Generative Knowledge Graph Construction: A Review

GENKGC
GENERATIVE KNOWLEDGE GRAPH CONSTRUCTION

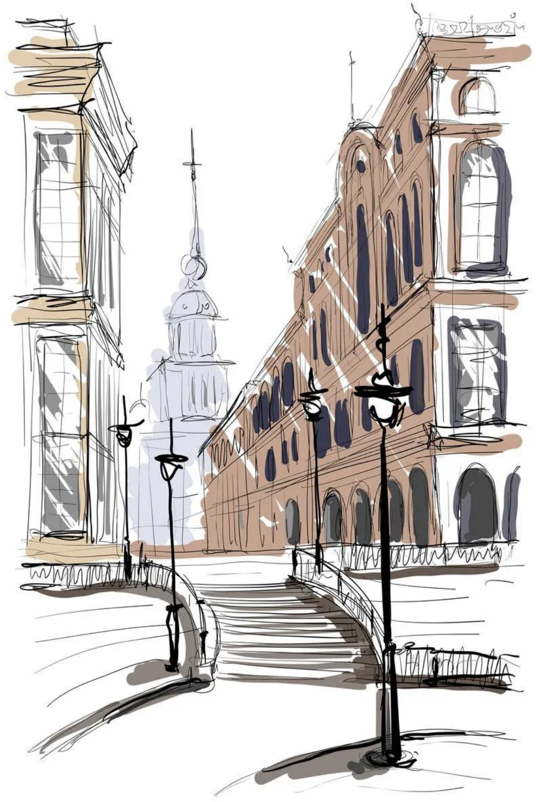
Hongbin Ye, Ningyu Zhang

Zhejiang University

Public Repository: https://github.com/zjunlp/Generative_KG_Construction_Papers

Outline

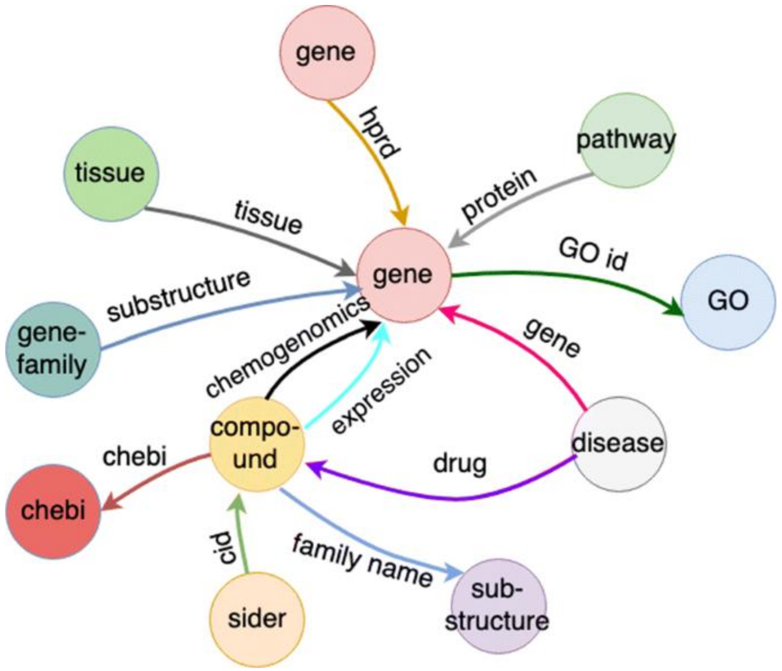
- Introduction
- Preliminary on Knowledge Graph Construction
- Discrimination and Generation Methodologies
- Taxonomy of Generative Knowledge
- Theoretical Insight
- Empirical Analysis
- Future Directions
- Conclusion



Neural network-based entity recognition, relationship extraction and other **information extraction** techniques need to strictly follow the predefined **Schema knowledge** specification

Text2Knowledge

Language Pretrain
BERT、GPT…..



