The Effects of Hallucinations in Synthetic Training Data for Relation Extraction

Steven **Rogulsky**[1], Nicholas **Popovic**[1] and Michael **Färber**[2]

[1]*Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
[2]*TU Dresden & ScaDS.AI, Dresden, Germany*

Abstract

Relation extraction is crucial for constructing knowledge graphs, with large high-quality datasets serving as the foundation for training, fine-tuning, and evaluating models. Generative data augmentation (GDA) is a common approach to expand such datasets. However, this approach often introduces hallucinations, such as spurious facts, whose impact on relation extraction remains underexplored. In this paper, we examine the effects of hallucinations on the performance of relation extraction on the document and sentence levels. Our empirical study reveals that hallucinations considerably compromise the ability of models to extract relations from text, with recall reductions between 19.1% and 39.2%. We identify that relevant hallucinations impair the model's performance, while irrelevant hallucinations have a minimal impact. Additionally, we develop methods for the detection of hallucinations to improve data quality and model performance. Our approaches successfully classify texts as either 'hallucinated' or 'clean,' achieving high F1-scores of 83.8% and 92.2%. These methods not only assist in removing hallucinations but also help in estimating their prevalence within datasets, which is crucial for selecting high-quality data. Overall, our work confirms the profound impact of relevant hallucinations on the effectiveness of relation extraction models.

1. Introduction

Relation extraction is an important step in extracting structured information from text documents, such as news articles, publications, patents, and websites, building the basis for knowledge graph construction. High-quality datasets play a crucial role in this process [1, 2, 3], as they form the basis for training, fine-tuning, and evaluating relation extraction models. Additionally, the amount of data they contain has a significant impact on the achieved results [4]. However, creating large datasets with high quality typically requires human annotation, which is expensive and slow. Although heuristic methods such as distant supervision can produce larger datasets, they often lack quality [5]. An alternative is *Generative Data Augmentation (GDA)*, a technique for synthetically expanding datasets by generating new data samples (here: texts and extracted triples). It can generate datasets that are much larger, more diverse, and less expensive than traditional human annotations without directly collecting new data [6]. In the context of relation extraction, GDA has been widely used in combination with pre-trained language models such as BERT and GPT [7, 8, 9, 10, 11].
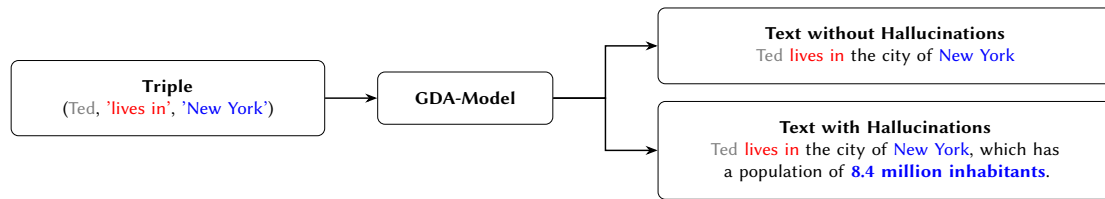
Figure 1: Text generated by Generative Data Augmentation (GDA) with and without hallucinations.

Despite its advantages, GDA often leads to *hallucinations* in the text, where the content deviates from the information in the input, as, for instance, additional facts are generated (see Figure 1). This issue commonly occurs in generative language models [12, 13, 14]. If a language model is trained on a dataset with incorrect annotations due to hallucinations, the effectiveness of the relation extraction method may be compromised – to which degree is unknown –, potentially reducing the accuracy of extracted triples as the model might not learn to capture all necessary information. Although the phenomenon of hallucinations is well-recognized, and the use of LLMs to generate training datasets is increasing, the specific effects of hallucinations on relation extraction have not been thoroughly investigated.

In this paper, we examine the impact of hallucinations in synthetic training data on relation extraction, considering several hallucination types on the document and sentence level. Our research focuses on two primary questions: **RQ1: Can we detect a significant influence of hallucinations on the relation extraction model's performance?** We address this question by evaluating the performance of models trained on datasets with varying levels of hallucinations, aiming to understand the presence and impact of different hallucination types. **RQ2: Can hallucinations be reliably detected?** To address this question, we develop and evaluate approaches for hallucination detection.

Our findings reveal substantial declines in dataset quality and model performance due to hallucinations, with recall decreases ranging from 19.1% to 39.2%. This indicates that hallucinations notably compromise the ability of models to extract relations from texts. In this context, it is crucial to differentiate between relevant and irrelevant hallucinations. The former significantly affects performance, while the latter has a minimal impact. Furthermore, we develop two methods for identifying and eliminating hallucinations, achieving F1-scores of 83.8% and 92.2%. These methods not only remove hallucinations but also assist in estimating their prevalence.

Overall, our contributions in this paper are as follows:[1]

- **Analyzing the Impact of Hallucinations on Model Performance:** We determine the effect of hallucinations on relation extraction models by training them on datasets with different levels of hallucinations and analyzing the performance discrepancies observed.
- **Classifying Hallucinations:** We categorize hallucinations into relevant and irrelevant types, and examine their impacts on datasets.
- **Detecting Hallucinations:** We evaluate language model-based methods for automatically detecting hallucinations.

---

[1]Our source code is available at https://github.com/BigPanda042/Relation-Extraction-Hallucination-Study.

2. Related Work

In this section, we first look at related work on creating synthetic training data. In the second part, we look at noisy data and hallucinations and how to recognize them.

