usage/linguistic-features#vectors-similarity>, last accessed: 2023-03-06.

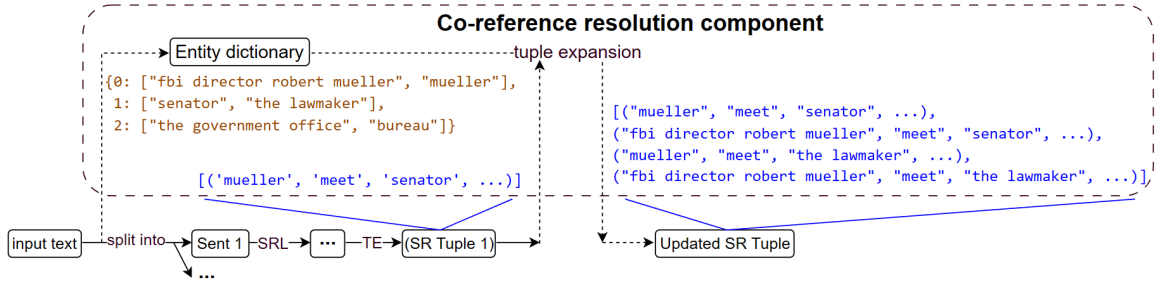


Figure 3: Example of the tuple expansion step through co-reference resolution. In addition to the original SR tuple, we add tuples with all possible permutations of the surface forms of mentioned entities.

In addition to the experimental analysis, we also study the behavior of **SRLScore** through a number of ablation experiments and a brief error analysis.

4.1 Evaluation Datasets

QAGS (Wang et al., 2020). The dataset comprises of two separate splits: the first contains 235 instances collected from the test split of CN-N/DailyMail (Nallapati et al., 2016), where each instance contains a source article and a model-generated summary using the bottom-up approach by Gehrmann et al. (2018). A secondary set contains 239 further instances from the test split of XSUM (Narayan et al., 2018), with generated summaries sampled from BART (Lewis et al., 2020).

SummEval (Fabbri et al., 2021). It includes synthetic summaries from 16 different abstractive and extractive models of 100 randomly selected articles from the test split of CNN/DailyMail. Unlike QAGS, which collected annotations from MTurk², each SummEval sample was evaluated by five crowd-sourced annotators and three experts. For each summary, judges were asked to evaluate the coherence, consistency, fluency and relevance. For our evaluation, we use the expert ratings with regard to factual consistency as the gold score, based on the recommendation by Fabbri et al. (2021).

4.2 Evaluation Metrics and Significance

In line with prior work, we evaluate metrics by computing Pearson correlation (denoted as ρ) and Spearman correlation (denoted as s) between model predictions and human reference ratings. Given the limited size of all considered evaluation datasets, we further test results for significance using permutation tests (Riezler and Maxwell, 2005; Deutsch et al., 2021), following the recommendation of Dror et al. (2018). In all tables, [†] denotes

²<https://www.mturk.com/>, last accessed: 2023-03-06.

a significance level of 0.05 ($p < 0.05$) and [‡] a level of 0.01 ($p < 0.01$). When testing significance against several systems, we further apply Bonferroni correction of significance levels (Dunn, 1961).

4.3 Implementation

We use AllenNLP (Gardner et al., 2018), specifically version 2.1.0, to extract semantic role labels. AllenNLP implements a BERT-based SRL tagger (Shi and Lin, 2019), with some modifications. The output of AllenNLP uses PropBank convention (Palmer et al., 2005; Bonial et al., 2012; Pradhan et al., 2022), which lists for each verb its permitted role labels using numbered arguments (*ARG0*, *ARG1*, ...) instead of names, due to the difficulty of providing a small, predefined list of semantic roles that is sufficient for all verbs. Since numbered arguments are meant to have a verb-specific meaning (Yi et al., 2007), this implies that our mapping between numbered arguments and semantic roles may not always be consistent. The exact mapping used in our experiments is detailed in Appendix A. For co-reference, we similarly use the model provided by AllenNLP (Lee et al., 2017), which matches the output format of the SRL tagger.

All experiments were carried out on a system with an Intel Xeon Silver 4210 CPU, two TITAN RTX GPUs (24 GB GPU VRAM each) and 64 GB of main memory. We run inference for the SRL model and co-reference component on separate GPUs.

