Enhancing Factual Consistency of Abstractive Summarization

Chenguang Zhu[1], **William Hinthorn**[1], **Ruochen Xu**[1], **Qingkai Zeng**[2], **Michael Zeng**[1], **Xuedong Huang**[1], **Meng Jiang**[2]
[1] Microsoft Cognitive Services Group
[2] University of Notre Dame
{chezhu,wihintho,ruox,nzeng,xdh}@microsoft.com
{qzeng,mjiang2}@nd.edu

Abstract

Automatic abstractive summaries are found to often distort or fabricate facts in the article. This inconsistency between summary and original text has seriously impacted its applicability. We propose a fact-aware summarization model FASUM to extract and integrate factual relations into the summary generation process via graph attention. We then design a factual corrector model FC to automatically correct factual errors from summaries generated by existing systems. Empirical results[1] show that the fact-aware summarization can produce abstractive summaries with higher factual consistency compared with existing systems, and the correction model improves the factual consistency of given summaries via modifying only a few keywords.

1 Introduction

Text summarization models aim to produce an abridged version of long text while preserving salient information. Abstractive summarization is a type of such models that can freely generate summaries, with no constraint on the words or phrases used. This format is closer to human-edited summaries and is both flexible and informative. Thus, there are numerous approaches to produce abstractive summaries (See et al., 2017; Paulus et al., 2017; Dong et al., 2019; Gehrmann et al., 2018).

However, one prominent issue with abstractive summarization is factual inconsistency. It refers to the *hallucination* phenomenon that the summary sometimes distorts or fabricates the facts in the article. Recent studies show that up to 30% of the summaries generated by abstractive models contain such factual inconsistencies (Kryściński et al., 2019b; Falke et al., 2019), raising concerns about the credibility and usability of these systems.

[1]We provide the prediction results of all models at https://github.com/zcgzcgzcg1/FASum/.

Article	**Real Madrid ace Gareth Bale** treated himself to a Sunday evening BBQ... The Welsh wizard was ... **scoring twice** and assisting another in an impressive victory... **Cristiano Ronaldo scored five goals** against Granada on Sunday ...
BOTTOMUP	... **The Real Madrid ace** scored five goals against Granada on Sunday. The Welsh wizard was in impressive form for...
SEQ2SEQ	...**Gareth Bale scored five** and assisted another in an impressive win in Israel...
FASUM (Ours)	...**Gareth Bale scored twice** and helped his side to a sensational 9-1 win. **Cristiano Ronaldo scored five goals** against Granada on Sunday...

Table 1: Example article and summary excerpts from CNN/DailyMail dataset.

Table 1 demonstrates an example article and excerpts of generated summaries. As shown, the article mentions that Real Madrid ace Gareth Bale scored twice and Cristiano Ronaldo scored five goals. However, both BOTTOMUP (Gehrmann et al., 2018) and SEQ2SEQ wrongly states that Bale scored five goals. Comparatively, our model FASUM generates a summary that correctly exhibits the fact in the article. And as shown in Section 4.6.1, our model achieves higher factual consistency not just by making more copies from the article.

On the other hand, most existing abstractive summarization models apply a conditional language model to focus on the token-level accuracy of summaries, while neglecting semantic-level consistency between the summary and article. Therefore, the generated summaries are often high in token-level metrics like ROUGE (Lin, 2004) but lack factual consistency. In view of this, we argue that a robust abstractive summarization system must be equipped with factual knowledge to accurately summarize the article.

In this paper, we represent facts in the form of knowledge graphs. Although there are numerous

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 718–733
June 6–11, 2021.

Article	The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. This year, **Israel's Memorial Day** commemoration is for bereaved family members such as Braham. "Now I truly understand everyone who has lost a loved one," Braham said. Her husband, **Philippe Braham**, was one of 17 people killed in January's terror attacks in Paris... As Israel mourns on the nation's remembrance day, **French Prime Minister** Manuel Valls announced after his weekly Cabinet meeting that French authorities had foiled a terror plot...
BOTTOMUP	**Valerie Braham** was one of 17 people killed in January 's terror attacks in Paris. **France's memorial day** commemoration is for bereaved family members as Braham. **Israel's Prime Minister** says the terror plot has not been done.
Corrected by FC	**Philippe Braham** was one of 17 people killed in January's terror attacks in Paris. **Israel's memorial day** commemoration is for bereaved family members as Braham. **France's Prime Minister** says the terror plot has not been done.

Table 2: Example excerpts of an article from CNN/DailyMail and the summary generated by BOTTOMUP. Factual errors are marked in red. The correction made by our model FC are marked in green.

efforts in building commonly applicable knowledge graphs such as ConceptNet (Speer et al., 2017), we find that these tools are more useful in conferring commonsense knowledge. In abstractive summarization for contents like news articles, many entities and relations are previously unseen. Plus, our goal is to produce summaries that do not conflict with the facts in the article. Thus, we propose to extract factual knowledge from the article itself.

We employ the information extraction (IE) tool OpenIE (Angeli et al., 2015) to extract facts from the article in the form of relational tuples: (subject, relation, object). This graph contains the facts in the article and is integrated in the summary generation process.

Then, we use a graph attention network (Veličković et al., 2017) to obtain the representation of each node, and fuse that into a transformer-based encoder-decoder architecture via attention. We denote this model as the Fact-Aware Summarization model, FASUM.

In addition, to be generally applicable for all existing summarization systems, we propose a Factual Corrector model, FC, to help improve the factual consistency of any given summary. We frame the correction process as a seq2seq problem: the input is the original summary and the article, and the output is the corrected summary. FC has the same architecture as UniLM (Dong et al., 2019) and initialized with weights from RoBERTa-Large (Liu et al., 2019). We finetune it as a denoising autoencoder. The training data is synthetically generated via randomly replacing entities in the ground-truth summary with wrong ones in the article. As shown in Table 2, FC makes three corrections, replacing the original wrong entities which appear elsewhere in the article with the right ones.

In the experiments, we leverage an independently trained BERT-based (Devlin et al., 2018) factual consistency evaluator (Kryściński et al., 2019b). Results show that on CNN/DailyMail, FASUM obtains 0.6% higher fact consistency scores than UNILM (Dong et al., 2019) and 3.9% higher than BOTTOMUP (Gehrmann et al., 2018). Moreover, after correction by FC, the factual score of summaries from BOTTOMUP increases 1.4% on CNN/DailyMail and 0.9% on XSum, and the score of summaries from TCONVS2S increases 3.1% on XSum. We also conduct human evaluation to verify the effectiveness of our models.

We further propose an easy-to-compute model-free metric, relation matching rate (RMR), to evaluate factual consistency given a summary and the article. This metric employs the extracted relations and does not require human-labelled summaries. Under this metric, we show that our models can help enhance the factual consistency of summaries.

