EMNLP 2024

Learning to Extract Structured Entities Using Language Models

Haolun Wu*, Ye Yuan*, Liana Mikaelyan,
Alexander Meulemans†, Xue Liu, James Hensman, Bhaskar Mitra

McGill University Mila Microsoft ETH zürich

*Equal contribution with random order; †Work done during an internship at Microsoft Research.

Background: Information Extraction

- Information extraction refers to a broad family of tasks that aim to extract structured information from unstructured text.

- Previous formulations of information extraction have predominantly centered around the extraction of ***<subject, relation, object>*** triplets.

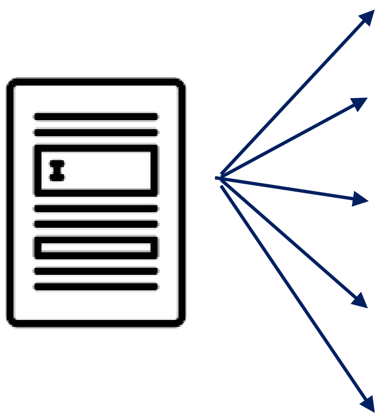
"Bill Gates is an American businessman. Gates is famous for co-founding the software giant Microsoft, a multinational technology corporation headquartered in Redmond."

<Bill Gates, is, American Businessman>
<Gates, co-founding, Microsoft>
…
<tech corporation, headquartered in, Redmon>

- The evaluations leverage conventional metrics, such as **precision**, **recall**, and **F1** scores.

Limitation & Motivation

Text paragraph

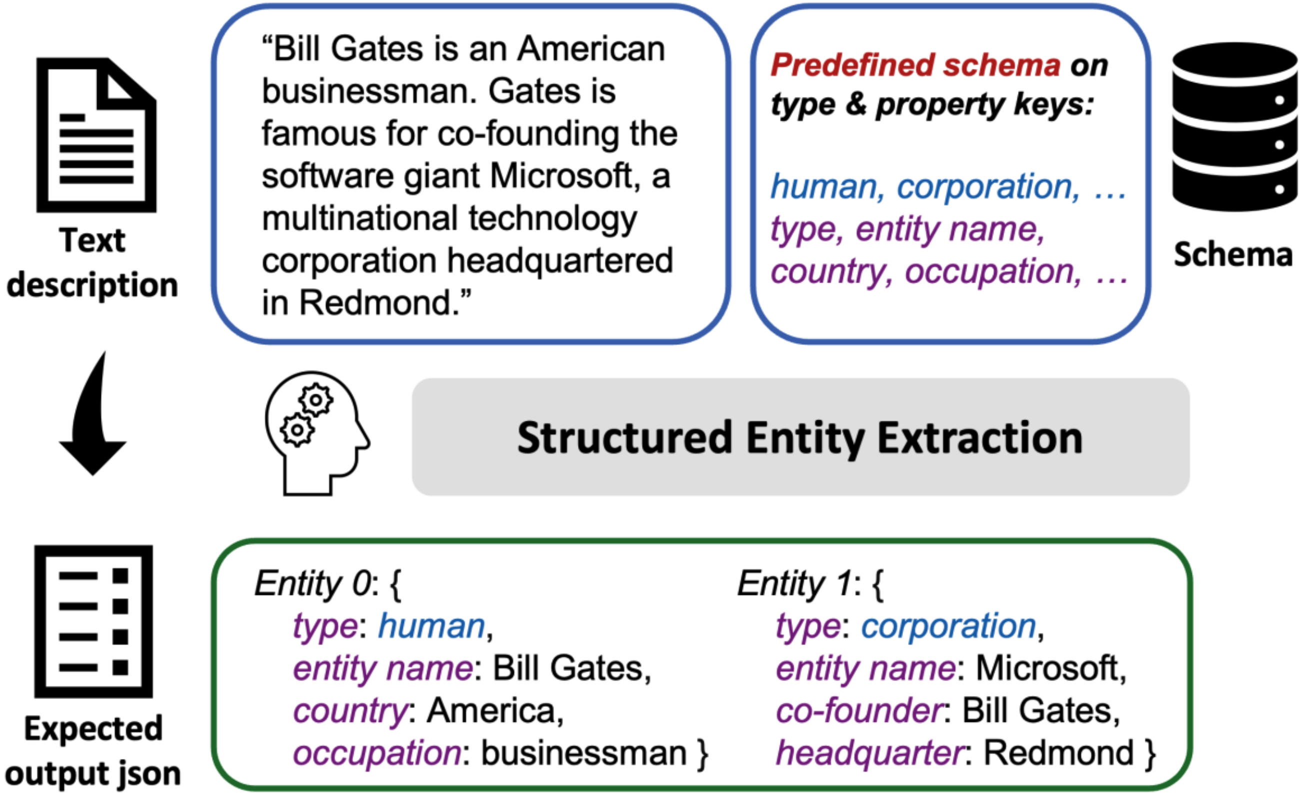


Entities		System A	System B
	Associated with 5 triplets as the subject	5	1
	Associated with 1 triplet as the subject	0	1
	Associated with 1 triplet as the subject	0	1
	Associated with 1 triplet as the subject	0	1
	Associated with 1 triplet as the subject	0	1

The precision and recall are the same, but the text comprehension is very different.