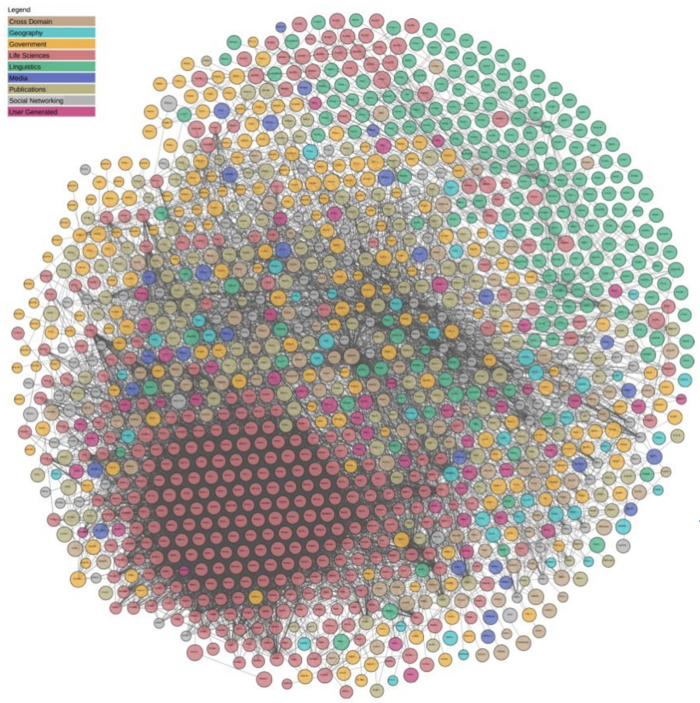
isSpouseOf

Jack is married to the microbiologist known as **Dr. Germ** in the USA.