2 Related Work

2.1 Abstractive Summarization

Abstractive text summarization has been intensively studied in recent literature. Rush et al. (2015) introduces an attention-based seq2seq model for abstractive sentence summarization. See et al. (2017) uses copy-generate mechanism that can both produce words from the vocabulary via a generator and copy words from the article via a pointer. Paulus et al. (2017) leverages reinforcement learning to improve summarization quality. Gehrmann et al. (2018) uses a content selector to over-determine phrases in source documents that helps constrain the model to likely phrases. Zhu et al. (2019) defines a pretraining scheme for summarization and produces a zero-shot abstractive summarization model. Dong et al. (2019) employs different masking techniques for both NLU and NLG tasks,

resulting in the UNILM model. Lewis et al. (2019) employs denoising techniques to help generation tasks including summarization.

2.2 Fact-Aware Summarization

Entailment models have been used to evaluate and enhance factual consistency of summarization. Li et al. (2018) co-trains summarization and entailment and employs an entailment-aware decoder. Falke et al. (2019) proposes using off-the-shelf entailment models to rerank candidate summary sentences to boost factual consistency.

Zhang et al. (2019b) employs descriptor vectors to improve factual consistency in medical summarization. Cao et al. (2018) extracts relational information from the article and maps it to a sequence as an additional input to the encoder. Gunel et al. (2019) employs an entity-aware transformer structure for knowledge integration, and Matsumaru et al. (2020) improves factual consistency of generated headlines by filtering out training data with more factual errors. In comparison, our model utilizes the knowledge graph extracted from the article and fuses it into the generated text via neural graph computation.

To correct factual errors, Dong et al. (2020) uses pre-trained NLU models to rectify one or more wrong entities in the summary. Concurrent to our work, Cao et al. (2020) employs the generation model BART (Lewis et al., 2019) to produce corrected summaries.

Several approaches have been proposed to evaluate a summary's factual consistency (Kryściński et al., 2019a; Goodrich et al., 2019; Maynez et al., 2020). Zhang et al. (2019a) employs BERT to compute similarity between pairs of words in the summary and article. Wang et al. (2020); Durmus et al. (2020) use question answering accuracy to measure factual consistency. Kryściński et al. (2019b) applies various transformations on the summary to produce training data for a BERT-based classification model, FactCC, which shows a high correlation with human metrics. Therefore, we use FactCC as the factual evaluator in this paper.

3 Model

3.1 Problem Formulation

We formalize abstractive summarization as a supervised seq2seq problem. The input consists of a pairs of articles and summaries: $\{(X_1, Y_1), (X_2, Y_2), ..., (X_a, Y_a)\}$. Each article is tokenized into $X_i = (x_1, ..., x_{L_i})$ and each summary is tokenized into $Y_i = (y_1, ..., y_{N_i})$. In abstrative summarization, the model-generated summary can contain tokens, phrases and sentences not present in the article. For simplicity, in the following we will drop the data index subscript. Therefore, each training pair becomes $X = (x_1, ..., x_m), Y = (y_1, ..., y_n)$, and the model needs to generate an abstrative summary $\hat{Y} = (\hat{y}_1, ..., \hat{y}_{n'})$.

3.2 Fact-Aware Summarizer

We propose the Fact-Aware abstractive Summarizer, FASUM. It utilizes the seq2seq architecture built upon transformers (Vaswani et al., 2017). In detail, the encoder produces contextualized embeddings of the article and the decoder attends to the encoder's output to generate the summary.

To make the summarization model fact-aware, we extract, represent and integrate knowledge from the source article into the summary generation process, which is described in the following. The overall architecture of FASUM is shown in Figure 1.

3.2.1 Knowledge Extraction

To extract important entity-relation information from the article, we employ the Stanford OpenIE tool (Angeli et al., 2015). The extracted knowledge is a list of tuples. Each tuple contains a subject (S), a relation (R) and an object (O), each as a segment of text from the article. In the experiments, there are on average 165.4 tuples extracted per article in CNN/DailyMail (Hermann et al., 2015) and 84.5 tuples in XSum (Narayan et al., 2018).

3.2.2 Knowledge Representation

We construct a knowledge graph to represent the information extracted from OpenIE. We apply the Levi transformation (Levi, 1942) to treat each entity and relation equally. In detail, suppose a tuple is (s, r, o), we create nodes s, r and o, and add edges s—r and r—o. In this way, we obtain an undirected knowledge graph $G = (V, E)$, where each node $v \in V$ is associated with text $t(v)$. During training, this graph G is constructed for each batch individually, i.e. there's no shared huge graph. One benefit is that the model can take unseen entities and relations during inference.

We then employ a graph attention network (Veličković et al., 2017) to obtain embedding $\boldsymbol{e}_j$ for each node v_j. The initial embedding of v_j is given by the last hidden state of a bidirectional LSTM

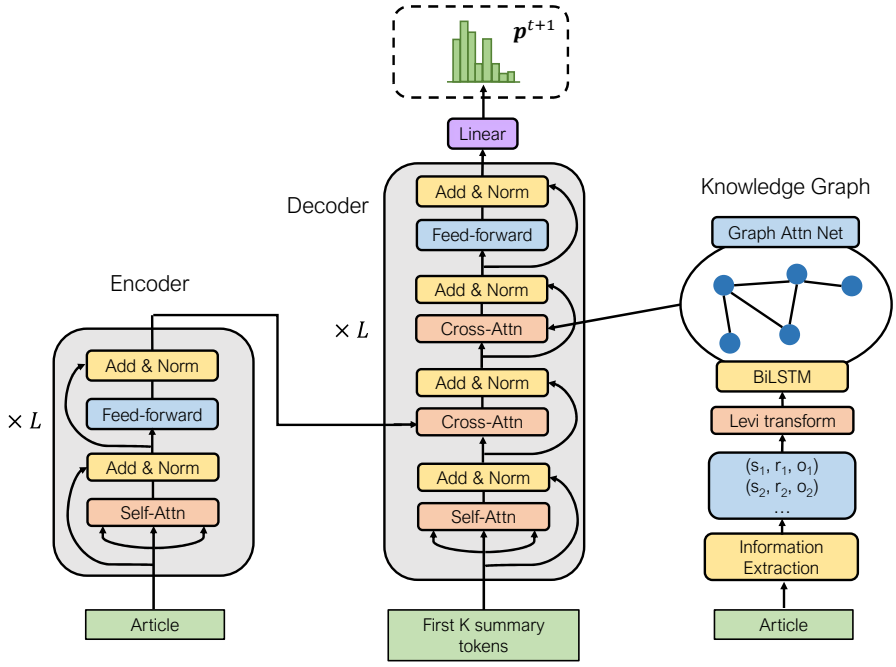

Figure 1: The model architecture of FASUM. It has L layers of transformer blocks in both the encoder and decoder. The knowledge graph is obtained from information extraction results and it participates in the decoder's attention.

applied to $t(v_j)$. In the experiment, we employ 2 graph attention layers.

