Unsupervised Fact Error Correction Modeling by Using Span-Level Contrastive Learning

Yuqing Lan, Zhenghao Liu✉, Yu Gu✉, Xinze Li, and Ge Yu

Northeastern University, Shenyang, Liaoning 110819, China
liuzhenghao@mail.neu.edu.cn, guyu@mail.neu.edu.cn

Abstract. Fact Error Correction (FEC) aims to identify and rectify factual errors in claims, thereby enhancing the accuracy and reliability of information. Existing work primarily focuses on unsupervised FEC methods to improve model faithfulness and ensure the factual consistency of outputs to the provided evidence. These methods identify informative spans within a given claim, mask these spans to generate questions, and then employ a Question Answering (QA) model to answer these questions, ultimately generating corrected claims. However, current unsupervised FEC models overlook the valuable hints provided by these informative spans, which contain semantic and type constraints crucial for error correction. To address this issue, we propose the method for training a Fact Error Correction model, named CorrectFEC, to correct factual errors more effectively. Specifically, CorrectFEC generates claim-evidence pairs using QA datasets, employs a vanilla T5 model to predict masked spans, and subsequently applies contrastive training to Pretrained Language Models (PLMs) to correct factual errors by leveraging the span prediction results. As a result, CorrectFEC can correct errors in claims by fully incorporating the semantic hints from these informative spans during inference. Our experimental results on the FEVER and SCIFACT datasets demonstrate that the CorrectFEC model outperforms existing unsupervised FEC models, achieving a 5% improvement in the SARI score. The code is available at https://github.com/NEUIR/CorrectFEC.

Keywords: Fact Error Correction · Contrastive Learning · Pretrained Language Model.

1 Introduction

Fake news [13], rumors [9], and unverified machine-generated texts [39, 38] harm individuals and communities, influence public policy and decision-making. The automated fact-checking [19, 40, 2, 23, 5] has been widely studied, significantly improving verification efficiency and reducing human workload in identifying fake information [28, 20, 37, 6]. To introduce interpretability into fact verification, researchers have developed the Factual Error Correction (FEC) task [27, 3, 10, 11, 1], which aims to correct factual errors rather than merely identifying them.

Existing work [12] builds an unsupervised FEC system to conduct span-level FEC and focuses more on the faithfulness of the FEC system instead of fluency

and grammatical that can be better guaranteed by the generative models. Faithfulness refers to the accuracy with which the FEC model's corrections reflect intended information updates. Such an FEC system usually employs a three-step pipeline. It first extracts informative spans that may contain factual errors, then synthesizes questions for querying these spans, and employs the QA models to answer the span-based questions. Finally, these question-answer pairs are used to generate and rerank candidate claims. Nevertheless, this approach may overlook important correction constraints associated with the identified spans when generating candidate claims solely based on question-answer pairs.

This paper introduces an error fact correction model CorrectFEC. To enhance the FEC ability of PLMs, we utilize the QA data to synthesize claim-evidence pairs [22] and then identify the spans that may contain factual errors [12]. CorrectFEC introduces a span-level contrastive learning method to train PLMs to correct these extracted spans. Specifically, we conduct a span-prediction base FEC model (PredFEC), which masks the informative spans and employs the vanilla T5 [25] model to predict these masked spans. These predicted spans of PredFEC are regarded as hard negatives to contrastively train the FEC ability of the CorrectFEC model. Such a training method helps CorrectFEC learn to correct the error spans and incorporate correction hints from these spans.

Our experimental results on both FEVER [32] and SCIAFCT [33] datasets demonstrate the effectiveness of CorrectFEC, which achieves more than a 5% SARI score than our main baseline ZeroFEC model [12]. It can broaden its effectiveness in the unsupervised FEC pipeline by conducting more accurate error correction operations. CorrectFEC shows stronger generalization ability than the LM-based FEC model and can learn from the predicted results of the language based model to further improve the FEC performance. Our further analyses demonstrate that MFEC reduces the uncertainty in predicting the golden span and enhances the FEC performance by calibrating the factual error correction of PLMs, particularly for synonym and negation errors.

