GenRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models

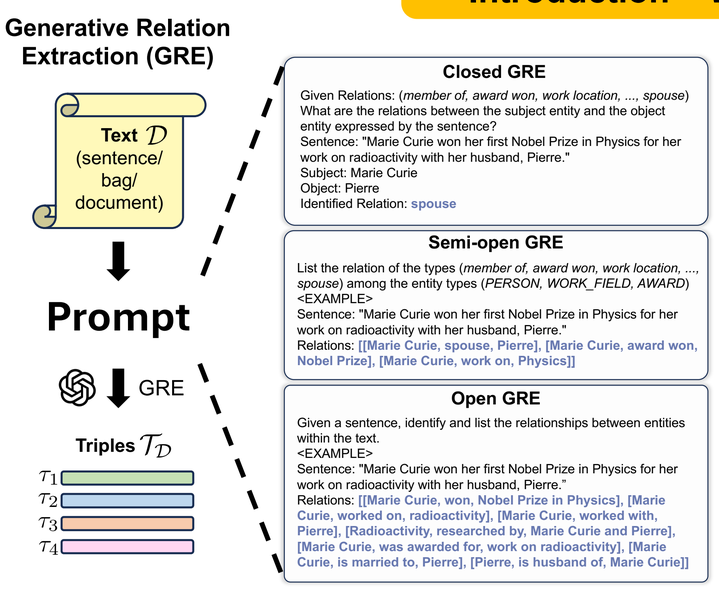
Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, Jiawei Han
University of Illinois Urbana-Champaign

PAPER

CODE

NAACL'24

Introduction – Why Current Evaluation Metrics for Generative Relation Extraction is Inadequate?



Three types of Generative Relation Extraction

I. Text	"Peter Munk , founder and chairman of Barrick Gold in Toronto , has warned that an exodus of head offices to other countries will cause , among other things , lower levels of charitable donations and fewer opportunities for skilled workers ."
II. Ground Truth	[Peter Munk, place lived, Toronto], [Barrick Gold, advisors, Peter Munk], [Barrick Gold, location, Toronto], [Barrick Gold, company, Peter Munk], [Barrick Gold, founders, Peter Munk], [Peter Munk, company, Barrick Gold], [Barrick Gold, place lived, Toronto]
III. Predefined Relation Types:	(*administrative_divisions, advisors, capital, children, company, contains, country, county_seat, ethnicity, featured_film_locations, founders, geographic_distribution, location, locations, majorshareholders, nationality, neighborhood_of, place_founded, place_lived, place_of_birth, place_of_burial, place_of_death, religion*)
IV. Predefined Entity Types:	(*business, company, country, deceasedperson, ethnicity, event, film, location, neighborhood, people, person, region, time, us_county*)

Closed GRE

Input: I, III, and entity pairs in II.

Output:
[Peter Munk, place founded, Toronto]
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Peter Munk, founder of, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)

Evaluation:
Tranditional: *P*: 71.4, *R*: 28.6, *F1*: 40.8
GREScores:
TS: 3.6, *US*: 66.7, *FS*: 85.7, *GS*: 100, *CS*: 57.1

Semi-open GRE

Input: I, III, and IV.

Output:
[Peter Munk, advisors, Barrick Gold] (*CS*)
[Peter Munk, founders, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)
[Peter Munk, warning, exodus]
[head offices, location, other countries],
[exodus, cause, lower levels of charitable donations and fewer opportunities for skilled workers] (*FS*, *GS*)

Evaluation:
Tranditional: *P*: 16.7, *R*: 14.2, *F1*: 15.4
GREScores:
TS: 22.1, *US*: 100.0, *FS*: 50.0, *GS*: 85.6, *CS*: 71.4

Open GRE

Input: I (text only).

Output:
[Peter Munk, founder of, Barrick Gold] (*FS*, *CS*)
[Peter Munk, chairman of, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, located in, Toronto] (*FS*, *CS*)
[Peter Munk, based in, Toronto] (*FS*)
[Peter Munk, warn, effects of exodus of head offices] (*FS*)
[exodus of head offices, will cause, lower levels of charitable donations] (*FS*)
[exodus of head offices, will cause, fewer opportunities for skilled workers] (*FS*)

Evaluation:
Tranditional: *P*: 0, *R*: 0, *F1*: 0
GREScores:
TS: 44.9, *US*: 80.0, *FS*: 100.0, *GS*: 100.0, *CS*: 57.1

The outcomes of using Precision/Recall/F1 to evaluate closed, semi-open, and open GRE methods. The most informative/human-preferred results are scored extremely low.

