Are Factuality Checkers Reliable? Adversarial Meta-evaluation of Factuality in Summarization

Yiran Chen[*], **Pengfei Liu**[♯,*], **Xipeng Qiu**[†]
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
2005 Songhu Road, Shanghai, China
♯Carnegie Mellon University
{yrchen19,xpqiu}@fudan.edu.cn
{pliu3}@cs.cmu.edu

Abstract

With the continuous upgrading of the summarization systems driven by deep neural networks, researchers have higher requirements on the quality of the generated summaries, which should be not only fluent and informative but also factually correct. As a result, the field of factual evaluation has developed rapidly recently. Despite its initial progress in evaluating generated summaries, the *meta-evaluation* methodologies of factuality metrics are limited in their *opacity*, leading to the insufficient understanding of factuality metrics' relative advantages and their applicability. In this paper, we present an adversarial meta-evaluation methodology that allows us to (i) diagnose the fine-grained strengths and weaknesses of 6 existing top-performing metrics over 24 diagnostic test datasets, (ii) search for directions for further improvement by data augmentation. Our observations from this work motivate us to propose several calls for future research. We make all codes, diagnostic test datasets, trained factuality models available: https://github.com/zide05/AdvFact.

1 Introduction

With the rapid development of neural networks in text summarization (Liu and Lapata, 2019; Liu, 2019; Zhong et al., 2019; Zhang et al., 2019; Lewis et al., 2019; Zhong et al., 2020; Liu and Liu, 2021), especially the use of contextualized pre-trained models (Devlin et al., 2019; Lewis et al., 2019), the state-of-the-art performance, measured by automated metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) has been constantly updated. However, although these systems can generate informative, and fluent summaries, they suffer from the problem of making factual errors–generating incorrect facts that can not be supported by the source document (Cao et al., 2018a).

Among this background, a large body of recent works (Wang et al., 2020a; Kryscinski et al., 2020; Durmus et al., 2020; Cao et al., 2020) are trying to search for new automated metrics that can assess the factuality of generated summaries due to the fact that existing metrics (e.g., ROUGE) are not correlated well with factual consistency (Maynez et al., 2020; Goyal and Durrett, 2020).

Generally, the process of designing these evaluation metrics w.r.t factuality is commonly formulated into different forms of NLP tasks, ranging from *text entailment* (Falke et al., 2019; Kryscinski et al., 2020) at sentence level or more fine-grained level (Goyal and Durrett, 2020) to *question answering* (Wang et al., 2020a; Durmus et al., 2020). Improving the understanding of these factuality metrics with diverse paradigms is critical for further metric improvement. However, the evaluation methodologies of factuality metrics are limited in their *opacity*–they are opaque to their results, which are usually holistic scores (e.g., accuracy) and not interpretable. Specifically, different from traditional non-learnable metrics like ROUGE, whose scores are relatively straightforward to interpret, e.g., lower ROUGE-2 Recall implies fewer bigrams from reference summaries are covered by generated summaries, there are diverse factors that could lead to lower score of factuality metrics (e.g., entity replacement, number inference). However, most of existing meta-evaluation strategies fail to tell (i) which types of factual errors the metric evaluated at hand are better at identifying, (ii) on which categories the error recognition ability of factuality metrics can not be well generalized. As a result, (1) the relative advantages between a better- and worse-performing systems w.r.t factuality are unclear. (2) the lack of understanding of factuality metrics' applicability reduces their reliability, and users may take the risk of over-estimating their

[*] These two authors contributed equally.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2082–2095
November 7–11, 2021.

generalization ability so as to apply them to inappropriate evaluation samples. (3) it's unclear how to improve the metric further.

Thus instead of further pursuing a new method, we take a step back to understand the shortcomings of existing metrics. We present an adversarial meta-evaluation framework which can perform fine-grained evaluation of factuality metrics. Methodologically, we (i) first conduct *error analysis* of existing state-of-the-art factuality metrics, (ii) define effective *adversarial transformations* based on the results of *error analysis*. We (iii) construct *diagnostic examples* by applying adversarial transformations to test datasets with different distributions and then diagnose existing top-scoring factuality metrics. (iv) We finally show that, the technique of *data augmentation*, driven by adversarial transformations, can increase the diversity of training samples, making factuality metrics more robust and reliable.

Our contributions can be summarized as follows: (1) We figure out several representative errors made by the existing top-performing factuality metrics (§4.2), inspiring the direction for further improvement. (2) We propose effective adversarial transformations that can either be applied to test set for model diagnosis (§5) or applied to training set for data augmentation (§6.2), by which we further improve the performance of current checkers. (3) We propose a fine-grained meta-evaluation methodology for factuality metrics and re-evaluate existing top-performing metrics to assess their relative strengths and weaknesses. (4) We call for a more fine-grained and interpretable meta-evaluation of factuality metrics for future research. As a first step, we released our constructed diagnostic test sets with various characteristics, as well as augmented training data and more robust factuality metrics.

2 Related Work

Factuality in Text Summarization Recent studies on factuality of text generation revolve around *metric design* and *system optimization*. Regarding the metric perspective, researchers formulate the design of automated metrics w.r.t factuality as different problems: text entailment over sequential (Kryscinski et al., 2020; Goyal and Durrett, 2021a) or tree (Goyal and Durrett, 2020, 2021a) structures; question answering (Wang et al., 2020a; Durmus et al., 2020) and sequence labeling (Zhao et al., 2020a). Concurrent to our work, Pagnoni et al. (2021a) constructs human annotated test sets for factuality metrics while using a different typology. Additionally, their method is difficult to be used as automatic data augmentation. Other works aim to learn factuality-aware summarization systems, which can be achieved by leveraging open information extraction and dependency parsing (Cao et al., 2018b; Zhu et al., 2020). Chen et al. (2020) explore how factuality metrics are influenced by *domain shift* and conclude that out-of-domain systems can even surpass in-domain systems in terms of factuality and factuality checkers like *FactCC* is limited in predictive power of positive samples.