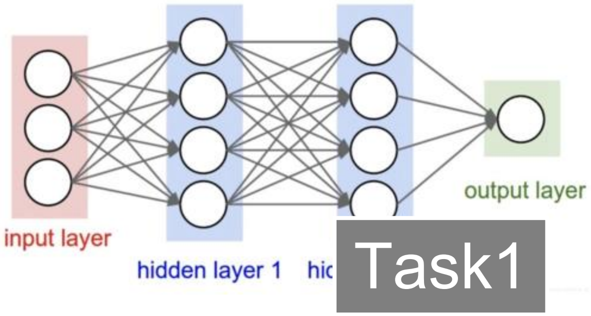
Entity Pair

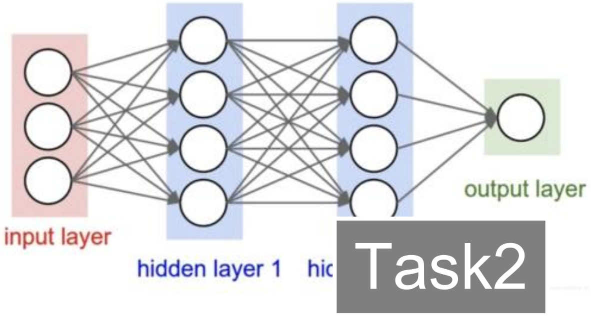
Introduction

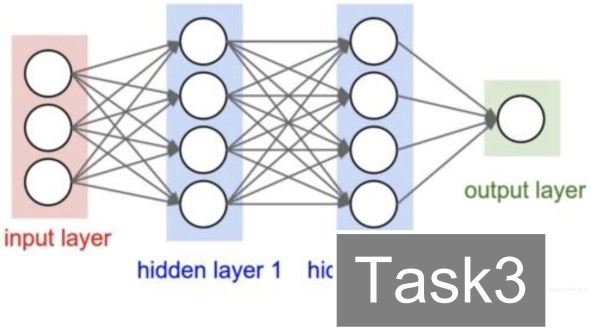
Poorly adapted to different tasks



Knowledge Graph Construction







Entity Discovery

Named Entity Recognition

Error propagation

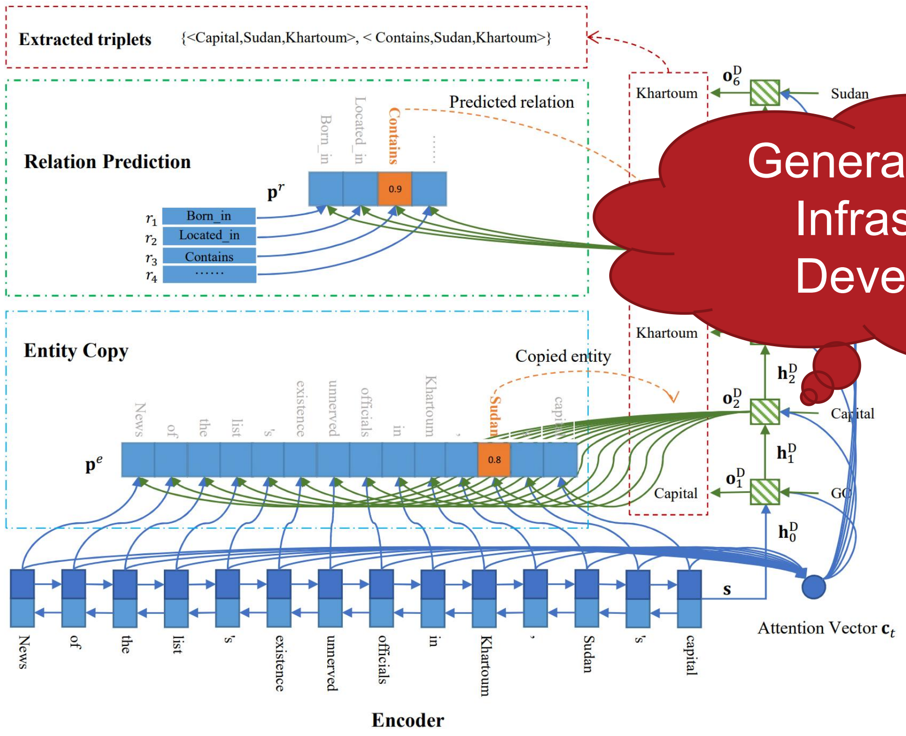
Entity Linking

Error propagation

Relation Extraction

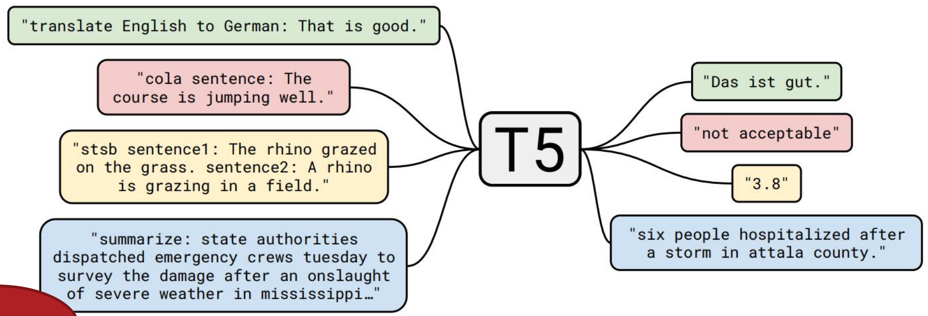