3.2.3 Knowledge Integration

The knowledge graph embedding is obtained in parallel with the encoder. Then, apart from the canonical cross attention over the encoder's outputs, each decoder block also computes cross-attention over the knowlege graph nodes' embeddings:

$$\alpha_{ij} = \text{softmax}_j(\beta_{ij}) = \frac{\exp(\beta_{ij})}{\sum_{j \in V} \exp(\beta_{ij})} \quad (1)$$

$$\beta_{ij} = \boldsymbol{s}_i^T \boldsymbol{e}_j, \quad (2)$$

$$\boldsymbol{u}_i = \sum_{j \in V} \alpha_{ij} \boldsymbol{e}_j, \quad (3)$$

where $\{\boldsymbol{e}_j\}_{j=1}^{|V|}$ are the final embeddings of the graph nodes, and $\{\boldsymbol{s}_i\}_{i=1}^{t}$ are the decoder block's representation of the first t generated tokens.

3.2.4 Summary Generation

We denote the final output of the decoder as $\boldsymbol{z}_1, ..., \boldsymbol{z}_t$. To produce the next token y_{t+1}, we employ a linear layer $\boldsymbol{W}$ to project $\boldsymbol{z}_t$ to a vector of the same size of the dictionary. And the predicted distribution of y_{t+1} is obtained by:

$$\boldsymbol{p}^{t+1} = \sigma(\boldsymbol{W} \boldsymbol{z}_t) \quad (4)$$

During training, we use cross entropy as the loss function $\mathcal{L}(\theta) = -\sum_{t=1}^{n} \boldsymbol{y}_t^T \log(\boldsymbol{p}^t)$, where $\boldsymbol{y}_t$ is the one-hot vector for the t-th token, and θ represent the parameters in the network.

3.3 Fact Corrector

To better utilize existing summarization systems, we propose a Factual Corrector model, FC, to improve the factual consistency of any summary generated by abstractive systems. FC frames the correction process as a seq2seq problem: given an article and a candidate summary, the model generates a corrected summary with minimal changes to be more factually consistent with the article.

While FASum has a graph attention module in the transformer, preventing direct adaptation from pre-trained models, the FC model architecture adopts the design of the pre-trained model UniLM (Dong et al., 2019). We initiailized the model weights from RoBERTa-Large (Liu et al., 2019). The finetuning process is similar to training a denoising autoencoder. We use back-translation and entity swap for synthetic data generation. For example, an entity in the ground-truth summary is randomly replaced with another entity of the same type from the article. This modified summary and the article is sent to the corrector to recover the original summary. In the experiments, we gener-

ated 3.0M seq2seq data samples in CNN/DailyMail and 551.0K samples in XSum for finetuning. We take 10K samples in each dataset for validation and use the rest for training.

During inference, the candidate summary from any abstractive summarization system is concatenated with the article and sent to FC, which produces the corrected summary.

4 Experiments

4.1 Datasets

We evaluate our model on benchmark summarization datasets CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018). They contain 312K and 227K news articles and human-edited summaries respectively, covering different topics and various summarization styles.

4.2 Implementation Details

We use the Huggingface's (Wolf et al., 2019) implementation of transformer in BART (Lewis et al., 2019). We also inherit their provided hyperparameters of CNN/DailyMail and XSum for the beam search. The minimum summary length is 56 and 11 for CNN/Daily Mail and XSum, respectively. The number of beams is 4 for CNN/DailyMail and 6 for XSum.

In FASUM, both the encoder and decoder has 10 layers of 10 heads for attention. Teacher forcing is used in training. We use Adam (Kingma and Ba, 2014) as the optimizer with a learning rate of 2e-4.

The bi-LSTM to produce the initial embedding of graph nodes has a hidden state of size 64 and the graph attention network (GAT) has 8 heads and a hidden state of size 50. The dropout rate is 0.6 in GAT and 0.1 elsewhere.

We use the subword tokenizer SentencePiece (Kudo and Richardson, 2018). The dictionary is shared across all the datasets. The vocabulary has a size of 32K and a dimension of 720.

The correction model FC follows the UniLM (Dong et al., 2019) architecture initialized with weights from RoBERTa-Large (Liu et al., 2019). We fine-tune the model for 5 epochs with a learning rate of 1e-5 and linear warmup over the one-fifths of total steps and linear decay. During decoding, it uses beam search with a width of 2, and blocks trigram duplicates. The batch size during finetuning is 24. More details are presented in the Appendix.

4.3 Metrics

To evaluate factual consistency, we re-implemented and trained the FactCC model (Kryściński et al., 2019b). The model outputs a score between 0 and 1, where a higher score indicates better consistency between the input article and summary. The training of FactCC is independent of our summarizer so no parameters are shared. More details are in the Appendix.

We also employ the standard ROUGE-1, ROUGE-2 and ROUGE-L metrics (Lin, 2004) to measure summary qualities. These three metrics evaluate the accuracy on unigrams, bigrams and the longest common subsequence. We report the F1 ROUGE scores in all experiments. And the ROUGE-L score on validation set is used to pick the best model for both FASUM and FC.

4.4 Baselines

The following abstractive summarization models are selected as baseline systems. TCONVS2S (Narayan et al., 2018) is based on topic modeling and convolutional neural networks. BOTTOMUP (Gehrmann et al., 2018) uses a bottom-up approach to generate summarization. UNILM (Dong et al., 2019) utilizes large-scale pretraining to produce state-of-the-art abstractive summaries. We train the baseline models when the predictions are not available in their open-source repositories.

4.5 Results

As shown in Table 3, our model FASUM[2] outperforms all baseline systems in factual consistency scores in CNN/DailyMail and is only behind UNILM in XSum. In CNN/DailyMail, FASUM is 0.6% higher than UNILM and 3.9% higher than BOTTOMUP in factual score. Statistical test shows that the lead is statistically significant with p-value smaller than 0.05. The higher factual score of UNILM among baselines corroborates the findings in Maynez et al. (2020) that pre-trained models exhibit better factuality. But our proposed knowledge graph component can help the train-from-scratch FASUM model to excel in factual consistency.

We conduct ablation study to remove the knowledge graph component from FASUM, resulting in the SEQ2SEQ model. As shown, there is a clear drop in factual score: 2.8% in CNN/DailyMail and 0.9% in XSum. This proves that the constructed

[2]We have put code and all the generated summaries of all models in the supplementary materials.

knowledge graph can help increase the factual correctness of the generated summaries.