Adversarial Evaluation of NLP Systems Adversarial evaluation has been extensively explored in many NLP tasks recently. The adversarial challenge sets have been introduced into tasks of natural language inference (Naik et al., 2018) question answering (Jia and Liang, 2017), machine translation (Burlot and Yvon, 2017) and language model (Marvin and Linzen, 2018) to examine system drawbacks. More recently, Gardner et al. (2020) introduces the concept of "contrast set" and proposes to use it to measure the generalization of different NLP systems. Instead of adversarially evaluate an NLP system, we perform an adversarial meta-evaluation of evaluation metrics.

Meta-evaluation for Automated Metrics Meta-evaluation aims to evaluate the reliability of automated metrics based on their correlation with human judgments (Graham, 2015; Peyrard, 2019; Bhandari et al., 2020). Most existing works perform meta-evaluation on metrics that measure semantic equivalence, such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). Yuan et al. (2021) more recently propose BARTScore and meta evaluate it on multiple evaluation perspectives. By contrast, in this paper, we focus on the evaluation of factuality metrics using our constructed diagnostic test sets. Concurrent with our work, Goyal and Durrett (2021b); Pagnoni et al. (2021b) also look into the error patterns of existing factuality checkers.[1]

[1]We encourage readers to read these works as well to obtain more interesting observations.

3 Preliminaries

3.1 Definition of Factuality

Although researchers have slightly different definitions of factuality (Maynez et al., 2020; Kryscinski et al., 2020). In this paper, we consider factuality as how well *generated summaries* are supported by *source documents* without using any external knowledge. A factual error happens when generated summaries contain salient facts (Kryscinski et al., 2020) that can not be inferred from source documents. The summary sentences that need to be verified are also called *claims* below to keep consistent with the field of fact verification (Zhou et al., 2019; Schuster et al., 2019; Liu et al., 2020).

Models	Type	Train data
MNLIBERT	NLI-S	MNLI
MNLIROBERTA	NLI-S	MNLI
MNLIELECTRA	NLI-S	MNLI
DAE	NLI-A	PARANMT-G
FACTCC	NLI-S	CNNDM-G
FEQA	QA	QA2D, SQuA

Table 1: The model types and training data of factuality metrics. NLI-A and NLI-S represent NLI-based metrics defining facts as dependency arcs and span respectively. PARANMT-G and CNNDM-G mean the automatically generated training data from PARANMT (Wieting and Gimpel, 2018) and CNN/DailyMail (Nallapati et al., 2016) (referred to as CNNDM in the rest of the paper).

3.2 Factuality Metrics

There are two major task formulations of factuality metrics: natural language inference (NLI) and question answering (QA). Model types and training data are summarized in Tab. 1.

3.2.1 NLI-based Metrics

NLI-based metrics consider factual consistency as a natural language inference problem, the core idea of which is to infer if facts from generated summaries can be entailed by its source documents. Specifically, different metrics have diverse definitions of facts.

FactCC Kryscinski et al. (2020) defines facts as salient spans in source documents and proposes to use a weakly-supervised method to learn a model-based factuality metric.

Dependency-level Entailment (DAE) Goyal and Durrett (2020) define facts as dependency arcs and propose DAE formulation to identify factual errors in a more fine-grained manner.

NLI transferred models Following Falke et al. (2019), we train different factuality checkers (MNLIBERT, MNLIROBERTA and MNLIELECTRA) based on BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2019) on MNLI dataset (Williams et al., 2018). The neutral class samples are deleted in the dataset for fair comparison following Goyal and Durrett (2020).

3.2.2 QA-based Metrics

The basic idea behind QA-based metrics is whether similar answers can be replied when we ask the same question to a generated summary S and its source document D (Durmus et al., 2020; Wang et al., 2020b). In practice, we use the recently proposed FEQA (Durmus et al., 2020).

FEQA It first generates questions based on summary, and answers the questions based on source document and summary separately. Mismatching answers indicate an inconsistency between document and summary, on the other hand, matching answers reveal consistency.

Eval. set	Dataset type	#Sys.	#Sam.	Nov.(%)
FaccTe	CNNDM	10	503	54.0
QagsC	CNNDM	1	504	28.6
RankTe	CNNDM	3	1072	52.5
FaithFact	XSum	5	2332	99.2

Table 2: Statistics of different human annotation datasets for meta-evaluating factuality metrics. Dataset type means the dataset that source document and summary belong to. Here, #Sys. and #Sam. represent the number of summarization systems that the output summaries come from and the test set size respectively. Nov. (abbreviation of novelty) means the proportion of trigrams in claims that don't exist in source documents.

3.3 Existing datasets for Meta-evaluation

To get a holistic overview of factuality metrics performances, we collect four different human judgment datasets that can be used to meta-evaluate the correctness of factuality metrics. They are `FaccTe` (Kryscinski et al., 2020), `QagsC` (Wang et al., 2020a), `RankTe` (Falke et al., 2019) and `FaithFact` (Maynez et al., 2020). Each sample of the evaluation sets is composed of one document, one summary sentence (claim), and a human annotated label that represents the factuality consistency between the document and summary. The detailed statistics of evaluation sets are showed in Tab. 2. As it shows, the claims of `FaccTe`, `QagsC`

and RankTe are the outputs from summarizers on CNNDM dataset. However, FaithFact includes faithfulness annotations[2] of five summarization systems outputs on XSum. It is included to measure the generalization ability of factuality metrics in domain different from CNNDM.