2 Related Work

Earlier work mainly focuses on the fact verification (FV) task, which aims to assess the veracity of the claims. To build the FV model, some researchers adopt Natural Language Inference (NLI) models [8, 4] for predicting claim labels. Thriving on the advantages of PLMs, some works also verify claims without any evidence [15] or encode claim-evidence pairs using PLMs for fact verification [32, 33, 16, 30]. Additionally, graph-based modeling [19, 40, 29] is further used to establish the relationships to conduct fine-gained reasoning. However, these fact verification models cannot correct these fact errors in claims, stimulating the development of the Fact Error Correction (FEC) task.

The first FEC dataset [31] builds upon the FEVER dataset [32], which facilitates numerous studies adopting the supervised training method. Some researchers explore distant supervision approaches that leverage rich resources from the FV tasks. The sentence modification model [27] proposes a two-stage

framework where a fact-checking neutrality classifier guides the masker model to mask the conflict spans in claims. A corrector model then generates correct claims by filling in these masked spans to align with the evidence. VENCE [3] proposes an iterative correction framework that edits claim spans and employs the FV model to validate modifications, gradually converging to factually consistent outputs. Furthermore, some advances in the FEC task focus on data synthesis approaches. CompEdit [7] trains a perturber module to synthesize error claims by introducing conflicting entities into raw claims and then trains the FEC model to delete these entities for correct claim generation. LIFE [11] firstly uses a mask model to identify the span of the claim and then trains a corrupt model to synthesize error claims for FEC training. PivotFEC [10] leverages ChatGPT's capabilities to generate claims containing factual errors, synthesizing diverse training data for FEC models. Different from these supervised FEC methods, ZeroFEC [12] explores the unsupervised FEC method, which uses a QA model to answer the span-based question and then re-ranks the generated claim candidates to select the correct claim. Our method also adopts an unsupervised approach that utilizes predictions from a base FEC model as hard negative examples to enhance correction accuracy.

3 Methodology

This section introduces a FEC model, CorrectFEC, that is trained to correct fact errors. We first describe the preliminary of the unsupervised FEC pipeline (Sec. 3.1). Then we describe how we use the span-level contrastive learning to enhance the unsupervised FEC model (Sec. 3.2).

3.1 Preliminary of Unsupervised FEC

As shown in Figure 1, with the given claim c and the evidence piece e, the unsupervised FEC models aim to correct the factual errors in the claim. To tailor the FEC models to correct factual errors in an unsupervised manner, we follow existing work [12] to build the pipeline to correct factual errors, including span identification, factual error correction, and correction scoring.

Span Identification. Firstly, we extract information spans that may contain errors in the claim c. We use Spacy to identify the noun chunks and named entities and then utilize Stanza to extract the nouns, verbs, adjectives, adverbs, noun phrases, verb phrases, and negation terms.

Factual Error Correction. Then we mask the extracted span s^i in the claim $c(s^i \rightarrow [\text{MASK}])$ and ask the language model to fill in the masked span according to the evidence e:

$$P(s^i_{\text{FEC}} \mid c, e, s^i, \theta_{\text{PLM}}) = \text{PLM}(c(s^i \rightarrow [\text{MASK}]), e), \tag{1}$$

where θ_{PLM} indicates the parameter of the pre-trained language model. We use the generated span s^i_{FEC} to replace the original span s^i. Then the claim candidates can be generated $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_n\}$, where $\tilde{c}_i = c(s^i \rightarrow s^i_{\text{FEC}})$.

