LasUIE: Unifying Information Extraction with Latent Adaptive Structure-aware Generative Language Model

Hao Fei[1], Shengqiong Wu[1], Jingye Li[2], Bobo Li[2], Fei Li[2], Libo Qin[1], Meishan Zhang[3], Min Zhang[3], Tat-Seng Chua[1]

1. Sea-NExT Joint Lab, National University of Singapore

2. Wuhan University 3. Harbin Institute of Technology (Shenzhen)

TL;DR

We propose a latent adaptive structure-aware generative language model for universal information extraction.

▶ Introduction

Universally modeling all typical information extraction tasks (UIE) with one generative language model (GLM) has revealed great potential by the latest study, where various IE predictions are unified into a linearized hierarchical expression under a GLM. Syntactic structure information, a type of effective feature which has been extensively utilized in IE community, should also be beneficial to UIE. In this work, we propose a novel structure-aware GLM, fully unleashing the power of syntactic knowledge for UIE. A heterogeneous structure inductor is explored to unsupervisedly induce rich heterogeneous structural representations by post-training an existing GLM. In particular, a structural broadcaster is devised to compact various latent trees into explicit high-order forests, helping to guide a better generation during decoding. We finally introduce a task-oriented structure fine-tuning mechanism, further adjusting the learned structures to most coincide with the end-task's need. Over 12 IE benchmarks across 7 tasks our system shows significant improvements over the baseline UIE system. Further in-depth analyses show that our GLM learns rich task-adaptive structural bias that greatly resolves the UIE crux, the long-range dependence issue and boundary identifying.

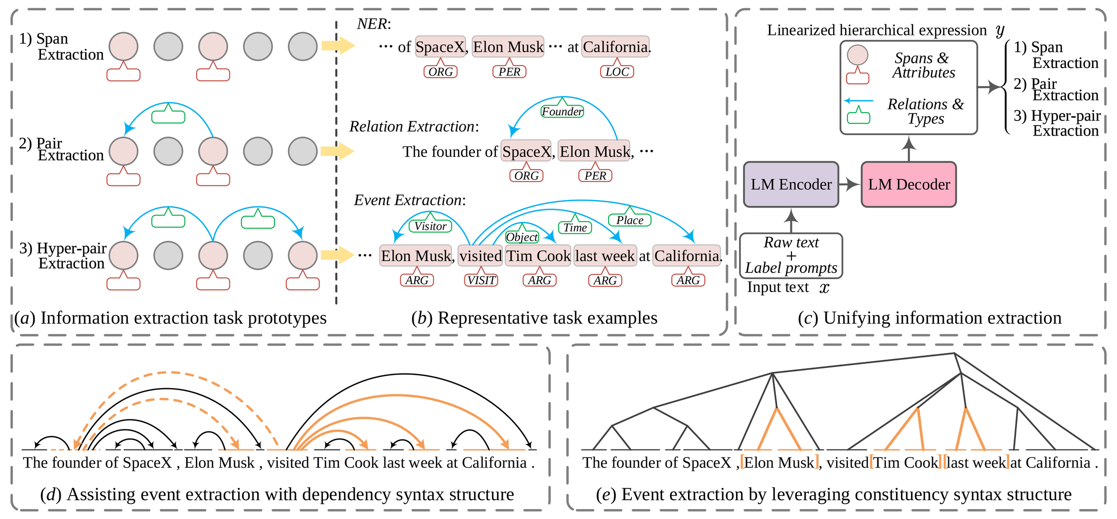

Figure 1: We reduce all the IE tasks into three prototypes (a) with representative examples (b). We unify all IEs with an encoder-decoder GLM (c). Both syntactic dependency (d) and constituency structure (e) plays a key but distinct role in IE, where the former helps solve long-range dependence problem and the latter benefits boundary detection issue.