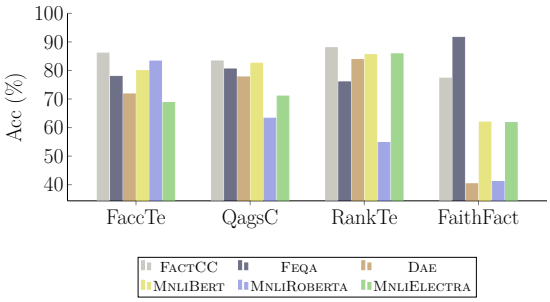

Figure 1: The overall accuracy performance of six representative factuality checkers.

4 Meta-Evaluation

4.1 Holistic Meta-evaluation

Fig. 1 illustrates meta-evaluation results of six factuality checkers on four human judgment sets. We can observe[3] that:
(1) FACTCC has achieved the best performance in most of the test sets except in FaithFact. The reason for this is that the claims in this set are highly paraphrased (its novelty is 99.2% in Tab. 2) thus will mislead FACTCC which is trained on less abstractive claims (CNNDM-G as shown in Tab. 1).
(2) FEQA underperforms FACTCC most of the time.
(3) With the same pre-trained model (ELECTRA), DAE outperforms MNLIELECTRA in FaccTe and QagsC. However, DAE with dependency information doesn't show constant superiority over MNLIELECTRA in all evaluation sets.

4.2 Fine-grained error analysis

Setup and Error Typology To get a more fine-grained understanding of factuality checkers and define the upper bound of the difficulty for the task, we choose FACTCC as the representative factuality checker (for its superior performance as described in §4.1) and perform error analysis on it. We examine 140 samples[4] that the checker fails to predict correctly in FaccTe and QagsC, and divide the reasons into diverse categories. Examples are presented in Tab. 3. Notably, there could be multiple error reasons for one mispredicted sample.

- **R1: VANs replacement**: the checker is hard to detect **V**erb, **A**djective and **N**oun replacements (e.g., antonym, synonym) thus producing the wrong prediction. Here noun represents noun or noun phrase excluding entity.
- **R2: Numerical inference** the checker obtains worse performance when verifying samples that require numerical inference (e.g., date). Similar results are also observed in (Zhao et al., 2020b).
- **R3: Entity coreference**: a slight change of person name or replacing the pronoun with its reference name will mislead factuality checker which suggests the lack of entity coreference resolution ability.
- **R4: Missing details**: when the claim lacks some detailed information (e.g., location), the checker tends to predict it as inconsistent though it is not. While this is frequently occurring in the scenario of summarization when the summarizer only extracts the most important information.
- **R5: Paraphrase** The more complex paraphrase patterns (e.g., complex reorder, passive-active transformation, sentence fusion and so on) other than simple token replacement or omission that cause the model to make wrong predictions.
- **R6: Background knowledge** The checker is fragile when extra knowledge is required.
- **R7: Truncate** The checker truncates long documents and will ignore the information of evidence sentences in later part of documents, therefore making wrong judgment.
- **R8: Wrong label** Incorrect annotated label.
- **R9: Others** Other reasons.

Analysis of Error Reasons As presented in Tab. 3, VANs replacement and Missing details account for a large proportion in all error reasons. It is because verb, adjective and noun (besides entity) replacement and detail omission are not included in the training data for FACTCC. Moreover, misclassifications that caused by paraphrase are account for 11.8%, which lies in the lack of paraphrase for training data of FACTCC as the only paraphrase pattern is introduced by backtranslation (Edunov et al., 2018). While entity and number swap are included in negative sample construction in (Kryscinski et al., 2020), FACTCC still makes wrong prediction facing samples requiring entity coreference resolution and numerical inference.

[2]Factuality annotation is not included because it needs out of context knowledge to make the judgment.
[3]The more detailed observation can be found in appendix.
[4]We have released these samples in our Github repository.

Typology	Source document	Claim	Ratio
R1: VANs replacement (inco → co)	...Japanese court issued a landmark injunction **halting** plans to restart two nuclear reactors in a western prefecture...	japanese court **orders** to restart two nuclear reactors in a western prefecture.	12.4%
R2: Numerical inference (co → inco)	...**On October 31**, 2014, the Italian government announced the end of "Mare Nostrum" ...	the italian government announced the end of "mare nostrum" **in 2014**.	1.3%
R3: Entity coreference (co → inco)	...Ahmed Farouq didn't have the prestige.....Before that, **Farouq** was the deputy emir of al Qaeda....	**ahmed farouq** was the deputy emir of al qaeda in the indian subcontinent.	17.0%
R4: Missing details (co → inco)	...Phil Rudd**, the drummer for legendary hard rock band AC/DC,** has pleaded guilty to charges of...	rudd has pleaded guilty to threatening to kill and possession of drugs in a court.	31.4%
R5: Paraphrase (inco → co)	...A police motorcycle stopped the rest of the pack, before organisers of the 151-mile race slowed the leaders to allow the pack to catch up...	Leaders of the tour de france were stopped by police as they crossed a railway line to avoid a train.	11.8%
R6: Background knowledge (co → inco)	Scientists from **harvard** medical school have discovered a way of turning stem cells into killing machines ...	Scientists in the **us** have developed a stem cell therapy for brain tumours.	0.7%
R7: Truncate (co → inco)	**[>512]**...Ben was slated for a clinical trial with an experimental drug....	ben was slated for a clinical trial with an experimental drug.	3.3%
R8: Wrong label (inco → co)	...**The man who spent six years as spokesman for the Glazer family** has written an enlightening account of his time with the Manchester United chiefs...	**Manchester united's unpopular owners** has written an enlightening account of his time with the manchester united chiefs.	9.8%
R9: Others (inco → co)	These days we are increasingly using outdoor space for the occasional barbecue or to relax in a hot tub **rather than** for tending flowers.	these days we are increasingly using outdoor space for tending flowers.	12.4%

