GREScores: Evaluate Knowledge Extraction Ability of LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated significant capabilities in a variety of tasks, particularly in natural language processing. Recent research has highlighted the exceptional ability of LLMs in generative relation extraction (GRE), where they efficiently extract complex relationships, often outperforming fine-tuned models in limited-sample scenarios. Due to the LLMs' flexibility in defining entities and relationships, traditional metrics for relation extraction are inadequate for accurately evaluating the quality of extracted knowledge triples. This situation underscores the urgent need for a specialized, multi-dimensional evaluation metric for GRE. Consequently, we propose GREScores, a novel metric designed to evaluate GRE performance from multiple aspects, including distribution similarity, factualness, completeness, uniqueness, and granularity. We have also benchmarked the GRE capabilities of current leading LLMs using GREScores, providing a comprehensive analysis of their performance in this advanced field.

1 Introduction

The digital age has ushered in an era of unprecedented data generation, with approximately 80% of the Internet's data available in text format (Dai and Maitra, 2020). Natural Language Processing (NLP) is at the forefront of harnessing this wealth of information, with relation extraction (RE) serving as a cornerstone for converting unstructured text into structured, actionable knowledge. Traditional relation extraction techniques have relied heavily on predetermined patterns and statistical models, which, while effective in certain contexts, lack the dynamism to deal with the fluidity of natural language.

Generative Relation Extraction (GRE) represents a significant evolution in the field of NLP. By leveraging the advanced capabilities of Large Language Models (LLMs) like the GPT series, GRE offers a more flexible, context-aware approach to relation extraction. These models can intuitively generate text and identify complex relationships within it, without the need for predefined patterns or extensive training datasets. This marks a departure from traditional methods, enabling a broader capture of relationships and embracing the subtle nuances of language expression.

The advantages of GRE over traditional RE are manifold. GRE systems are highly efficient, capable of producing structured knowledge without the extensive and often labor-intensive data labeling and feature engineering required by traditional RE (Jiang, 2023). Moreover, the inherent understanding of context by LLMs allows GRE to process a wide array of texts with minimal domain-specific tuning, making it a versatile tool across different knowledge domains (Naik et al., 2023; Li et al., 2023b; Jiang et al., 2023b). (TODO)

However, the versatility of GRE introduces complexities in evaluation (Wadhwa et al., 2023). Traditional metrics, designed to assess precision and recall based on exact matches, struggle to accommodate the generative outputs of LLMs. Such outputs often express the same semantic information in varied forms, challenging the applicability of conventional evaluation methods (Taillé et al., 2020).

To navigate these challenges, we introduce GREScores, a multifaceted evaluation framework specifically devised for the generative approach of GRE. GREScores systematically evaluates the performance of LLMs across multiple dimensions:

- *Topical Similarity*: It assesses the alignment of the extracted triples with the topical content of the source text, ensuring that the information is contextually relevant.
- *Uniqueness*: It ensures the diversity of information by evaluating the extracted triples for redundancy and similarity, promoting the richness of the content extracted.

- *Factualness*: It gauges the veracity of the extracted triples, which is critical for the credibility of the GRE system.
- *Granularity*: It examines the specificity of the information captured by the triples, ensuring an optimal level of detail.
- *Completeness*: Functioning like recall, it measures the extent to which the GRE captures the range of information present in the source text, compared to gold standard.

This paper delves into the fundamentals of GRE, elucidates the intricacies of GREScores, and validates its efficacy in benchmarking the performance of contemporary LLMs. Our contributions not only shed light on the potential of LLMs in GRE but also establish a new benchmark for evaluating relation extraction in the era of generative AI.

2 Preliminaries

Before discussing in detail of our approach, we define a few key concepts and notations below.