Event Extraction

浙江大学 ZHEJIANG UNIVERSITY

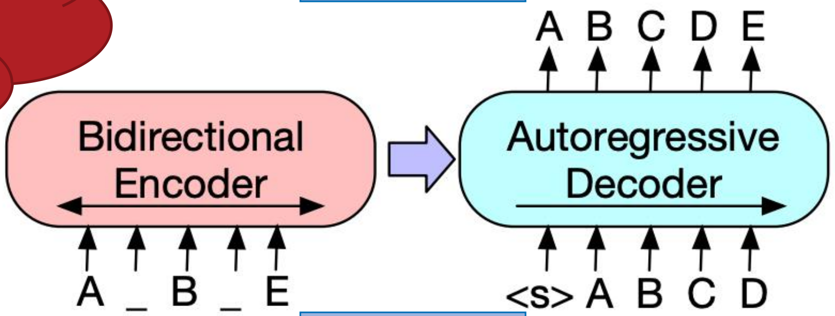


Generative Model Infrastructure Development

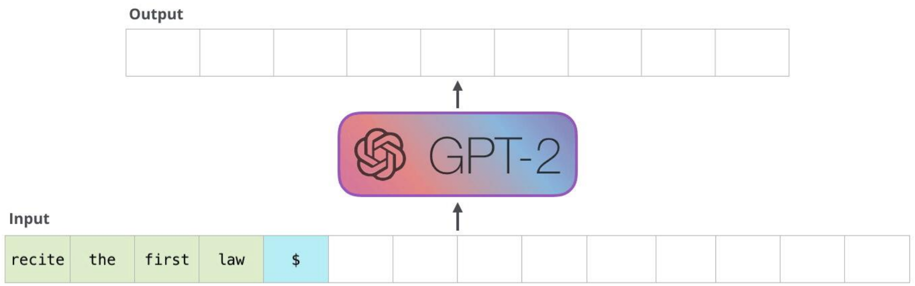
Early work: CopyRE



T5



BART



GPT

Preliminary on Knowledge Graph Construction

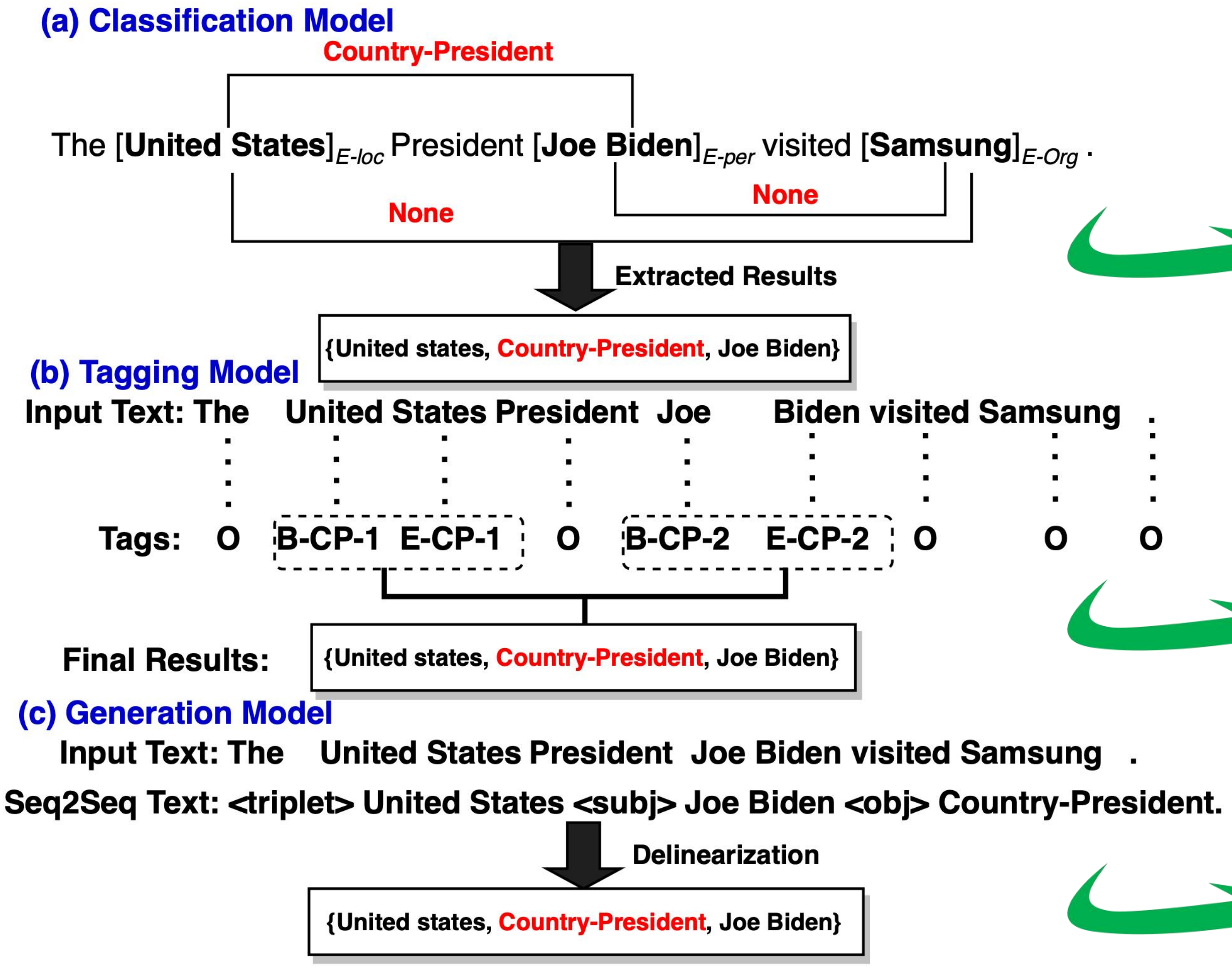
浙江大学 ZHEJIANG UNIVERSITY

Input sentence: "Steve Jobs and Steve Wozniak co-founded Apple in 1977."