Table 3: Error reasons with their corresponding examples and the ratio of them. The bold span is corresponding to the error reason. co → inco represents the gold label is factually correct while checker misclassifies it as factually incorrect (inco → co means the opposite). [>512] means there are more than 512 subwords before this position.

5 Construction of Diagnostic Set

It is not realistic to produce large scale human annotated test sets with multiple error reasons observed above. As a consequence, former work (Hidey et al., 2020) and (Naik et al., 2018) construct diagnostic test sets automatically. In this section, we first introduce automatic rule-based transformation methods based on error analysis (§5.1). Then we construct 24 diagnostic test sets based on three types of baseline test sets.

5.1 Adversarial Transformations

We introduce four types of automatic transformation methods corresponding to the R1-4 error reasons in error analysis (§4.2). Paraphrasing (R5) is not included here for it is hard to produce simply with rule, thus we introduce it in another way–using gold references as claims in §5.2. The rest four error reasons are either too hard for models (R6, R9) or correspond to systematic error (R7) or lie in annotation error (R8), and also will not be included here. The adversarial transformation examples are shown in Tab. 4.

R1: Antonym Substitution We first use Stanza (Qi et al., 2020) to do Part-of-Speech tagging and then use WordNet wrapped in NLTK (Bird et al., 2009) package to find antonyms for verb and adjective. Negative samples are produced by replacing the original word with its antonyms. The reason we do not include synonyms replacement is that simply replacing word with its synonyms can introduce factual error and cause the gold label ambiguous.

R2: Numerical Editing FACTCC exhibits worse performance when it needs numerical reasoning to derive the result as §4.2 shows, which motivates us to design a numerical adversarial transformation. Specifically: (1) to produce negative samples, we replace numerical entity [5] with a randomly chosen entity of the same type in source document and guarantee the transformed claim differs from the origin. On the other hand, we also add preposition (e.g.,"after") before date and timing type entities while adding "more than" and "less than" before other types of numerical entities; (2) For positive samples, we change the number or date [6] and add "before", "after", "more than" and "less than" properly (e.g., "in 2019" to "two years before 2021"). We include more complex negative and positive transformations for numerical inference compared with Kryscinski et al. (2020).

R3: Entity Replacement At the phase of error analysis, we discover FACTCC fails to understand the equivalence between named entities referring to

[5] NER is also performed by Stanza.

[6] We use Python wrapper of SUTime (Chang and Manning) to identify the exact year, month and day to change the date.

Adv Trans.	Type	Transformed Claim
R1: AntoSub	verb	poolside : guests enjoyed the sunny weather as they waited for the show to ~~commence~~ → end .
	adj.	on monday , children will flock from every state to decorate eggs on the south lawn of the ~~white~~ → black house .
R2: NumEdit	pos	silk flowers and a sign saying ' pray for justice ' adorn the highway 34 bridge on the edge of alsea bay in waldport , oregon , in a picture taken ~~in october 2002~~ → before May, 2003 .
	neg	silk flowers and a sign saying ' pray for justice ' adorn the highway 34 bridge on the edge of alsea bay in waldport , oregon , in a picture taken ~~in october 2002~~ → in 2011 .
R3: EntRep	pos	actor ~~isaiah washington~~ → isaiah tweeted : ' okay , watching the #walterscott video was horrible , but i think the brave person who captured the murder is a hero and a godsend #truthdom . '
	neg	actor ~~isaiah washington~~ → michelle williams tweeted : ' okay , watching the #walterscott video was horrible , but i think the brave person who captured the murder is a hero and a godsend #truthdom . '
R4: SynPrun	prepo.	the queen and the duke of edinburgh appeared in good spirits as they arrived to a red carpet ~~at the event~~ .
	clause	the mystery hero ~~who raced to the edge of a cliff and pulled a driver from his precariously-balanced car~~ has been identified as a 29-year - old man who fled the scene to go to work .

Table 4: Adversarial transformations corresponding to error reasons R1-4 in §4.2. "Type" here means subtype of adversarial transformations. Specifically, we display verb and adjective *antonym substitution* for AntoSub. Also, factual consistent and inconsistent samples (pos and neg) are displayed for NumEdit and EntRep. Lastly, prepo. and clause mean the omission of preposition phrases and sub-clauses.

the same person. Thus, we produce positive examples by replacing PERSON named entity with its subtoken (e.g., replace `Isaiah washington` with `Isaiah`). Negative samples are produced by replacing the entity with a randomly chosen entity of the same type from the source document. Here we prevent the new entity from being substring of the origin entity and vise versa. Another type of negative transformation is replacing part of PERSON entity with different one. The transformation in (Kryscinski et al., 2020) doesn't include positive samples as well as PERSON entity editing as negative samples.