Definition 1 (Source Document) *A source document $\mathcal{D}$ is a piece of free-text, which can be a sentence, a passage, or a document.*

Definition 2 (Extracted Triples) *A triple $\tau = \langle s|r|o\rangle$ is a structure formatting a piece of free-text into a subject s, a relation r, and an object o. Example: For a sentence "Alice lives in Champaign.", "Alice" is the subject, "live in" is the relation, and "Champaign" is the object. Together, they form a triple $\langle Alice|live\_in|Champaign\rangle$. We define $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, ...]$ as a list of triples extracted from the source document $\mathcal{D}$.*

Definition 3 (Generative Relation Extraction) *Generative Relation Extraction uses a generative large language model (LLM) to extract relational triples from a source document $\mathcal{D}$. The model functions on an autoregressive basis at the token level, expressed as $P(x_t|x_1, x_2, \ldots, x_{t-1}, \mathcal{D})$, where x_t represents the t^{th} token in the output sequence. The process generates a sequence of tokens that are structured into triples $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, \ldots, \tau_m]$.*

3 GREScores

We introduce GREScores, an automated multi-aspect evaluation metric for generative relation extraction. GREScores are composed of the following sub-scores: (1) Distribution Similarity Score, (2) Uniqueness Score, (3) Factualness Score, (4) Granularity Score, and (5) Completeness Score,. Each of these scores will be elaborated upon in the subsequent subsections.

3.1 Topical Similarity Score

We compute topical similarity score (TS) to measure the information abundance of the extracted triples $\mathcal{T}_\mathcal{D}$ compared to the source text $\mathcal{D}$. Here, we employ a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), an algorithm that represents each document as a blend of a certain number of topics, for topic modeling. We concatenate the elements in each triple so that $\mathcal{T}^\Delta_\mathcal{D} = [\tau'_1, \tau'_2, ...] = [s_1 \oplus r_1 \oplus o_1, s_2 \oplus r_2 \oplus o_2, ...]$. TS is computed as:

$$t(\mathcal{D}, \mathcal{T}^\Delta_\mathcal{D}) = e^{-\sum_{i=1}^{K} LDA(\mathcal{D})_i \cdot \log\left(\frac{LDA(\mathcal{D})_i}{LDA(\mathcal{T}^\Delta_\mathcal{D})_i}\right)} \quad (1)$$

which is based on the KL-divergence of two topical distributions. A higher TS indicates that the extracted triples closely align with the topical content of the source document, reflecting effective and relevant information extraction, while a lower TS suggests that the extracted triples may be missing key topical elements from the source.

3.2 Uniqueness Score

The Uniqueness Score (US) assesses the diversity of the extracted triples $\mathcal{T}_\mathcal{D}$ in GRE, emphasizing the importance of extracting varied and distinct relationships. Given $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, \ldots, \tau_n]$, with each triple τ_i encoded into a vector $\mathbf{v}_i$ using word embeddings, the US is computed as follows:

$$u(\mathcal{T}_\mathcal{D}) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(1 - \text{CosSim}(\mathbf{v}_i, \mathbf{v}_j)^2\right) \quad (2)$$

where $\text{CosSim}(\mathbf{v}_i, \mathbf{v}_j)$ is the cosine similarity between the vector representations of triples τ_i and τ_j. However, instead of using the raw cosine similarity scores, we square these scores. Squaring the cosine similarity emphasizes larger differences between the embeddings, effectively stretching the distribution of the Uniqueness Scores. A higher US still indicates greater diversity among the triples, while a lower score suggests more similarity and potential redundancy. However, with this adjustment, the US is more sensitive to variations in the embeddings, potentially providing a more nuanced view of the uniqueness of the information contained in the triples.