Generating Synthetic Data. Several data augmentation approaches have been proposed. Feng et al. [15] differentiate between three main types: (1) **Rule-based approaches** use algorithms to modify existing real-world datasets. Techniques such as synonym replacement, random insertion, swapping and deletion are used to significantly increase the volume of training data [16, 3, 17]. (2) **Sample interpolation**, also known as **Mixed Sample Data Augmentation**, [18] interpolates data points to create more diverse and robust datasets for training language models [19, 20, 21, 22, 23]. Both approaches are limited by the fact that they are based on existing datasets. As a result, they are not able to introduce completely new features or vary the data types significantly, such as the relation types for relation extraction tasks. This can lead to the persistence of existing biases in the original datasets [15]. (3) **Model-based approaches**, referred to as **Generative Data Augmentation (GDA)**, overcome these limitations. They are able to generate completely new and specific data points, independent of existing datasets. For example, the Control Prefixes model [24] is characterized by the generation of text data from structured knowledge graphs using the WebNLG dataset [25, 26]. Other notable implementations include the use of pretrained language models (PLMs), such as GPT-3.5, which have been successfully used to improve performance on relation extraction tasks [27, 2, 28, 6].

Josifoski et al. [6] developed a large synthetic dataset named Wiki-cIE for closed information extraction, utilizing GPT-3.5 with prompt engineering. This dataset, containing 1.8 million data points, serves as a robust alternative to both distantly supervised and directly supervised datasets in terms of size and quality. It is positioned closely in scale to the largest distantly supervised dataset, REBEL [5]. Importantly, the Wiki-cIE dataset offers enhanced quality, especially in the distribution of relation types and the accuracy of text annotations. Josifoski et al. demonstrate that relation extraction models trained on Wiki-cIE significantly outperform those trained on REBEL, attributing this advantage to the superior quality of their synthetic dataset. However, they do not specify which particular attributes of the datasets contribute to these performance differences. A notable quality difference is in the accuracy of the text annotations, suggesting that this aspect may be a critical factor in the observed improvements in model performance.

Detecting and Assessing Noisy Data. Corrupted or noisy data, characterized by issues such as incorrect labels, affects language model training [29, 30, 31, 32, 33, 34]. Several strategies have been developed to address noisy data in datasets. Techniques include resampling [35], loss reweighting [29], and label correction [36]. Additionally, some approaches advocate training models using noise-robust loss functions [30, 31], with a notable recent development being a noise-robust re-weighting framework [2]. While these methods effectively mitigate the impact of noisy data or reduce its presence, they do not specifically explore the influence of hallucinations within synthetic training data on the performance of relation extraction models.

Analyzing Hallucinations. Ji et al. [13] provide an overview on hallucinations, including relevant data-to-text use cases. The authors distinguish between two types of hallucinations: intrinsic and extrinsic. **Intrinsic hallucinations** are false information in texts that contradict the annotations, while **extrinsic hallucinations** consist of additional information in the texts

that is not supported by the annotations. There exist several approaches to detect both types of errors. Typical textual similarity metrics such as BLEU or ROUGE are unsuitable for the detection of hallucinations [37, 38]. Other approaches can be divided into statistical and model-based methods. Statistical approaches [39, 40, 41] focus primarily on lexical information, i.e., the specific words used, and therefore cannot adequately take syntactic or semantic variations into account. Thus, the more relevant alternatives are model-based approaches. Liu et al. [42] use named entity recognition to extract the entities from a text and compare them with those in the annotated table. The number of hallucinations is then based on the difference between annotated and found entities. Dušek and Kasner [43] have developed an approach that uses a natural language based inference method. It compares the input data and the output text in both directions and can thus detect omissions or hallucinations. The last methods to be mentioned are the language model-based approaches by Filipova [32] and Tian et al. [44]. However, these methods provide results that either focus on table-to-text generation or are not precise enough for our needs. While the presented methods contribute to the task of detecting hallucinations, none of them examines the exact influence of hallucinations on training performance or attempts to differentiate between different types of hallucinations.

3. Evaluation

The concept of **hallucinations** lacks a universally accepted definition [13, 32, 45, 12, 40]. Figure 1 provides an example of a hallucination. In this scenario, a GDA model, tasked with generating text from the input triple ('Ted', 'lives in', 'New York'), should ideally produce 'Ted lives in the city of New York.' Instead, the model might extend this to 'Ted lives in the city of New York, which has a population of 8.4 million inhabitants.' This addition introduces an unsupported triple ('New York', 'has', '8.4 million inhabitants'), which is a hallucination.

Formally, we define hallucinations H as the set difference $H = TC \setminus T$ between the set of triples T and the triples TC that are actually generated in the text S.

We differentiate between **relevant and irrelevant hallucinations** [24, 46] in relation extraction models, as illustrated in Figure 2. Relevant hallucinations occur when the text expresses triples with relation types that are relevant (i.e., included in the schema) but absent from the annotations. For example, if a model is trained exclusively to detect birth dates in texts, only triples related to birth dates are considered relevant. Conversely, irrelevant hallucinations involve relations that the model is designed to ignore, as they do not pertain to its trained focus.

In the following, we first analyze the influence of hallucinations on synthetic training datasets for relation extraction. We then consider the automatic detection of hallucinations in relation extraction datasets.

