Evaluating the Factual Consistency of Abstractive Text Summarization

Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher
Salesforce Research
{kryscinski,bmccann,cxiong,rsocher}@salesforce.com

Abstract

The most common metrics for assessing summarization algorithms do not account for whether summaries are factually consistent with source documents. We propose a weakly-supervised, model-based approach for verifying factual consistency and identifying conflicts between source documents and generated summaries. Training data is generated by applying a series of rule-based transformations to the sentences of source documents. The factual consistency model is then trained jointly for three tasks: 1) predict whether each summary sentence is factually consistent or not, 2) in either case, extract a span in the source document to support this consistency prediction, 3) for each summary sentence that is deemed inconsistent, extract the inconsistent span from it. Transferring this model to summaries generated by several neural models reveals that this highly scalable approach outperforms previous models, including those trained with strong supervision using datasets from related domains, such as natural language inference and fact checking. Additionally, human evaluation shows that the auxiliary span extraction tasks provide useful assistance in the process of verifying factual consistency. We also release a manually annotated dataset for factual consistency verification, code for training data generation, and trained model weights at https://github.com/salesforce/factCC.

1 Introduction

The goal of text summarization is to transduce long documents into a shorter form that retains the most important aspects from the source document. Common approaches to summarization are *extractive* (Dorr et al., 2003; Nallapati et al., 2017) where models directly copy salient parts of the source document into the summary, *abstractive* (Rush et al., 2015; Paulus et al., 2017) where the important parts are paraphrased to form novel sentences, and *hybrid* (Gehrmann et al., 2018), combining the two methods by employing specialized extractive and abstractive components.

Advancements in neural architectures (Cho et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), transfer learning (McCann et al., 2017; Devlin et al., 2018), and availability of large-scale supervised datasets (Nallapati et al., 2016; Grusky et al., 2018) allowed deep learning-based approaches to dominate the field. State-of-the-art solutions utilize self-attentive Transformer blocks (Liu, 2019; Liu and Lapata, 2019), attention and copying mechanisms (See et al., 2017; Cohan et al., 2018), and multi-objective training strategies (Guo et al., 2018; Pasunuru and Bansal, 2018), including reinforcement learning techniques (Kryściński et al., 2018; Dong et al., 2018; Wu and Hu, 2018).

Despite significant efforts made by the research community, there are still many challenges limiting progress in summarization: insufficient evaluation protocols that omit important dimensions, such as factual consistency, noisy datasets that leave the task underconstrained, and strong, domain-specific layout biases in the data that dominate training signal (Kryściński et al., 2019).

We address the problem of verifying factual consistency between source documents and generated summaries: a factually consistent summary contains only statements that are entailed by the source document. Recent studies show that up to 30% of summaries generated by abstractive models contain factual inconsistencies (Cao et al., 2018; Goodrich et al., 2019; Falke et al., 2019; Kryściński et al., 2019). Such high levels of factual inconsistency render automatically generated summaries virtually useless in practice.

The problem of factual consistency is closely related to natural language inference (NLI) and fact

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9332–9346,
November 16–20, 2020.

Source article fragments	
(CNN) The mother of a quadriplegic man who police say was left in the woods for days cannot be extradited to face charges in Philadelphia until she completes an unspecified "treatment," Maryland police said Monday. The Montgomery County (Maryland) Department of Police took Nyia Parler, 41, into custody Sunday (...)	(CNN) The classic video game "Space Invaders" was developed in Japan back in the late 1970's – and now their real-life counterparts are the topic of an earnest political discussion in Japan's corridors of power. Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that (...)
Model generated claims	
Quadriplegic man Nyia Parler, 41, left in woods for days can not be extradited.	Video game "Space Invaders" was developed in Japan back in 1970.

Table 1: Examples of factually incorrect claims output by summarization models. Green text highlights the support in the source documents for the generated claims, red text highlights the errors made by summarization models.

checking. Current NLI datasets (Bowman et al., 2015; Williams et al., 2018) focus on classifying logical entailment between short, single sentence pairs, but verifying factual consistency requires the entire source document. Fact checking focuses on verifying facts against the whole of available knowledge, whereas factual consistency checking focuses on adherence of facts to information provided by a source document without guarantee that the information is true.

We propose a novel, weakly-supervised BERT-based (Devlin et al., 2018) model for verifying factual consistency, and we add specialized modules that explain which portions of both the source document and generated summary are pertinent to the model's decision. Training data is generated from source documents by applying a series of rule-based transformations that were inspired by error-analysis of neural summarization model outputs. Through human evaluation we show that the explanatory modules that augment our factual consistency model provide useful assistance to humans as they verify the factual consistency between a source document and generated summaries. Together with this manuscript we release a manually annotated dataset for factual consistency verification, code for training data generation, and trained model weights at `https://github.com/salesforce/factCC`.

2 Related Work

This work builds on prior research in factual consistency in text summarization and natural language generation. Goodrich et al. (2019) proposed an automatic, model-dependent metric for evaluating the factual accuracy of generated text. Facts are represented as *subject-relation-object* triplets and factual accuracy is defined as the precision between facts extracted from the summary and source document. Despite positive results, the authors highlighted remaining challenges, such as its inability to adapt to negated relations or relation names expressed by synonyms.