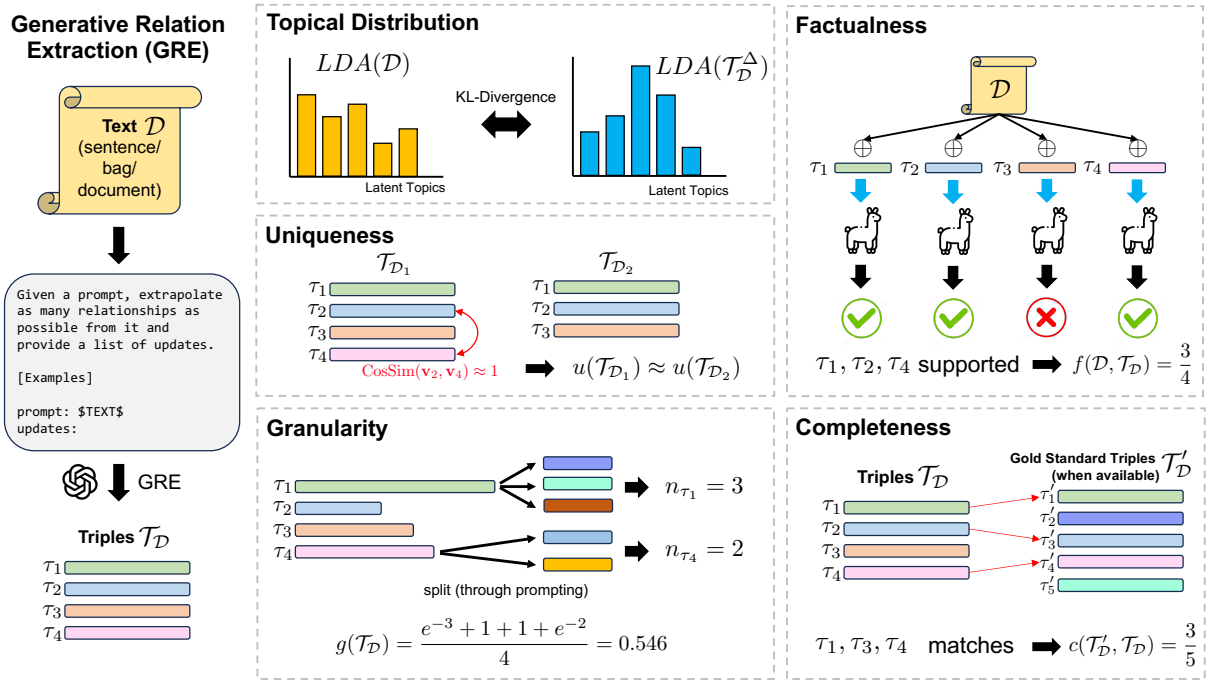

Figure 1: **GREScore for the evaluation of generative relation extraction (GRE).** ***Left***: An example showing the GRE process to extract triples $\mathcal{T}_\mathcal{D}$ from a source text $\mathcal{D}$ through prompting generative large language model. ***Right***: illustration of sub-scores contained in GREScore regarding: Topical Distribution (§3.1), Uniqueness (§3.2), Fatualness (§3.3), Granularity (§3.4), and Completeness (§3.5).

3.3 Factualness Score

The Factualness Score (FS) measures the degree to which the extracted triples $\mathcal{T}_\mathcal{D}$ are supported by the source text $\mathcal{D}$. It addresses the problem of hallucinations (Zhang et al., 2023), where LLMs generate information that is not present or implied in the source text. Inspired by previous works (Min et al., 2023; Jiang et al., 2021), FS is determined through a triple-wise verification where each triple τ from $\mathcal{T}_\mathcal{D}$ is individually assessed for its entailment with facts in the source document $\mathcal{D}$:

$$f(\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\tau \in \mathcal{T}_\mathcal{D}} [\![\tau \text{ is supported by } \mathcal{D}]\!] \quad (3)$$

where $[\![\tau \text{ is supported by } \mathcal{D}]\!]$ represents an indicator function that yields 1 if the triple is factual and 0 otherwise. A higher FS indicates that a larger proportion of the extracted triples are factually accurate with respect to the source text, whereas a lower score indicates a higher occurrence of hallucinated or unsupported information. This metric is essential for ensuring the trustworthiness of the information extracted by the generative model. Following (Min et al., 2023), we use LLaMA-7B (Touvron et al., 2023a) trained on Super Natural Instructions as the evaluator for FS.

3.4 Granularity Score

The Granularity Score (GS) evaluates the level of detail of the extracted triples $\mathcal{T}_\mathcal{D}$ from the source text $\mathcal{D}$. It is based on the premise that triples should capture the optimal granularity of information—not too coarse. The GS aims to penalize triples that are overly broad and could be further split into more precise statements. The process involves an assessment of each triple's potential to be split into more granular sub-triples. This can be performed by prompting an LLM to evaluate if a given triple can be divided into additional, more specific triples. The number of possible splits is represented by n_τ for each triple τ.