R4: Syntactic Pruning *Syntactic pruning* is used to produce positive examples with detail omitted. Despite using dependency parsing, we choose constituency parsing to disentangle the summary sentence for it is more suitable to capture clauses and phrases. To produce positive examples, clauses with label "S" and "SBAR" and prepositional phrases with label "PP" are deleted based on the assumption that the lack of sub-clause will not affect the factual consistency.

5.2 Diagnostic Datasets

We construct 24 diagnostic datasets[7] based on three types of base test sets as follows: Besides only using sentences in ***source document*** (`DocAsClaim`) as input to transformation as previous work (Kryscinski et al., 2020) does, we propose to use another two base test sets: ***gold summary*** (`RefAsClaim`) and ***generated summary*** (`FaccTe`, `QagsC`,`RankTe` and `FaithFact`) to serve as input to the adversarial transformation. Reasons are: (i) the diagnostic set constructed based on reference summaries corresponds to the error reason R5 in §4.2, which is a more challenging test set for factuality checkers due to its more complex paraphrase patterns. (ii) the distribution of generated summaries will be more closed to summaries verified by factuality checkers in real scenarios (e.g., generated summaries from BART). Finally we obtain 6 base test sets and 24 diagnostic test sets (4 adversarial transformations on every base test set).

5.3 Quality Examination

In order to explore the reliability of the automatically generated diagnostic test sets, we conduct human examination on whether the generated claim is *grammatically correct* and maintains *correct label*. This is carried out on 50 randomly chosen samples for each type of adversarial transformation. Results [8] show that all the diagnostic sets are grammatically correct (ratio around 85%) and possess correct factuality labels (ratio higher than 90%).

6 Experiment

6.1 Re-evaluation on Diagnostic Datasets

Antonym Substitution The performances of checkers drop when tested in AntoSub as Tab. 5

[7] We have released the datasets on our Github repository. And the detailed information of it is included in the appendix.

[8] The detailed results are shown in appendix.

Evaluation Set	DocAsClaim					RefAsClaim					FaccTe				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
MNLIBERT	76.48	-48.01	-46.77	-38.22	+3.41	77.10	-37.34	-43.57	-37.08	-3.08	79.92	-45.74	-56.81	-43.78	+8.24
MNLIROBERTA	92.85	-80.49	-69.49	-61.15	+0.74	52.08	+0.17	-3.25	-1.06	-0.99	83.30	-66.14	-52.30	-48.53	+8.54
MNLIELECTRA	79.67	-53.42	-47.61	-40.59	+0.54	74.23	-41.04	-39.33	-36.18	-0.28	68.79	-22.67	-29.96	-26.97	+0.60
DAE	67.02	-32.18	-28.13	-24.58	+2.40	77.69	-52.27	-45.44	-44.10	+0.83	71.77	-47.59	-36.82	-36.77	-2.79
FEQA	81.04	-53.26	-42.35	-34.85	-8.93	36.93	+35.75	+26.10	+31.31	-1.94	77.93	-48.53	-35.60	-27.70	-8.26
FACTCC	72.54	-37.62	-10.52	+10.75	-4.36	40.62	+22.58	+31.98	+40.99	-3.92	86.08	-73.09	-30.93	+0.51	-10.98
Evaluation Set	**QagsC**					**RankTe**					**FaithFact**				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
MNLIBERT	82.54	-65.52	-67.58	-56.57	+15.47	85.54	-57.41	-59.97	-48.25	+0.57	61.92	-22.53	-25.19	-26.83	+1.64
MNLIROBERTA	63.29	-24.61	-25.73	-23.03	+4.52	54.76	-7.13	-10.02	-5.48	+5.05	41.12	-23.21	-9.49	-11.30	+43.63
MNLIELECTRA	71.03	-47.26	-40.46	-33.92	+9.31	85.82	-65.71	-59.71	-54.27	+1.22	61.75	-23.18	-0.05	-18.77	+0.11
DAE	77.73	-69.43	-57.73	-47.12	+14.84	83.86	-70.98	-57.60	-53.35	+1.14	40.31	-13.86	-7.33	-18.38	+26.64
FEQA	80.52	-59.56	-44.10	-44.34	-2.52	76.00	-42.89	-29.44	-22.28	-10.07	91.59	+2.90	-1.16	-6.50	-87.35
FACTCC	83.33	-61.81	-27.56	-0.96	-7.55	87.97	-68.59	-26.98	-3.22	-10.38	77.32	-12.03	+14.17	+3.38	-38.34

Table 5: Adversarial Evaluation Results. The first column of every subtable represents the factuality checker performance in the original test set (gray). The rest four columns represent four types of diagnostic test sets, the value of which is the difference between model accuracy in diagnostic and original test set. Here we don't use balanced accuracy because AntoSub and SynPrun only possess negative samples and positive samples respectively. The positive value implies the performance increases when evaluated in the diagnostic test set while the negative value does the opposite (red). Here `DocAsClaim` and `RefAsClaim` represent two evaluation set with document sentences and summary reference sentences as claims respectively.

shows (nearly all entry values of AntoSub columns are negative).

However, FEQA and FACTCC obtain obvious performance improvement in the AntoSub diagnostic set of `RefAsClaim`. It is because claims in `RefAsClaim` original set are highly paraphrased which will mislead the checkers to produce negative labels and cause lower accuracy. While *Antonym Substitution* introduces factual inconsistent samples, thus instead, model performance improves. Models transferred from MNLI and DAE are more robust to samples with highly paraphrased claims.

Numerical Editing Nearly all factuality checkers get worse performance with NumEdit transformation (almost all results of NumEdit columns are negative in Tab. 5). Even FACTCC is not the exception though it may possess numerical inference ability to some extend. **It emphasizes the importance to improve numerical inference ability for factuality checkers.** However, FEQA and FACTCC get better performances when tested in NumEdit diagnostic set of `RefAsClaim` because the *numerical editing* transformation introduces more negative samples (reason is similar as described above).