A parallel line of research focused on improving factual consistency of summarization models by exploring different architectural choices and strategies for both training and inference. In Falke et al. (2019), the authors proposed re-ranking potential summaries based on factual correctness during beam search. The solution used textual entailment (NLI) models to score summaries by means of the entailment probability between all source document-summary sentence pairs. The summary with the highest aggregate entailment score was used as the final output of the summarization model. The authors concluded that out-of-the-box NLI models do not transfer well to the task of factual correctness. In Cao et al. (2018), the authors proposed a novel, dual-encoder architecture that in parallel encodes the source documents and all the facts contained in them. During generation, the decoder attends to both the encoded source and facts which, according to the authors, forces the output to be conditioned on the both inputs. Human evaluation showed that the proposed technique substantially lowered the number of errors in generated single-sentence summaries.

The synthetic data generation process proposed as part of our approach is based on prior work done in the domains of data augmentation and weakly-supervised learning. Wei and Zou (2019) proposed an augmentation framework aimed at boosting performance of text classification models. The authors used 4 text transformations to synthesize data: synonym replacement, random insertion, random swap, random deletion,

and showed increased performance of classifiers on 5 downstream tasks, both for convolutional and recurrent neural models. In (Sennrich et al., 2015; Edunov et al., 2018) the authors introduced and analyzed the effects of using backtranslation based data augmentation on the performance of machine translation models, while Iyyer et al. (2018) used the mentioned transformation to synthesize training data for a paraphrase generation network. Meng et al. (2018) investigated a two step approach for training text classification models on weakly-supervised data, which includes pretraining models on fully synthetic data.

3 Methods

A careful study of the outputs of state-of-the-art summarization models provided us with valuable insights about the specifics of factual errors made during generation and possible means of detecting them. Primarily, checking factual consistency on a *sentence-sentence* level, where each sentence of the summary is verified against each sentence from the source document, is insufficient. Some cases might require a longer, multi-sentence context from the source document due to ambiguities present in either of the compared sentences. Summary sentences might paraphrase multiple fragments of the source document, while source document sentences might use certain linguistic constructs, such as coreference, which bind different parts of the document together. In addition, errors made by summarization models are most often related to the use of incorrect entity names, numbers, and pronouns. Other errors such as negations and common sense error occur less often. Taking these insights into account, we propose and test a *document-sentence* approach for factual consistency checking, where each sentence of the summary is verified against the entire body of the source document.

3.1 Training data

Currently, there are no training datasets for factual consistency checking. Creating a large-scale, high-quality dataset with supervision collected from human annotators is expensive and time consuming. We consider an alternative approach to acquiring training data that is highly scalable.

Considering the state of summarization, in which the level of abstraction of generated summaries is low and models mostly paraphrase single sentences and short spans from the source (Kryściński et al., 2018; Zhang et al., 2018), we propose using a synthetic, weakly-supervised dataset for the task at hand. Our data creation method requires an unannotated collection of source documents in the same domain as the summarization models that are to be checked. Examples are created by first sampling single sentences, later referred to as *claims*, from the source documents. Claims then pass through a set of textual transformations that output novel sentences with both *positive* and *negative* labels. Though transformations are applied to single sentences, we found that, in keeping with our aforementioned observations of model-generated summaries, the process of verifying their consistency often requires referring to the entire document. A detailed description of the data generation function is presented in Figure 1. The benefit of using a synthetic dataset is that it allows for creation of large volumes of data at a marginal cost. The data generation process also allows to collect additional metadata that can be used in the training process. In our case, the metadata contains information about the original location of the extracted claim in the source document and the locations in the claim where text transformations were applied.

Our data generation process draws inspiration from data augmentation and adversarial example generation techniques in NLP (Iyyer et al., 2018; Wu et al., 2019; Zhang et al., 2019; Wei and Zou, 2019). The proposed process incorporates both semantically invariant ($\mathcal{T}^+$), and variant ($\mathcal{T}^-$) text transformations to generate novel claims with `CORRECT` and `INCORRECT` labels accordingly. This work uses the following transformations:

Paraphrasing A paraphrasing transformation covers cases where source document sentences are rephrased by the summarization model. Paraphrases were produced by backtranslation using Neural Machine Translation systems (Iyyer et al., 2018). The claim sentence was translated to an intermediate language and translated back to English yielding a semantically-equivalent sentence with minor syntactic and lexical changes. *French*, *German*, *Chinese*, *Spanish*, and *Russian* were used as intermediate languages. These languages were chosen based on the performance of recent NMT systems with the expectation that well-performing languages could ensure better translation quality.

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Table 2: Examples of text transformations used to generate training data. Green and red text highlight the changes made by the transformation. Paraphrasing is a semantically invariant transformation, Sentence negation, entity, pronoun, and number swaps are semantically variant transformation.

We used the Google Cloud Translation API [1] for translations.

Entity and Number swapping To learn how to identify examples where the summarization model used incorrect numbers or entities during generation we used the Entity and Number swapping transformation. An NER system was applied to both the claim and source document to extract all entities. To generate a novel, semantically changed claim, an entity in the claim sentence was replaced with an entity from the document. Both of the swapped entities were chosen at random while ensuring that they were unique. Extracted entities were divided into two groups, *named entities*, covering person, location and institution names, and *numbers*, such as dates and all other numeric values. Entities were swapped within their respective groups. We used the SpaCy NER tagger (Honnibal and Montani, 2017).