The Granularity Score for the extracted triples $\mathcal{T}_\mathcal{D}$ is calculated using the formula:

$$g(\mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\tau \in \mathcal{T}_\mathcal{D}} e^{-n_\tau} \quad (4)$$

where e^{-n_τ} is the exponential decay function based on the number of splits n_τ, which assigns a lower score to triples that can be split into more sub-triples (indicating they are too broad or general). Therefore, a lower Granularity Score indicates that the triples could be broken down further, while a higher score suggests that the triples are at an appropriate level of specificity.

3.5 Completeness Score

The Completeness Score (CS) evaluates how comprehensively the extracted triples $\mathcal{T}_\mathcal{D}$ cover the information present in the source text $\mathcal{D}$. This score is analogous to the recall metric in information retrieval and is particularly important when gold standard triples $\mathcal{T}'_\mathcal{D}$ are available for comparison. CS is assessed by determining the proportion of gold standard triples that are successfully captured by the extracted triples. For each gold standard triple τ', we find the best matching triple τ from $\mathcal{T}_\mathcal{D}$, using cosine similarity of their embeddings as the matching criterion. If the cosine similarity exceeds a specified threshold ϕ, the triple τ is considered a match. CS is then computed as:

$$c(\mathcal{T}'_\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{|\{\tau' \in \mathcal{T}'_\mathcal{D} | \exists \tau \in \mathcal{T}_\mathcal{D}, \text{sim}(\tau, \tau') \geq \phi\}|}{|\mathcal{T}'_\mathcal{D}|} \tag{5}$$

where $\text{sim}(\tau, \tau') = \text{CosSim}(emb(\tau), emb(\tau'))$ calculates the cosine similarity between the embeddings of the extracted triple and the gold standard triple. The threshold ϕ is pre-defined to determine the acceptable level of similarity for a match. A higher CS indicates that the extracted triples effectively capture the complete range of information as represented by the gold standard, while a lower CS points to potential gaps in the extracted knowledge.

4 Experiments

4.1 Datasets

We evaluate the following RE datasets, focusing on GRE performance using their test sets, which feature meticulous human annotations.

CDR (Li et al., 2016). A *document-level* RE dataset comprising 1,500 PubMed abstracts. The dataset is divided evenly for training, development, and testing. Each abstract has been meticulously annotated to mark binary interactions between chemical compounds and disease entities.

DocRED (Yao et al., 2019a). A *document-level* RE dataset derived from Wikipedia and Wikidata, featuring 5,053 Wikipedia documents with 132,375 entities and 56,354 relational facts. It includes human annotations for entity mentions, coreferences, and intra- and inter-sentence relations, along with supporting evidence.

NYT10m & **Wiki20m** (Han et al., 2019). Two *bag-level*[1] RE datasets sourced from New York Times and Wikipedia, respectively. Both datasets have manually annotated test sets.

TACRED (Zhang et al., 2017) & **Wiki80** (Han et al., 2019): Two *sentence-level* RE datasets. TACRED includes 106,264 examples from newswire and web texts, covering 41 relation types, using TAC KBP challenge data and crowdsourcing. Wiki80, sourced from FewRel (Han et al., 2018b), contains 80 relations with 56,000 instances from Wikipedia and Wikidata.

For the document-level, bag-level, and sentence-level datasets, we randomly select 200, 500, and 800 samples, respectively, from their test sets.

4.2 Models

We test the GRE capabilities of the following LLMs that can understand instructions. **LLAMA Family** (Touvron et al., 2023b,c): LLAMA-2-7B, LLAMA-2-70B, Vicuna-1.5-7B, Vicuna-1.3-33B, and WizardLM-70B (Xu et al., 2023). **GPT Family** (Brown et al., 2020): text-davinci-003, GPT-3.5-Turbo (1106), GPT-3.5-Turbo-Instruct, GPT-4, and GPT-4-Turbo. **Others**: GALACTICA (Taylor et al., 2022), Mistral-7B-Instruct (Jiang et al., 2023a), and OpenChat-3.5 (Wang et al., 2023). We also test the human annotations (provided as the labels in the datasets) with our GREScores.