Knowledge Graph Construction

- **Named Entity Recognition**
  - 'Steve Job', 'Steve Wozniak' ⇒ PERSON,
  - 'Apple' ⇒ ORG
- **Relation Extraction**
  - ⟨Steve Job, Apple⟩ => founder
- **Event Extraction**
  - Event type => Business Start-Org Event trigger => 'co-founded'
  - Event Arguments:
  - (Steve Jobs, Steve Wozniak) => AGENT and Apple => ORG
- **Entity Linking**
  - Steve Job => Steven Jobs (Q19837)
  - Apple => Apple (Q312)
- **Knowledge Graph Completion**
  - ⟨Steve Job, create, ?⟩ => Apple, NeXT Inc. and Pixar

Discrimination and Generation Methodologies



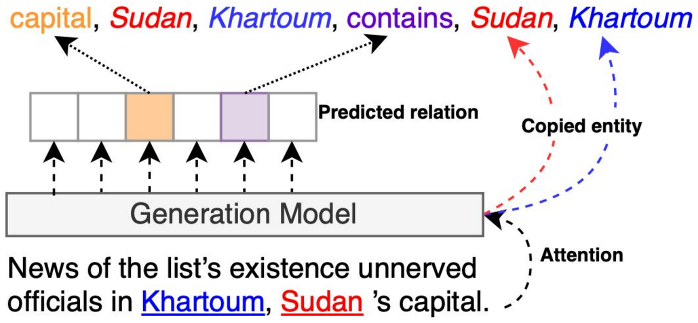
Maximize the data likelihood:

$$p_{cls}(t|x) = \prod_{(s,r,o)\in t_j} p\left((s,r,o) \mid x_j\right)$$

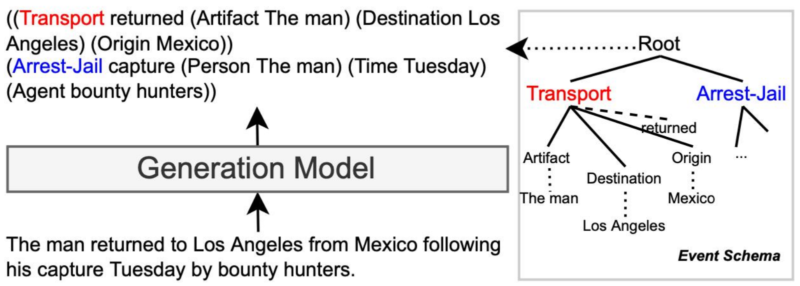
$$p_{tag}(y \mid x) = \frac{\exp(h_i, y_i)}{\sum_{\mathbf{y}'\in R} \exp\left(\exp(h_i, y_i')\right)}$$

$$p_{gen}(y \mid x) = \prod_{i=1}^{\mathrm{len}(y)} p_{gen}\left(y_i \mid y_{<i}, x\right)$$

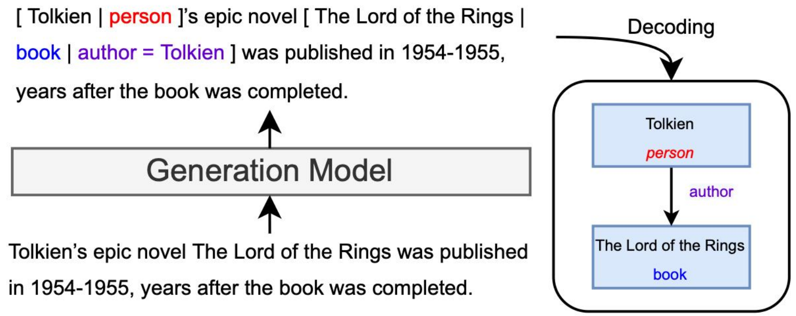
Discrimination and Generation Methodologies