3.1. Evaluating the Effects of Hallucinations

3.1.1. Influence of Relevant Hallucinations on Document Level

In this subsection, we focus on evaluating the impact of *relevant* hallucinations on *document-level* relation extraction.
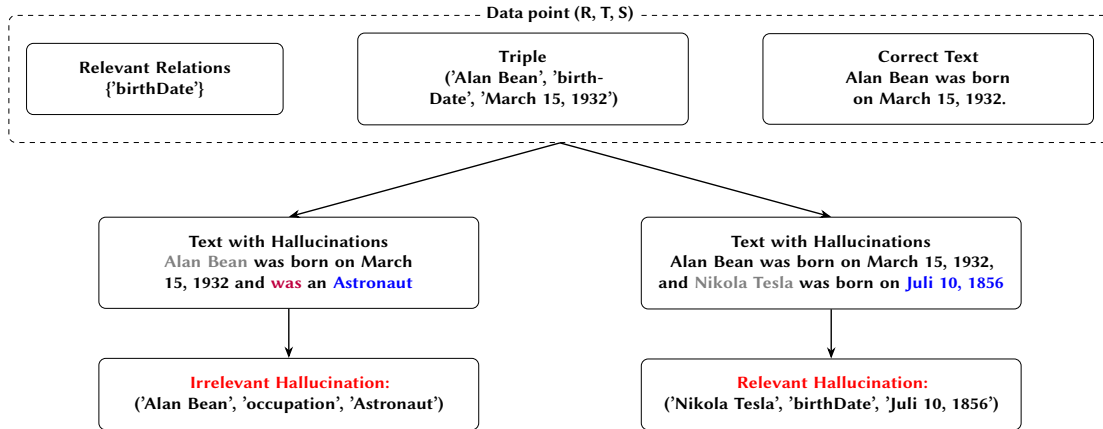
Figure 2: Examples of hallucinations, showing the distinction between relevant hallucinations (e.g., relation type 'birthDate') and irrelevant hallucinations (e.g., relation type 'occupation').
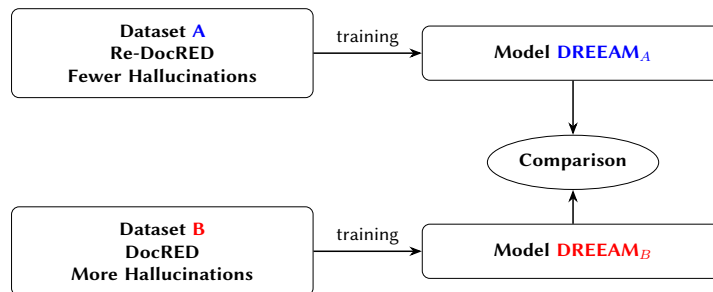

Figure 3: Illustration of the approach used for testing the influence of hallucinations at document-level with the datasets A and B and the language model approach DREEAM.

Table 1

Key characteristics of Re-DocRED and DocRED.

	Re-DocRED			**DocRED**	
	Train	Dev	Test	Train	Dev
Documents	3,053	500	500	3,053	1,000
Avg. Triples	28.1	34.6	34.9	12.5	12.3
Avg. Sentences	7.9	8.2	7.9	7.9	8.1

Datasets. As presented in Figure 3, we select two datasets:

1. **Dataset A** is characterized by fewer hallucinations. Specifically, we employ the Re-DocRED dataset [33], as is known for its extensive use and minimal irrelevant content.
2. **Dataset B** contains a significant presence of relevant hallucinations. We use the DocRED dataset [47], an earlier version of the Re-DocRED dataset known for its incomplete annotations and the consequent prevalence of relevant hallucinations.

Table 2
Precision, recall, and F1-score of the DREEAM model trained on A and B and evaluated on A^{test} or B^{test}. Each value is the average of five configurations.

Model	Test Dataset	Precision	Recall	F1
DREEAM_A	A^{test}	89.87	68.79	77.93
DREEAM_B	A^{test}	94.86	29.59	45.11
DREEAM_B	B^{test}	67.51	59.79	63.42

The differences between dataset A, Re-DocRED, and B, DocRED, are outlined in Table 1.

Relation Extraction Model. We select the DREEAM model [48], which has achieved top performance on the DocRED and Re-DocRED datasets [49, 50]. This model is optimized for compatibility with both datasets, thereby obviating the need for further modifications.

The model is initially trained on Dataset A and B to produce two tailored versions: DREEAM_A and DREEAM_B. These models are then evaluated on the respective test portions of the datasets and benchmarked against the findings of Ma et al. [48]. Although the standard practice for DocRED involves using a development dataset for parameter tuning and testing, we adopt it as our test dataset. The ultimate comparison of DREEAM_A and DREEAM_B's performance is conducted using the same test dataset A^{test}, which is free of hallucinations [33].

Evaluation Results and Discussion: Table 2 shows the evaluation results, revealing a significant discrepancy between the two model configurations. Notably, the recall differs strongly, which is also reflected in the F1-score. In the case of relation extraction, the recall measures the ratio of correctly extracted triples compared to all relevant triples that should have been extracted. Since DREEAM_B was trained on data where the triples in the annotation do not accurately reflect the text's triples, it learned that not all triples must be extracted to obtain a correct solution. This results in a lower recall, as expected.

The precision, however, surprisingly increases for DREEAM_B when evaluated on A^{test}, compared to the evaluation on B^{test}, to an even higher value than for DREEAM_A. This difference is most likely due to the wrong test dataset. The model most likely extracted true positives, but since the test dataset is incorrect, those correct triples were not present in the test annotation and counted as false positives. Those false positives became true positives through the correct A^{test}, and the precision increased. Nevertheless, this cannot explain why the precision increased further than the precision of DREEAM_A. One potential reason is that DREEAM_B tends to extract fewer triples than DREEAM_A. DREEAM_B was trained on a dataset with generally fewer triples in T but the same texts and thus learned to extract fewer triples. Another possibility presents the relation type distribution. In A, the number of underrepresented relation types may have increased, or new, more difficult ones to extract correctly may have been introduced.

3.1.2. Influence of Relevant Hallucinations on Sentence Level

Datasets. We now require datasets on the sentence-level. We use the WebNLG [26] dataset for A, a widely used knowledge graph-to-text dataset [25, 24, 8, 51]. Based on an own analysis and to the best of our knowledge, the dataset is free of relevant hallucinations [52].

Table 3

Precision, recall and F1-score of PFN models trained on datasets with relevant hallucinations B_{MT} and B_{MT} compared to the original dataset A. Each value is the average of five trained configurations evaluated on A^{test}.

Model	Precision	Recall	F1
PFN_A	93.07	92.97	93.02
$PFN_{B_{MT}}$	91.41	73.87	81.70
$PFN_{B_{LT}}$	93.12	72.95	81.40

Table 4

Differences between models trained on the datasets A and B on sentence-level (B_{MT} results used for B) and document-level.