Pronoun swapping To learn how to find incorrect pronoun use in claims we used a pronoun swapping transformation. First, all gender-specific pronouns were first extracted from the claim. Next, a randomly chosen pronoun was swapped with a different one from the same pronoun group to ensure syntactic correctness, i.e. a possessive pronoun could only be replaced with another possessive pronoun. New sentences were considered semantically variant.

[1] https://cloud.google.com/translate/

Sentence negation To give the consistency checking model the ability to handle negated sentences we used a sentence negation transformation. First, a claim was scanned in search of auxiliary verbs. To switch the meaning, a randomly chosen auxiliary verb was replaced with its negation. Positive sentences would be negated by adding *not* or *n't* after the verb, negative sentences would be switched by removing the negation.

Noise injection Given that verified summaries are generated by neural networks, they should be expected to contain certain types of noise. In order to make the factual consistency model robust to such generation errors, training examples were injected with noise. For each token in a claim the decision was made whether noise should be added at the given position with a preset probability. If noise should be injected, the token was randomly duplicated or removed from the sequence. Examples of all transformations are shown in Table 2.

3.2 Development and test data

Apart from the synthetic training set, separate, manually annotated, validation and test sets were created. Both of the annotated sets used summaries output by state-of-the-art summarization models. Each summary was split into sentences and all *(document, sentence)* pairs were annotated by the authors of this work. Since the focus was to collect data for verifying the factual consistency of summarization models, any unreadable sentences

Require:
S - set of source documents
$\mathcal{T}^+$ - set of semantically invariant transformations
$\mathcal{T}^-$ - set of semantically variant transformations

function GENERATE_DATA($\mathcal{S}, \mathcal{T}^+, \mathcal{T}^-$)
  $\mathcal{D} \leftarrow \emptyset$ ▷ set of generated data points
  for doc **in** $\mathcal{S}$ **do**
    $doc\_sents \leftarrow$ sentence_tokenizer(doc)
    $sent \leftarrow$ choose_random(doc_sents)
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, sent, +)\}$
    for fn **in** $\mathcal{T}^+$ **do**
      $new\_sent \leftarrow$ fn($doc, sent$)
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, +)\}$
    end for
  end for

  for $example$ **in** $\mathcal{D}$ **do**
    $doc, sent, \_ \leftarrow example$
    for fn **in** $\mathcal{T}^-$ **do**
      $new\_sent \leftarrow$ fn($doc, sent$)
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, -)\}$
    end for
  end for
  return $\mathcal{D}$
end function

Figure 1: Procedure to generate synthetic training data. $\mathcal{S}$ is a set of source documents, $\mathcal{T}^+$ is a set of semantically invariant text transformations, $\mathcal{T}^-$ is a set of semantically variant text transformations, $+$ is a positive label, $-$ is a negative label.

caused by poor generation were not labeled. The validation set consists of 931 examples, the test set contains 503 examples. The model outputs used for annotation were provided by the authors of papers: Hsu et al. (2018); Gehrmann et al. (2018); Jiang and Bansal (2018); Chen and Bansal (2018); See et al. (2017); Kryściński et al. (2018); Li et al. (2018); Pasunuru and Bansal (2018); Zhang et al. (2018); Guo et al. (2018).

Effort was made to collect a larger set of annotations through crowdsourcing platforms, however the inter-annotator agreement and general quality of annotations was too low to be considered reliable. This aligns with the conclusions of (Falke et al., 2019), where the authors showed that for the task of factual consistency the inter-annotator agreement coefficient κ reached 0.75 only when 12 annotations were collected for each example. This in turn yields high annotations costs that our approach aims to circumvent.

3.3 Models

Considering the significant improvements in natural language understanding (NLU) tasks (including NLI) coming from using pre-trained Transformer-based models [2], we decided to use BERT (Devlin et al., 2018) as the base model for our work. An *uncased*, *base* (110M params) BERT architecture was used as the starting checkpoint and fine-tuned on the generated training data. The source document and claim sentence were fed as input to the model and two-way classification (CONSISTENT/INCONSISTENT) was done using a single-layer classifier based on the [CLS] token. We refer to this model as the factual consistency checking model (**FactCC**).

We also trained a version of FactCC with additional span selection heads using supervision of start and end indices for selection and transformation spans in the source and claim. Span selection heads allow the model not only to classify the consistency of the claim, but also highlight spans in the source document that contain the support for the claim and spans in the claim where a possible mistake was made. We refer to this model as the factual consistency checking model with explanations (**FactCCX**).

Model	CNN/DailyMail		XSum	
	Accuracy *(weighted)*	F1-score	Accuracy *(weighted)*	F1-score
BERT+MNLI	51.39	0.86	**59.92**	0.58
BERT+FEVER	52.07	**0.88**	55.23	0.26
FactCC (ours)	72.65	0.86	54.11	**0.73**
FactCCX (ours)	**72.88**	0.87	53.05	0.60

Table 3: Performance of models evaluated by means of weighted (class-balanced) accuracy and F1 score on the manually annotated test set of CNN/DailyMail (this work) and XSum (Maynez et al., 2020).