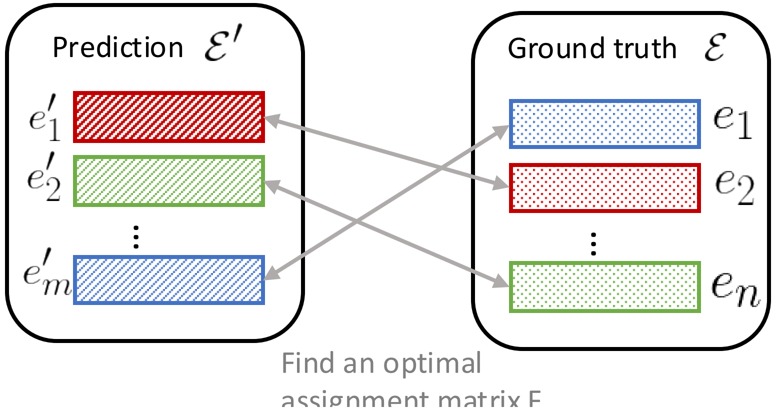
Structured Entity Extraction

- To facilitate diverse evaluations and provide more holistic perspectives, we propose Structured Entity Extraction (SEE), an **entity-centric** formulation of information extraction.



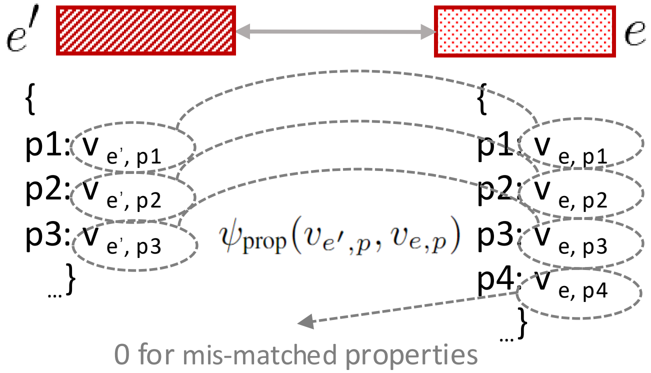
Approximate Entity Set OverlaP (AESOP) Metric

Phase 1: Optimal Entity Assignment



$$\mathbf{F} = \arg\max_{\mathbf{F}} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{F}_{i,j} \cdot \mathbf{S}_{i,j}$$

Phase 2: Pairwise Entity Comparison



$$\psi_{\text{ent}}(e', e) = \bigotimes_{p \in \mathcal{P}} \psi_{\text{prop}}(v_{e',p}, v_{e,p})$$

Overall: AESOP Metric

Focuses on the entity-level and more flexible to include different level of normalization:

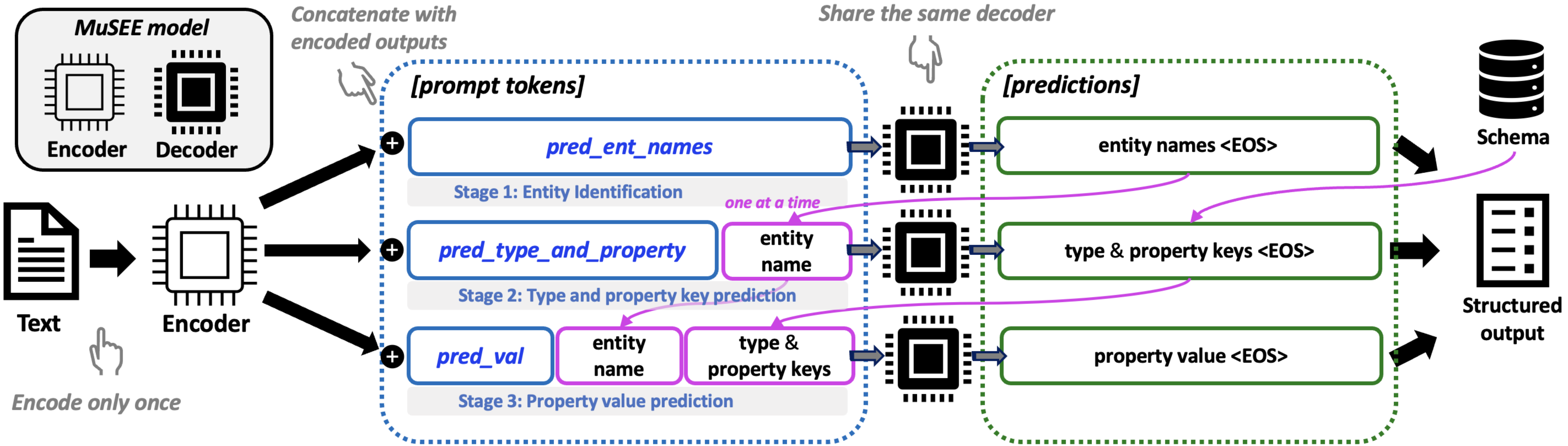
$$\Psi(\mathcal{E}', \mathcal{E}) = \frac{1}{\mu} \bigoplus_{i,j}^{m,n} \mathbf{F}_{i,j} \cdot \psi_{\text{ent}}(\vec{E'}_i, \vec{E}_j)$$