Relation Extraction Task	Dataset	Precision	Recall	F1
Sentence-Level	A	93.07	92.97	93.02
	B	91.41	73.87	81.70
	Δ	1.66	19.10	11.32
Document-Level	A	89.87	68.79	77.93
	B_{MT}	94.86	29.59	45.11
	Δ	5.01	39.20	32.82

The first variant for B, B_{MT}, is created to ensure direct comparability to the document-level datasets used above and to prevent biases by controlling the creation process. To accomplish this, we delete one triple from each of A's data points, given that at least two triples are present. We randomly select which triple to delete to avoid any bias regarding the position of missing information. This ensures that the same texts are kept in both A and B_{MT} but with different annotations. In total, we delete 28.1% of all triples, corresponding to a 39.0% hallucination rate (calculated by dividing the total number of triples in all the texts by the total number of triples in all the annotations).

Additionally, we create B_{LT} to ensure that measured differences cannot solely be attributed to B_{MT} just having fewer triples in the annotation. We add the text (of an unrelated data point) that contains no identical triples in the annotation to each data point of B_{LT}. This way, we can include relevant information in the text without altering the annotations.

Relation Extraction Model. We use the state-of-the-art PFN model [46, 53]. Our initial step involved a preliminary experiment similar to the one described in Section 3.1.1. We adapted PFN to dataset A, resulting in PFN_A, and assessed its performance on the A^{test} dataset. The F1-score differences were minor, averaging less than 0.6% in variation, which we deemed acceptable given the unknown variance in Yan et al.'s results. Subsequently, we fine-tuned PFN on datasets B_{MT} and B_{LT}, producing $PFN_{B_{MT}}$ and PFN_{BLT}, respectively. Both models were then evaluated against the original test dataset.

Evaluation Results and Discussion. Table 3 reveals performance differences between PFN_A and $PFN_{B_{MT}}$. Specifically, recall diminishes by an average of 19.1%, while precision
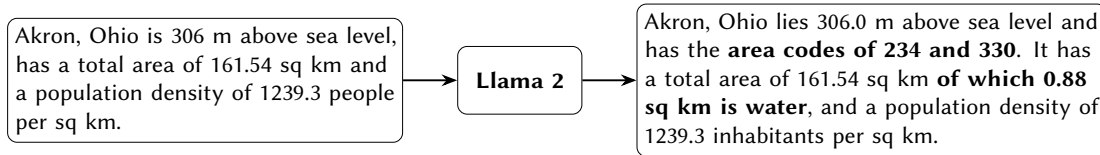
Figure 4: Illustration of how LLAMA2-13b-chat may introduce hallucinations into a text.

declines by 2.29%, a decrease deemed statistically significant through a paired t-test at a 95% confidence level. Regarding $\text{PFN}_{B_{LT}}$, recall is similarly reduced by 19.98%. Conversely, there is a marginal increase in precision of 0.06%. Consequently, the F1-scores for $\text{PFN}_{B_{MT}}$ and $\text{PFN}_{B_{LT}}$ decrease by 11.32% and 11.62%, respectively.

These findings underline that the persistence of triples in longer texts within B_{LT} does not counterbalance the reduced training data volume. As discussed in Section 3.1.1, only the variation in hallucination rates across datasets explains the altered recall rates.

Differences remain substantial in recall between document-level and sentence-level extraction, as summarized in Table 4. Document-level recall decreases nearly twice as much in absolute terms and three times in relative terms compared to sentence-level, primarily due to differing hallucination rates. Table 1 shows that Re-DocRED contains three times more triples per annotation than DocRED, a stronger contrast than observed between A and B_{MT}. Yet, without control over document-level dataset creation, a definitive causality cannot be verified here.

Contrary to expectations, precision varies significantly across the experiments. Notably, a 5% difference at the document-level, as indicated in Table 2, diverges from the sentence-level findings between PFN_A and $\text{PFN}_{B_{MT}}$. This discrepancy suggests potential document-specific effects or dataset variances not previously accounted for. Given the controlled modifications in B_{MT}, these results are considered more reliable, highlighting a distinct decrease in precision between PFN_A and $\text{PFN}_{B_{MT}}$.

3.1.3. Differences Between Relevant and Irrelevant Hallucinations

In this subsection, we evaluate the influence of relevant and irrelevant hallucinations on the sentence level.

Dataset. We keep A the same as it can serve as the dataset with fewer hallucinations. On the other hand, B needs to contain *irrelevant* hallucinations instead of *relevant* ones. We also create another test dataset for testing whether the newly trained models only extract the first part of a text and ignore the rest. To that end, we use the chat version of LLAMA2 [54] to add irrelevant hallucinations to each of A's data points. The LLM takes text as input and returns the same text but with additional information. To create additional dataset variants, we adjust the prompt by adding or removing specific instructions. This allows us to partly control the amount and type of information added. In total, we produce five modified WebNLG datasets, with the only difference being the prompt we use for the creation process.

New Test Dataset. We modify the test dataset to assess if text length affects model performance. Each data point in A^{test} is altered by fusing two data points (i.e., concatenating S

Table 5

Precision, recall, and F1-score of PFN models trained on datasets with irrelevant hallucinations (B) compared to the original dataset (A), both evaluated on the original A^{test} dataset. B consists of the average and standard deviation of five different datasets, with five model configurations trained on each of them.

Model	Precision	Recall	F1
PFN_A	93.07	92.97	93.02
$PFN_{B_{fused}}$	92.55 ±0.26	90.99 ±1.15	91.76 ±0.65

Table 6

Precision, recall, and F1-score of PFN trained on A and B, both evaluated on the new test dataset.

Model	Precision	Recall	F1
PFN_A	65.14	88.49	75.02
PFN_B	63.74	86.99	73.56

and merging T), creating a test set with longer texts and more triples per data point without adding new hallucinations.

Language Model. We fine-tune PFN on each of the five modified WebNLG datasets.

Evaluation Results on Original Test Dataset. The results are presented in Table 5. Since all the results of the modified WebNLG datasets are similar to each other, we present with B_{fused} an average of all five (all results are in our repository). The evaluation on the original A^{test} dataset shows that the recall drops for an average of 1.98% and the F1-score for 1.26% (statistically significant differences using the paired t-test and a confidence interval of 0.95).