Evaluate Semi-Open and Open GRE with Hard-Matching Precision/Recall/F1

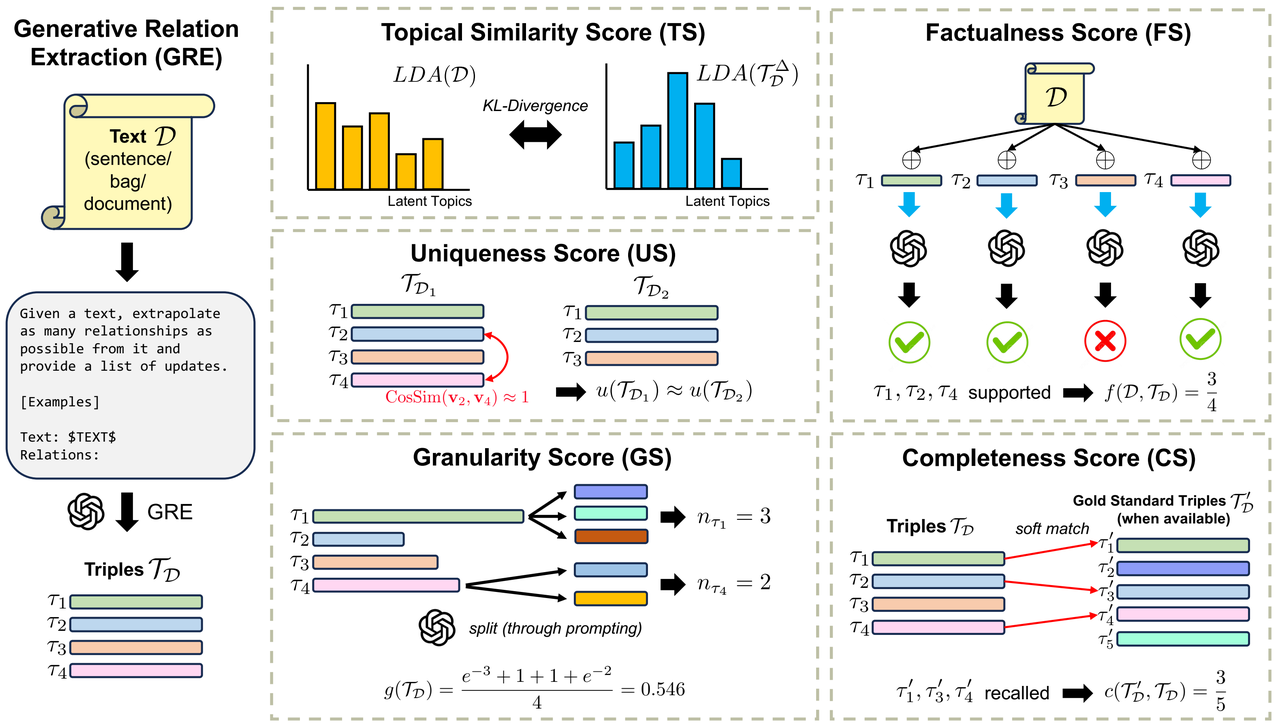
	CDR				NYT10m			
	C	**S**	**O**	**GT**	**C**	**S**	**O**	**GT**
#tri	10.1	6.8	16.1	10.1	1.4	2.9	5.8	1.4
#tok	6.6	4.0	8.3	5.8	4.6	2.0	7.0	4.5
P	58.8	1.1	0.4	-	29.3	5.2	0.0	-
R	58.7	0.8	0.7	-	26.6	12.7	0.0	-
F1	58.8	0.7	0.5	-	27.5	6.5	0.0	-

(P: Precision, R: Recall)

Traditional metrics are insufficient to evaluate the performance of (semi-)open **Generative Relation Extraction (GRE)**

There is a need for an automated multi-dimensional evaluation framework for GRE.