We report scores of all system and baseline variants across a single random seed only. Since we are comparing provided "plug-and-play" metrics, it is reasonable to assume that these are the primary choice for others evaluating their own datasets. Particularly for **SRLScore**, we further note that due to the system design, no fine-tuning or training is necessary. The only parameters varied during the experiments are thus the argument weights, which we describe in the following section.

Metrics	QAGS-CNN/DM		QAGS-XSUM		SummEval		Avg.
	ρ	s	ρ	s	ρ	s	
ROUGE-1 (F1)	0.34	0.32	-0.01	-0.05	0.13	0.14	0.15
BLEU	0.13	0.33	0.08	0.03	0.09	0.14	0.10
METEOR	0.33	0.36	0.06	0.01	0.12	0.14	0.17
BARTScore	0.65	0.57	0.00	0.02	0.27	0.26	0.31
BARTScore _{cnn}	0.73	0.68	0.19	0.18	0.35	0.32	0.42
BARTScore _{cnn+para}	0.69	0.62	0.07	0.07	0.42	0.37	0.39
CoCo _{span}	0.64	0.55	0.22	0.20	0.40	0.35	0.42
CoCo _{sent}	0.68	0.59	0.16	0.14	0.39	0.35	0.41
ClozE-R _{en_core_web_trf} *	0.66	-	0.32	-	0.47	-	0.48
ClozE-R _{confidence} *	0.65	-	0.29	-	0.48	-	0.47
SRLScore _{base}	0.67	0.59	0.20	0.18	0.43	0.33	0.43
SRLScore _{coref}	0.65	0.58	0.27	0.26	0.43	0.32	0.45
SRLScore _{coref-optimized}	-	-	0.33	0.33	-	-	-

Table 1: Pearson (ρ) and Spearman (s) correlation of metrics with human ratings on the evaluated datasets. Bold scores indicate highest absolute values. For **SRLScore** variants, we report highest scores across all similarity functions. No significant differences were found between the correlation scores of factuality-specific metrics.

*: results were taken from the respective paper, as there is no existing code to reproduce their results as of now.

4.4 System Variants

We compare with a number of generic automatic evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Besides, we also consider several metrics specifically developed for factuality estimation, which have reported prior state-of-the-art correlation. Wherever possible, we reproduce scores with the official scripts provided by authors. Comparison is done with three variants of BARTScore (Yuan et al., 2021), two variants of CoCo (Xie et al., 2021), and two variants of ClozE (Li et al., 2022). For more details on reproducibility, see Appendix B. We chose each variant such that the highest self-reported scores of each paper on all evaluated datasets are considered.

For our own method, SRLScore_{base} represents a default setting, assigning equal weights $w_i = \frac{1}{7}$ to all attributes (*agent, negation, relation, patient, recipient, time, location*); the respective similarity function (exact match, spaCy vector, or ROUGE similarity) is chosen to maximize dataset-specific performance (see results of Table 2). SRLScore_{coref} uses the same weights, with co-reference enabled. We further provide model ablations to test various specifications of our models. As we could not find a implementation based on the original tuple extraction approach by Goodrich et al. (2019), we introduce SRLScore_{openie} and SRLScore_{goodrich} as approximations of their method. Here, fact tuples are reduced to (agent, relation, patient) triplets

(with equal weights $w_i = \frac{1}{3}$). We note that this is not a true equivalence to the original method, although "[i]n most English sentences the subject is the agent" (Bates and Macwhinney, 1982); in reality, a broader variety of roles in the subject position may be encountered. The same applies for our mapping between *object* and the *patient* role. However, by using the same upstream labeling tool (i.e., the SRL model provided by AllenAI), we may more accurately compare the algorithmic scoring methods, independent of the annotation accuracy. We argue that our SRL-based modeling of relationship triplets allows for a better generalization beyond Wikipedia, which Goodrich et al. were using in their own experiments.

The difference of SRLScore_{openie} and SRLScore_{goodrich} lies in the implemented scoring function, where the OpenIE variant employs our own scoring algorithm, SRLScore_{goodrich} uses the preliminary filtering step defined in Goodrich et al. (2019). We do not apply a co-reference system in either one of the two ablation settings. Finally, SRLScore_{coref-optimized} illustrates the possibility of adapting our method to a particular dataset. For this variant, we optimize available hyperparameters (weights, scoring function, co-reference) in order to obtain the highest possible scores.