Evaluation Results on New Test Dataset. The results are presented in Table 6. They indicate a similar difference in the performance of PFN_A and PFN_B between the evaluation on the altered and the original test dataset. The recall difference is not statistically significant while the f1-score and precision are significant (using a paired t-test and confidence interval of 0.95).

Discussion. The presented findings indicate that whether we extensively increase the information content, keep it a bit shorter, or create more similar information, the different prompts and irrelevant hallucinations have only a minor impact on the trained relation extraction models. Through the evaluation on the new test dataset, we observed that the small differences between PFN_A and $PFN_{B_{fused}}$ cannot be attributed to the fact that $PFN_{B_{fused}}$ learned to ignore the back part of the natural language texts (which contains the newly added hallucinations) and only extracts triples from the front part. This is evident in the statistically insignificant difference between the recall of PFN_A and PFN_B evaluated on the new test dataset. The significant results in precision and F1-score are not further relevant for us.

Overall, there is a minor impact of irrelevant hallucinations in relation extraction models because these models are trained to prioritize and extract only those relations classified under relevant relation types, effectively disregarding all others categorized as irrelevant relation types. Thus, irrelevant hallucinations are systematically ignored during the training process.

The observed differences in model performance, despite expectations, may stem from two potential factors that require further investigation. The first possibility is that Llama 2 occa-

sionally introduces relevant relations in what is mostly uncontrolled information. The second possibility is an increase in errors by the relation extraction model due to processing a larger volume of text, regardless of its relevance. Further experiments are needed in this regard.

Based on these results and explanations, we can confirm the assumption that relevant hallucinations in (synthetic) training data have a much stronger impact on the performance of relation extraction models trained on them than irrelevant hallucinations. Therefore, irrelevant hallucinations can mostly be neglected regarding the influence on training performance of relation extraction models. This also means that when creating datasets or improving annotation quality, removing relevant hallucinations should be the priority.

3.2. Evaluating Hallucination Detection

We consider two approaches of hallucination detection, as outlined in the following.

3.2.1. Named Entity Recognition-Based Hallucination Detection

A first approach for hallucination detection was suggested by Liu et al. [42] and involves named entity recognition (NER). This approach extracts entities from a text and compares them to the entities in the corresponding triples. Entities found in the text but absent from the triples are identified as hallucinations.

Dataset: The sentence-level WebNLG dataset version v3.0 [55] serves as the basis for this work. The dataset includes annotations with one to seven triples.

Model: We decided to use the widely used SpaCy model given its wide usage and solid performance. Through preliminary tests, we can confirm the theory that the entities extracted by the model from S are often correct but not equivalent to those in the annotation T. This can result in cases where, for example, 'Alan Bean' is in T but only 'A. Bean' is extracted from S, which essentially means the same thing. To solve this problem, we use the sentence similarity model *all-mpnet-base-v2* [56] to compare the extracted and annotated entities.

Evaluation Setup: We use the precision, recall, and F1-score for the evaluation. We classify 'hallucination-free' texts that are correctly accepted as true positives.

For our experiments, we utilize 3,000 data points sampled from D. For each data point, we randomly select one correct text and one hallucination to maintain a balanced ratio between the two. The hallucination text can contain one to six hallucinated triples. Additionally, we conduct a hyperparameter sweep across all acceptance thresholds ranging from 0.05 to 0.95 (inclusive) in 0.05 increments to find the best-performing threshold for the sentence similarity model.

Evaluation Results: Figure 5 shows a climbing precision and falling recall with increasing threshold. Given those trends, the F1-score increases with a higher threshold up to 0.55. After this, it falls until the end. At the peak of 0.55, the precision, recall, and F1-score are 85.34, 82.25, and 83.76%, respectively. Given this, the overall results obtained from the tests seem satisfactory. Out of all the texts classified as 'clean,' around 85.34% were correctly identified as clean. Similarly, among all the tested clean texts, 82.25% were accurately classified as 'clean.' With that performance, the approach can be used to detect hallucinations and provide an approximate understanding of the amount of hallucinations in datasets.
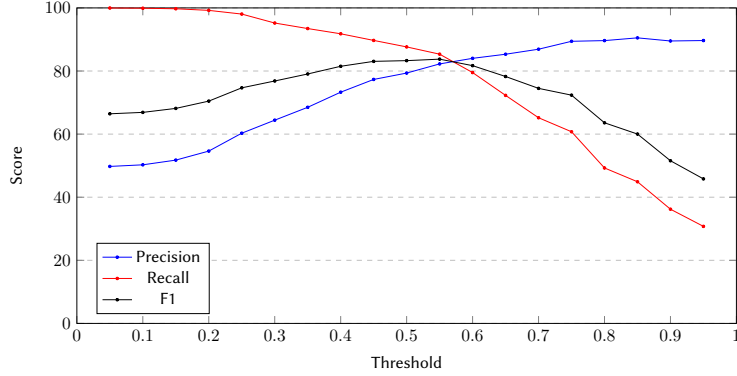
Figure 5: Precision, recall, and F1-score of the NER metric at varying acceptance thresholds. The highest F1-score is achieved at a threshold of 0.55.

The presented approach has several limitations. One limitation is that the equivalence between extracted and annotated entities depends on the sentence similarity model, making it unclear how many entities were incorrectly accepted or rejected. A fixed threshold is also needed to define equivalence, with the best results found at 0.55, indicating significant differences between extracted entities and annotations. This complicates model evaluation, and the issue can vary across datasets due to differing annotation formats. A potential solution is to use textual entailment instead of sentence similarity to assess entity matches.

3.2.2. Textual Entailment Approach for Hallucination Detection

Another approach, inspired by Dušek and Kasner [43], uses an entailment model M to check if a sentence S contains the same information as a set of triples T. The triples $t \in T$ are combined into a single sentence using conjunctions. If M classifies T as not entailed, it indicates hallucinations; if classified as entailed, S is considered hallucination-free.