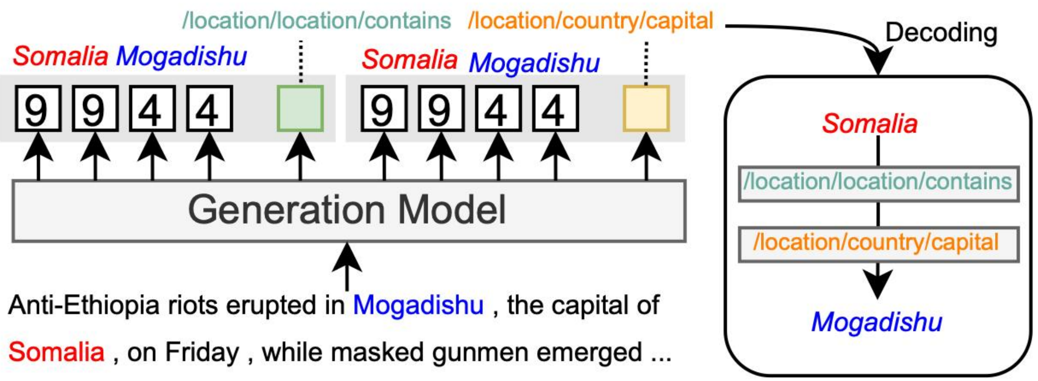
Copy-based Sequence



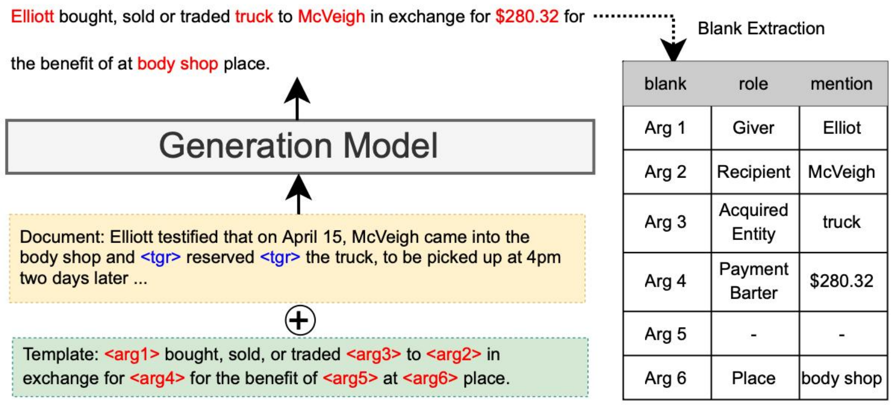
Structure-linearized Sequence



Label-augmented Sequence

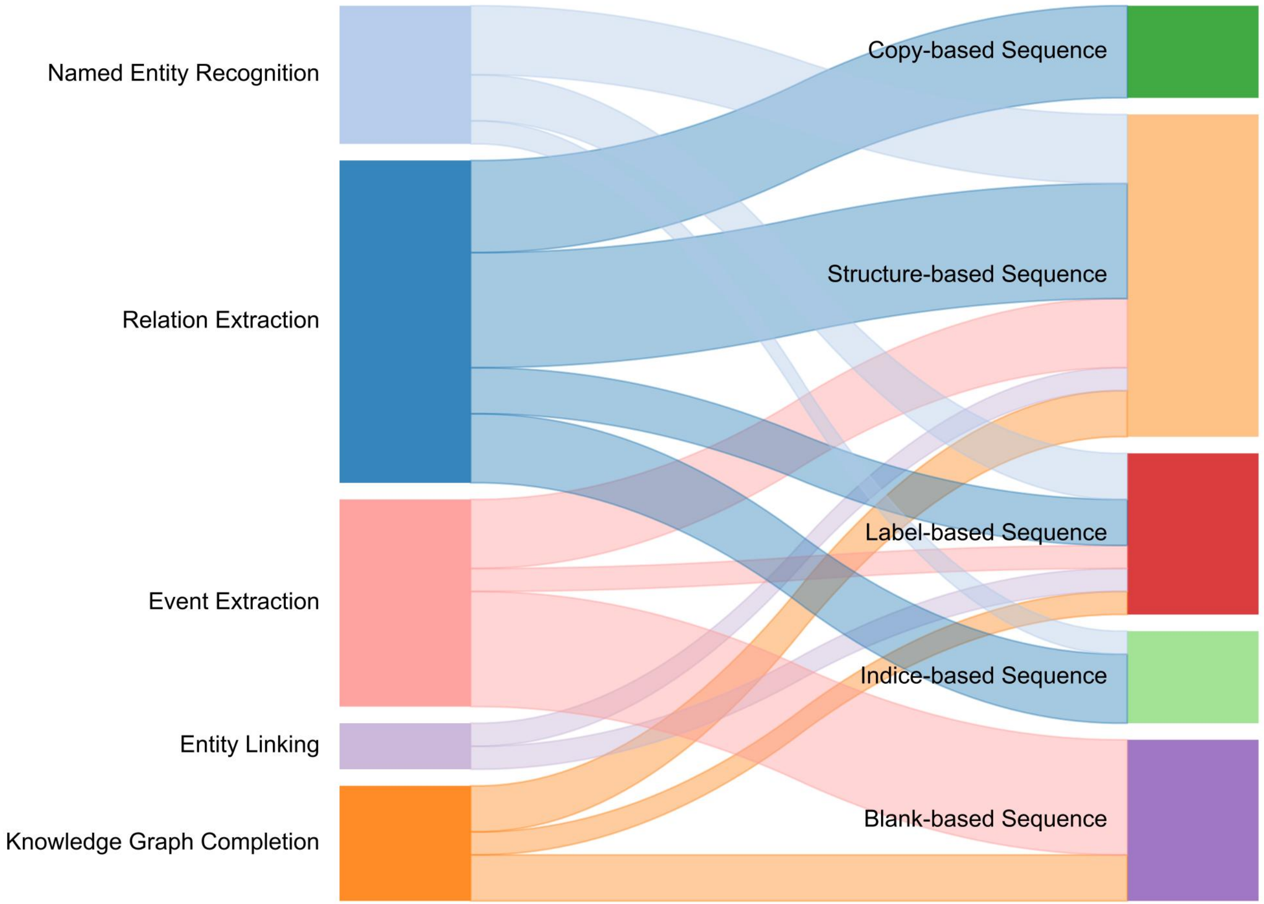


Indice-based Sequence



Blank-based Sequence

Taxonomy of Generative Knowledge



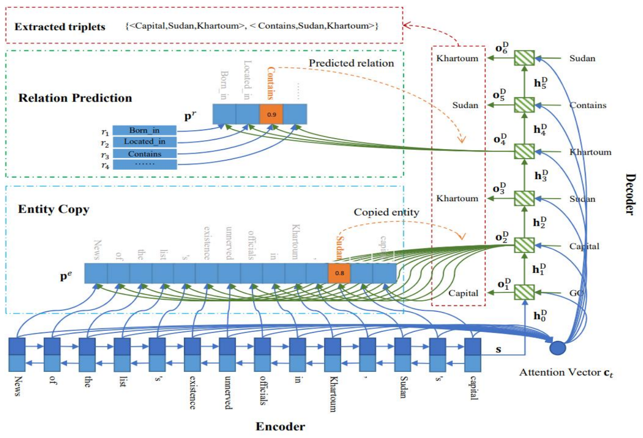
Sankey diagram of knowledge graph construction tasks with different generative paradigms