4.5 Main Results

The central evaluation results with recommended default settings are shown in Table 1. In almost all cases, specialized factuality metrics show higher

correlation than generic summarization evaluation metrics (ROUGE-1, BLEU and METEOR). Notably, despite the high increase in absolute scores, we do not always detect a significant level of improvement between factuality-specific metrics and generic metrics, particularly on QAGS-XSUM; we will discuss further implications of this in more detail later. When testing our own method, $\text{SRLScore}_{\text{base}}$, against generic metrics, we find strongly significant improvements only for Pearson correlation of QAGS-CNN/DM and SummEval, as well as Spearman correlation on SummEval ($p < 0.01$, with Bonferroni correction).

It should be further noted that $\text{BARTScore}_{\text{cnn}}$ and CoCo results use BART models (Lewis et al., 2020) that were fine-tuned on the CNN/DailyMail corpus (respectively a variant fine-tuned on XSUM for CoCo on QAGS-XSUM); this may shift the results in favor of these methods for the particular dataset. In comparison, **SRLScore** does not make such assumptions, which may indicate a potentially stronger generalization to unseen datasets.

The results in Table 1 also show that there are no significant differences between any of the factuality-specific metrics (**SRLScore**, BARTScore , and CoCo), particularly after applying Bonferroni correction for the comparison against several methods. These insights open up discussions about the current claims of "state-of-the-art" performance, which may not be easily distinguished on the current evaluation datasets. We admit that there is likely no trivial solution to this (besides further annotations), as the main problem seems to stem from the high variance on small sample sizes.

4.6 Ablation Study

Given the limited expressiveness of the generic result evaluation, we perform a series of ablation studies on **SRLScore**, to support the individual algorithmic choices made in our method.

Extending Tuple Attributes. We investigate the assumption that semantic representations of sentences are usually far more complicated than the simplistic view of (*agent, relation, patient*) triplets, and the fact that errors may involve further roles. To this end, we compared $\text{SRLScore}_{\text{openie}}$, using a triplet representation, against $\text{SRLScore}_{\text{base}}$ with seven roles. The results in Table 2 confirm that extending tuples to cover more semantic roles is effective across datasets and metrics; $\text{SRLScore}_{\text{base}}$

Metrics		QCNNDM		QXSUM		SummE	
		ρ	s	ρ	s	ρ	s
SRLScore _{openie}	Exact	0.59	0.51	0.09	0.09	0.34	0.28
	ROUGE	0.62	0.56	0.07	0.07	0.41	0.32
	SpaCy	0.59	0.53	0.13	0.10	0.37	0.32
SRLScore _{base}	Exact	0.61	0.54	0.14	0.15	0.37 [†]	0.31 [‡]
	ROUGE	0.67	0.59	0.15 [†]	0.13	0.43[†]	0.33
	SpaCy	0.63	0.55	0.20	0.18	0.40 [†]	0.34[†]

Table 2: Comparison of **SRLScore** with a simplified triplet representation ($\text{SRLScore}_{\text{openie}}$). Extending the fact tuples strictly improves correlation with human ratings across all similarity functions. Significance markers indicate improvements over the same similarity function of the openie variant.

Weight Setting	QCNNDM		QXSUM		SummE	
	ρ	s	ρ	s	ρ	s
Static weights	0.59	0.49	0.09	0.09	0.38	0.28
Dynamic weights	0.67	0.59	0.20	0.18	0.43	0.33

Table 3: Correlation scores of $\text{SRLScore}_{\text{base}}$ with and without weight re-normalization enabled.

scores consistently better than $\text{SRLScore}_{\text{openie}}$, with significant improvements primarily on SummEval (the largest considered dataset).

Performance of Similarity Functions. Also seen in Table 2 is the difference in scores across various similarity functions. **SRLScore** achieves generally higher correlation when using vector (spaCy) or ROUGE similarity over exact matching, although not to a significant degree. These observations can be attributed to the hypothesis that abstractive entity references will not be detected by exact matching. Also note that results on QAGS-XSUM are particularly affected by this, which shows higher levels of abstraction than CNN/DM-derived resources (Wang et al., 2020; Pagnoni et al., 2021). This is also visible for the $\text{SRLScore}_{\text{coref}}$ variant, as seen in Table 1, which can further improve the matching of re-formulations.