Experiments

- **Datasets:**
  - Repurpose the **NYT**, **CoNLL04**, and **REBEL** datasets.
  - Introduce **Wikidata-based**, crafted using an approach similar to REBEL but with two primary distinctions: *(i) property values are not necessarily entities*; *(ii) simplify the entity types by consolidating them into broader categories based on the Wikidata taxonomy graph*.

- **Baselines:**
  - **LM-JSON**: fine-tune a pre-trained language model to generate JSON directly.
  - Repurpose **GEN2OIE**, **IMoJIE**, and **GenIE** to structured entity extraction task.
  - All methods utilize **T5**-Base and T5-Large as the backbone models.

Multi-stage Structured Entity Extraction Model (MuSEE)



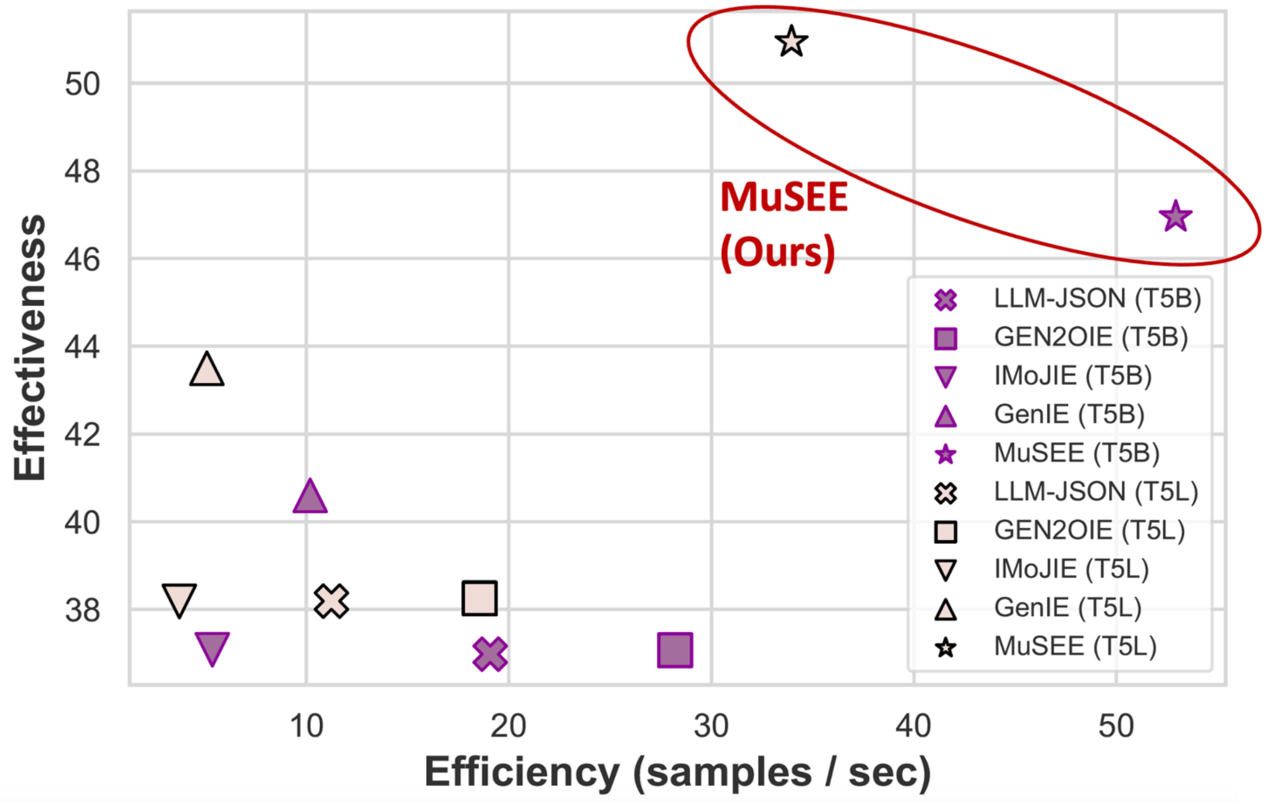
Experiments – Effectiveness Comparison

Model	REBEL				NYT				CoNLL04				Wikidata-based			
	AESOP	Precision	Recall	F1	AESOP	Precision	Recall	F1	AESOP	Precision	Recall	F1	AESOP	Precision	Recall	F1
LM-JSON (T5-B)	41.91	38.33	**51.29**	43.87	66.33	73.10	52.66	61.22	68.80	61.63	48.04	53.99	36.98	43.95	29.82	35.53
GEN2OIE (T5-B)	44.52	35.23	40.28	37.56	67.04	72.08	53.02	61.14	68.39	62.35	42.20	50.26	37.07	40.87	28.37	33.55
IMoJIE (T5-B)	46.11	34.10	48.61	40.08	63.86	72.28	48.99	58.40	63.68	52.00	42.62	46.85	37.08	41.61	28.23	33.64
GenIE (T5-B)	48.82*	**57.55**	38.70	46.28*	79.41*	87.68	**73.24**	**79.81**	74.74*	72.49*	59.39	65.29	40.60*	50.27*	**29.75**	37.38
MuSEE (T5-B)	**55.24**	56.93	42.31	**48.54**	**81.33**	**88.29**	72.21	79.44	**78.38**	**73.18**	**60.28**	**66.01**	**46.95**	**53.27**	29.33	**37.99**
LM-JSON (T5-L)	45.92	39.49	40.82	40.14	67.73	73.38	53.22	61.69	68.88	61.50	47.77	53.77	38.19	43.24	31.63	36.54
GEN2OIE (T5-L)	46.70	37.28	41.12	39.09	68.27	73.97	53.32	61.88	68.52	62.76	43.31	51.16	38.25	41.23	28.54	33.77
IMoJIE (T5-L)	48.13	38.55	**49.73**	43.43	65.72	73.46	50.03	59.52	67.31	53.00	43.44	47.75	38.18	41.74	30.10	34.98