Key points:

1. We propose a latent adaptive structure-aware generative language model for UIE (namely LasUIE).
2. We reduce UIE into three uniform prototypes, upon which we transform the UIE into generative paradigm with an encoder-decoder GLM, predicting the linearized hierarchical expression, i.e., spans&attributes, relations&types, as shown in Fig. 1(c)).
3. We adopt a three-stage of LM training procedure, where an additional structure-aware post-training is added between the pre-training and fine-tuning stages for structure learning.
4. We design a heterogeneous structure inductor (HSI) module, where two heterogeneous syntactic structures are simultaneously measured and automatically learned. With HSI, our GLM during post-training performs unsupervised syntax induction based on unlabeled texts without relying on external syntax parses or any annotation labor.
5. We further enhance the utility of syntax by introducing a structural broadcaster (SB) module. SB compacts multiple varying latent trees from different encoding attention heads into an explicit constituency-like and a dependency-like forest respectively. During each decoding step, two heterogeneous syntactic forests are utilized to produce high-order features at global level for guiding better content generation.
6. Finally, during the prompt-based fine-tuning stage we perform task-oriented structure adaptive tuning to narrow the gaps between the induced syntactic and task-specific structures. With policy gradient we dynamically adjust the attributes of two heterogeneous structures according to the feedback of end task performance.

▶ Unsupervised Structure-aware Post-training

The overall framework is built upon a Transformer-based encoder-decoder GLM, based on which we additionally add 1) a heterogeneous structure inductor module at top of the encoder for learning structure, 2) a structural broadcaster module between GLM encoder and decoder for enhancing the structural feature utility. Fig. 2 shows the overall framework of our proposed LasUIE.

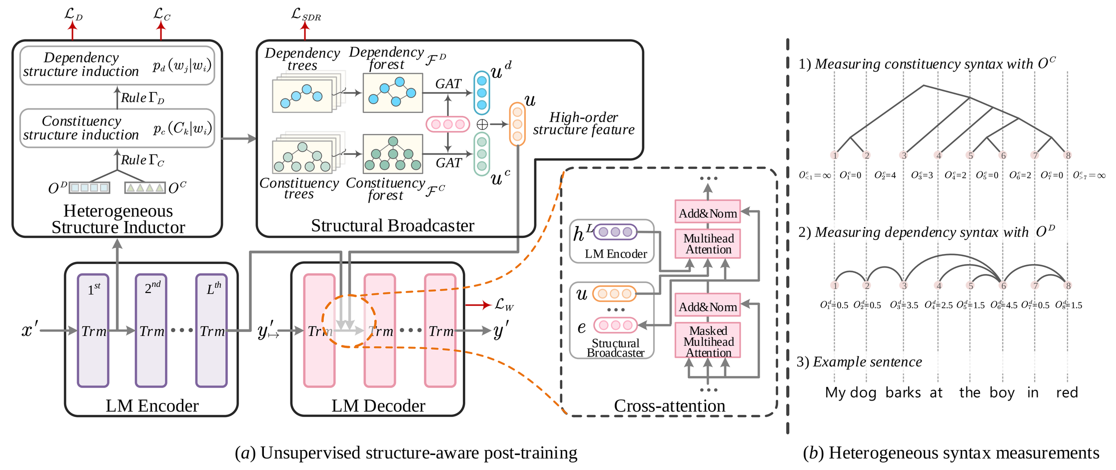

Figure 2: Overall LasUIE framework.

Heterogeneous structure inductor module generates both constituency and dependency structures via two heterogeneous syntax measurements Fig. 2(b).

Structural broadcaster module compacts multiple varying latent trees from different encoding attention heads into an explicit constituency-like and a dependency-like forest respectively.

▶ Task-oriented Structure Fine-tuning

Finally, during the prompt-based fine-tuning stage we perform task-oriented structure adaptive tuning to narrow the gaps between the induced syntactic and task-specific structures. With policy gradient we dynamically adjust the attributes of two heterogeneous structures according to the feedback of end task performance.