Dynamic Weight Re-Normalization. We next analyze the contribution of our dynamic weighting scheme through removing the weight re-normalization W_{norm} and instead defaulting to a static weighting on $\text{SRLScore}_{\text{base}}$. Results in Table 3 demonstrate that re-distributing static weights dynamically to present roles is very effective, however, results show no statistical significance.

Scoring Method	QCNNDM		QXSUM		SummE	
	ρ	s	ρ	s	ρ	s
SRLScore _{goodrich}	0.45	0.38	0.05	0.07	0.29	0.24
SRLScore _{openie}	0.62[†]	0.56[†]	0.13	0.10	0.41[‡]	0.32[†]

Table 4: Results of the ablation experiment comparing the scoring method by Goodrich et al. (2019) with our proposed scheme, based on triplet representations.

SRLScore		BARTScore		
base	coref	base	cnn	cnn+para
2.35	19.32	0.22	0.23	0.23

Table 5: Average processing time (in seconds) per instance in QAGS-CNN/DM. SRLScore uses ROUGE similarity. BARTScore is run with a batch size of 4.

Ablation of Goodrich Scoring Method. We finally examine the performance of our scoring system against the partial matching approach of Goodrich et al. For fairness, we compare results on the reduced triplet sets. SRLScore_{openie} uses the presented weighting function, SRLScore_{goodrich} implements an equivalent scoring to Goodrich et al. Results in Table 4 show that the presented scoring algorithm performs better than the scores determined by Goodrich’s approach on different datasets, in most instances to a significant degree.

Performance of Co-reference Resolution System. Results in Table 1 reveal that the co-reference system is not always improving scores, particularly on the CNN/DailyMail-derived datasets. However, the use of co-reference resolution will significantly increase the processing time, as shown in Table 5. This is expected, given that there are now more fact tuples due to the *tuple expansion*; since the presented scoring method requires the comparison of each fact tuple in the summary against *all* input text tuples. We further compare the runtime against BARTScore, which only requires a single forward-pass through a neural net and can be batched easily, resulting in a 10x speed-up. In contrast, SRLScore requires construction and comparison the fact tuples, which are the main contributors for slower inference times.

4.7 Error Analysis

To better understand the limitations of our presented methods, we examine a number of instances manually, particularly those where there are large

differences between model-generated scores and human annotations on QAGS-XSUM. Table 6 shows two instances, where SRLScore respectively predicts a much higher and lower factuality score than human annotators. Notably, human raters tend to drastically reduce factuality scores in the presence of even a single mistake (what we refer to as *strike-out scoring*). In comparison, SRLScore and other factuality metrics tend to be more heavily influenced by the correctness of the *majority* of attributes, which can be seen as a *bottom-up scoring* (scores are built up from an initial factuality of zero instead of deducing from an initial score of one). On the other hand, highly abstractive samples, which retain factuality according to human raters, may pose a challenge for tuple-based SRLScore. In the second example of Table 6, synonymous expressions like *step down* instead of *resign* cause low predicted similarity; potential solutions could be found in verb sense disambiguation (Brown et al., 2011, 2022).

5 Conclusion and Future Directions

In this work, we presented a semantically consistent metric for estimating the factual truthfulness of two pieces of text: we applied our presented metric to the problem of text summarization evaluation, and demonstrated that it performs on par with existing approaches. In fact, we find that due to the small sample sizes of evaluation datasets, there are no significant differences between any of the considered state-of-the-art factuality estimation metrics. Our approach strikes with its relative simplicity and interpretability due to the intermediate representation of "fact tuples", which makes it possible for human annotators to review how or why system decisions were made. Furthermore, we have demonstrated the suitability of our approach over more naive tuple-based scoring methods through a series of ablation experiments, which also show the adaptability of our method to particular unseen settings by simply adjusting a series of parameters.

