**Paper 1:**

**FACTKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge**

**src:**[**https://arxiv.org/pdf/2305.08281v2.pdf**](https://arxiv.org/pdf/2305.08281v2.pdf)

This paper discusses a program called FACTKB, a tool that evaluates if automatically created summaries are factually correct. At the moment, the tools we have to do this task still make mistakes, especially when it comes to identifying entities (specific items, objects, or people) and relations (ways these entities relate to each other) in new areas or topics.

To improve on this, the researchers who wrote this paper created FACTKB, which works across various topics and is particularly good at identifying entities and relations. It's based on language models (computational structures that predict how human language works) that have been trained using facts from external sources of knowledge.

In order to train these models, the researchers used three approaches based on (1) facts specific to entities, (2) facts extracted from supplementary information related to entities, and (3) facts created using a method called knowledge base walks.

The researchers concluded that FACTKB topped other factuality evaluation models in terms of performance, showing its effectiveness in news summarization and scientific literature. They found it could better detect wrong facts relating to entities and their relationships in summaries and could be easily applied to various topics.

The paper highlights the importance of making sure document summaries are not only informative but also factually accurate. However, it's often tricky to evaluate factual accuracy, especially in specialized fields like science or law. Current methods have limitations because they rely on neural classifiers trained with limited data. To address this, FACTKB is introduced as a new evaluation model that focuses on improving accuracy with entities and relationships. It achieves this by using language models enhanced with factual knowledge from external sources. The paper details three training strategies to strengthen FACTKB's ability to understand entities, incorporate supporting evidence, and improve overall reasoning. The evaluation shows that FACTKB performs significantly better than existing methods, particularly in spotting errors related to entities and relationships. It's also adaptable across different fields, making it a valuable tool for checking summary accuracy.

**FACTKB methodology**

The FACTKB methodology is designed to enhance the understanding of 'facts' or 'truths' in language models (LMs). Here's a simple explanation of its processes:

* **Factuality Pretraining:** This process works with knowledge bases (KBs), which are rich sources of facts. The goal is to use KBs as 'fact teachers' to improve the language model's understanding of entities and relations.

It utilizes three strategies:

**Strategy 1 - Entity Wiki:** This strategy works with the practice of predicting missing connections in KBs based on available facts. The language model is trained to anticipate masked entities or relations in the facts provided by the KBs. This trains LMs to infer facts from surrounding information and punishes unsupported claims about entities and relations.

**Strategy 2 - Evidence Extraction:** This strategy is about enhancing the model’s ability to judge facts based on adequate evidence. Here, a triple (ei, rk, ej) is selected randomly from the KB. The first paragraph of the Wikipedia description of ei is used as auxiliary knowledge. Then, a sentence is formed using these two, creating a corpus of triples mixed with auxiliary knowledge. This corpus aids FACTKB in selecting evidence from documents to support its factuality evaluation.

**Strategy 3 - Knowledge Walk:** This strategy is meant to improve the understanding of multi-hop claims. The idea is to have a randomly picked entity and selected further from its direct neighborhood, forming one-hop triple {e(0), r(0,1), e(1)}. This process is repeated several times, creating a corpus used for factuality pretraining.

* **LM Training:** After the factuality pretraining, the language models are then trained using the proposed methods, refining the fact-enhanced LM on a factuality error detection dataset.

Through these methods, FACTKB is expected to improve in understanding and evaluating the factuality of entities and relations in a document.

**FACTKB training**

The FACTKB training process begins with initializing FACTKB with encoder-based Language Models (LMs). Each of the three factuality pretraining strategies is then individually applied to FACTKB in a process involving the masked language modeling approach. The goal is to evaluate the effectiveness of each strategy. This results in LMs that are enhanced with the ability to better represent facts, entities, and relationships.

Following that, FACTKB undergoes a fine-tuning process using human-annotated factual error detection datasets. This is done in a sequence classification setting, where "Summary" and "Document" serve as the input. The model produces either a "Factual" or "Non-Factual" label.

The [CLS] token is used during the classification process. This process results in FACTKB, which is an entailment-based factuality evaluation model capable of classifying machine-generated summaries as either factual or non-factual.

**Data and Experiment**

The experimentation begins with training the model using a dataset from YAGO, a knowledge base built from Wikidata for factuality pretraining. The refining of the model, however, is done using the FactCollect dataset which collects human annotations concerning factual errors from various sources such as CNN, Daily Mail, and BBC, and consolidates them into a single dataset. For each pair of summary and article, a FACTUAL or NON-FACTUAL label is assigned.