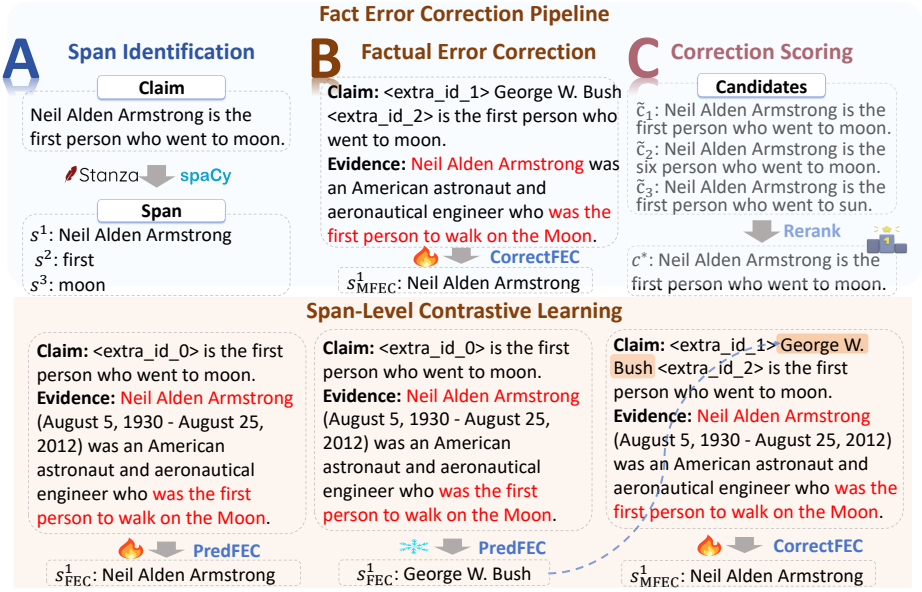


Fig. 1: The Illustration of the CorrectFEC Based Unsupervised FEC Pipeline.

Correction Scoring. Finally, we follow previous work [12, 34] to calculate the correction score $SCORE_{\tilde{c}_i}$ of each candidate correction $\tilde{c}_i$ and select the claim c^* that has the highest score:

$$SCORE_{\tilde{c}_i} = (\text{DocNLI}(\tilde{c}_i, e) + \text{ROUGE-1}(\tilde{c}_i, c)), \tag{2}$$

$$c^* = \arg\max_{\tilde{c}_i \in \tilde{C}} (SCORE_{\tilde{c}_i}), \tag{3}$$

where the DocNLI score is calculated using the Roberta-large [18] based DocNLI model [36], which provides the document-level NLI score between each candidate claim $\tilde{c}_i$ and the evidence e. The ROUGE-1 aims to control the candidate claim $\tilde{c}_i$ is more similar to the raw claim c.

3.2 Enhancing Unsupervised Fact Error Correction by Using Span-Level Contrastive Learning

This subsection describes the Factual Error Correction method. We first introduce the method of training the FEC module using unsupervised data and then describe the method of span-level contrastive learning.

Training PredFEC Using Unsupervised Data. To tailor language models for the FEC task, we collect unsupervised training data from the question-answering (QA) dataset, which is also used in previous work of fact verification [22]. Specifically, we covert the query q and answer a to the claim c by using the claim generation model [12]. Then the claim can be verified by the

query-related passage, which is regarded as the evidence e in the FEC task. Finally, we can get the unsupervised training data $D = \{(c_1, e_1), \ldots, (c_K, e_K)\}$, which consists of the pairs of synthesized claims and corresponding evidence. It is evident that all claims are supported by the given evidence.

To train the unsupervised FEC module, we mask the span identified by the spacy and stanza, and then ask the language models to fill in the masked content. Given the masked claim $c(s^i \rightarrow$ <extra_id_0>) and the evidence e, the PredFEC module aims to use the pretrained language model, such as T5 [25], to generate the corrected span s^i_{PredFEC}:

$$P(s^i_{\text{PredFEC}} \mid c, e, s_i, \theta_{\text{PredFEC}}) = \text{T5}(c(s^i \rightarrow \text{<extra\_id\_0>}), e). \tag{4}$$

Correct Fact Errors via Span-Level Contrastive Learning. The LM-based FEC model (Eq. 4) may overlook the knowledge in raw claims, causing unnecessary span corrections. Thus, the PredFEC model is asked to correct the information span s_i of the raw claim during the inference. It can be regarded as a hint of the purpose of claim writing. To tailor PLMs to a span correction manner, we use the language model (Eq. 4) to synthesize some negative spans during training, which teaches the CorrectFEC model to better capture the semantics from vanilla spans to better correct factual errors.