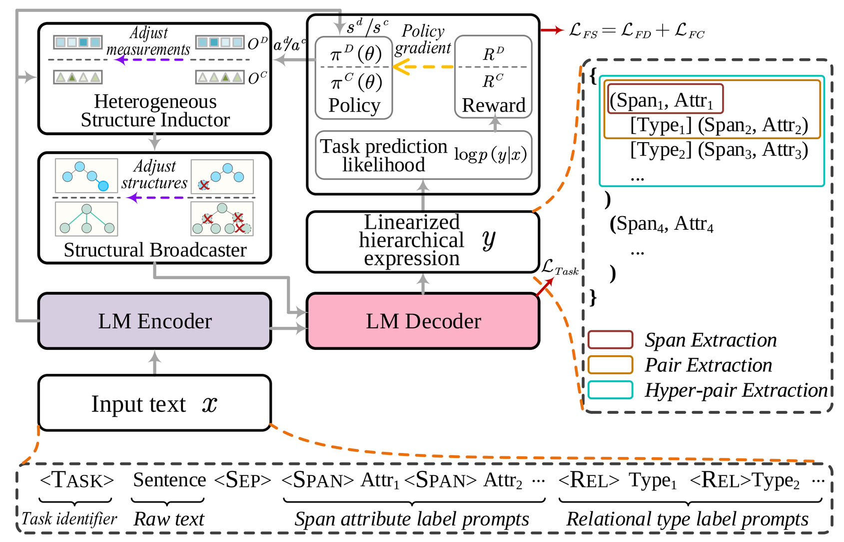

Figure 3: Fine-tuning our GLM with structure adaptive learning.

▶ Experiments

(1) Main Results

LasUIE consistently outperforms the baseline UIE and other SoTA models on all tasks in both two learning scenarios under both the Large or Base T5 initiations.

Task&Data	Span Extraction				Pair Extraction					Hyper-pair Extraction			Avg.
	NER				RE			AOP	ASTE	ORL	SRL	EE	
	CoNLL03	OntoNote	ACE04	ACE05	CoNLL04	NYT	ACE05	Res14	Res14	MPQA	CoNLL12	ACE05	
• Separate IE													
M1 SoTA*	93.2	91.9	86.8	84.7	73.6	92.7	65.6	69.3	73.6	53.0	73.5	48.3	75.5
M2 GEN-T5	91.0	89.1	84.3	83.0	69.4	90.3	60.2	62.5	71.8	49.8	69.3	43.7	72.0
M3 +*DepSyn*	91.5	89.5	84.9	83.4	70.3	91.8	62.4	64.3	72.6	51.5	70.8	45.5	73.2
M4 +*ConSyn*	92.1	90.0	85.3	83.8	69.8	90.9	61.5	63.1	72.3	50.7	70.1	44.3	72.8
M5 +*Dep&ConSyn*	92.3	90.4	85.3	84.0	71.2	92.1	63.3	66.0	73.0	51.8	71.3	46.2	73.9
• Unified IE													
M6 UIE*†	93.0	/	**86.9**	85.8	75.0	/	66.0	/	74.5	/	/	/	/
M7 UIE*	92.1	/	86.5	85.5	73.1	93.5	64.7	/	/	/	/	/	/
M8 **LasUIE*** (Ours)	**93.2**	**93.0**	86.8	**86.0**	**75.3**	**94.2**	**66.4**	**73.6**	**75.2**	**57.8**	**76.3**	**51.7**	**77.4**
M9 UIE	91.4	89.7	85.0	83.5	70.5	91.0	61.6	65.8	72.8	50.8	70.2	44.6	73.1
M10 +*DepSyn*	91.8	90.0	85.3	83.7	71.2	92.0	62.9	67.6	73.5	52.0	71.5	46.4	74.0
M11 +*ConSyn*	92.0	90.5	85.6	84.0	70.8	91.3	62.1	66.1	73.1	51.3	71.0	45.2	73.6
M12 +*Dep&ConSyn*	92.3	90.7	85.8	84.5	71.7	92.4	63.4	68.2	73.7	53.6	72.6	47.0	74.6
M13 **LasUIE** (Ours)	**92.6**	**92.0**	**86.3**	**85.0**	**73.2**	**93.0**	**64.4**	**70.2**	**74.8**	**56.0**	**74.7**	**49.0**	**75.9**
M14 w/o SB	92.0	90.7	85.5	84.2	71.5	91.8	62.9	68.3	73.4	54.7	73.4	47.7	74.6
M15 w/o $\mathcal{L}_{SDR}$	92.2	91.6	86.2	84.8	72.8	92.4	64.1	70.0	74.4	55.5	74.0	48.6	75.6
M16 w/o $\mathcal{L}_{FS}$	92.4	91.4	85.9	84.7	71.8	92.0	63.6	69.1	73.6	54.2	73.0	47.1	74.9

Figure 4: Overall IE performances by different methods.