Entity Replacement Similar to *numerical Editing*, the *entity replacement* transformation also tends to mislead six factuality checkers as nearly all values of EntRep columns in Tab. 5 are negative. Although FACTCC is trained with data that also includes *entity replacement* transformation, it still obtains worse performance in EntRep diagnostic test sets of `QagsC` and `RankTe`. This implies the incompleteness of *entity replacement* in (Kryscinski et al., 2020). It shows the same pattern as AntoSub when models are tested in EntRep diagnostic sets of `RefAsClaim` and the reason is similar as described above.

Syntactic Pruning The diagnostic test sets of SynPrun can lead to more performance drop when the base test sets are `RefAsClaim` and `FaccTe` because the last columns of these subtables get more negative values. Transformation of this type will be more confusing when the claims are highly paraphrased.

As observed in Tab. 5, models transferred from MNLI dataset and DAE are more robust when *syntactic pruning* are introduced, while FACTCC and FEQA are constantly misled by SynPrun diagnostic test sets. This can be attributed to the lack of highly paraphrased claims in FACTCC training set. DAE tends to extract dependency triples of summary and make prediction based on them, thus is more robust when evaluated in SynPrun diagnostic sets. As for models transferred from MNLI, it may because the training set of MNLI already possesses pattern of detail omission and the trained models have the capability to recognize it.

Takeaways (1) Most factuality checkers obtain poor performance in AntoSub and NumEdit diagnostic sets, which suggests that current factuality metrics are not faithful when dealing with *antonym substitution* and *numerical editing* samples. (2)

Evaluation Set	DocAsClaim					RefAsClaim					FaccTe				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
FactCC	72.54	34.92	62.02	83.29	68.18	40.62	63.20	72.60	81.61	36.70	86.08	12.99	55.15	86.59	75.10
FactCC$_{sub}$	**78.24**†	27.44	60.34	80.28	74.99	54.17	48.05	66.15	78.91	53.85	**88.27**	8.96	52.23	82.05	86.12
FactCC$^{adv}_{sub}$	77.06	**86.00**†	**90.16**†	**87.69**†	**80.00**†	**58.08**†	**80.99**†	**86.19**†	**83.39**†	**61.40**†	88.07	**80.45**†	**86.99**†	**87.27**†	**96.73**†
FactCC$^{ref}_{sub}$	82.92†	22.44	59.20	77.85	78.59	78.09†	27.37	60.30	71.11	78.11	88.67	4.93	51.07	82.05	90.20
FactCC$^{ref-adv}_{sub}$	81.87	71.58†	83.69†	84.17†	80.88†	75.12	82.73†	85.31†	86.15†	78.32	88.87	69.70†	88.35†	92.73†	96.73†
Evaluation Set	**QagsC**					**RankTe**					**FaithFact**				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
FactCC	83.33	21.52	55.77	82.37	75.78	87.97	19.38	60.99	84.75	77.59	**77.32**	65.29	**91.49**	80.70	38.98
FactCC$_{sub}$	82.74	16.03	54.63	79.04	83.48	90.11	13.18	59.85	82.53	83.15	65.44	46.83	79.79	83.33	**50.85**
FactCC$^{adv}_{sub}$	**85.32**†	**80.03**†	**88.78**†	**84.23**†	**97.72**†	**91.42**†	**79.77**†	**84.58**†	**87.09**†	**96.67**†	69.85†	**80.99**†	90.43†	**90.35**†	45.76
FactCC$^{ref}_{sub}$	84.92	10.69	53.50	78.29	86.32	91.32	7.59	57.10	81.62	90.00	49.01	33.88	76.60	75.44	64.41†
FactCC$^{ref-adv}_{sub}$	86.71	73.42†	90.89†	87.94†	94.02†	92.72†	74.79†	89.01†	89.05†	93.33†	62.95†	87.33†	88.30†	88.60†	55.93

Table 6: The adversarial training accuracy results. Cells in bold means the highest score among FACTCC, FactCC$_{sub}$ and FactCC$^{adv}_{sub}$. While cells in red means highest score among FactCC$^{ref}_{sub}$ and FactCC$^{ref-adv}_{sub}$. † and † indicate the difference between FACTCC$^{adv}_{sub}$, FactCC$_{sub}$ and FactCC$^{ref-adv}_{sub}$, FactCC$^{ref}_{sub}$ is significant.

FACTCC can handle *entity replacement* diagnostic sets to some extent, but can not maintain the performance constantly over all EntRep sets. (3) MNLIBERT, MNLIROBERTA, MNLIELECTRA and DAE are more reliable to deal with highly paraphrased claims and are more robust to *syntactic pruning* transformation.

6.2 Data Augmentation

Besides utilizing adversarial transformation to construct test sets, it can also be used to create more training data, i.e., data augmentation, to improve the model performance. Here we choose FACTCC to conduct adversarial training[9] due to the excellent performance of FACTCC in § 4.1.