4 Experiments

4.1 Experimental Setup

Training data was generated as described in Section 3.1 using news articles from the CNN/DailyMail (Nallapati et al., 2016) dataset as source documents. 1,003,355 training examples were created, out of which 50.2% were labeled as negative (INCONSISTENT) and the remaining 49.8% were labeled as positive (CONSISTENT). Models were evaluated in two settings: 1) with summaries from models trained on the CNN/DailyMail (Nallapati et al., 2016) dataset, which contains longer and more extractive reference summaries, and 2) with summaries from models trained on the XSum (Narayan et al.,

[2] http://gluebenchmark.com/leaderboard

Article
(CNN) Blues legend B.B. King was hospitalized for dehydration, though the ailment didn't keep him out for long. King's dehydration was caused by his Type II diabetes, but he "is much better," his daughter, Claudette King, told the Los Angeles Times. The legendary guitarist and vocalist released a statement thanking those who have expressed their concerns. "I'm feeling much better and am leaving the hospital today," King said in a message Tuesday. Angela Moore, a publicist for Claudette King, said later in the day that he was back home resting and enjoying time with his grandchildren. "He was struggling before, and he is a trouper," Moore said. "He wasn't going to let his fans down." (...)
Claim
Angela Moore was back home resting and enjoying time with his grandchildren.

Table 4: Example of a test pair correctly classified as *incorrect* and highlighted by our explainable model. Orange text indicates the span of the source documents that should contain support for the claim. Red text indicates the span of the claim that was selected as incorrect.

	Sentence Pair Ranking	
Model	Incorrect	Δ
Random	50.0%	-
DA (Falke et al., 2019)	42.6%	-7.4
InferSent (Falke et al., 2019)	41.3%	-8.7
SSE (Falke et al., 2019)	37.3%	-12.7
ESIM (Falke et al., 2019)	32.4%	-17.6
BERT (Falke et al., 2019)	35.9%	-14.1
BERT+FEVER (ours)	34.3%	-15.7
BERT+MNLI (ours)	32.1%	-17.9
FactCC (ours)	**30.0%**	**-20.0**

Table 5: Percentage of incorrectly ordered sentence pairs using different consistency prediction models and crowdsourced human performance on the dataset.

2018) dataset, which contains single-sentence, highly abstractive reference summaries. The CNN/DailyMail validation and test sets were manually annotated, as described in Section 3.2, while the XSum test data was collected by Maynez et al. (2020)

Models were implemented using the Transformers library (Wolf et al., 2019) written in PyTorch (Paszke et al., 2017). Models were trained for 10 epochs using batch size of 12 examples and learning rate of `2e-5`. Experiments were conducted on 8 Nvidia V100 GPUs, training took 23 hours on average. Best model checkpoints were chosen based on the performance on the validation set, final model performance was evaluated on the test set.

4.2 Results

To understand whether datasets for related tasks transfer to the task of verifying factual consistency of summarization models, we trained factual consistency checking models on the MNLI entailment data (Williams et al., 2018) and FEVER fact-checking data (Thorne et al., 2018). For fair comparison, before training, we removed examples assigned to the neutral class from both of the datasets. Table 3 shows the performance of trained models evaluated by means of class-balanced accuracy and F1 score. Both FactCC and FactCCX models substantially outperform classifiers trained on the MNLI and FEVER datasets when evaluated on the CNN/DailyMail test set. However, on the more abstractive XSum data, results show a reverse trend with the MNLI model achieving highest performance. Considering that both MNLI and FEVER datasets contain abstractive, human-written claims, while our data generation pipeline lacks multi-sentence paraphrasing transformations, such outcome was expected. In the case of more extractive outputs, the results suggests priority of weak-supervision in-domain over strong-supervision in adjacent domains for factual consistency checking.

To compare our models with other NLI models for factual consistency checking, we conducted the sentence ranking experiment described by Falke et al. (2019) using the test data provided by the authors. In this experiment an article sentence is paired with two claim sentences, *positive* and *negative*. The goal is to see how often a model assigns a higher probability of being correct to the positive rather than the negative claim. Results are presented in Table 5. Despite being trained in a *(document, sentence)* setting, our model transfers well to the *(sentence-sentence* setting and outperforms all other NLI models, including BERT fine-tuned on the MNLI dataset. We were unable to replicate the summary re-ranking experiment because the experimental test setup does not rely on data that can be reused.

Considering that strongly supervised data is

not available for factual consistency verification, proxy datasets must be used to train automatic verification models. Empirical results presented in this section suggest that when verifying less abstractive domains, it is more beneficial to train on weakly-supervised, but in-domain data, rather than to rely on the models ability to transfer knowledge from strongly supervised datasets in related domains. The experiments also highlight the necessity of extending the data generation pipeline with more abstractive, multi-sentence paraphrasing transformations as part of future work.

In addition to improved performance, using synthetic data allows to train models with explainable components, such as FactCCX. Examples of span selections generated by the model are show in Table 4. The test set consists of model-generated summaries that do not have annotations for quantifying the quality of spans returned by FactCCX. Instead, span quality is measured through human evaluation and discussed in Section 5.