4.3 Implementation Details

4.4 Performance Comparison

4.5 Case Study

[1] A "bag" of sentences that share the same entity pair. https://opennre-docs.readthedocs.io

	CDR							DocRED						
	#triples	***#tok/tri***	***TS***	***US***	***FS***	***GS***	***CS***	***#triples***	***#tok/tri***	***TS***	***US***	***FS***	***GS***	***CS***
Ground Truth														
Vicuna-7B	6.79	8.37	57.8	55.1	84.7	31.8	29.7	7.43	9.94	23.1	67.6	93.4	37.7	34.6
Vicuna-33B	6.41	10.47	73.0	67.5	97.3	30.5	33.0	10.76	9.81	34.7	73.5	97.2	42.0	43.2
LLaMA-2-7B	5.55	6.74	48.6	48.1	62.0	29.5	23.8	2.7	3.22	12.8	25.8	34.0	20.7	14.1
LLaMA-2-70B	10.75	8.14	74.8	67.9	96.6	48.9	51.4	13.81	8.72	39.2	73.7	97.3	51.8	45.6
WizardLM-70B	10.18	7.78	65.4	53.4	76.4	29.2	36.6	5.84	3.6	24.3	27.5	37.9	18.3	16.5
text-davinci-003	12.69	8.26	76.7	69.6	96.8	44.1	51.7	15.25	8.48	40.1	73.2	97.6	49.5	51.6
GPT-3.5-Turbo-Inst.	16.12	8.28	77.6	68.3	96.8	43.1	53.8	17.79	8.85	47.8	72.6	98.1	46.3	51.1
GPT-3.5-Turbo	11.22	11.4	81.7	69.5	98.2	33.0	41.5	14.96	9.89	50.4	72.5	98.5	42.1	44.5
GPT-4	14.29	9.34	81.7	68.7	97.9	-	49.3	17.82	8.71	48.6	74.0	98.6	-	52.8
GPT-4-Turbo	18.63	8.5	82.1	68.1	96.8	-	52.6	21.52	8.7	50.0	71.4	97.6	-	54.6
Mistral-7B-Inst.	14.22	9.11	69.0	66.9	93.5	-	35.9	11.26	9.6	30.2	75.7	94.1	-	35.0
Zephyr-7B-Beta	25.85	8.83	49.1	69.7	70.1	-	34.1	18.62	8.55	27.9	74.3	94.7	-	43.1
Galactica-30B	0.2	0.3	4.1	1.7	-	-	0.7	0.0	0.0	8.6	0.0	0.0	-	0.0
OpenChat-3.5	8.55	12.6	78.7	69.6	-	-	38.0	15.35	8.85	39.7	74.6	-	-	49.6

Table 1: **GRESCORES evaluation on *document-level* datasets.** Scores (%) are averaged across documents. *#triples* and *#tok/tri* denote the number of triples per document and the number of tokens per triple, respectively.