The model is configured with some specific resource limits and parameters: a corpus size of 100,000, a masking probability of 15%, and a knowledge walk length of 5. The model is then pretrained for 5 epochs and fine-tuned for a maximum of 50 epochs. Three types of factuality pretraining are conducted.

Then comes the evaluation stage which consists of in-domain and out-of-domain evaluations. The in-domain evaluation is focused on the news media. The FactCollect dataset and the FRANK benchmark serve as the basis for this evaluation. The FRANK benchmark provides human judgments on the factual consistency of model-generated summaries, categorizing various factual errors.

However, as summary systems are used across various domains (news, social media, scientific literature, etc.), the model is also evaluated for its ability to provide reliable factuality scores across these varied domains. For this, three datasets from the scientific literature domain are used: CovidFact, HealthVer, and SciFact. CovidFact collects claims from a subreddit and verifies them against relevant scientific literature and Google search results. HealthVer consists of claims from TREC-COVID and verified against the CORD-19 corpus. SciFact includes claims from biomedical literature and verifies them against the abstract of the cited paper.

Finally, the performance of FACTKB is compared with existing factuality evaluation models like QAGS, QUALS, DAE, FalseSum, SummaC, FactCC, and FactGraph to understand its relative effectiveness.

**Result:**

In the results, FACTKB was tested against other models on the FactCollect dataset and it outperformed all others by achieving 3.8 BACC points higher on average across different dataset settings. This suggests FACTKB's capability and the positive effect of using knowledge bases and factuality pretraining for factuality evaluation.

The model was also tested on the FRANK benchmark to see how it correlates with human judgment. FACTKB achieved high correlation, making it a practical approach for evaluating the factual consistency of generated news summaries.

When FACTKB was evaluated on out-of-domain datasets from scientific literature, it outperformed existing factuality metrics by 4.1 BACC points on average across the datasets. This suggests that the factuality pretraining strategies applied in FACTKB can effectively adapt to different domains, making the model more robust and therefore suitable in various applications.

In contrast, the existing models only performed slightly better than random factuality scores on these scientific literature datasets, suggesting their limitations in generalizing to other domains. The results shows that FACTKB exceeded performance expectations in both in-domain and out-of-domain settings, validating its effectiveness and robustness in different applications.

**Analysis:**

Upon analysis the study investigates where FACTKB has made improvements, its compatibility with various knowledge bases (KBs) and language models (LMs), and the impact of different parameters.

* Where did FACTKB Improve? FACTKB was found to be very effective in identifying 'semantic frame' errors, which involve entities and relations, a sign that the knowledge base-based factuality pretraining strategies are working. FACTKB also performed well in identifying other types of factual errors, making it a comprehensive tool for factual error detection.
* KB and LM Compatibility FACTKB can work with various combinations of language models and knowledge bases. The researchers tested six different LMs and KBs with FACTKB, and the results showed improved factual error detection in all cases. It was noted that certain models and bases performed better than others, suggesting that specific combinations could be more effective.
* Simplicity Study Compared to other approaches, FACTKB requires less complex setup and fewer resources. FACTKB does not require additional processing steps like processing dependency structures and doesn't require multiple iterations. This makes it a more straightforward and lightweight method for factuality evaluation.
* Parameter Analysis The researchers studied the effects of different parameters on the effectiveness of the model, which included the pretraining corpus size, the number of pretraining epochs, and knowledge walk length. For instance, they found that pretraining with larger corpus sizes or too many epochs could be counterproductive. Similarly, a knowledge walk length of 5 was found to provide a moderate amount of compositionality in the factuality pretraining corpora, showing the best results.

The analysis showed that FACTKB has made significant improvements in factual error detection, can work with various language models and knowledge bases, and is a more straightforward and lightweight method for factuality evaluation. The performance of FACTKB can be influenced by different parameters such as corpus size, number of pretraining epochs, and knowledge walk length.

**Conclusion:**

The study introduces FACTKB, a new and simple way to evaluate the factuality of information using language models. The unique aspect of this approach is that it uses external knowledge bases to improve the representation of entities and relations in these models. By applying three pretraining strategies called 'entity wiki', 'evidence extraction', and 'knowledge walk', they aim to enhance the model's understanding of facts.

Detailed experiments show that FACTKB performs better than existing methods in evaluating factuality within the same domain of news media and across different domains, such as scientific literature. It also aligns well with human judgments on truthfulness and is especially proficient in detecting semantic frame errors, which are types of factual errors related to entities and relations.

Thus, FACTKB serves as a user-friendly and adaptable metric for factuality, which will aid research on producing factually consistent summaries.