It's worth noticing that the ROUGE metric does not always reflect the factual consistency, sometimes even showing an inverse relationship, a phenomenon observed in multiple studies (Kryściński et al., 2019a; Maynez et al., 2020). For instance, although BOTTOMUP has 0.69 higher ROUGE-1 points than FASUM in CNN/DailyMail, there are many factual errors in its summaries, as shown in the human evaluation. On the other hand, to make sure the improved factual correctness of our models is not achieved by simply copying insignificant information from the article, we conduct analysis on abstractiveness in Section 4.6.1 and human evaluation in Section 4.6.3.

Furthermore, the correction model FC can effectively enhance the factual consistency of summaries generated by various baseline models, especially when the original summary has a relatively low factual consistency. For instance, on CNN/DM, the factual score of BOTTOMUP increases by 1.4% after correction. On XSum, after correction, the factual scores increase by 0.2% to 3.1% for all baseline models. Interestingly, FC can also boost the factual consistency of our FASUM model. Furthermore, the correction has a rather small impact on the ROUGE score, and it can improve the ROUGE scores of most models in XSum dataset.

We check and find that FC only makes modest modifications necessary to the original summaries. For instance, FC modifies 48.3% of summaries generated by BOTTOMUP in CNN/DailyMail. These modified summaries contain very few changed tokens: 94.4% of these corrected summaries contain 3 or fewer new tokens, while the summaries have on average 48.3 tokens.

In the appendix of supplementary materials, we show several examples of summaries given by FASUM and corrected by FC to demonstrate the improved factual consistency of summarization.

4.6 Insights

4.6.1 Novel n-grams

It has been shown in Durmus et al. (2020) that less abstractive summaries are more factual consistent with the article. Therefore, we inspect whether our models boost factual consistency simply by copying more portions of the article.

On XSum's testset, we compute the ratio of novel n-grams in summaries that do not appear in the article. Figure 2 shows that FASUM achieves the closest ratio of novel n-gram compared with reference summaries, and higher than BOTTOMUP and UNILM. This demonstrates that FASUM can produce highly abstractive summaries while ensuring factual consistency.

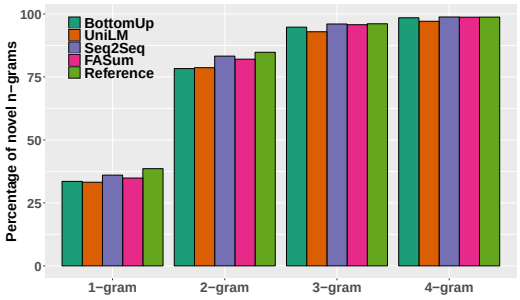

Figure 2: Percentage of novel n-grams for summaries in XSum test set.

4.6.2 Relation Matching Rate

While the factual consistency evaluator FactCC (Kryściński et al., 2019b) is based on pre-trained models, it requires finetuning on articles and labelled summaries. Furthermore, we empirically find that the performance of FactCC degrades when it is finetuned on one summary dataset and used to evaluate models on another dataset.

Therefore, in this subsection, we design an easy-to-compute model-free factual consistency metric, which can be used when ground-truth summaries are not available.

As the relational tuples in the knowledge graph capture the factual information in the text, we compute the precision of extracted tuples in the summary. In detail, suppose the set of the relational tuples in the summary is $R_s = \{(s_i, r_i, o_i)\}$, and the set of the relational tuples in the article is R_a. Then, each tuple in R_s falls into one of the following three categories:

1. **Correct hit** (C): $(s_i, r_i, o_i) \in R_a$;

2. **Wrong hit** (W): $(s_i, r_i, o_i) \notin R_a$, but $\exists o' \neq o_i, (s_i, r_i, o') \in R_a$, or $\exists s' \neq s_i, (s', r_i, o_i) \in R_a$;

3. **Miss** (M): Otherwise.

We define two kinds of relation matching rate

Model	100×Fact Score	ROUGE-1	ROUGE-2	ROUGE-L
CNN/DailyMail				
BOTTOMUP	83.9	41.22	18.68	38.34
Corrected by FC	85.3* (↑1.4%)	40.95	18.37	37.86
UNILM	87.2	**43.33**	**20.21**	**40.51**
Corrected by FC	87.0 (↓0.2%)	42.75	20.07	39.83
SEQ2SEQ	85.0	41.03	18.04	37.93
FASUM	87.8*	40.53	17.84	37.40
Corrected by FC	**88.1** (↑0.3%)	40.38	17.67	37.23
XSum				
BOTTOMUP	78.0	26.91	7.66	20.01
Corrected by FC	78.9* (↑0.9%)	28.21	8.00	20.69
TCONVS2S	79.8	31.89	11.54	25.75
Corrected by FC	82.9* (↑3.1%)	32.44	11.83	26.02
UNILM	83.2	42.14	**19.53**	34.13
Corrected by FC	**83.4** (↑0.2%)	**42.18**	**19.53**	**34.15**
SEQ2SEQ	80.6	31.44	10.91	24.69
FASUM	81.5	30.28	10.03	23.76
Corrected by FC	81.7 (↑0.2%)	30.20	9.97	23.68

Table 3: Factual consistency score and ROUGE scores on CNN/DailyMail and XSum test set. *p-value< 0.05.

Model	$\text{RMR}_1 \uparrow$	$\text{RMR}_2 \uparrow$	NLI ↓
UNILM	60.0	39.6	10.2
FC correction	61.4	40.7	10.0
SEQ2SEQ	53.8	32.2	11.2
FASUM	65.0	46.0	9.3
FC correction	**67.0**	**47.4**	**8.3**

Table 4: Average relation matching rate (RMR, Eq. 6) and NLI contradictory ratio between article and summary in CNN/DailyMail test set. The arrow indicates whether larger or smaller value means better result.

(RMR) to measure the ratio of correct hits:

$$\text{RMR}_1 = 100 \times \frac{C}{C+W} \quad (5)$$

$$\text{RMR}_2 = 100 \times \frac{C}{C+W+M} \quad (6)$$

Note that this metric is different from the ratio of overlapping tuples proposed in Goodrich et al. (2019), where the ratio is computed between the ground-truth and the candidate summary. Since even the ground-truth summary may not cover all the salient information in the article, we choose to compare the knowledge tuples in the candidate summary directly against those in the article. An additional advantage of our metric is that it does not require ground-truth summaries to be available.