	NYT10m							Wiki20m						
	#triples	***#tok/tri***	***TS***	***US***	***FS***	***GS***	***CS***	***#triples***	***#tok/tri***	***TS***	***US***	***FS***	***GS***	***CS***
Ground truth														
Vicuna-1.5-7B	3.09	7.79	42.0	60.6	80.0	49.4	40.3	3.02	7.47	48.3	72.9	50.0	55.8	55.0
Vicuna-1.3-33B	4.69	7.23	47.8	69.4	75.1	55.2	48.7	4.06	6.99	49.8	73.0	84.4	59.7	62.7
LLAMA-2-7B	1.01	0.0	21.3	0.0	86.7	3.8	4.7	3.13	6.3	37.9	63.3	73.4	-	49.0
LLAMA-2-70B	4.97	6.94	45.4	69.1	81.7	63.5	56.0	4.08	6.87	45.2	69.6	88.1	66.1	64.9
WizardLM-70B	4.39	4.21	30.5	38.6	43.9	32.7	31.3	3.56	5.62	43.1	59.4	67.3	47.9	52.7
text-davinci-003	4.87	7.13	50.6	71.1	85.8	60.0	58.0	3.73	8.16	51.8	78.2	91.3	59.0	63.3
GPT-3.5-Turbo-Inst.	5.81	7.02	54.2	70.1	84.0	62.5	58.8	4.79	7.67	54.0	75.3	90.1	62.2	60.2
GPT-3.5-Turbo	4.13	6.2	43.3	55.9	68.2	42.4	32.7	3.55	7.7	48.2	61.0	85.2	49.8	47.7
GPT-4	5.14	7.39	56.2	70.7	89.0	-	57.1	3.76	8.07	59.0	74.1	93.2	-	59.9
GPT-4-Turbo	5.29	7.81	58.1	70.2	89.6	-	57.7	4.16	7.58	56.4	75.8	92.4	-	64.6
Mistral-7B-Inst.	5.66	7.4	40.6	63.9	75.4	-	38.0	3.98	6.88	43.3	67.8	83.6	-	55.1
Zephyr-7B-Beta	7.8	7.2	36.5	67.9	-	-	48.5	5.23	6.75	40.3	71.8	-	-	60.7
Galactica-30B	8.31	8.65	29.7	70.6	-	-	33.5	5.98	8.39	35.3	68.5	-	-	52.2
OpenChat-3.5	5.22	7.2	54.0	70.7	-	-	58.3	4.27	6.96	57.5	77.2	-	-	64.6

Table 2: **GRESCORES evaluation on *bag-level* datasets.** Scores (%) are averaged across bags. *#triples* and *#tok/tri* denote the number of triples per bag and the number of tokens per triple, respectively.

	TACRED							Wiki80						
	#triples	*#tok/tri*	*TS*	*US*	*FS*	*GS*	*CS*	*#triples*	*#tok/tri*	*TS*	*US*	*FS*	*GS*	*CS*
Ground Truth														
Vicuna-7B	2.55	8.65	40.4	56.6	75.6	50.3	77.9	2.38	7.91	41.3	59.8	81.0	-	50.7
Vicuna-33B	4.3	7.31	44.3	68.6	71.0	58.5	82.2	3.81	7.23	47.3	72.7	79.9	60.2	62.5
LLAMA-2-7B	2.79	6.25	36.7	55.6	66.9	57.2	79.9	2.36	5.76	25.8	58.1	60.4	-	41.4
LLAMA-2-70B	4.07	6.39	40.8	67.5	74.5	67.2	84.7	3.68	6.62	41.5	70.4	82.4	65.6	61.7
WizardLM-70B	2.07	2.85	23.3	28.5	28.0	24.7	69.8	2.08	3.21	25.6	32.8	36.6	27.3	28.2
text-davinci-003	4.39	7.08	56.1	69.9	84.0	63.4	86.7	4.03	6.82	59.2	75.4	89.2	64.0	67.5
GPT-3.5-Turbo-Inst.	5.03	6.99	58.6	69.4	81.6	63.8	86.0	4.39	6.88	60.2	74.4	88.7	63.9	67.0
GPT-3.5-Turbo	3.94	6.82	52.7	61.2	76.4	52.1	82.4	3.38	6.31	50.9	60.9	75.6	48.1	47.4
GPT-4	4.25	7.49	59.1	68.7	87.6	-	86.4	3.96	7.1	65.4	75.3	92.3	-	64.0
GPT-4-Turbo	4.43	7.79	58.5	68.3	88.6	-	87.9	4.01	7.57	61.9	74.6	92.8	-	65.8
Mistral-7B-Inst.	4.7	7.05	43.9	55.2	71.0	-	82.3	3.56	7.76	44.6	60.4	-	-	54.1
Zephyr-7B-Beta	5.38	7.64	36.4	70.0	-	-	80.1	4.51	7.77	43.2	72.5	-	-	60.4
Galactica-30B	8.46	8.86	33.4	74.4	-	-	75.5	5.58	7.17	35.0	68.7	-	-	52.3
OpenChat-3.5	4.26	7.13	50.7	69.8	-	-	86.6	3.97	7.02	53.8	74.2	-	-	65.2

Table 3: **GRESCORES evaluation on *sentence-level* datasets.** Scores (%) are averaged across sentences. *#triples* and *#tok/tri* denote the number of triples per sentence and the number of tokens per triple, respectively.