Task&Data	Span Extraction				Pair Extraction					Hyper-pair Extraction			Avg.
	NER				RE			AOP	ASTE	ORL	SRL	EE	
	CoNLL03	OntoNote	ACE04	ACE05	CoNLL04	NYT	ACE05	Res14	Res14	MPQA	CoNLL12	ACE05	
• 1-shot													
UIE†	**46.4**	/	/	/	22.1	/	/	/	/	/	/	/	/
GEN-T5+*Dep&ConSyn*	27.2	20.4	14.8	17.6	8.2	25.7	10.8	12.8	10.8	1.1	6.5	1.5	13.1
UIE+*Dep&ConSyn*	30.3	23.6	17.5	20.7	12.8	26.7	14.3	16.7	13.0	2.8	14.0	3.8	16.4
LasUIE	39.4	**47.6**	**38.5**	**44.7**	**25.7**	**45.0**	**26.7**	**30.0**	**38.4**	**18.9**	**32.8**	**23.7**	**34.3**
• 10-shot													
UIE†	73.9	/	/	/	52.4	/	/	/	/	/	/	/	/
GEN-T5+*Dep&ConSyn*	67.4	64.7	49.2	52.8	45.6	50.8	37.4	19.7	17.8	5.4	18.7	12.2	36.8
UIE+*Dep&ConSyn*	69.5	68.4	52.8	54.1	51.8	56.0	43.8	22.5	26.1	10.5	23.2	17.6	41.4
LasUIE	**74.0**	**78.3**	**60.3**	**65.3**	**55.0**	**67.7**	**46.1**	**42.4**	**48.8**	**25.4**	**45.8**	**27.1**	**53.0**
• 1% data													
UIE†	**82.8**	/	/	/	30.8	/	/	/	/	/	/	/	/
GEN-T5+*Dep&ConSyn*	79.5	72.4	58.3	61.7	17.8	35.8	15.4	15.3	15.3	3.3	10.7	3.4	32.4
UIE+*Dep&ConSyn*	80.6	73.2	60.4	63.8	23.5	40.4	22.7	20.6	18.5	5.3	17.6	10.2	36.4
LasUIE	82.1	**84.5**	**65.7**	**70.1**	**32.0**	**53.6**	**34.2**	**34.8**	**41.7**	**21.0**	**39.8**	**25.7**	**48.8**
• 10% data													
UIE†	89.6	/	/	/	59.2	/	/	/	/	/	/	/	/
GEN-T5+*Dep&ConSyn*	89.0	84.0	71.3	68.8	52.4	80.4	45.7	56.0	59.7	22.4	50.7	26.7	58.9
UIE+*Dep&ConSyn*	89.3	85.8	72.1	70.6	54.9	82.5	47.6	58.3	62.6	27.4	54.3	31.7	64.4
LasUIE	**91.6**	**89.3**	**83.6**	**81.7**	**60.8**	**86.0**	**50.5**	**63.0**	**66.7**	**36.0**	**58.4**	**38.4**	**67.2**

Figure 5: Performances on low-resource settings by IE models.