To learn the parameters of CorrectFEC model $\theta_{\text{CorrectFEC}}$, we use the unsupervised data $D = \{(c_1, c(s^i \rightarrow s_{\text{PredFEC}}), e_1), \ldots\}$ to train it. In which, the span generated by the PredFEC model is replaced by the original span. Thus, the span $s^i_{\text{CorrectFEC}}$ can consider the parameters of the PredFEC model θ_{PredFEC}. According to the Bayes formula, we can approximately calculate the probability $P(s^i_{\text{CorrectFEC}} \mid c, e, s_i, \theta_{\text{CorrectFEC}}, \theta_{\text{PredFEC}})$ using the following equation:

$$P(s^i_{\text{CorrectFEC}} \mid s^i_{\text{PredFEC}}, c, e, s_i, \theta_{\text{CorrectFEC}}) \cdot P(s^i_{\text{PredFEC}} \mid c, e, s_i, \theta_{\text{PredFEC}}). \tag{5}$$

Then we can respectively calculate the LM base span prediction probability $P(s^i_{\text{PredFEC}} \mid c, e, s_i, \theta_{\text{PredFEC}})$ and the LM enhanced span correction probability $P(s^i_{\text{CorrectFEC}} \mid s^i_{\text{PredFEC}}, c, e, s_i, \theta_{\text{CorrectFEC}})$. Specifically, the span prediction probability can be calculated using Eq. 4. And then span correction probability can be calculated by incorporating the generation result s_{PredFEC}:

$$\begin{aligned} &P(s^i_{\text{CorrectFEC}} \mid s^i_{\text{PredFEC}}, c, e, s^i, \theta_{\text{CorrectFEC}}) \\ &= \text{T5}(c(s^i \rightarrow \text{<extra\_id\_1>} s^i_{\text{PredFEC}} \text{<extra\_id\_2>}), e). \end{aligned} \tag{6}$$

4 Experimental Methodology

This section describes the datasets, evaluation metrics, baselines, and implementation details of our experiments.

Datasets. We follow ZeroFEC [12] and use FEVER [31] and SCIFACT [33] datasets for evaluation. The claims that do not have corresponding evidence are removed. The unsupervised dataset that we used is built based on the NQ dataset [14, 24]. We use the QA2claim model[1] to convert the QA pair.

[1] https://huggingface.co/khhuang/zerofec-qa2claim-t5-base

Table 1: Overall Performance. The best results are highlighted in bold.

	FEVER				SCIFACT			
	SARI	Rouge-2	GPT	NLI	SARI	Rouge-2	GPT	NLI
Fine-tune	48.61	65.63	2.85	78.48	40.42	74.62	2.75	82.74
CompEdit [7]	59.64	62.88	2.93	74.56	71.22	85.19	2.75	88.41
PivotFEC [10]	61.40	71.49	**4.23**	87.16	54.54	73.76	**3.41**	87.70
ZeroFEC [12]	69.12	70.19	3.84	86.31	67.15	78.79	3.25	89.28
CorrectFEC	**74.80**	**75.84**	3.85	**87.37**	**75.94**	**86.51**	3.20	**92.94**

Evaluation Metrics. The main evaluation metric is SARI. We calculate the SARI with the huggingface evaluate library. ROUGE-2 and Sentence-BERT (NLI) [26] measure the bi-grams overlap and the natural language inference score between the generated claim and the ground truth, respectively. The GPT score measures the factual consistency between the evidence and the generated claim. It follows gpteval [17] and uses GPT-4o-mini for evaluation.