In our opinion, there are two key challenges concerning the effective deployment of SRLScore. The current implementation still suffers from impractically long runtimes for longer input texts. Notably, however, both the tuple generation and comparison stages can be parallelized and we are currently working on improving the compute effi-

	Sample Text	Extracted Fact Tuples	Human SRLScore	
Input	Former England fast bowler Chris Tremlett has announced his retirement ...	(Former England fast bowler chris tremlett, announce, his retirement, ...)	0	0.87
Summary	Former England seamer James Tremlett has announced his retirement ...	(Former England seamer james tremlett, announce, his retirement, ...)		
Input	The head of Japanese advertising group Dentsu is to step down following the suicide of an employee ...	(The head of japanese advertising group dentsu, step, ..., following the suicide of an employee, ...)	1	0.10
Summary	The chief executive of Japanese advertising firm Dentsu will resign after a worker killed herself ...	(The chief executive of japanese advertising firm dentsu, resign, ..., after a worker killed herself, ...), (a worker, killed, herself, ...)		

Table 6: Examples from the QAGS-XSUM dataset where the majority vote of human ratings differs strongly from **SRLScore**'s predicted factuality. Colored text segments highlight the position of relevant facts, where red text indicates a factual discrepancy between input and summary segments.

ciency of our method. Secondly, we have seen a general trend that factuality estimation metrics are scoring differently from human annotators, who are putting heavy emphasis on a *completely* factual summary instead. We suspect that adopting a similar *strike-out scoring* for estimation may better correlate with human ratings, although it will require sufficiently accurate taggers to ensure correct recognition of all entities.

Limitations

While the presented method exhibits stable correlation with human judgments on some of the evaluated datasets, it still exhibits instances under which it will predict opposing factuality scores. It should therefore be considered an *addition* to human evaluation, but at this point not fully replace it.

We also want to point out that the underlying summarization datasets that were used to compare human ratings on are known for their own set of limitations, particularly being fairly extractive in nature. This plays well with **SRLScore**'s estimation of matching between individual tuples extracted from single sentences; on the other hand, if summary texts contain facts derived from multiple source sentences (or undergo otherwise complex structural changes), fact tuples may be insufficient in their current form.

Another limitation is the expressiveness of results on the fairly small human-annotated datasets. Here, statistically significant differences can rarely be obtained. However, we are to our knowledge the first to demonstrate this insight about (significant) differences between existing methods, which we consider a particularly useful insight for future work. We further want to point out that our method was only evaluated on English datasets; we argue that it

can be applied to other languages, given a similarly performing SRL labeling model. In practice, however, the existence of available models is currently limited for non-English languages.

Ethics Statement

The paper considers the automated analysis of factuality in generated text. While we see no imminent risk in the development of our presented method, we want to point to the explicitly spelled out limitations of the current method (see the previous section). The blind application of factuality metrics could be considered harmful in instances where the predicted scores are differing strongly from human ratings. We therefore recommend that factuality metrics should be employed purely as a *complementary* evaluation, and never directly replace analysis with humans in the loop.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. The work of Jing Fan is supported by a scholarship of the China Scholarship Council (CSC).

References

- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Elizabeth Bates and Brian Macwhinney. 1982. Functionalist approaches to grammar. *Child Language: The State of the Art*.

- Claire Bonial, Jena Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 48.
- Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. [Semantic representations for NLP using verbnet and the generative lexicon](#). *Frontiers Artif. Intell.*, 5:821697.
- Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. 2011. [VerbNet class assignment as a WSD task](#). In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. [Re-examining system-level correlations of automatic summarization evaluation metrics](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.
- Esin Durmus, He He, and Mona Diab. 2020. [FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. [Ranking generated summaries by correctness: An interesting but challenging application for natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.
- Tim Fischer, Steffen Remus, and Chris Biemann. 2022. [Measuring faithfulness of abstractive summaries](#). In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. [AllenNLP: A deep semantic natural language processing platform](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.
- Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. [Bottom-up abstractive summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.
- Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. [Assessing the factual accuracy of generated text](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM.
- Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.
- Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. [End-to-end neural coreference resolution](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