As the original training data of FACTCC has more than 100 million samples, we first subsample 50 million data to train FACTCC$_{sub}$. Moreover, we add 34,912 adversarial training data to the subsampled set and train another checker called FACTCC$^{adv}_{sub}$. Also, we investigate whether introducing references as claims to the training set will enhance model performance. We include references as claims and make negative transformations in (Kryscinski et al., 2020) on them to train FACTCC$^{ref}_{sub}$. Lastly, adversarial transformation based on reference is also included and the trained model calls FACTCC$^{ref-adv}_{sub}$. The analysis results of them in different baseline and diagnostic test sets are showed in Tab. 6, from which we can draw several conclusions:

Subsampling doesn't mean performance decrease. Compared with the original FACTCC that trained from more than 100 million data, the subsampling version FACTCC$_{sub}$ with 50 million training data performs better when tested in the original test set of `DocAsClaim`, `RefAsClaim`, `FaccTe` and `RankTe` in Tab. 6.

Adversarial data augmentation improves model performance on both original and diagnostic test sets most of time. As shown in Tab. 6, FACTCC$^{adv}_{sub}$ outperforms FACTCC and FACTCC$_{sub}$ in original test sets of `RefAsClaim`, `QagsC` and `RankTe`. Moreover, FACTCC$^{adv}_{sub}$ shows significantly[10] superior performance on the diagnostic test sets because nearly all cells in the line of FACTCC$^{adv}_{sub}$ are bold on diagnostic test sets.

Adding reference as augmented training data can improve model performance to some extend. FACTCC$^{ref}_{sub}$ performs better than FACTCC$_{sub}$ in all origin evaluation set except in `FaithFact`. When introducing adversarial training set, the performances are significantly improved in Tab. 6, especially when tested in diagnostic test sets (nearly all cells of row FACTCC$^{ref-adv}_{sub}$ are red).

7 Implications and Future Directions

In this paper, we present an adversarial meta-evaluation methodology driven by our fine-grained analysis, which not only allows us to re-evaluate existing top-performing factuality metrics, diagnosing their limitations, but also instructs us to further improve current metrics by data augmentation. Based on what we have explored and observed in this work, we suggest following potentially promising future directions:

(1) *Knowledge-guided factuality metric*: One error reason in §4.2 is the lacking of extra knowledge reference ability for factuality metrics. It would

[9]The detailed model information is presented in appendix.

[10]We carry out bootstrap pair-wise significance test with significance rate 0.05.

be promising to explore the effectiveness of external knowledge like knowledge base (Bordes et al., 2013), citation graph (Lo et al., 2020) (for scientific summarization).
(2) *Long document Modeling:* Lengths of most of summarization documents are over 512, which brings great challenge for pretrain based factuality metrics (R7 in §4.2). Various methodologies (e.g., first retrieval then verification (Zhou et al., 2019)) should be put forwards to deal with the problem.
(3) *Fine-grained meta-evaluation and more diverse human judgments*: To reliably evaluate factuality metrics, human judgments over diverse distribution are needed. Moreover, fine-grained meta-evaluation for metrics is beneficial to further identify their drawbacks and suggest future directions.

Acknowledgments

We thanks Ming Zhong and all reviewers for their valuable comments and helpful suggestions. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106702) and National Natural Science Foundation of China (No. 62022027).

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018a. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions.

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv*, pages arXiv–2005.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Tobias Falke, Leonardo FR 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.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021a. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Tanya Goyal and Greg Durrett. 2021b. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu. 2019. Fine-tune BERT for Extractive Summarization.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021a. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021b. 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.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3410–3416.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020b. 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.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

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.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020a. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020b. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*.

A Appendix

A.1 Experimental Setup

The training sets and trained models can be found in our github: `https://github.com/zide05/AdvFact`. Here we introduce the training process and model details below:

FactCC We use the trained FACTCC model in Kryscinski et al. (2020) as the origin FACTCC. Also we train other four versions namely FACTCC$_{sub}$, FACTCC$^{adv}_{sub}$, FACTCC$^{ref}_{sub}$ and FACTCC$^{ref-adv}_{sub}$. All four checkers are trained with the code by Kryscinski et al. (2020) on 4 TITAN Xp for 15 epochs. The training batch size for each gpu is 8 and the optimizer is AdamW with initial learning rate 2e-5. The code url can be found in https://github.com/salesforce/factCC.

DAE We use DAE in Goyal and Durrett (2020) and the ELECTRA based DAE which trained on training set consists of paraphrase data, synonym data and hallucination data are included. The trained model and code can be found in https://github.com/tagoyal/dae-factuality.

NLI transferred models We train three NLI transferred models (MNLIBERT, MNLIROBERTA and MNLIELECTRA) on MNLI dataset (Williams et al., 2018) and the samples with neutral label are deleted for fair comparison. Every model is trained on 4 TITAN Xp for 15 epochs. We choose the AdamW as optimizer and set the learning rate to 2e-5. The training batch size for each gpu is 8. The code and the trained checkpoints can be found in our github https://github.com/zide05/AdvFact.

FEQA The trained FEQA in (Durmus et al., 2020) are used in this paper and the checkpoints and codes can be found in https://github.com/esdurmus/feqa.

A.2 Experimental Results

Detailed information for baseline and diagnostic datasets. We introduce the basic information for the baseline datasets in Tab. 7. The more detailed statistics for baseline and diagnostic datasets are displayed in Tab. 11.

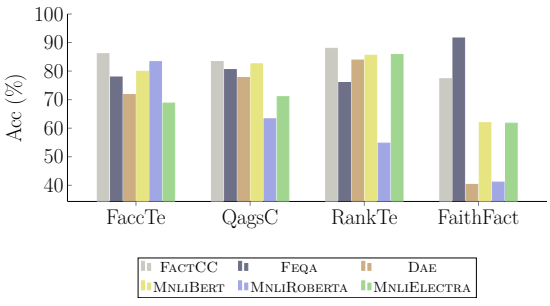

Figure 2: The overall accuracy performance of six representative factuality checkers.