**Paper 2**

**Enhancing Semi-Supervised Learning for Extractive Summarization with an LLM-based pseudolabeler**

**src:**[**https://arxiv.org/pdf/2311.09559.pdf**](https://arxiv.org/pdf/2311.09559.pdf)

The paper addresses the challenge of creating brief summaries of text, particularly when there is limited labeled data available. The researchers propose a technique where they use a language model called GPT-4 to select the best pseudolabels or artificially created labels for summarization. They evaluate this technique on three types of datasets: TweetSumm, WikiHow, and ArXiv/PubMed. The results show that the use of GPT-4 improves summary quality significantly across these datasets, proving that the method is highly effective. They also found that this method requires fewer unlabeled examples to work effectively.

As digital content is growing exponentially, tools that can automatically summarize text are becoming crucial. For instance, companies with customer service chats could use these tools to improve the user experience. However, the challenge is that in real-life situations, there's often not enough labeled data to train a fully supervised summarization model. This mandates the exploration of semi-supervised techniques that can work with limited labeled data.

This study specifically focuses on extractive summarization, where the summary is built by copying parts of the source document. While pretraining language models is important for this, the study suggests that careful selection of pseudolabels during the training phase is crucial.

The paper introduces a new prompt-based strategy that uses GPT-4 to select the best pseudolabels. The method is tested across three types of datasets and results show that the strategy effectively improves the performance of a standard teacher-student learning algorithm. The study is the first of its kind to demonstrate that having a good selection and pseudolabeling strategy can lead to improvements in summarization, on par with developing a better pretrained model.

**Methodology**

Here, how a self-supervised summarization model is created by applying a standard teacher-student training process and using GPT-4 for pseudolabel generation is explained. Firstly, an initial 'teacher' model is trained using limited labeled data. This teacher then generated pseudolabels (artificial labels) for the unlabeled dataset. Top 50 pseudolabels are selected based on the teacher’s confidence level. To quantify the teacher model's confidence in a given pseudolabel, the average predicted probability of each sentence is calculated in the pseudolabel being part of the summary.

Next, using GPT-4, a numerical rating to each of these top 50 pseudolabels are provided based on certain criteria. These criteria included the summary being concise, covering key points from the original text, and being an extractive summary meaning it should directly use sentences from the input conversation without any modifications.

Finally, the top 5 pseudolabels with the highest GPT-4 score are selected, then used GPT-4 to generate new pseudolabels for unlabeled examples which were then added to the training data for the next cycle. The same steps were then repeated in each cycle to continually refine the model.

**Experimental Setup**

In the experimental setup, three datasets were used:

* **TweetSumm**, which contains customer service chat data
* **Wikihow**, which contains articles with corresponding summaries
* **ArXiv/PubMed**, a collection of scientific articles with their abstracts as the summaries.

Initially models were trained on a small subset of labelled examples from each dataset, and the rest of the data was treated as unlabeled. The trained model were to generate pseudolabels (artificial labels) for the unlabeled examples. 50 pseudolabels were selected based on the model's confidence in its labelling.

Then, GPT-4 was used to give a score to each pseudolabel by considering several criteria such as the conciseness of the summary, coverage of key points from the text, and whether it is an extractive summary (directly using sentences from the conversation without modification). The top 5 pseudolabels with the highest GPT-4 scores were added to the training set for the next cycle.

This process was repeated over multiple training cycles until they had a final set of 300 examples (50 labeled and 250 pseudolabels). The models' performance was evaluated using ROUGE scores, which consider the overlap of n-grams (continuous sequence of n items) between the predicted and actual summaries.

Finally, the researchers compared their approach with different baselines including a teacher-student learning framework, self-supervised pre-training, the original PreSumm model, and a random model selection approach.

**Results:**

Different models and strategies were compared and it was found that when more labels (or data) are used for training, models produce better summaries. However, this improvement was not so noticeable in WikiHow and ArXiv/PubMed datasets because the model's labels were created using a matching technique that might not be the best way to create labels.

Among the semi-supervised models (which learn from both labelled and unlabeled data), all selection strategies they tested did better than a model trained with random pseudolabels (artificial labels). They observed that using GPT-4 to rate pseudolabels improved the performance for all datasets, suggesting that the summarization model can generate some high-quality labels and that GPT-4 is beneficial for determining those high-quality labels.

It was concluded that the performance of their model, PreSumm, aligns poorly with the actual distribution of the data, possibly due to having only a small labelled dataset. Hence, depending on GPT-4 to rate generated pseudolabels can be a beneficial strategy, especially in situations where there is little labelled data available.