5 Analysis

To further understand performance of our proposed models, we conducted human-based experiments and manually inspected model outputs.

5.1 Human Studies

Experiments using human annotators demonstrated that the span highlights returned by FactCCX are useful tools for researchers and crowdsource workers manually assessing the factual consistency of summaries. For each experiment, examples were annotated by 3 human judges selected from English-speaking countries. These experiments used 100 examples sampled from the manually annotated CNN/DM test set. Data points were sampled to ensure an equal split between `CONSISTENT` and `INCONSISTENT` examples.

To establish whether model-generated highlighted spans in the article and claim are helpful for the task of factual consistency checking, we hired human annotators to complete the mentioned task. Each of the verified *document-sentence* pairs was augmented with the highlighted spans output by FactCCX. Judges were asked to evaluate the correctness of the claim and instructed to use the provided segment highlights as suggestions. After the annotation task, judges were asked whether the highlighted spans were helpful for solving the task. The helpfulness of *article* and *claim* highlights were evaluated separately. The left part of Table 6 presents the results of the survey. A combined number of 91.75% annotators found the article highlights at least somewhat helpful, 81.33% of annotators found the claim highlights at least somewhat helpful. To ensure that low-quality judges do not bias the scores, we applied different data filters to the annotations: *Raw Data* considered all submitted annotations, *Golden Aligned* only considered annotations where the annotator-assigned label aligned with the author-assigned label for the example, *Majority Aligned* only considered examples where the annotator-assigned aligned with the majority-vote label assigned for the example by all judges. As shown in Table 6, filtering the annotations does not yield substantial changes in the helpfulness assessment.

Despite instructing the annotators to consider the provided highlights only as a suggestion when solving the underlying task, the annotators perception of the task could have been biased by the model-highlighted spans. To check how well the model-generated span highlights align with an unbiased human judgement, we repeated the previous experiment with only one change: model-generated highlights were not displayed to the annotators. The annotators were asked to solve the underlying task and highlight the spans of the source and claim that they found adequate. Using the annotations provided by the judges, we computed the overlap between the model-generated spans and unbiased human spans. Results are shown in the right part of Table 6. The overlap between spans was evaluated using two metrics. *Accuracy* is based on a binary score of whether the entire model-generated span was contained within the human-selected span. *F1* score is computed between the tokens of the model-generated span and the tokens of the human-selected span. Results show 65.33% and 65.66% *accuracy* and 0.6207 and 0.6650 *F1* for the *article* and *claim* highlights, respectively. We again applied different data filters to understand how the quality of annotations affects the score. We found that in this case, *accuracy* and *F1* score were higher in the *Majority Aligned* than in the case of using *Raw Data*, and performance increases dramatically in the *Majority Aligned*. This suggests that when model-generated highlights are not provided, the task is less constrained and requires more careful preci-

	Model Highlight Helpfulness			Model-Annotator Highlight Overlap	
Annotation subset	Helpful	Somewhat Helpful	Not Helpful	Accuracy	F1 score
Article Highlights					
Raw Data	79.21%	12.54%	8.25%	65.33%	0.6207
Golden Aligned	77.73%	12.66%	9.61%	74.87%	0.7161
Majority Aligned	81.11%	11.48%	7.41%	69.88%	0.6679
Claim Highlights					
Raw Data	64.44%	16.89%	18.67%	65.66%	0.6650
Golden Aligned	67.28%	16.05%	16.67%	80.54%	0.8190
Majority Aligned	67.17%	16.67%	16.16%	69.48%	0.6992

Table 6: Quality of spans highlighted in the *article* and *claim* by the FactCCX model evaluated by human annotators. The left side shows whether the highlights were considered helpful for the task of factual consistency annotations. The right side shows the overlap between model generated and human annotated highlights. Different rows show how the scores change depending on how the collected annotations are filtered.

sion on the part of judges.

To further understand the affects of providing model-generated highlights to annotators, we ran two factual consistency annotation tasks designed to test annotation efficiency. In the first, highlights were provided to the annotators. In the second, annotators did not receive highlights. In both, we measured the average time spent by an annotator on the task and the inter-annotator agreement of annotations. Results are shown in Table 7. When completing the task with highlights, annotators were 21% faster, and the inter-annotator agreement, measured with Fleiss' κ, increased by 38%.

Crowdsourcing experiments support the hypothesis that model-generated highlights from FactCCX can play a valuable role in supporting human-based factual consistency checking.

	Task without model highlights	Task with model highlights
Average work time (sec)	224.89	178.34
Inter-annotator agreement (κ)	0.1571	0.2526

Table 7: Annotation speed and inter-annotator agreement measured for factual consistency checking with and without assisting, model generated highlights.

5.2 Qualitative Study

To better understand the limitations of our proposed approach, we manually inspected examples that were misclassified by our models. The majority of errors were related to commonsense mistakes made by summarization models. Humans can easily spot such errors, but they are difficult to capture with transformations necessary to generate data for weak supervision.