Table 4 displays the result of this metric in CNN/DailyMail's testset. As shown, FASUM achieves the highest precision of correct hits under both measures. And there is a considerable boost from the knowledge graph (FASUM vs SEQ2SEQ): 11.2% in RMR_1 and 13.8% in RMR_2. And the correction from the FC model can further improve the metric for both FASUM and UNILM.

We also compute factual consistency via natural language inference models following Maynez et al. (2020). We use the BERT-Large model finetuned on MNLI dataset (Williams et al., 2018) provided by fairseq (Ott et al., 2019). The model predicts the relationship between the article and summary to be one of the following: entailment, neutral and contradiction. We report the ratio of contradiction as predicted by the model in Table 4. As shown, FASUM achieves the lowest ratio and FC helps further reducing conflicting facts in generated summaries.

4.6.3 Human Evaluation

We conduct human evaluation on the factual consistency and informativeness of summaries. We randomly sample 100 articles from the test set of CNN/DailyMail. Then, each article and summary pair is labelled by 3 people from Amazon Mechanical Turk (AMT) to evaluate the factual consistency and informativeness. Each labeller gives a score in each category between 1 and 3 (3 being perfect). The kappa-ratio between reviewer scores is 0.32 for

Model	Factual Score	Informativeness
BOTTOMUP	2.32	2.23
UNILM	2.65	**2.45**
SEQ2SEQ	2.59	2.30
FASUM	**2.74***	2.42

Table 5: Human evaluation results of summaries for 100 randomly sampled articles in CNN/DailyMail test set. *p-value< 0.05.

BOTTOMUP is better	FC is better	Same
15.0%	42.3%	42.7%
UNILM is better	FC is better	Same
20.4%	31.2%	48.4%

Table 6: Human evaluation results for side-by-side comparison of factual consistency on 100 randomly sampled articles from CNN/DailyMail test set where FC makes modifications.

factual consistency and 0.28 for informativeness.

Here, factual consistency indicates whether the summary's content is faithful with respect to the article; informativeness indicates how well the summary covers the salient information in the article.

As shown in Table 5, our model FASUM achieves the highest factual consistency score, higher than UNILM and considerably outperforming BOTTOMUP. We conduct a statistical test and find that compared with UNILM, our model's score is statistically significant with p-value smaller than 0.05 under paired t-test. In terms of informativeness, our model is comparable with UNILM and outperforms BOTTOMUP. Finally, without the knowledge graph component, the SEQ2SEQ model generates summaries with both less factual consistency and informativeness.

To assess the effectiveness of the correction model FC, we conduct a human evaluation of side-by-side summaries. In CNN/DailyMail, we randomly sample 100 articles where the summaries generated by BOTTOMUP are modified by FC. 3 labelers are asked whether the original or the corrected version is factually more correct. We collect all the feedbacks and compute the ratio of judgements for each case. To reduce bias, we randomly shuffle the two versions of summaries. We conduct similar evaluation on UNILM.

As shown in Table 6, the corrected summaries are significantly more likely to be judged as more factually correct for both baseline models. For example, 42.3% of the judgements think the corrected summaries are factually more correct, 42.7% think the corrected version neither improves nor worsens the factual consistency, while only 15.0% think that the corrected version becomes worse than the original BOTTOMUP summary. Therefore, FC can help boost the factual consistency of summaries from given systems.

Finally, to evaluate the quality of the relation matching rate (RMR), we compute the correlation coefficient γ between the factual score given by human labelers and the RMR value. The result shows that $\gamma = 0.43$, indicating observable relationship between RMR and human evaluation results.

5 Conclusion

In this paper, we extract factual information from the article to be represented by a knowledge graph. We then integrate this factual knowledge into the process of producing summaries. The resulting model FASUM enhances the ability to preserve facts during summarization, demonstrated by both automatic and human evaluation. We also present a correction model, FC, to rectify factual errors in candidate summaries. Furthermore, we propose an easy-to-compute model-free metric, relation matching rate, to measure factual consistency based on the overlapping ratio of relational tuples.

For future work, we plan to integrate knowledge graphs into pre-training for more accurate and factually consistent summarization. Moreover, we will combine the internally extracted knowledge graph with an external knowledge graph (e.g. ConceptNet) to enhance the commonsense capability of summarization models.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *ACL 2015*, pages 344–354.

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. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. 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 preprint arXiv:2005.03754*.

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.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

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*, pages 166–175.

Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. Mind the facts: Knowledge-boosted coherent abstractive text summarization. In *Proceedings of The Workshop on Knowledge Representation and Reasoning Meets Machine Learning in NIPS 2019*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, pages 1693–1701.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Friedrich Wilhelm Levi. 1942. Finite geometrical systems.

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*.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

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

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*.

Kazuki Matsumaru, Takase Sho, and Okazaki Naoaki. 2020. Improving truthfulness of headline generation. *arXiv preprint arXiv:2005.00882*.

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

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. *arXiv preprint arXiv:1808.08745*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2019. Make lead bias in your favor: A simple and effective method for news summarization. *arXiv preprint arXiv:1912.11602*.

Model	Incorrect
Random	50.0%
BERT (Falke et al., 2019)	35.9%
ESIM (Falke et al., 2019)	32.4%
FactCC (Kryściński et al., 2019b)	30.0%
FactCC (our version)	**26.8%**

Table 7: Percentage of incorrectly ordered sentence pairs using different consistency prediction models in CNN/DailyMail, using data from Falke et al. (2019).

A Implementation details

For hyperparameter search, we tried 4 layers with 4 heads, 6 layers with 6 heads and 10 layers with 10 heads.

There're 108.3M parameters in the FASUM model and it takes 2 hours (CNN/DailyMail) / 0.5 hours (XSum) for 4 v100 GPUs to train 1 epoch. The batch size is set to 48 for both datasets.

On validation datasets, FASUM achieves ROUGE-1 41.08%, ROUGE-2 18.35% and ROUGE-L 37.95% on CNN/DailyMail, and it achieves ROUGE-1 30.28%, ROUGE-2 10.09% and ROUGE-L 23.85% on XSum.

B Factual Consistency Evaluator

To automatically evaluate the factual consistency of a summary, we leverage the FactCC model (Kryściński et al., 2019b), which maps the consistency evaluation as a binary classification problem, namely finding a function $f : (A, C) \longrightarrow [0, 1]$, where A is an article and C is a summary sentence defined as a claim. $f(A, C)$ represents the probability that C is factually correct with respect to the article A. If a summary S is composed of multiple sentences $C_1, ..., C_k$, we define the factual score of S as: $f(A, S) = \frac{1}{k} \sum_{i=1}^{k} f(A, C_i)$.

To generate training data, we adopt backtranslation as a paraphrasing tool. The ground-truth summary is translated into an intermediate language, including French, German, Chinese, Spanish and Russian, and then translated back to English. Together with the original summaries, these claims are used as positive training examples. We then apply entity swap, negation and pronoun swap to generate negative examples (Kryściński et al., 2019b).