GenIE (T5-L)	50.06*	**58.00**	42.56	**49.09**	79.64*	84.82*	**75.69**	80.00	72.92*	**77.75**	55.64*	64.86	43.50*	**54.05**	30.98	**39.38**
MuSEE (T5-L)	**57.39**	57.11	42.89	48.96	**82.67**	**89.43**	73.32	**80.60**	**79.87**	74.89	**60.72**	**67.08**	**50.94**	53.72	**31.12**	39.24

- Our MuSEE model consistently outperforms other baselines in terms of AESOP metric across all datasets.
- It does not invariably surpass the baselines across all the traditional metrics of precision, recall, and F1 score.

Experiments – Efficiency Comparison

- We choose the Wikidata-based dataset to illustrate.
- The effectiveness is measured by AESOP metric.
- MuSEE is generally more effective and efficient than baselines.



Experiments

- We randomly select 400 test passages from the Wikidata-based dataset, and generate outputs of our model MuSEE and the strongest baseline GenIE.
- Human evaluators are presented with a passage and two randomly flipped extracted sets of entities with properties.
- Evaluators are then prompted to choose the output they prefer or express no preference based on three criteria, ***Completeness***, *Correctness*, and ***(no) Hallucinations***.

	Human Evaluation			Quantitative Metrics			
	Complete.	Correct.	Halluc.	AESOP	Precision	Recall	F1
MuSEE prefer	61.75	59.32	57.13	61.28	45.33	37.24	40.57

Conclusion & Future Work

- **Contribution:**
  - An **entity-centric** formulation of the information extraction task.
  - An evaluation metric, **Approximate Entity Set OverlaP (AESOP)**, with more flexibility tailored for assessing structured entity extraction.
  - A new model leveraging the capabilities of LMs, improving the effectiveness and efficiency for structured entity extraction.

- **Future Work:**
  - Consider scenarios where **a property’s value might consist of a set**, such as varying “names”. Adapting our method to accommodate these scenarios presents a promising research direction.

Thanks for your attention!

Paper

Code