Baselines. CompEdit [7] trains a perturber to generate longer claims with the entities not in the evidence. Then, CompEdit is trained to compress and remove irrelevant entities. PivotFEC [10] uses ChatGPT to inject errors into claims, generating claim pairs containing factual errors to train the FEC model. ZeroFEC [12] asks QA models to answer the claim-based question and generates the corrected claim candidates by span-level FEC. We also implement the Fine-tune model, which trains the T5-base model [25] using the unsupervised training dataset of CorrectFEC. It feeds the claim-evidence pairs to T5 and asks it to directly generate the correct claims.

Implementation Details. During training, we set the max number of epochs to 5, batch size to 8, and the gradient accumulation step to 4. The AdamW optimizer is used for optimization. The learning rate is 5e-5 and weight decay is 1e-8. During inference, we set the beam search width to 4 and the maximum length to 20.

5 Experimental Results

We conduct experiments to evaluate CorrectFEC's effectiveness in this section.

5.1 Overall Performance

The FEC performance of different models is shown in TABLE 1.

Compared to ZeroFEC [12], CorrectFEC achieves a significant improvement in the SARI score by training to directly generate the correct span. Such a method can avoid additional transforming processes between questions and claims of ZeroFEC, which is more efficient. Besides, the CorrectFEC model shows its effectiveness by achieving much better GPT scores. The improvement demonstrates that CorrectFEC can effectively capture the semantics from evidence to

Table 2: Ablation Study.

	FEVER				SCIFACT			
	SARI	Rouge-2	GPT	NLI	SARI	Rouge-2	GPT	NLI
Zero-Shot	70.72	63.23	2.89	81.53	71.90	80.99	2.93	88.99
PredFEC	73.47	74.97	**3.90**	86.81	69.48	84.10	3.11	90.69
CorrectFEC	**74.80**	**75.84**	3.85	**87.37**	**75.94**	**86.51**	**3.20**	**92.94**

correct the span by training using the synthesized unsupervised data. Furthermore, we use the unsupervised data to train the T5-base model [25], conducting the Fine-tune model. The Fine-tune model shows a much lower performance than ZeroFEC. It shows that the factual reasoning relationship behind QA pairs is less effective in helping PLMs correct factual errors. Moreover, it achieves a higher 8.79% SARI improvement on the SCIFACT dataset, which shows its generalization ability to broaden the effectiveness to the science-specific domain.

5.2 Ablation Study

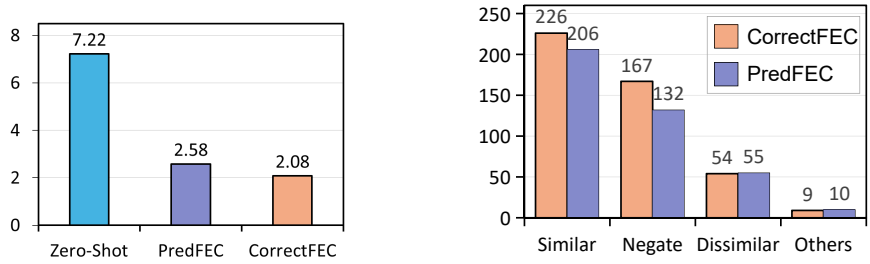
As shown in TABLE 2, we further explore the effectiveness of the ablation modules in our CorrectFEC model.

The Zero-Shot model inherits the vanilla checkpoint of T5-base [25] and the span of the claim is replaced with <extra_id_0>. Both the CorrectFEC and PredFEC models are trained on unsupervised claim-evidence data generated from QA pairs. The PredFEC model uses the mask language modeling method for correction, while CorrectFEC highlights the span to incorporate the knowledge of raw claims to help correct factual errors. In the result, the PredFEC model outperforms the Zero-Shot model, which demonstrates that continuous pretraining language models to learn the factual alignment between the synthesized claim and evidence is effective to guild the PredFEC model to find more clues from evidence to fill in the masked span in the claim. After contrastive learning in span generated by the PredFEC model, the CorrectFEC model can incorporate the knowledge of the predFEC model and be regarded as a kind of hint to help it identify and correct factual errors.