Our analysis also showed that the proposed models fail to correctly classify examples where the verified claim is highly abstractive. This is especially true when the claim finds support in multiple spans distant from each other in the source document, as mostly found in the XSum dataset. Additionally, the current set of transformations do not adequately capture temporal inconsistencies or incorrect coreference. Nonetheless, the current transformations yield models already useful to humans by their own judgment; this analysis points toward key areas for future work. Correct and incorrect model predictions are presented in Appendix A.

6 Conclusions

We introduced a novel approach for factual consistency checking of summaries generated by abstractive neural models. In our approach, models are trained to perform factual consistency checking on the *document-sentence* level, which allows them to handle a broader range of errors in comparison to previously proposed *sentence-sentence* approaches. Models are trained using artificially generated, weakly-supervised data created based on insights coming from the analysis of errors made by state-of-the-art summarization models. Quantitative studies showed that on less abstractive domains, such as CNN/DailyMail news articles, our proposed approach outperforms other models trained on existing textual entailment and fact-checking data, motivating our use

of weak-supervision over transfer learning from related domains. Experiments with human annotators showed that our proposed approach, including an explainable factual consistency checking model, can be a valuable tool for assisting humans in factual consistency checking. Shortcomings of our approach explained in Section 5.2 can serve as guidelines for future work. We hope that this work will encourage continued research into factual consistency checking of abstractive summarization models.

Acknowledgements

We thank Nitish Shirish Keskar, Dragomir Radev, Ben Krause, and Wenpeng Yin for reviewing this manuscript and providing valuable feedback, and Shashi Narayan for help with experiments on the XSum dataset.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. 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.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *ACL (1)*, pages 675–686. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *HLT-NAACL*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *CoRR*, abs/1808.09381.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *EMNLP*, pages 4098–4109. 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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. http://spacy.io.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1875–1885.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *EMNLP*, pages 4067–4077. Association for Computational Linguistics.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *CoRR*, abs/1908.08960.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *EMNLP*, pages 1808–1817. Association for Computational Linguistics.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with structural regularization. In *EMNLP*, pages 4078–4087. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NIPS*, pages 6297–6308.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. *CoRR*, abs/1809.01478.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Ramesh Nallapati, Bowen Zhou, Çağlar Gülçehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Proceedings of SIGNLL Conference on Computational Natural Language Learning*.

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*, Brussels, Belgium.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. *CoRR*, abs/1804.06451.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. In *ICLR*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of EMNLP*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. *CoRR*, abs/1803.05355.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. 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)*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art natural language processing.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT contextual augmentation. In *Computational Science - ICCS 2019 - 19th International Conference, Faro, Portugal, June 12-14, 2019, Proceedings, Part IV*, pages 84–95.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. 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*.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018. On the abstractiveness of neural document summarization. In *EMNLP*, pages 785–790. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1298–1308.

A Model predictions

Example predictions made by the FactCC model.

Example 1

Label: CONSISTENT
Prediction: INCONSISTENT

Article (CNN/DM)

(cnn) james best, best known for his portrayal of bumbling sheriff rosco p. coltrane on tv's "the dukes of hazzard," died monday after a brief illness. he was 88. best died in hospice in hickory, north carolina, of complications from pneumonia, said steve latshaw, a longtime friend and hollywood colleague. although he'd been a busy actor for decades in theater and in hollywood, best didn't become famous until 1979, when "the dukes of hazzard's" cornpone charms began beaming into millions of american homes almost every friday night. for seven seasons, best's rosco p. coltrane chased the moonshine-running duke boys back and forth across the back roads of fictitious hazzard county, georgia, although his "hot pursuit" usually ended with him crashing his patrol car. although rosco was slow-witted and corrupt, best gave him a childlike enthusiasm that got laughs and made him endearing. his character became known for his distinctive "kew-kew-kew" chuckle and for goofy catchphrases such as "cuff 'em and stuff 'em!" upon making an arrest. among the most popular shows on tv in the early '80s, "the dukes of hazzard" ran until 1985 and spawned tv movies, an animated series and video games. several of best's "hazzard" co-stars paid tribute to the late actor on social media. (...)

Claim

"hazzard" ran from 1979 to 1985 and was among the most popular shows on tv.

Example 2

Label: CONSISTENT
Prediction: INCONSISTENT

Article (CNN/DM)

(cnn) the attorney for a suburban new york cardiologist charged in what authorities say was a failed scheme to have another physician hurt or killed is calling the allegations against his client "completely unsubstantiated." appearing Saturday morning on cnn's "new day," randy zelin defended his client, dr. anthony moschetto, who faces criminal solicitation, conspiracy, burglary, arson, criminal prescription sale and weapons charges in connection to what prosecutors called a plot to take out a rival doctor on long island. "none of anything in this case has any evidentiary value," zelin told cnn's christi paul. "it doesn't matter what anyone says, he is presumed to be innocent." moschetto, 54, pleaded not guilty to all charges wednesday. he was released after posting $2 million bond and surrendering his passport. zelin said that his next move is to get dr. moshetto back to work. "he's got patients to see. This man, while he was in a detention cell, the only thing that he cared about were his patients. and amazingly, his patients were flooding the office with calls, making sure that he was ok," zelin said. (...)

Claim

a lawyer for dr. anthony moschetto says the charges against him are baseless.