GenRES Framework for Multi-Dimensional Evaluation of Generative Relation Extraction



Topical Similarity Score (TS) – How much content of the source text is covered by the relationships extracted (by comparing triples* to the source text)"

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

Uniqueness Score (US) – "How many unique relationships are extracted (by comparing similarity within the extracted triples)"

$$u(\mathcal{T}_\mathcal{D}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (\text{CosSim}(\mathbf{v}_i, \mathbf{v}_j) < \phi)$$

Factualness Score (FS) – "How factual the extracted triples are, referring to the source text"

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

Granularity Score (GS) – "How atomic the extracted triples are (by asking LLM to split each triple)"

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

Completeness Score (GS) – "How many ground truth relations are predicted (by computing soft matching recall)"

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

Benchmarking LLMs' Capabilities of Generative Relation Extraction Using GenRES

Sentence-Level Performance

		TACRED							Wiki80						
		#tri	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***	***#tri***	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***
	Ground Truth	1.4	4.6	15.8	92.7	87.0	94.9	100	1.0	5.8	5.9	100	90.1	84.4	100
LLaMA	Vicuna-7B	2.6	8.7	40.4	85.0	**75.6**	58.9	36.2	2.4	7.9	41.3	76.8	81.0	61.7	36.6
	Vicuna-33B	4.3	7.3	**44.3**	75.5	71.0	69.2	47.2	3.8	7.2	**47.3**	62.1	79.9	73.8	46.8
	LLaMA-2-7B	2.8	6.3	36.7	85.3	66.9	71.2	37.8	2.4	5.8	25.8	69.8	60.4	**76.9**	31.4
	LLaMA-2-70B	4.1	6.4	40.8	79.3	74.5	**76.8**	**56.4**	3.7	6.6	41.5	64.8	**82.4**	**76.9**	**49.4**
	WizardLM-70B	2.1	2.9	23.3	**90.7**	28.0	72.1	9.8	2.1	3.2	25.6	**84.9**	36.6	74.4	21.4
GPT	text-davinci-003	4.4	7.1	56.1	79.8	84.0	72.8	58.6	4.0	6.8	59.2	65.3	89.2	74.9	51.9
	GPT-3.5-Turbo-Inst.	5.0	7.0	58.6	80.5	81.6	72.6	58.6	4.4	6.9	60.2	69.3	88.7	**75.4**	**54.8**
	GPT-3.5-Turbo	3.9	6.8	52.7	81.1	76.4	67.5	39.7	3.4	6.3	50.9	**69.5**	75.6	68.9	36.0
	GPT-4	4.3	7.5	**59.1**	80.4	87.6	69.1	57.8	4.0	7.1	**65.4**	66.2	92.3	74.2	47.8
	GPT-4-Turbo	4.4	7.8	58.5	**82.6**	**88.6**	**73.2**	**63.4**	4.0	7.6	61.9	69.4	**92.8**	74.5	47.1
others	Mistral-7B-Inst.	4.7	7.1	43.9	78.6	71.0	65.5	41.2	3.6	7.8	44.6	67.8	83.9	67.7	38.5
	Zephyr-7B-Beta	5.4	7.6	36.4	78.6	65.8	72.0	44.9	4.5	7.8	43.2	68.1	77.8	74.2	42.6
	Galactica-30B	8.5	8.9	33.4	43.9	57.5	64.1	30.9	5.6	7.2	35.0	47.9	63.1	73.3	38.4
	OpenChat-3.5	4.3	7.1	**50.7**	**80.8**	**80.4**	**72.1**	**60.0**	4.0	7.0	**53.8**	**69.7**	**88.7**	**74.9**	**50.6**