Following Kryściński et al. (2019b), we finetune the $\text{BERT}_{\text{BASE}}$ model using the same hyperparameters to finetune FactCC. We concatenate the article and the generated claim together with special tokens [CLS] and [SEP]. The final embedding of [CLS] is used to compute the probability that the claim is entailed by the article content.

As shown in Table 7, on CNN/Daily Mail, our reproduced model achieves better accuracy than that in Kryściński et al. (2019b) on the human-labelled sentence-pair-ordering data (Falke et al., 2019). Thus, we use this evaluator for all the factual consistency assessment tasks in the following.[3]

C Examples

Table 8, 9 and 10 show examples of CNN/DailyMail articles and summaries generated by our model and several baseline systems. The factual errors in the summary are marked in red, the correct facts in the summaries of FASUM are marked in green and the corresponding facts are marked in bold in the article.

As shown, while baseline systems like BOTTOMUP and UNILM achieve high ROUGE scores, they are susceptible to factual errors. For instance, in Article 5, both BOTTOMUP and SEQ2SEQ wrongly state that Rickie Fowler accused Alexis. In fact, Alexis, Rickie's girlfriend, was accused by an online hater. In Article 1, UNILM mistakenly summarizes that Arsenal lost 4-1 where in fact Arsenal won 4-1 against Liverpool.

In comparison, our proposed fact-aware summarizer FASUM could faithfully summarize the salient information in the article. And it can reorganize the phrasing instead of merely copying content from the article.

Table 11 and Table 12 show examples of CNN/DailyMail articles, summaries generated by BOTTOMUP/UNILM and the corrected version by FC. As shown, our correction model can select the wrong entities and replace them with correct ones. For instance, in Article 1, BOTTOMUP's summary states that Rual Castro, who appears elsewhere in the article, is the President of Venezuela, while FC correctly replaces it with Nocolas Maduro. In Article 4, UNILM wrongly attributes the statement to Scott's lawyer (probably because "Scott" appears closer to the statement in the article), while it was actually said by Slager's lawyer. This error is corrected by FC.

[3]We use the same setting and train another evaluator for XSum dataset.

Article 1	Arsenal target Raheem Sterling is **a more talented player** than Gunners star Mesut Ozil, according to Michael Owen. **The former Liverpool forward made the claim** before Brendan Rodgers' men lost 4-1 at the Emirates on Saturday ... Ozil bends home Arsenal's second goal in their **4-1 win against Liverpool** at the Emirates...
Reference	Arsenal are interested in signing Liverpool's Raheem Sterling this summer. The 20-year-old has rejected a new PS100,000-a-week deal at Anfield. Michael Owen believes **Sterling is more talented than Mesut Ozil**. Adrian Durham: Sterling would be earning the same as Balotelli if he signed PS100,000-a-week deal at Liverpool... Durham: Arsenal only turn it on when the pressure is off. The German scored in **Arsenal's 4-1 win** against Liverpool on Saturday...
BOTTOMUP	Raheem Sterling made the claim before Brendan Rodgers' men lost 4-1 at the Emirates on Saturday. Bt Sport pundit Owen believes there should not be a debate about whether Sterling is more talented than Ozil. **The former Liverpool forward** is a better player than Mesut Ozil.
UNILM	Michael Owen believes Raheem Sterling is a better player than Mesut Ozil. Owen made the claim before **Arsenal lost 4-1** at the Emirates on Saturday. Owen's opinion drew criticism on social media but the former Liverpool forward has refused to back down. Owen said :' Is there a debate? Not even close' CLICK HERE for all the latest Arsenal news.
FASUM	Michael Owen says there **should not be a debate** about whether Raheem Sterling is more talented than Mesut Ozil. The former Liverpool forward made the claim before **Brendan Rodgers' men lost 4-1** at the Emirates on Saturday. Owen repeated his claim that the wantaway forward is a superior player than Ozil made.
SEQ2SEQ	Michael Owen made the claim before Liverpool's men lost 4-1 at the Emirates on Saturday. **Michael Owen believes there should be a debate** about whether Sterling is more talented than Ozil. The ex-England striker has refused to back down, even after Sterling's poor display in north London.
Article 2	**Yaya Toure was accused by Jamie Carragher** of ducking out of the way of Jason Puncheon's free-kick which killed off Manchester City's hopes of salvaging something from Selhurst Park. Toure, who is **paid PS200,000-a-week**, was **part of a five-man wall** when **Puncheon stepped up** to double Crystal Palace's lead...
Reference	Crystal Palace beat Manchester City 2-1 at Selhurst Park on Monday night. Jason Puncheon's free-kick **doubled Palace's lead** in the second half. **Yaya Toure** was accused of ducking out of the way of the winning goal. Sportsmail's **Jamie Carragher:** 'The rest of them are desperate to be hit with the ball in the face and it's poor from Yaya Toure' Gary Neville: City have a mentality problem... they cannot sustain success.
BOTTOMUP	Yaya Toure was accused by Jamie Carragher out of the way of Jason Puncheon's free-kick at Selhurst Park on Monday. Toure, who is paid 200,000-a-week, was part of a **wall with Crystal Palace's lead.** Carragher said the midfielder did not do enough to prevent the winner.
UNILM	Yaya Toure was accused by Sportsmail's Jamie Carragher of ducking out of the way of Jason Puncheon's free-kick at Selhurst Park on Monday. The midfielder was part of a five-man wall when Puncheon stepped up to double Crystal Palace's lead. Toure is paid 200,000-a-week a **by the Premier League champions**.
FASUM	**Yaya Toure was accused by Jamie Carragher** of ducking out of the way of Jason Puncheon's free-kick. Fernandinho and Vincent Kompany look keen to get their head to the ball but Toure appears to duck. **Toure paid PS200,000-a-week**, was **part of a five-man wall** when Puncheon stepped up to **double Crystal Palace's lead.**
SEQ2SEQ	**Jamie Carragher was accused** of ducking out of the way of Jason Puncheon's free-kick which killed off Manchester City's chances at Selhurst Park on Monday Night. Toure is paid PS200,000-a-week, was part of a five-man wall when Puncheon stepped up to double Palace's lead. Sportsmail columnist Jamie Carragher feels City midfielder did not do enough to prevent winner.