5 Related Works

5.1 Relation Extraction

Traditional RE. Traditional RE is centered around a predefined set of relations and is typically categorized based on its granularity: sentence-level, bag-level, and document-level. Sentence-level RE methods focus on learning paradigms and extracting relations between two entities mentioned within a single sentence (Doddington et al., 2004; Xu et al., 2016; Wei et al., 2020; Zhou et al., 2016). Bag-level RE, on the other hand, deals with scenarios where an entity pair may appear multiple times in different sentences, with a substantial likelihood that some of these sentences convey the relation between the entity pair (Lin et al., 2016; Han et al., 2018a; Ye and Ling, 2019; Zeng et al., 2015). Document-level RE seeks to predict entity relations across multiple sentences (Xu et al., 2022; Yao et al., 2019b; Christopoulou et al., 2019). In this paper, we assess the performance of generative models on these three levels of traditional RE datasets, highlighting the potential of generative models in addressing Traditional RE tasks.

Open RE. In contrast to Traditional RE, Open RE is not constrained by predefined relation types. It aims to uncover new relation types from unsupervised open-domain corpora and can be broadly categorized into two approaches: tagging-based and clustering-based. Tagging-based methods view Open RE as a sequence labeling problem and extract relational phrases composed of words from sentences (Jia et al., 2019; Cui et al., 2018; Stanovsky et al., 2018). Clustering-based Open RE approaches, on the other hand, extract rich features for relations using external linguistic tools and cluster semantic patterns into distinct relation types (Zhou et al., 2023; Marcheggiani and Titov, 2016; ElSahar et al., 2017).

Generative RE. Generative models have exhibited significant promise in the field of RE (Wadhwa et al., 2023). Before the era of LLMs, researchers such as Ni et al. (2022); Paolini et al. (2021); Cabot and Navigli (2021) utilized sequence-to-sequence models like BART (Lewis et al., 2020) for extracting relation triplets. With the rise of LLMs, Wadhwa et al. (2023) have demonstrated the remarkable capabilities of LLMs such as GPT-3 (Brown et al., 2020) and FLAN-T5 (Chung et al., 2022) in RE tasks using generative LLMs. Generative RE employing LLMs consistently outperforms traditional RE methods by a significant margin.

5.2 Evaluation Metrics for Text Generation

Multi-aspect Metrics : UniEval (Zhong et al., 2022), GPTScore (Fu et al., 2023), G-Eval (Liu et al., 2023), other similar works (Gao et al., 2023; Li et al., 2023a) ...

Single-aspect Metrics : Factualness: FActScore (Min et al., 2023) ...

6 Conclusions

We presented GREScores, an automated multi-aspect evaluation metric for generative relation extraction (GRE).

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2370–2381. Association for Computational Linguistics.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4924–4935. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 407–413. Association for Computational Linguistics.

Fan Dai and Ranjan Maitra. 2020. Practical text analytics: Maximizing the value of text data. *Technometrics*, 62(2):1–286.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Hady ElSahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frédérique Laforest. 2017. Unsupervised open relation extraction. In *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 10577 of *Lecture Notes in Computer Science*, pages 12–16. Springer.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. 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): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018a. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2236–2245. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.

Shengbin Jia, Shijia E, and Yang Xiang. 2019. Supervised neural models revitalize the open relation extraction. *CoRR*, abs/1908.01761.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Pengcheng Jiang. 2023. Txbkg: Transforming any pdfs into knowledge graphs.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023b. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ruosen Li, Teerth Patel, and Xinya Du. 2023a. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Xue Li, Fina Polat, and Paul Groth. 2023b. Do instruction-tuned large language models help with relation extraction?

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Trans. Assoc. Comput. Linguistics*, 4:231–244.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2023. Care: Extracting experimental findings from clinical literature. *arXiv preprint arXiv:2311.09736*.