Dataset: Since we evaluate a new approach on the same task as in the previous Section 3.2.1, we do not need to adjust the dataset and can continue to use D.

Model: For this task, we focus on the *roberta-large-mnli* [57] and *deberta-v2-xlarge-mnli* [58] models. Both models perform well on SQuAD 1.1/2.0 and various GLUE benchmarks.

Classifier: An entailment model can be used to test whether sentence S2 is part of sentence S1 or if the content of S1 implies the content of S2. We design the model to test for any hallucinations in S compared to T, from each data point of a dataset.

The initial step involves pre-processing the triples, typically formatted as 'Entity_1 | relation | Entity_2' or as a three-element list. Here, we receive the input in the former format and replace all occurrences of '_' and ' | ' with spaces. Next, the goal is to create a sentence, ST, that encompasses all triples from T and accurately conveys T's informative value. This is done by combining the pre-processed triples $t \in T$ into a single sentence, linked by the conjunction 'and.' Finally, M verifies whether S is entailed in ST. If the result is 'entailed,' S is deemed correct; otherwise, 'neutral' or 'contradictory' results signal the presence of hallucinations.

Evaluation Setup: We test M on 4,000 sampled data points from D. Each sample comprises an annotation T and two texts, $S1 \in h$ and $S2 \in c$, while $h, c \subset D$. That means that two tests

Table 7

Comparison of NER-based and textual entailment approaches for detecting hallucinations in relation extraction tasks.

Approach	Precision	Recall	F1
NER with *SpaCy*	85.34	82.25	83.76
NLI with *roberta-large-mnli*	88.29	93.45	90.79
NLI with *deberta-v2-xlarge-mnli*	**90.05**	**94.35**	**92.15**
NLI with *Dušek and Kasner* [43]	77.20	79.60	78.40

have to be conducted for each data point, one for each text. This results in 8,000 classifications. The evaluation is otherwise similar to the NER approach from the previous section.

Evaluation Results: The results are presented in Table 7. Both tested models outperformed the SpaCy model by 7.03% to 8.39% in F1-score using a straightforward sentence creation procedure for ST. The deberta-v2-xlarge-mnli model outperforms the roberta-large-mnli model by 1.36% in F1-score, with the most significant difference in recall, while precision increases only slightly. When comparing our best hallucination detection approach to the method used by Dušek and Kasner [43], our approach performs significantly better, achieving a 13.75% higher F1-score. However, this comparison should be considered with caution, as their study used an older version of the WebNLG dataset with many incorrect annotations.

An F1-score of 92.15% demonstrates that it is possible to reliably classify each data point in a dataset as either containing hallucinations or being hallucination-free. With high recall and precision, most clean texts are correctly identified as such, and the error rate for texts wrongly identified as hallucination-free is under 10%. This performance allows for the effective detection (and removal) of hallucinations in datasets, thereby significantly improving annotation quality.

Despite the superior results compared to Dušek and Kasner [43], who also relied on natural language inference, the comparison must be approached with caution due to their use of an outdated WebNLG version with incorrect annotations affecting their outcomes. Furthermore, this approach is limited to the detection of hallucinations and to sentence-level datasets, similar to the constraints discussed for the NER metric.

4. Conclusion

In this paper, we analyzed the impact of hallucinations in synthetic training data on relation extraction tasks. Our evaluation revealed significant performance declines with recall reductions between 19.1% and 39.2%. This indicates that hallucinations notably compromise the ability of models to accurately extract relations from texts. We identified a distinction between relevant and irrelevant hallucinations, noting that the former significantly impairs performance, while the latter has a minimal impact. Additionally, we developed methods for the detection (and thus mitigation) of hallucinations to improve data quality and, thus, model performance. Our approaches, successfully classified texts as either 'hallucinated' or 'clean,' with notable F1-scores of 83.8% and 92.2%. In the future, we will analyze the impact of hallucinations in datasets for other NLP tasks, such as entity and event extraction.

References

[1] A. Hernández-García, P. König, Data augmentation instead of explicit regularization, CoRR abs/1806.03852 (2018). URL: http://arxiv.org/abs/1806.03852. arXiv:1806.03852.

[2] J. Gao, R. Pi, Y. Lin, H. Xu, J. Ye, Z. Wu, W. Zhang, X. Liang, Z. Li, L. Kong, Self-guided noise-free data generation for efficient zero-shot learning, in: Proceedings of the Eleventh International Conference on Learning Representations, ICLR'23, 2023.

[3] J. Wei, K. Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP'19, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.

[4] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do not have enough data? deep learning to the rescue!, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI'20, AAAI Press, 2020, pp. 7383–7390. URL: https://doi.org/10.1609/aaai.v34i05.6233. doi:10.1609/AAAI.V34I05.6233.

[5] P. H. Cabot, R. Navigli, REBEL: relation extraction by end-to-end language generation, in: Findings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP'21, 2021, pp. 2370–2381. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.204.

[6] M. Josifoski, M. Sakota, M. Peyrard, R. West, Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP'23, Singapore, 2023, pp. 1555–1574. URL: https://aclanthology.org/2023.emnlp-main.96. doi:10.18653/v1/2023.emnlp-main.96.

[7] L. F. R. Ribeiro, M. Schmitt, H. Schütze, I. Gurevych, Investigating Pretrained Language Models for Graph-to-Text Generation, in: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, NLP4ConvAI@ACL'21, Online, 2021, pp. 211–227. URL: https://aclanthology.org/2021.nlp4convai-1.20. doi:10.18653/v1/2021.nlp4convai-1.20.

[8] Q. Wang, S. Yavuz, X. V. Lin, H. Ji, N. F. Rajani, Stage-wise fine-tuning for graph-to-text generation, in: Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, JUli 5-10, 2021, Association for Computational Linguistics, 2021, pp. 16–22. URL: https://doi.org/10.18653/v1/2021.acl-srw.2. doi:10.18653/V1/2021.ACL-SRW.2.