Article 3	Real Madrid ace Gareth Bale treated himself to a Sunday evening BBQ after ... a sensational victory. The Welsh wizard was ... **scoring twice** and assisting another in an impressive victory in Israel... he **opened the scoring in Real Madrid's impressive 9-1 victory**... His goal in the **25th minute** may have got the ball rolling ... **Cristiano Ronaldo scored five goals** against Granada on Sunday ...
Reference	Welsh Wizard Gareth Bale opened the scoring in Sunday's stunning win. Real Madrid put nine goals past Granada to keep pressure on Barcelona. Bale treated himself to a BBQ in the Spanish sun following victory. Cristiano Ronaldo scored five goals in the sensational team performance.
BOTTOMUP	Gareth Bale scored twice in Real Madrid's 9-1 victory against Granada. **The Real Madrid ace** scored five goals against Granada on Sunday. The Welsh wizard was in impressive form for the international break.
UNILM	Gareth Bale scored twice as Real Madrid beat Granada 9-1. The Welsh winger scored for Wales away on international duty. Bale Tweeted:'Unbelievable result this afternoon, great team performance. Now for a BBQ' Real Madrid bounced back from El Clasico defeat by Barcelona with a thumping win. Cristiano Ronaldo scored five goals against an eight-minute hat-trick.
FASUM	Real Madrid beat Granada 9-1 at the Bernabeu on Sunday evening. **Gareth Bale scored twice** and helped his side to a sensational 9-1 win. **Cristiano Ronaldo scored five goals** against Granada on Sunday. The Welsh winger Tweeted:'Unbelievable result this afternoon, great team performance now for a BBQ'
SEQ2SEQ	Gareth Bale scored twice as Real Madrid beat Granada 9-1 at the Bernabeu on Sunday. The Welsh ace opened the scoring with a goal against Granada in the 25th minute. **Gareth Bale scored five** and assisted another in an impressive win in Israel. Bale Tweeted:' Unbelievable result this afternoon, great team performance. Now for a BBQ'
Article 4	***Former* Huddersfield centre Josh Griffin scored a try and kicked three goals** as improving **Salford secured back-to-back Super League wins** with an 18-12 triumph at the John Smith's Stadium. **Griffin and Ben Jones-Bishop scored tries** in the first half, although Jack Hughes gave the Giants hope. **Carl Forster's try and Griffin's penalty** made the game safe... **Salford Red Devils** secured a second win?...
Reference	Salford sealed consecutive Super League wins with victory at Huddersfield. Josh Griffin inspired Red Devils with a try and three kicked goals. Ben Jones-Bishop and Carl Forster scored the other tries for the visitors.
BOTTOMUP	Josh Griffin and Ben Jones-Bishop scored tries in the first half. Salford Red Devils win a second win in a row with a 18-12 victory at Huddersfield. **Carl Forster scored a try and kicked three goals** for Salford.
UNILM	Former Huddersfield centre Josh Griffin scored a try and kicked three goals. Ben Jones-Bishop scored tries in the first half. Carl Forster's try and Griffin's penalty made the game safe.
FASUM	**Former Huddersfield centre Josh Griffin scored a try and kicked three goals** as Salford secured back-to-back Super League wins with an 18-12 win at John Smith's Stadium. Ben Jones-Bishop scored tries in the first half but Jack Hughes gave the Giants hope. **Carl Forster's try and Griffin's penalty** made the game safe, though.
SEQ2SEQ	**Huddersfield centre** Josh Griffin scored a try and kicked three goals. Griffin scored tries in the first half for the Giants. The game was twice delayed early on when **Kevin Locke for Huddersfield's** Brett Ferres were carried off injured.

Table 9: Continuation of Table 8.

Article 5	**Rickie Fowler has earned extra brownie points** with his growing legion of fans after calling out an online hater who accused his girlfriend of being a 'gold digger.' The golfer's girlfriend, bikini model **Alexis Randock, had posted a photo** on her Instagram account of her and sister Nicole on the beach last week... an **abusive commenter known as 'fatalsplash'** who was quick to judge and accused Alexis of being a 'gold digger.' ... Fowler told the hater to **'get your facts straight'** ...
Reference	Rickie Fowler responded to online troll who abused his girlfriend. Bikini model Alexis Randock had posted a photo on her Instagram account. Randock was pictured on the beach alongside her sister Nicole. A troll called 'fatalsplash' then accused Alexis of being a 'gold digger' Fowler told the hater to 'get your facts straight' following comment.
BOTTOMUP	**Rickie Fowler accused Alexis** of being a 'gold digger' who did n't have to work. **The golfer was criticised** by an online troll for not working due to her relationship with Fowler. Fowler has earned extra brownie points with his legion of fans.
UNILM	Rickie Fowler 's girlfriend, bikini model Alexis Randock, posted a photo on her Instagram account of her and sister Nicole on the beach last week. An abusive commenter called 'fatalsplash' accused Alexis of being a 'gold digger'
FASUM	Rickie Fowler has earned extra brownie points with his growing legion of fans. The golfer's girlfriend, bikini model **Alexis Randock, had posted a photo** on her Instagram account. The photo was shared with her 25,000 followers, which included an abusive commenter known as 'fatalsplash' who was quick to judge and accused Alexis of being a 'gold digger'
SEQ2SEQ	Golfer's girlfriend posted a picture of her and sister Nicole Randock on Instagram. The picture shared with her 25,000 followers was**shared by her sister** on the beach last week. The **golfer was quick to judge and accused Alexis** of being a **'gold digger' troll**. He responded by **posting the caption** 'fatalsplash' accusing her of being 'a digger who didn't work for it'
Article 6	The VII Summit of the Americas was supposed to be all about the symbolic handshake between the United States and Cuba... Heads of state from 35 countries in the Western Hemisphere have met every three years...**Obama and Castro would come face-to-face**... **Venezuelan President Nicolas Maduro** recently accused the United States of trying to topple his government...
Reference	U.S., Venezuelan relations threaten to overshadow Obama, Castro meeting. Venezuelan President says United States moved to oust him; he has the support of the Cuban foreign minister.
BOTTOMUP	Heads of state from 35 countries have met every three years to discuss economic, social or political issues since the summit in 1994. **Venezuela's President Raul Castro** has been criticized for human rights violations. The **U.S. government says** the summit of the Americas is a "triangle of tension."
UNILM	Summit of the Americas was supposed to be all about the symbolic handshake between the U.S. And Cuba. **The tide changed when** President Obama and Castro announced that more than five decades of Cold War rivalry was ending. Venezuelan President Nicolas Maduro recently accused the U.S. Of trying to topple his government...
FASUM	The summit of the Americas was supposed to be all about the symbolic handshake between the u.s. and Cuba. Diplomats from both countries began negotiations **to establish embassies in Havana and Washington**. President Nicolas Maduro recently accused the United States of trying to topple his government.
SEQ2SEQ	The VII summit of the Americas was supposed to be all about the symbolic handshake between the United States and Cuba. Relations between the u.s. and Cuba have been the wrench in diplomatic machinery, with some Latin American leaders threatening not to attend. President Barack Obama and Castro **face-to-face embassies** in Havana and Washington.