- 7871–7880, Online. Association for Computational Linguistics.
- Yiyang Li, Lei Li, Qing Yang, Marina Litvak, Natalia Vanetik, Dingxin Hu, Yuze Li, Yanquan Zhou, Dongliang Xu, and Xuanyu Zhang. 2022. [Just cloze! A fast and simple method for evaluating the factual consistency in abstractive summarization.](#) *CoRR*, abs/2210.02804.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries.](#) In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using GPT-4 with better human alignment.](#) *CoRR*, abs/2303.16634.
- Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. [Semantic role labeling: An introduction to the special issue.](#) *Comput. Linguistics*, 34(2):145–159.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. [Looking beyond sentence-level natural language inference for question answering and text summarization.](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. [Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.](#) In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence rnns and beyond.](#) In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 280–290. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. [Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics.](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. [The Proposition Bank: An annotated corpus of semantic roles.](#) *Computational Linguistics*, 31(1):71–106.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation.](#) In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O’gorman, James Gung, Kristin Wrightbettner, and Martha Palmer. 2022. [PropBank comes of Age—Larger, smarter, and more diverse.](#) In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.
- Stefan Riezler and John T. Maxwell. 2005. [On some pitfalls in automatic evaluation and significance testing for MT.](#) In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. 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.
- Peng Shi and Jimmy Lin. 2019. [Simple BERT models for relation extraction and semantic role labeling.](#) *CoRR*, abs/1904.05255.
- Maartje ter Hoeve, Julia Kiseleva, and Maarten de Rijke. 2020. [What makes a good summary? reconsidering the focus of automatic summarization.](#) *CoRR*, abs/2012.07619.
- Brian Thompson and Matt Post. 2020. [Automatic machine translation evaluation in many languages via zero-shot paraphrasing.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. [Asking and answering questions to evaluate the factual consistency of summaries.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. [Factual consistency evaluation for text summarization via counterfactual estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kun Xu, Han Wu, Linfeng Song, Haisong Zhang, Linqi Song, and Dong Yu. 2021. [Conversational semantic role labeling](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2465–2475.

Szu-ting Yi, Edward Loper, and Martha Palmer. 2007. [Can semantic roles generalize across genres?](#) In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 548–555. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BartScore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BertScore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Mapping of PropBank Arguments to Semantic Role Tuple Attributes

In our implementation, we extract sentence spans with label ARG0 as *agent* and spans with label ARG1 as *patient*. The extraction of *time* and *location* also does not pose any difficulties, because ARGM-TMP and ARGM-LOC are both given as modifiers that remain relatively stable across predicates (Jurafsky and Martin, 2009). However, as shown in Table 7, there is no one-to-one relationship between numbered arguments and the *recipient* role. For the sake of simplicity, we extracted elements with label ARG2 as *recipient*, because the probability that ARG2 correlates to *recipient* is the highest among all other possible roles (Yi et al., 2007).

ARG0 agent	ARG1 patient
ARG2 instrument, recipient, attribute	ARG3 starting point, recipient, attribute
ARG4 ending point	ARGM modifier

Table 7: Mapping between numbered arguments in PropBank and semantic roles (Bonial et al., 2012). Particularly the mapping of argument 2 makes simplifying assumptions about different verb forms.

B Reproducing Scores of Related Work

We use the official scripts provided by the authors of BARTScore³ and CoCo⁴. Unfortunately, no public implementation exists at the time of writing for the work of Li et al. (2022), which prevents significance testing against Cloze models. For the work by (Goodrich et al., 2019), we similarly found no publicly available implementation; however, we note their wikipedia-based training data for generating fact extractors is available online⁵.

When attempting to reproduce the scores of Xie et al. (2021), based on their own implementation, we encountered wildly differing scores compared to the values reported by the authors. Some results show drastic improvements from a reported Pearson correlation 0.58 to a reproduced score of 0.68, while other values dropped (e.g., on QAGS-XSUM, we see a reduction of scores from 0.24 to 0.16 in terms of Pearson correlation). For the sake of reproducibility, we have included the exact commands that were used to run the CoCo models in our repository.

On the other hand, all of our reproduced scores for BARTScore (Yuan et al., 2021) match the available self-reported results by the authors.

For significance testing, we use our own implementation of a permutation-based significance test, again included in the code repository. We fix the initial NumPy random seed to 256, and compute results over 10,000 iterations for each test.

³<https://github.com/neulab/BARTScore>, last accessed: 2023-02-01.

⁴https://github.com/xieyxclack/factual_coco, last accessed: 2023-03-16.

⁵<https://github.com/google-research-datasets/wikifact>, last accessed: 2023-05-17