Example 3

Label: INCONSISTENT
Prediction: CONSISTENT

Article (CNN/DM)

(cnn) north korea accused mexico of illegally holding one of its cargo ships wednesday and demanded the release of the vessel and crew. the ship, the mu du bong, was detained after it ran aground off the coast of mexico in july. mexico defended the move wednesday, saying it followed proper protocol because the company that owns the ship, north korea's ocean maritime management company, has skirted united nations sanctions. "because the company has avoided the sanctions imposed by the u.n. security council, the mexican government is acting on the basis of its international obligations as a responsible u.n. member state," the permanent mission of mexico to the united nations said. the security council blacklisted ocean maritime management in july, saying it "played a key role in arranging the shipment of concealed arms and related materiel" on another ship, the chong chon gang, which was detained by panama in 2013. but an myong hun, north korea's deputy ambassador to the united nations, said there was no reason to hold the mu du bong and accused mexico of violating the crew members' human rights by keeping them from their families.

Claim

north korea accused mexico of using one of its cargo ships.

Example 4

Label: CONSISTENT
Prediction: CONSISTENT

Article (CNN/DM)

(cnn) spoiler alert! it's not just women getting cloned. that was the big twist at the end of "orphan black's" second season. the kickoff to the new season leads the list of six things to watch in the week ahead. 1. "orphan black," 9 p.m. et, saturday, april 18, bbc america. the cloning cult sci-fi series remains one of the most critically acclaimed shows on tv, thanks in large part to the performance of tatiana maslany, who has taken on at least six roles on the show so far, including a newly introduced transgender clone. maslany told reporters this week that we can expect even more impressive scenes with multiple clones. "we like to push the boundaries of what we're able to do and the limits of those clone scenes," she said. (...)

Claim

critically acclaimed series "orphan black" returns.

Example 5

Label: INCONSISTENT
Prediction: INCONSISTENT

Article (CNN/DM)

boston (cnn) dzhokhar tsarnaev's bombs tore through their bodies: singeing flesh, shattering bones, shredding muscles and severing limbs. but on tuesday, jurors also began to hear about the holes his bombs left in the hearts of the survivors and the families of the dead. now that he has been found guilty on every count, the jury must decide whether boston marathon bomber tsarnaev, 21, should live or die for what he has done. this is the victim impact part of the case, and the testimony was heartbreaking. four young people are gone, and grief fills the spaces they once occupied. a father with a shock of white hair cried for the daughter he called "princess." "krystle was the light of my life. she was extremely smart, hardworking, beautiful, every father's dream. i miss her a lot," said william a. campbell sr., dabbing at his eyes as he described his daughter, a 29-year-old restaurant manager who was killed in the first blast at the 2013 boston marathon. she was the one who could round up the family and put on big celebrations, he said. (...)

Claim

dzhokhar tsarnaev, 21, was killed in the first blast at the 2013 boston marathon.

Example 6

Label: INCONSISTENT
Prediction: INCONSISTENT

Article (CNN/DM)

(the hollywood reporter) the author of a 2006 novel has accused the "avengers" director and "cabin" director drew goddard of stealing his idea. with just weeks until his box-office victory lap for "avengers: age of ultron," joss whedon is now facing a lawsuit accusing him of stealing the idea for the 2012 meta-horror movie the cabin in the woods. whedon produced and co-wrote the script for cabin with director drew goddard, a writer on whedon's "buffy the vampire slayer" and a fanboy favorite in his own right, with credits that include netflix's "daredevil" (and reportedly may soon include sony's upcoming spider-man projects). whedon and goddard are named as defendants, along with lionsgate and whedon's mutant enemy production company, in the complaint filed monday in california federal court. joss whedon slams 'jurassic world 'clip as "'70s - era sexist" in the complaint, peter gallagher (no, not that peter gallagher) claims whedon and goddard took the idea for "the cabin in the woods" from his 2006 novel "the little white trip: a night in the pines." he's suing for copyright infringement and wants $10 million in damages. gallagher is basing his claim on the works' similar premises: both feature a group of young people terrorized by monsters while staying at a cabin in what is revealed to be (spoiler alert) a horror-film scenario designed by mysterious operators. (...)

Claim

joss whedon claims whedon and goddard took the idea for "the cabin in the woods".

Example 7

Label: INCONSISTENT
Prediction: INCONSISTENT

Article (XSUM)

More than 5,300 bottles of alcohol were seized by the investigators in the southern city of Liuzhou. They also found packets of a white powder called Sildenafil, better known as the anti-impotence drug Viagra. Police in the Guangxi region are now investigating the two distillers. The Liuzhou Food and Drug Administration said (in Chinese) that the powder was added to three different types of 'baijiu' - a strong, clear spirit that is the most popular drink in China. They said the haul was worth up to 700,000 yuan. Doctors recommend that adults requiring prescription should take only one dose of Viagra a day, with a lower dose for those over the age of 65. China continues to face widespread food safety problems. In June, police in cities across China seized more than 100,000 tonnes of smuggled meat, some of which was more than 40 years old. The 2008 tainted milk scandal outraged the nation. Some 300,000 people were affected and at least six babies died after consuming milk adulterated with melamine.

Claim

police in southern china have seized more than 1,000 alcohol bottles and seized more than 1,200 bottles of contaminated milk, local media report.