Detailed holistic meta-evaluation Following conclusions can be drawn from the holistic meta-evaluation results in Fig. 2:

(1) FACTCC has achieved the best performance in most of the test sets except in FaithFact. The reason for this is that the claims in this set are highly paraphrased (novelty of it is 99.2% in Tab. 7) thus will mislead FACTCC which is trained on less abstractive claims (CNNDM-G as shown in Tab. 10).

(2) FEQA underperforms FACTCC most of the time. In FaithFact, however, FEQA gets higher accuracy. Because the claims in FaithFact are highly paraphrased, thus FEQA tends to label samples as factually inconsistent. On the other hand, the negative samples account for 92% in FaithFact. Thus the tendency of producing negative labels helps to improve the accuracy of FEQA.

(3) With the same pre-trained model ELECTRA, DAE outperforms MNLIELECTRA in FaccTe and QagsC. However DAE with dependency information doesn't show constant superiority over NLI based model MNLIELECTRA in all evaluation sets. It shows especially worse performance in FaithFact. Opposite to FEQA, DAE averages the factuality scores of all dependency arc triples as the claim-level factuality score, which is biased towards the label of factually correct. Therefore it will obtain lower accuracy in the test set with more negative samples.

Base Test Sets	Dataset type	Nov.	#Sys.
DocAsClaim	CNNDM	0 .0	0
RefAsClaim	CNNDM	77.7	0
FaccTe	CNNDM	54	10
QagsC	CNNDM	28.6	1
RankTe	CNNDM	52.5	3
FaithFact	XSum	99.2	5

Table 7: The basic statistics of baseline test sets. Dataset type means the dataset that source document and summary belong to. Here, CNNDM means CNN/DailyMail dataset. Nov.(%) means the proportion of trigrams in claims that don't exist in source documents. #Sys. represents the number of summarization systems that the output summaries come from.

Adversarial trained FACTCC model details. The detailed training set composition of adversarial trained FACTCC models are presented in Tab. 8.

Quality examination of diagnostic evaluation sets. Tab. 9 shows the ratio of generated claims

Models	Base	Adv_{base}	Ref	Adv_{ref}
FactCC	origin (100 m)	×	×	×
FactCC_{sub}	sub (50 m)	×	×	×
$\text{FactCC}^{adv}_{sub}$	sub (50 m)	✓	×	×
$\text{FactCC}^{ref}_{sub}$	sub (50 m)	✓	✓	×
$\text{FactCC}^{refadv}_{sub}$	sub (50 m)	✓	✓	✓

Table 8: Data augmented models and their corresponding training data set composition. Base and Ref represent the base training set and augmented data using references as claims. Adv_{base} and Adv_{ref} mean the adversarial augmented data based on the base training data and reference augmented data respectively.

that are *grammatically correct* and maintains *correct label*.

Trans.	CoLabel (%)	CoGrammar (%)
AntoSub	90	84
NumEdit	98	90
EntRep	96	92
SynPrun	90	82

Table 9: Quality examination of four diagnostic evaluation sets. "CoLabel" and "CoGrammar" represent the correctness rate of automatically generated labels and grammar.

Models	Type	Train data
MNLIBERT	NLI-S	MNLI
MNLIROBERTA	NLI-S	MNLI
MNLIELECTRA	NLI-S	MNLI
DAE	NLI-A	PARANMT-G
FACTCC	NLI-S	CNNDM-G
FEQA	QA	QA2D, SQuA

Table 10: The model types and training data of factuality metrics. NLI-A and NLI-S represent NLI-based metrics defining facts as dependency arcs and span respectively. PARANMT-G and CNNDM-G mean the automatically generated training data from PARANMT (Wieting and Gimpel, 2018) and CNN/DailyMail (Nallapati et al., 2016)

Evaluation Set	**DocAsClaim**					**RefAsClaim**					**FacTe**				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
# PosSam.	11490	0	2706	1936	9533	10000	0	2091	5537	4572	441	0	102	118	245
# NegSam.	0	26487	12477	4880	0	0	14131	9530	23221	0	62	670	413	322	0
# Sam.	11490	26487	15183	6816	9533	10000	14131	11621	28758	4572	503	670	515	440	245
AvgText	778.78	787.67	766.58	785.08	764.70	817.28	836.23	821.39	816.35	821.65	760.28	767.48	714.59	796.92	737.69
AvgClaim	23.32	28.31	29.08	28.58	23.55	14.45	16.17	16.92	15.81	12.70	16.75	20.12	19.98	18.47	16.45
Evaluation Set	**QagsC**					**RankTe**					**FaithFact**				
Transf.	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun	Origin	AntoSub	NumEdit	EntRep	SynPrun
# PosSam.	401	0	100	134	351	1001	0	212	201	540	183	0	8	16	118
# NegSam.	103	711	515	405	0	71	1646	1098	566	0	2149	363	86	98	0
# Sam.	504	711	615	539	351	1072	1646	1310	767	540	2332	363	94	114	118
AvgText	356.40	360.21	360.15	353.54	360.59	816.19	795.37	805.08	805.87	842.13	440.45	768.37	2385.57	1152.77	425.81
AvgClaim	17.99	22.62	21.21	20.30	17.74	17.29	20.46	21.68	20.04	18.01	21.08	22.42	24.93	23.37	16.33

Table 11: The detailed statistics of baseline (gray) and diagnostic test sets. # PosSam., # NegSam. and # Sam. represent the numbers of positive samples, negative samples and all samples respectively. AvgText and AvgClaim mean the average token length of texts and claims.