[9] J. Chen, Z. Yang, D. Yang, MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL'20, Online, 2020, pp. 2147–2157. URL: https://aclanthology.org/2020.acl-main.194. doi:10.18653/v1/2020.acl-main.194.

[10] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 296–310. URL: https://aclanthology.org/2021.naacl-main.28. doi:10.18653/v1/2021.naacl-main.28.

[11] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, W. Park, GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation, in: Findings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP'21, Punta Cana, Dominican Republic, 2021, pp. 2225–2239. URL: https://aclanthology.org/2021.findings-emnlp.192. doi:10.18653/v1/2021.findings-emnlp.192.

[12] H. Ye, T. Liu, A. Zhang, W. Hua, W. Jia, Cognitive Mirage: A Review of Hallucinations in Large Language Models, 2023. URL: http://arxiv.org/abs/2309.06794. doi:10.48550/arXiv.2309.06794.

[13] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, H. S. Chan, W. Dai, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Computing Surveys 55 (2023) 1–38. URL: http://arxiv.org/abs/2202.03629. doi:10.1145/3571730, arXiv:2202.03629.

[14] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation, CoRR abs/2307.03987 (2023). URL: https://doi.org/10.48550/arXiv.2307.03987. doi:10.48550/ARXIV.2307.03987.

[15] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A Survey of Data Augmentation Approaches for NLP, in: Findings of the Association for Computational Linguistics, ACL-IJCNLP'21, Virtual Event, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.

[16] Y. Li, T. Cohn, T. Baldwin, Robust Training under Linguistic Adversity, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL'17, Valencia, Spain, 2017, pp. 21–27. URL: https://aclanthology.org/E17-2004.

[17] J. Wei, C. Huang, S. Xu, S. Vosoughi, Text Augmentation in a Multi-Task View, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL'21, Virtual Event, 2021, pp. 2888–2894. URL: https://aclanthology.org/2021.eacl-main.252. doi:10.18653/v1/2021.eacl-main.252.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond Empirical Risk Minimization, 2018. URL: http://arxiv.org/abs/1710.09412. doi:10.48550/arXiv.1710.09412, arXiv:1710.09412.

[19] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, J. Choe, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV'19, IEEE, 2019, pp. 6022–6031. URL: https://doi.org/10.1109/ICCV.2019.00612. doi:10.1109/ICCV.2019.00612.

[20] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold Mixup: Better Representations by Interpolating Hidden States, in: Proceedings of the 36th International Conference on Machine Learning, ICML'19, 2019, pp. 6438–6447.

[21] H. Guo, Nonlinear Mixup: Out-Of-Manifold Data Augmentation for Text Classification, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 4044–4051. doi:10.1609/aaai.v34i04.5822.

[22] C. Beckham, S. Honari, V. Verma, A. M. Lamb, F. Ghadiri, R. D. Hjelm, Y. Bengio, C. Pal, On Adversarial Mixup Resynthesis, in: Advances in Neural Information Processing Systems, volume 32 of *NeurIPS'19*, 2019. URL: https://papers.nips.cc/paper/2019/hash/

f708f064faaf32a43e4d3c784e6af9ea-Abstract.html.

[23] D. Guo, Y. Kim, A. Rush, Sequence-Level Mixed Sample Data Augmentation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP'20, Virtual Event, 2020, pp. 5547–5552. URL: https://aclanthology.org/2020.emnlp-main.447. doi:10.18653/v1/2020.emnlp-main.447.

[24] J. Clive, K. Cao, M. Rei, Control Prefixes for Parameter-Efficient Text Generation (2022). URL: http://arxiv.org/abs/2110.08329. doi:10.48550/arXiv.2110.08329, arXiv:2110.08329.

[25] WebNLG, Papers with Code - WebNLG Dataset, 2024. URL: https://paperswithcode.com/dataset/webnlg.

[26] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, The WebNLG Challenge: Generating Text from RDF Data, in: Proceedings of the 10th International Conference on Natural Language Generation, INLG'17, Santiago de Compostela, Spain, 2017, pp. 124–133. URL: https://aclanthology.org/W17-3518. doi:10.18653/v1/W17-3518.

[27] X. Xu, Y. Zhu, X. Wang, N. Zhang, How to Unleash the Power of Large Language Models for Few-shot Relation Extraction?, in: Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing, SustaiNLP'23, Toronto, Canada (Hybrid), 2023, pp. 190–200. URL: https://aclanthology.org/2023.sustainlp-1.13. doi:10.18653/v1/2023.sustainlp-1.13.

[28] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, L. Kong, Zerogen: Efficient zero-shot learning via dataset generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP'22, Association for Computational Linguistics, 2022, pp. 11653–11669. URL: https://doi.org/10.18653/v1/2022.emnlp-main.801. doi:10.18653/V1/2022.EMNLP-MAIN.801.

[29] S. Thulasidasan, T. Bhattacharya, J. A. Bilmes, G. Chennupati, J. Mohd-Yusof, Combating label noise in deep learning using abstention, in: Proceedings of the 36th International Conference on Machine Learning, volume 97 of *ICML'19*, PMLR, 2019, pp. 6234–6243. URL: http://proceedings.mlr.press/v97/thulasidasan19a.html.

[30] X. Ma, H. Huang, Y. Wang, S. R. S. Erfani, J. Bailey, Normalized loss functions for deep learning with noisy labels, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *ICML'20*, 2020, pp. 6543–6553.

[31] Y. Liu, H. Guo, Peer loss functions: Learning from noisy labels without knowing noise rates, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *ICML'20*, 2020, pp. 6226–6236. URL: http://proceedings.mlr.press/v119/liu20e.html.

[32] K. Filippova, Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data, in: Findings of the 20210Conference on Empirical Methods in Natural Language Processing, EMNLP'20, Online, 2020, pp. 864–870. URL: https://aclanthology.org/2020.findings-emnlp.76. doi:10.18653/v1/2020.findings-emnlp.76.