5.3 Effectiveness of CorrectFEC in Different Testing Scenarios

As shown in Fig. 2a, the Zero-Shot model exhibits significantly higher logarithmic complexity compared to the PredFEC and CorrectFEC models. This highlights that our unsupervised training dataset effectively enhances the ability of PLMs to adapt to FEC tasks, enabling the model to have higher confidence in producing the golden span. Furthermore, CorrectFEC conducts lower log perplexity scores than the PredFEC model. This indicates that CorrectFEC is highly confident in producing the golden span with the help of the span generated by the PredFEC model [35, 21].

We then quantify the generated claims across various mutation types, which serve as templates for creating evidence-refuted claims by replacing spans in the



(a) Log Perplexity Scores of Different Models in Producing Golden Spans.

(b) The Number of Correct Claims on Different Mutation Types.

Fig. 2: The Effectiveness of CorrectFEC in Different Testing Scenarios.

original sentences. These mutations effectively represent different categories of factual errors that can be introduced into the claims. "Similar" and "Dissimilar" mutations substitute verbs and/or objects with alternatives from the same and different categories respectively, while "Negate" mutations invert the claim's meaning. The "Others" category contains three modification types: "Rephrase" which preserves the original meaning, "More Specific" and "More General" which makes the claim meaning more specific and more general respectively. CorrectFEC shows much better performance than the PredFEC model, particularly in the "Similar" and "Negate" categories. It demonstrates that the PredFEC model generates some confusing span candidates for teaching the CorrectFEC model, which are the crucial semantic hints to help CorrectFEC review the errors and generate more accurate correction results by reflecting the negated relationships and synonyms between the given claim and the corrected one.

6 CONCLUSIONS

This paper proposes the CorrectFEC model, which tailors the PLMs in the unsupervised FEC pipeline. The PredFEC trained based on our synthesized QA-generated unsupervised dataset is asked to generate some confusing spans. These spans can include rich and diverse error-correction instances and help teach CorrectFEC to learn more precise correction rules to generate more accurate correction results. By highlighting instead of masking spans, the model can identify the correlation between erroneous content and contextual information, leading to corrections that maintain higher coherence with the original text. Experimental results demonstrate that CorrectFEC not only successfully incorporates knowledge from PredFEC to improve correction accuracy but also broadens its effectiveness to some vertical domains.

Acknowledgments

This work is partly supported by the Natural Science Foundation of China under Grant (No. 62206042), the Joint Funds of Natural Science Foundation of Liaoning Province (No. 2023-MSBA-081), and the Fundamental Research Funds for the Central Universities under Grant (No. N2216017).

References

1. Adams, G., Shing, H.C., Sun, Q., Winestock, C., McKeown, K., Elhadad, N.: Learning to revise references for faithful summarization. ArXiv (2022)
2. Chen, J., Bao, Q., Sun, C., Zhang, X., Chen, J., Zhou, H., Xiao, Y., Li, L.: Loren: Logic-regularized reasoning for interpretable fact verification. In: AAAI (2022)
3. Chen, J., Xu, R., Zeng, W., Sun, C., Li, L., Xiao, Y.: Converge to the truth: Factual error correction via iterative constrained editing. In: AAAI (2023)
4. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: ACL (2017)
5. Deng, Z., Schlichtkrull, M., Vlachos, A.: Document-level claim extraction and decontextualisation for fact-checking. In: ACL (2024)
6. Eldifrawi, I., Wang, S., Trabelsi, A.: Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In: ACL (2024)
7. Fabbri, A.R., Choubey, P.K., Vig, J., Wu, C.S., Xiong, C.: Improving factual consistency in summarization with compression-based post-editing. ArXiv (2022)
8. Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., Gurevych, I.: UKP-athene: Multi-sentence textual entailment for claim verification. In: FEVER Workshop (2018)
9. Haouari, F., Elsayed, T., Suwaileh, R.: Aured: Enabling arabic rumor verification using evidence from authorities over twitter. In: ACL (2024)
10. He, X., Jin, A.L., Ma, J., Yuan, Y., Yiu, S.: Pivotfec: Enhancing few-shot factual error correction with a pivot task approach using large language models. In: EMNLP Findings (2023)
11. He, X., Zhang, Q., Jin, A.L., Ma, J., Yuan, Y., Yiu, S.M.: Improving factual error correction by learning to inject factual errors. In: AAAI (2024)
12. Huang, K.H., Chan, H.P., Ji, H.: Zero-shot faithful factual error correction. In: ACL (2023)
13. Huang, K.H., McKeown, K., Nakov, P., Choi, Y., Ji, H.: Faking fake news for real fake news detection: Propaganda-loaded training data generation. In: ACL (2023)
14. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019)
15. Lee, N., Bang, Y., Madotto, A., Fung, P.: Towards few-shot fact-checking via perplexity. In: NAACL-HLT (2021)
16. Li, X., Burns, G.A., Peng, N.: A paragraph-level multi-task learning model for scientific fact-verification. In: AAAI (2021)
17. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: EMNLP. pp. 2511–2522 (2023)