Bag-Level Performance

		NYT10m							Wiki20m						
		#tri	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***	***#tri***	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***
	Ground truth	1.4	4.5	8.7	69.3	84.1	93.1	100	2.0	6.3	4.4	21.2	88.7	85.1	100
LLaMA	Vicuna-7B	3.1	7.8	42.0	86.4	80.0	60.2	38.9	3.0	7.5	48.3	67.8	50.0	68.6	37.3
	Vicuna-33B	4.7	7.2	**47.8**	80.1	75.1	65.2	46.5	4.1	7.0	**49.8**	56.4	84.4	75.4	46.1
	LLaMA-2-7B	3.1	6.0	35.4	82.2	78.9	69.2	38.4	3.1	6.3	37.9	**73.8**	73.4	75.6	36.0
	LLaMA-2-70B	5.0	6.9	45.4	83.0	**81.7**	**71.8**	**52.4**	4.1	6.9	45.2	62.0	**87.1**	**78.4**	**50.2**
	WizardLM-70B	4.4	4.2	30.5	**88.9**	43.9	68.9	27.6	3.6	5.6	43.1	67.8	67.3	75.0	40.9
GPT	text-davinci-003	4.9	7.1	50.6	81.4	85.8	69.3	52.6	3.7	8.2	51.8	56.9	91.3	73.3	43.5
	GPT-3.5-Turbo-Inst.	5.8	7.0	54.2	83.0	84.0	**71.9**	53.4	4.8	7.7	54.0	60.3	90.1	78.9	43.8
	GPT-3.5-Turbo	4.1	6.2	43.3	82.3	68.2	62.8	29.8	3.6	7.7	48.2	61.8	80.2	72.7	32.5
	GPT-4	5.1	7.4	56.2	81.8	89.0	68.2	52.6	3.8	8.1	**59.0**	56.2	**93.2**	77.2	40.0
	GPT-4-Turbo	5.3	7.8	**58.1**	**84.2**	**89.6**	69.1	**53.7**	4.2	7.6	56.4	**62.0**	92.4	**81.2**	**52.7**
others	Mistral-7B-Inst.	5.7	7.4	40.6	77.6	75.4	62.9	36.5	4.0	6.9	43.3	57.0	83.6	69.9	40.1
	Zephyr-7B-Beta	7.8	7.2	36.5	80.8	64.9	**73.8**	47.0	5.2	6.8	40.3	**65.5**	75.5	**79.0**	45.9
	Galactica-30B	8.3	8.7	29.7	48.4	52.4	60.6	37.0	6.0	8.4	35.3	49.4	65.2	66.8	38.6
	OpenChat-3.5	5.2	7.2	**54.0**	**84.7**	**84.3**	69.7	**55.3**	4.3	7.0	**57.5**	61.8	**90.5**	76.0	**47.7**

Document-Level Performance

		CDR							DocRED						
		#tri	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***	***#tri***	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***
	Ground Truth	10.1	5.8	9.6	33.4	93.5	98.1	100	12.4	6.0	8.4	64.0	94.4	81.9	100
LLaMA	Vicuna-7B	6.8	8.4	57.8	86.9	84.7	44.6	30.7	7.4	9.9	23.1	81.9	93.4	46.8	28.3
	Vicuna-33B	6.4	10.5	73.0	89.2	**97.3**	38.4	32.0	10.8	9.8	34.7	82.8	97.2	49.6	36.9
	LLaMA-2-7B	5.6	6.7	48.6	92.0	62.0	44.9	25.7	2.7	3.2	12.8	93.3	34.0	60.6	12.1
	LLaMA-2-70B	10.8	8.1	**74.8**	87.6	96.6	**57.8**	**51.0**	13.8	8.7	**39.2**	82.6	**97.3**	**60.9**	**39.2**
	WizardLM-70B	10.2	7.8	65.4	**94.1**	76.4	46.2	32.6	5.8	3.6	24.3	**94.9**	37.9	56.7	12.8
GPT	text-davinci-003	12.7	8.3	76.7	87.2	96.8	**55.4**	44.3	15.3	8.5	40.1	84.2	97.6	59.8	46.2
	GPT-3.5-Turbo-Inst.	16.1	8.3	77.6	89.6	96.8	54.2	47.8	17.8	8.9	47.8	85.6	98.1	56.2	44.7
	GPT-3.5-Turbo	11.2	11.4	81.7	89.2	**98.2**	40.3	30.2	15.0	9.9	**50.4**	84.0	98.5	49.1	36.5
	GPT-4	14.3	9.3	81.7	91.0	97.9	49.1	46.3	17.8	8.7	48.6	82.8	**98.6**	59.6	47.3
	GPT-4-Turbo	18.6	8.5	**82.1**	**91.9**	96.8	53.1	**48.8**	21.5	8.7	50.0	**87.4**	97.6	**63.1**	**49.3**
others	Mistral-7B-Inst.	14.2	9.1	69.0	74.9	93.5	51.1	**40.0**	11.3	9.6	30.2	76.4	94.1	55.2	27.5
	Zephyr-7B-Beta	25.9	8.8	49.1	79.5	70.1	**57.7**	29.3	18.6	8.6	27.9	79.4	94.7	**64.7**	37.1
	Galactica-30B	0.2	0.3	4.1	1.1	0.9	44.4	0.0	0.0	0.0	8.6	0.0	0.0	0.0	0.0
	OpenChat-3.5	8.6	12.6	**78.7**	**91.9**	**97.4**	38.2	31.8	15.4	8.9	**39.7**	**82.1**	**98.1**	61.7	**43.4**