Jian Ni, Gaetano Rossiello, Alfio Gliozzo, and Radu Florian. 2022. A generative model for relation extraction and classification. *CoRR*, abs/2202.13229.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's stop incorrect comparisons in end-to-end relation extraction! *arXiv preprint arXiv:2009.10684*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging

framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1476–1488. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. 2022. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2920–2929. Association for Computational Linguistics.

Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1461–1470. ACL.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019a. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019b. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2810–2819. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1753–1762. The Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Jie Zhou, Shenpo Dong, Yunxin Huang, Meihan Wu, Haili Li, Jingnan Wang, Hongkui Tu, and Xiaodong Wang. 2023. U-CORE: A Unified Deep Cluster-wise Contrastive Framework for Open Relation Extraction. *Transactions of the Association for Computational Linguistics*, 11:1301–1315.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

A Templates Used for Prompting

B Implementation Details

C Post-processing

This is an appendix.

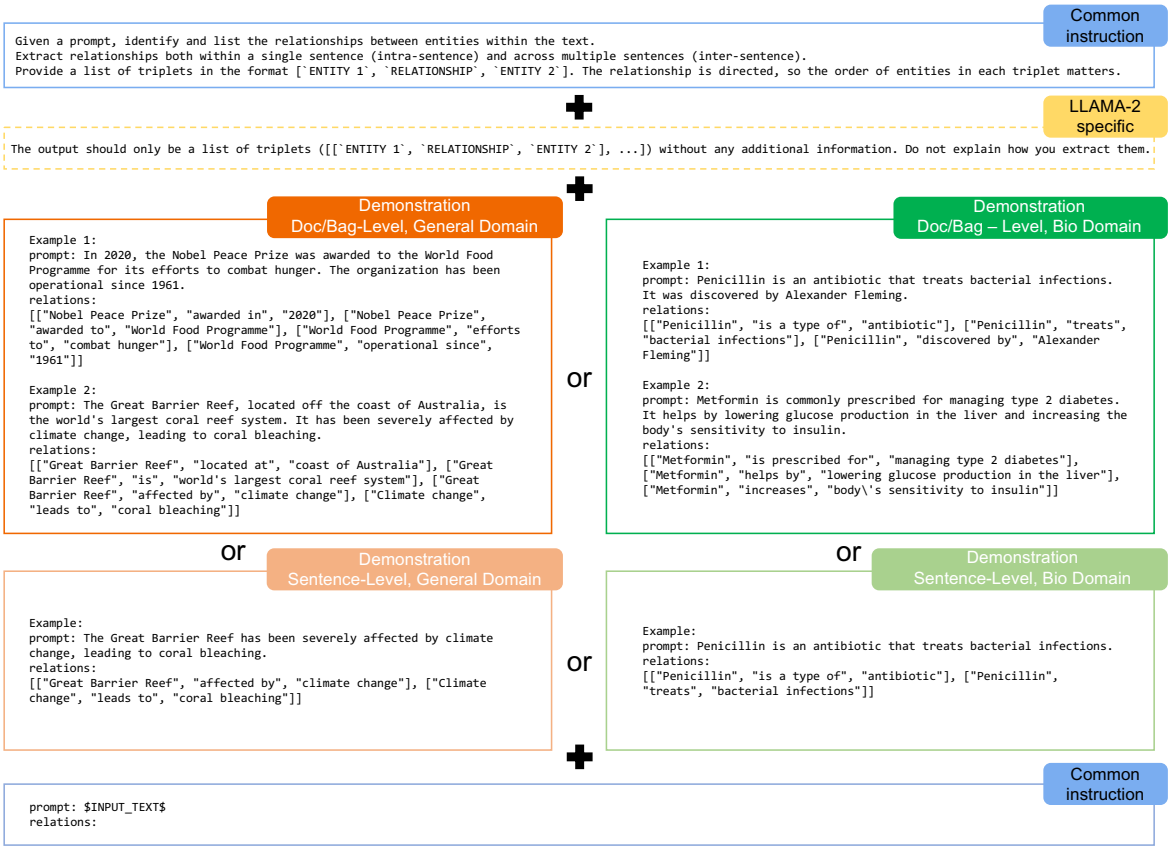

Figure 2: **Templates used for generative relation extraction.**