[33] Q. Tan, L. Xu, L. Bing, H. T. Ng, S. M. Aljunied, Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP'22, Abu Dhabi, United Arab Emirates, 2022, pp. 8472–8487. URL: https://aclanthology.org/2022.emnlp-main.580. doi:10.18653/v1/2022.emnlp-main.580.

[34] G. Stoica, E. A. Platanios, B. Poczos, Re-TACRED: Addressing Shortcomings of the TACRED

Dataset, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 13843–13850. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17631. doi:10.1609/aaai.v35i15.17631.

[35] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, M. Sugiyama, How does Disagreement Help Generalization against Label Corruption?, in: Proceedings of the 36th International Conference on Machine Learning, ICML'19, 2019, pp. 7164–7173. URL: https://proceedings.mlr.press/v97/yu19b.html.

[36] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S. Xia, S. N. R. Wijewickrema, J. Bailey, Dimensionality-driven learning with noisy labels, in: Proceedings of the 35th International Conference on Machine Learning, volume 80 of *ICML'18*, 2018, pp. 3361–3370. URL: http://proceedings.mlr.press/v80/ma18d.html.

[37] E. Reiter, A Structured Review of the Validity of BLEU, Computational Linguistics 44 (2018) 393–401. URL: https://doi.org/10.1162/coli_a_00322. doi:10.1162/coli_a_00322.

[38] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, 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, ACL'19, Florence, Italy, 2019, pp. 2214–2220. URL: https://aclanthology.org/P19-1213. doi:10.18653/v1/P19-1213.

[39] Z. Wang, X. Wang, B. An, D. Yu, C. Chen, Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL'20, 2020, pp. 1072–1086. URL: http://arxiv.org/abs/2005.00969. doi:10.18653/v1/2020.acl-main.101, arXiv:2005.00969.

[40] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval Augmentation Reduces Hallucination in Conversation, in: Findings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP'21, Punta Cana, Dominican Republic, 2021, pp. 3784–3803. URL: https://aclanthology.org/2021.findings-emnlp.320. doi:10.18653/v1/2021.findings-emnlp.320.

[41] M. Martindale, M. Carpuat, K. Duh, P. McNamee, Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation, in: Proceedings of Machine Translation Summit XVII, Dublin, Ireland, 2019, pp. 233–243. URL: https://aclanthology.org/W19-6623.

[42] T. Liu, X. Zheng, B. Chang, Z. Sui, Towards Faithfulness in Open Domain Table-to-text Generation from an Entity-centric View, volume 35 of *AAAI'21*, 2021, pp. 13415–13423. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17583. doi:10.1609/aaai.v35i15.17583.

[43] O. Dušek, Z. Kasner, Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference, in: Proceedings of the 13th International Conference on Natural Language Generation, ICNLG'20, Dublin, Ireland, 2020, pp. 131–137. URL: https://aclanthology.org/2020.inlg-1.19. doi:10.18653/v1/2020.inlg-1.19.

[44] R. Tian, S. Narayan, T. Sellam, A. P. Parikh, Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation, 2020. URL: http://arxiv.org/abs/1910.08684. doi:10.48550/arXiv.1910.08684, arXiv:1910.08684 [cs].

[45] F. Nie, J.-G. Yao, J. Wang, R. Pan, C.-Y. Lin, A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL'19, Florence, Italy, 2019, pp. 2673–2679.

URL: https://aclanthology.org/P19-1256. doi:10.18653/v1/P19-1256.

[46] Z. Yan, C. Zhang, J. Fu, Q. Zhang, Z. Wei, A Partition Filter Network for Joint Entity and Relation Extraction, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP'21, Online and Punta Cana, Dominican Republic, 2021, pp. 185–197. URL: https://aclanthology.org/2021.emnlp-main.17. doi:10.18653/v1/2021.emnlp-main.17.

[47] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A Large-Scale Document-Level Relation Extraction Dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL'19, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074. doi:10.18653/v1/P19-1074.

[48] Y. Ma, A. Wang, N. Okazaki, DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction (2023). URL: http://arxiv.org/abs/2302.08675. doi:10.48550/arXiv.2302.08675, arXiv:2302.08675.

[49] Papers with Code - DocRED Benchmark (Relation Extraction), 2024. URL: https://paperswithcode.com/sota/relation-extraction-on-docred.

[50] Papers with Code - ReDocRED Benchmark (Relation Extraction), 2024. URL: https://paperswithcode.com/sota/relation-extraction-on-redocred.

[51] A. Aghajanyan, D. Okhonko, M. Lewis, M. Joshi, H. Xu, G. Ghosh, L. Zettlemoyer, HTLM: hyper-text pre-training and prompting of language models, in: Proceedings of the 10th International Conference on Learning Representations, ICLR'22, 2022.

[52] X. Zeng, D. Zeng, S. He, K. Liu, J. Zhao, Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL'18, Melbourne, Australia, 2018, pp. 506–514. URL: https://aclanthology.org/P18-1047. doi:10.18653/v1/P18-1047.

[53] WebNLG, PapersRE with Code - WebNLG Benchmark (Relation Extraction), 2024. URL: https://paperswithcode.com/sota/relation-extraction-on-webnlg, available at https://paperswithcode.com/sota/relation-extraction-on-webnlg.

[54] H. Touvron, L. Martin, et al., Llama 2: Open foundation and fine-tuned chat models, CoRR abs/2307.09288 (2023). URL: https://doi.org/10.48550/arXiv.2307.09288. doi:10.48550/ARXIV.2307.09288, arXiv:2307.09288.

[55] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, Creating Training Corpora for NLG Micro-Planners, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 179–188. URL: https://aclanthology.org/P17-1017. doi:10.18653/v1/P17-1017.

[56] sentence-transformers/all-mpnet-base-v2 · Hugging Face, 2024. URL: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[57] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). URL: http://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692, arXiv:1907.11692.

[58] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: Proceedings of the 9th International Conference on Learning Representations, ICLR'21, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.