18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv (2019)
19. Liu, Z., Xiong, C., Sun, M., Liu, Z.: Fine-grained fact verification with kernel graph attention network. In: ACL (2020)
20. Ma, H., Xu, W., Wei, Y., Chen, L., Wang, L., Liu, Q., Wu, S.: Ex-fever: A dataset for multi-hop explainable fact verification. In: ACL (2024)
21. Ott, M., Auli, M., Grangier, D., Ranzato, M.: Analyzing uncertainty in neural machine translation. In: ICML (2018)
22. Pan, L., Chen, W., Xiong, W., Kan, M.Y., Wang, W.Y.: Zero-shot fact verification by claim generation. In: ACL (2021)
23. Park, E., Lee, J., Jeon, D.H., Kim, S., Kang, I., Na, S.: SISER: semantic-infused selective graph reasoning for fact verification. In: COLING (2022)
24. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: KILT: a benchmark for knowledge intensive language tasks. In: NAACL-HIT (2021)
25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**, 1–67 (2020)
26. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: EMNLP (2019)
27. Shah, D., Schuster, T., Barzilay, R.: Automatic fact-guided sentence modification. In: AAAI (2020)
28. Si, J., Zhao, Y., Zhu, Y., Zhu, H., Lu, W., Zhou, D.: CHECKWHY: Causal fact verification via argument structure. In: ACL (2024)
29. Soleimani, A., Monz, C., Worring, M.: BERT for evidence retrieval and claim verification. In: ECIR (2020)
30. Subramanian, S., Lee, K.: Hierarchical evidence set modeling for automated fact extraction and verification. In: EMNLP (2020)
31. Thorne, J., Vlachos, A.: Evidence-based factual error correction. In: ACL (2021)
32. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: NAACL-HLT (2018)
33. Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. In: EMNLP (2020)
34. Wan, D., Bansal, M.: Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. ArXiv (2022)
35. Wang, S., Liu, Y., Wang, C., Luan, H., Sun, M.: Improving back-translation with uncertainty-based confidence estimation. In: EMNLP (2019)
36. Yin, W., Radev, D., Xiong, C.: DocNLI: A large-scale dataset for document-level natural language inference. In: ACL Findings (2021)
37. Yue, Z., Zeng, H., Shang, L., Liu, Y., Zhang, Y., Wang, D.: Retrieval augmented fact verification by synthesizing contrastive arguments. In: ACL (2024)
38. Zhang, S., Yu, T., Feng, Y.: TruthX: Alleviating hallucinations by editing large language models in truthful space. In: ACL (2024)
39. Zhong, Z., Zhou, K., Mottin, D.: Harnessing large language models as post-hoc correctors. In: ACL Findings. pp. 14559–14574 (2024)
40. Zhou, J., Han, X., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: GEAR: Graph-based evidence aggregating and reasoning for fact verification. In: ACL (2019)