Table 10: Continuation of Table 8.

Article 1	The VII Summit of the Americas was supposed to be all about the symbolic handshake between the United States and Cuba. But insert Venezuela into the mix and Panama City, Panama, quickly turns into a "triangle of tension."... Cuba has historically been the wrench in the diplomatic machinery, with some Latin American leaders threatening not to attend the Summit of the Americas if the United States and Canada didn't agree to invite President Raul Castro... The much anticipated handshake between Obama and Castro would steal all the headlines if it wasn't for Cuba's strongest ally, Venezuela. Venezuelan President **Nicolas Maduro** recently accused the United States of trying to topple his government and banned former President George Bush...
BOTTOMUP	Heads of state from 35 countries have met every three years to discuss economic, social or political issues since the summit in 1994. Venezuela's President **Raul Castro** has been criticized for human rights violations. The u.s. government says the summit of the Americas is a "triangle of tension."
Corrected by FC	Heads of state from 35 countries have met every three years to discuss economic, social or political issues since the summit in 1994. Venezuela's President **Nicolas Maduro** has been criticized for human rights violations. The u.s. government says the summit of the Americas is a "triangle of tension."
Article 2	She's one of the hottest and most successful Latinas in Hollywood, but now Sofia Vergara is playing defense in a legal battle initiated by her ex-fiance: He wants to keep the two frozen embryos from their relationship, both female. The 42-year-old actress and star of the hit TV sitcom "Modern Family" split from businessman **Nick Loeb** in May 2014. Loeb is suing the Colombian-born actress in Los Angeles to prevent Vergara from destroying their two embryos conceived through in vitro fertilization in November 2013, according to published reports by New York Daily News and In Touch magazine...
BOTTOMUP	**Sofia Vergara** wants to keep the frozen embryos from their relationship, both female. He is suing the actress in Los Angeles to prevent Vergara from their embryos. The actress and star of the "Modern Family" split from Nick Loeb in May 2014.
Corrected by FC	**Nick Loeb** wants to keep the frozen embryos from their relationship, both female. He is suing the actress in Los Angeles to prevent Vergara from their embryos. The actress and star of the"Modern Family"split from Businessman Nick Loeb in May 2014.
Article 3	Volvo says it will begin exporting vehicles made in a factory in southwest China to the United States next month, the first time Chinese-built passenger cars will roll into American showrooms. Parent company **Geely Automobile**, which bought Volvo in 2010, is seeking to prove that a Chinese company can manage a global auto brand...
BOTTOMUP	**Volvo** is seeking to prove that a Chinese company can manage a global auto brand. The car will be one of four models produced in a manufacturing plant in Chengdu. China is largest market for car sales globally in 2009.
Corrected by FC	**Geely Automobile** is seeking to prove that a Chinese company can manage a global auto brand. The car will be one of four models produced in a manufacturing plant in Chengdu. China is largest market for car sales globally in 2009.

Table 11: Example articles, summaries and corrections made by FC in CNN/DailMail. The factual errors in the original summary are marked in red, the corrected part by FC are marked in green and the corresponding facts are marked in bold in the article.

Article 4	... Scott family attorney Chris Stewart said the man with Scott was a co-worker and friend... Slager has been fired and faces life in prison or the death penalty if convicted on a murder charge. Video shot by a witness shows the officer shooting Scott in the back as he runs away. **Slager** had told investigators he feared for his safety after he and Scott tussled over his Taser. **His lawyer**, Andy Savage, said Friday he "has not received the cooperation from law enforcement that the media has." ...
UNILM	"This is a heartbreaking tragedy for everyone in our community," Charleston mayor says. Scott family attorney says the man with Scott was a co-worker and friend. Michael Slager has been fired and faces life in prison or the death penalty if convicted on a murder charge. **Scott**'s lawyer says he "has not received the cooperation from law enforcement that the media has
Corrected by FC	"This is a heartbreaking tragedy for everyone in our community," Charleston mayor says. Scott family attorney says the man with Scott was a co-worker and friend. Michael Slager has been fired and faces life in prison or the death penalty if convicted on a murder charge. **Slager**'s lawyer says he "has not received the cooperation from law enforcement that the media has
Article 5	**Rangers boss Stuart McCall** has revealed he was warned about taking over at Ibrox by Walter Smith - but insists he has made the right decision. The former Light Blues midfielder steered his new side to their third victory on the bounce with Sunday's 2-1 win over Hearts. It was a crucial triumph that moves **Gers** back into second place in the Scottish Championship and looks to have restored vital momentum ahead of the end-of-season play-offs. **Stuart McCall says** he was warned not to take the Rangers job but is pleased that he took the risk in the end.
UNILM	Stuart McCall guided Rangers to a 2-1 win over Hearts on Sunday. Kenny Miller and Haris Vuckic scored the goals for the Gers at Ibrox. The win moved **Rangers** back into second place in the Scottish Championship. The **Gers** boss says he was warned not to take over by Walter Smith. But he says he is pleased that he took the risk in the end.
Corrected by FC	Stuart McCall guided Rangers to a 2-1 win over Hearts on Sunday. Kenny Miller and Haris Vuckic scored the goals for the Gers at Ibrox. The win moved **Gers** back into second place in the Scottish Championship. The **Rangers** boss says he was warned not to take over by Walter Smith. But he says he is pleased that he took the risk in the end.
Article 6	**Wasps** will appeal against the three-week ban issued to **Nathan Hughes**, but their protest will not enable the back row to play in Sunday's Champions Cup quarter-final against Toulon. The club have said in a statement they will appeal on the grounds the incident that left **Northampton wing George North** unconscious and resulted in a red card for Hughes was an accident and no foul play had occurred... **Wasps** are set to appeal the three-match ban handed to No 8 **Nathan Hughes** for knocking out **George North**... **Hughes** was given the suspension - to include three meaningful fixtures - after being **found guilty** of striking North in the 52-30 Aviva Premiership defeat at Franklin's Gardens on March 27...
UNILM	Wasps will appeal the three-match ban handed to No 8 Nathan Hughes. The **Northampton** winger was found guilty of striking George North. The incident occurred during **Northampton**'s 52-30 defeat at Franklin's Gardens. North suffered a third confirmed concussion in just four months.
Corrected by FC	Wasps will appeal the three-match ban handed to No 8 Nathan Hughes. The **Wasps** winger was found guilty of striking George North. The incident occurred during **Wasps**'s 52-30 defeat at Franklin's Gardens. North suffered a third confirmed concussion in just four months.

Table 12: Continuation of Table 11