Example 8

Label: CONSISTENT
Prediction: CONSISTENT

Article (XSUM)

The victim was queuing for food at the branch in St George's Street, Canterbury at about 02:15 GMT on Friday when the assault occurred. Investigating officers said three men entered the restaurant and began being noisy and bumping into people. It is believed one of the group then set light to the woman's hair. Officers have released CCTV images of three men they are keen to speak to regarding the attack. Det Sgt Barry Carr said: "Fortunately the fire was put out quickly and the victim was not seriously hurt, but things could clearly have turned out much worse. T̈his was a nasty and extremely dangerous thing to do, and I urge anyone who recognises the men in the CCTV images to contact me as soon as possible."

Claim

a woman was assaulted and assaulted in a mcdonald's restaurant in kent, police have said.

Example 9

Label: INCONSISTENT
Prediction: INCONSISTENT

Article (XSUM)

Jung won aboard Sam, who was a late replacement when Fischertakinou contracted an infection in July. France's Astier Nicolas took silver and American Phillip Dutton won bronze as GB's William Fox-Pitt finished 12th. Fox-Pitt, 47, was competing just 10 months after being placed in an induced coma following a fall. The three-time Olympic medallist, aboard Chilli Morning, produced a faultless performance in Tuesday's final show-jumping phase. But the former world number one's medal bid had already been ruined by a disappointing performance in the cross-country phase on Monday. He led after the dressage phase, but dropped to 21st after incurring several time penalties in the cross country. Ireland's Jonty Evans finished ninth on Cooley Rorkes Drift. Why not come along, meet and ride Henry the mechanical horse at some of the Official Team GB fan parks during the Rio Olympics? Find out how to get into equestrian with our special guide. Subscribe to the BBC Sport newsletter to get our pick of news, features and video sent to your inbox.

Claim

great britain's eventers missed out on a bronze medal at the rio olympics after losing in the dressage.

Example 10

Label: INCONSISTENT
Prediction: CONSISTENT

Article (XSUM)

A review for the Commission on Local Tax Reform said there was no ”magic bullet” to cure defects in the system. It said the council tax had built-in problems ”from day one” but a failure to modify it had stored up more difficulties for policy makers. The commission, set up by the Scottish government and council body Cosla, will report back later this year. Prof Kenneth Gibb, from the University of Glasgow, was asked to review different systems of local taxation across the world. He found that a tax on property was used by almost all OECD countries and was seen by academics as a ”good tax” because it was stable, difficult to avoid and could have a desirable impact on housing markets. But it also generated confusion with taxpayers unclear whether it was a tax on wealth or a charge for services such as refuse collection. Some felt it was unfair because it was not linked to current income. Prof Gibb noted that a local income tax, used by many countries, was generally perceived as fairer. But he found such a system created difficulties for local authorities because it meant their income fluctuated. There was also little opportunity to vary tax rates to reflect local priorities. He said: ”It is clear there is no magic bullet. ”Past experience from the UK and across the world shows that reform is always going to be difficult and will inevitably be bound up with the previous experiences and traumas of past reform. ”So whilst the current council tax has many deficiencies, change and reform is a major undertaking.” The commission now intends to hold a public consultation across Scotland before publishing its report in the autumn. A Scottish government spokesman said ministers consider the current council tax system ”as a whole to be unfair”. He added: ”That is why, along with our local government partners, we have established the cross-party Commission on Local Tax Reform to examine fairer alternatives. ”The Scottish government awaits the commission’s report, which is due in the Autumn.

Claim

the scottish government has been accused of ”unfairly unfair” by a watchdog after a report found that a council tax system was not stored

Example 11

Label: INCONSISTENT
Prediction: CONSISTENT

Article (XSUM)

The crash happened at about 14:15 BST on the B1191 at Thornton, near Woodhall Spa. Lincolnshire Police said the motorcyclist killed in the collision lived locally, but has not released any further details. The tractor driver was not injured. The force has appealed for witnesses to the collision to come forward. The B1191 was closed in both directions between the B1192 Tattershall Road junction in Woodhall Spa and the A158 Jubilee Way junction in Horncastle

Claim

a motorcyclist killed in a crash with a tractor and a tractor in lincolnshire has been named.

Example 12

Label: INCONSISTENT
Prediction: CONSISTENT

Article (XSUM)

The Australian, 21, beat world number 29 Querrey 6-4 6-4 in 53 minutes to progress to the second round. Kyrgios, ranked a career-high 12th in the world, won the Japan Open on Sunday and is closing in on the top 10. ”I was just a bit bored at times,” said Kyrgios, when asked why he was not his usual vocal self against Querrey. ”I was feeling very tired. It was just tough. I’m just tired so maybe I just wanted to get the job done.” Kyrgios said his success in Japan, and the travelling involved in playing at the Qi Zhong Stadium, an hour from Shanghai city centre, had taken its toll. ”I didn’t have the greatest sleep last night and obviously got in late the day before,” he said. ”The ride to the courts isn’t great either.” It was at the Shanghai Masters last year that Kyrgios was fined $1,500 for a foul-mouthed outburst, describing the tournament a ”circus”.

Claim

australia’s nick kyrgios said he was ”not afraid to sleep” after reaching the second round of the shanghai masters.