Observations

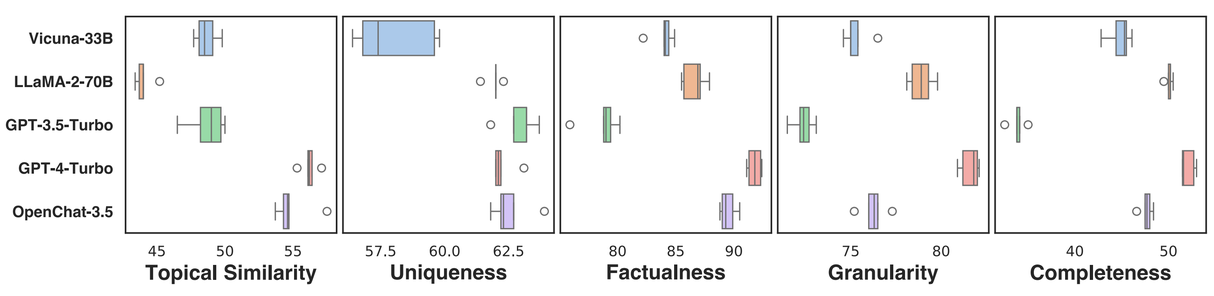
(1) LLaMA-2-70Bb, GPT-4-Turbo, and OpenChat-3.5 notably lead in performance. Small LLM OpenChat-3.5 (7B) achieves comparable or even better performance than large LLMs.

(2) High Completeness Score (CS) can indicate high Factualness Score (FS). This means human annotations are still valuable to evaluate GRE with our soft matching recall.

(3) A greater number of tokens per triple does not inherently result in a lower Granularity Score (GS). This suggests that the GS metric can encourage models to identify more atomic relationships rather than merely focusing on brevity.

(4) GPT-4-Turbo outperforms human labels on factualness.

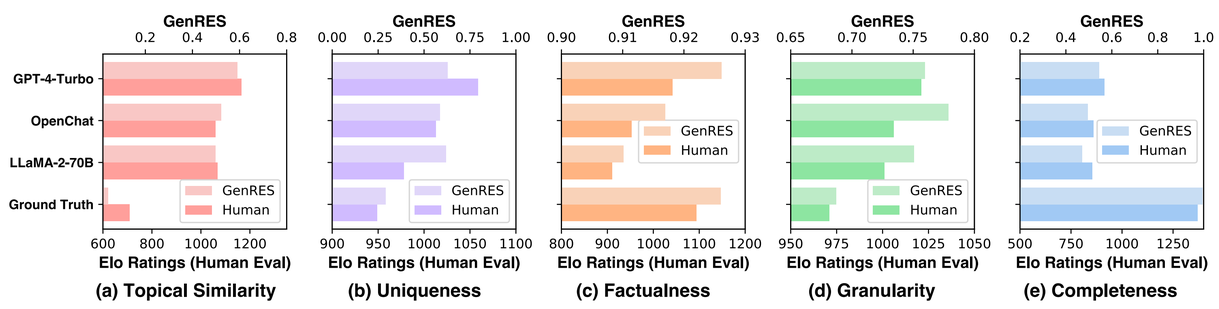
Robustness of GenRES on Evaluating Generative Relation Extraction



GRE performance of five LLMs on Wiki20m, each with 5 runs with random seeds.

This indicates the robustness of GenRES across different LLMs and metrics.

Alignment with Human Evaluation



Human Preference Evaluation vs GenRES Evaluation on 100 Wiki20m samples.

In most cases, GenRES aligns human evaluation well on generative relation extraction.

Thanks for your interest! Email pj20@illinois.edu if you have further questions!