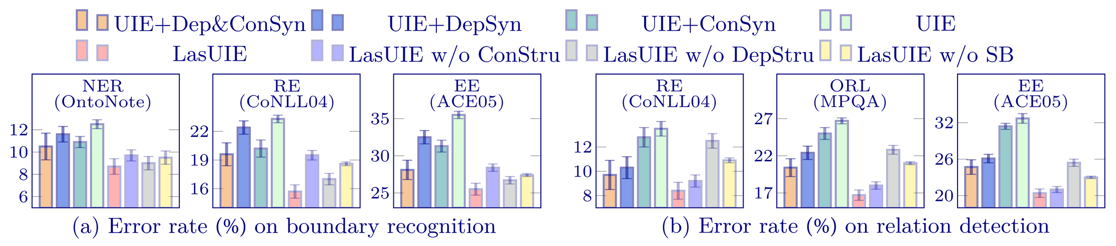

Figure 6: Error rates on boundary recognition and relation detection, respectively.

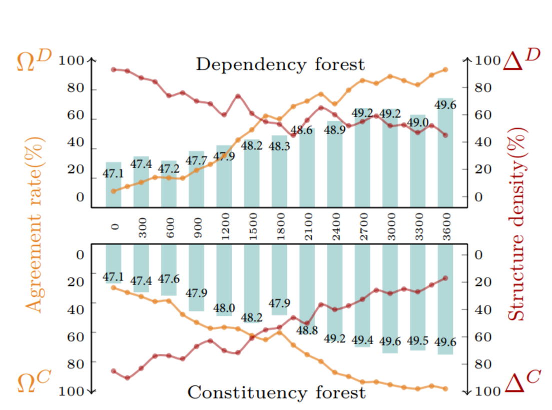

Figure 7: Trajectories of the changing structure agreement rates and densities during task-oriented structure fine-tuning, based on event extraction (ACE05). X-axis is the iteration steps for fine-tuning. Bars means the task performances (F1).

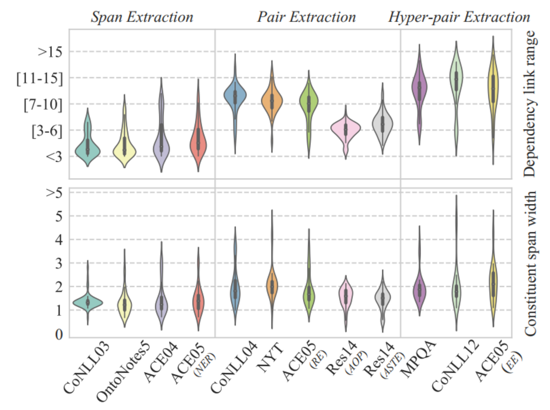

Figure 8: The distributions of the range of word-word dependency link (words) in forest $\mathcal{F}^D$ and the constituency phrasal span width (words) in forest $\mathcal{F}^C$ on each data.

(2) Analysis

★ Q1: Can fusing syntax structure knowledge into GLM contribute to UIE? **Answer**: Either in separate or unified IE setup, integrating additional linguistic syntax features into GLM improves IE performances.

★ Q2: What are the differences to integrate the constituency and dependency syntactic structure? *Answer*: On span extraction type IE (i.e., NER) the improvements from constituency syntax prevail, and the dependency type of structure features dominate the pair-wise tasks, i.e., (hyper-)pair extraction.

★ Q3: For UIE, is it more advanced for GLM to automatically learn latent structures than injecting external syntax parse trees? **Answer**: Yes, it is advanced for LMs to automatically learn latent structure information for better UIE.

★ Q4: Is it necessary to further fine-tune the structures in GLM for UIE? **Answer**: Yes, it is necessary to further fine-tune the structures in GLM for UIE.

▶ Conclusion

This work investigates developing a novel structure-aware generative language model (GLM) that learns rich heterogeneous syntactic structure representations for better unified information extraction (UIE). First, a well pre-trained GLM is taken as backbone to reach the goal of UIE, feeding with label prompt-based texts and predicting linearized hierarchical expressions that describe the actual IE target. During post-training, the proposed heterogeneous structure inductor automatically generates rich structure information without relying on any additional syntax annotation. A structural broadcaster then compacts various trees into forests for enhancing the structural feature utility and guiding better context generation. The learned structural knowledge is further fine-tuned on the in-house training data so as to adapt into the task-specific need. Extensive experiments and in-depth analyses demonstrate the efficacy of our system on improving the UIE.