Taxonomy of Generative Knowledge

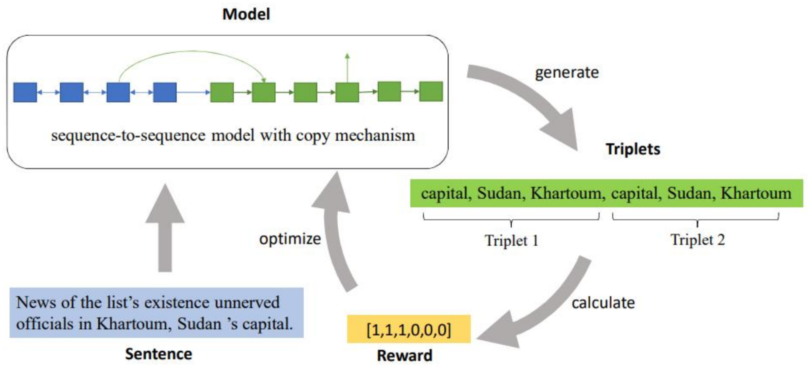
- Generative KGC Taxonomy
  - Generation Target
    - Copy-based Sequence: CopyRE (Zeng et al., 2018), CopyRRL (Zeng et al., 2019), CopyMTL (Zeng et al., 2020), TEMPGEN (Huang et al., 2021), Seq2rel (Giorgi et al., 2022)
    - Structure-based Sequence: Seq2Seq4ATE (Ma et al., 2019), Nested-seq (Straková et al., 2019), CGT (Zhang et al., 2021b; Ye et al., 2021), PolicyIE (Ahmad et al., 2021), Text2Event (Lu et al., 2021), HySPA (Ren et al., 2021), REBEL (Cabot and Navigli, 2021), SQUIRE (Bai et al., 2022), GenKGC (Xie et al., 2022), EPGEL (Lai et al., 2022), HuSe-Gen (Saha et al., 2022), UIE (Lu et al., 2022), DEEPSTRUCT (Wang et al., 2022), De-Bias (Zhang et al., 2022b), KGT5 (Saxena et al., 2022), KG-S2S (Chen et al., 2022a)
    - Label-based Sequence: ANL (Athiwaratkun et al., 2020a), GENRE (Cao et al., 2021), TANL (Paolini et al., 2021)
    - Indice-based Sequence: PNDec (Nayak and Ng, 2020), SEQ2SEQ-PTR (Rongali et al., 2020), GRIT (Du et al., 2021a), UGF for NER (Yan et al., 2021b), UGF for ABSA (Yan et al., 2021a)
    - Blank-based Sequence: COMET (Bosselut et al., 2019), BART-Gen (Li et al., 2021b), GTT (Du et al., 2021b), DEGREE (Hsu et al., 2022), ClarET (Zhou et al., 2022), GTEE (Liu et al., 2022), X-GEAR (Huang et al., 2022), PAIE (Ma et al., 2022)
  - Generation Tasks
    - Named Entity Recognition: Nested-seq (Straková et al., 2019), ANL (Athiwaratkun et al., 2020a), TANL (Paolini et al., 2021), HySPA (Ren et al., 2021), UGF for NER (Yan et al., 2021b), DEEPSTRUCT (Wang et al., 2022), De-Bias (Zhang et al., 2022b), UIE (Lu et al., 2022)
    - Relation Extraction: CopyRE (Zeng et al., 2018), CopyRRL (Zeng et al., 2019), PNDec (Nayak and Ng, 2020), CopyMTL (Zeng et al., 2020), CGT (Zhang et al., 2021b; Ye et al., 2021), TANL (Paolini et al., 2021), HySPA (Ren et al., 2021), TEMPGEN (Huang et al., 2021), REBEL (Cabot and Navigli, 2021), DEEPSTRUCT (Wang et al., 2022), UIE (Lu et al., 2022), Seq2rel (Giorgi et al., 2022)
    - Event Extraction: CGT (Zhang et al., 2021b; Ye et al., 2021), TANL (Paolini et al., 2021), BART-Gen (Li et al., 2021b), GTT (Du et al., 2021b), GRIT (Du et al., 2021a), Text2Event (Lu et al., 2021), DEGREE (Hsu et al., 2022), ClarET (Zhou et al., 2022), GTEE (Liu et al., 2022), X-GEAR (Huang et al., 2022), DEEPSTRUCT (Wang et al., 2022), PAIE (Ma et al., 2022), UIE (Lu et al., 2022)
    - Entity Linking: GENRE (Cao et al., 2021), EPGEL (Lai et al., 2022)
    - Knowledge Graph Completion: COMET (Bosselut et al., 2019), SQUIRE (Bai et al., 2022), GenKGC (Xie et al., 2022),, HuSe-Gen (Saha et al., 2022), ClarET (Zhou et al., 2022), KGT5 (Saxena et al., 2022), KG-S2S (Chen et al., 2022a)

Copy-based Sequence

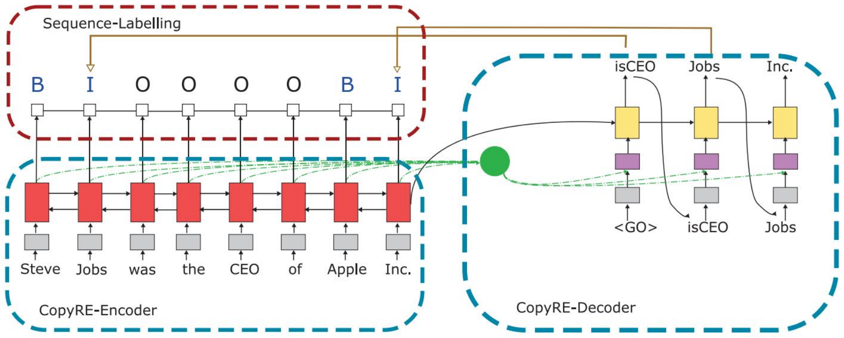
Directly copy entity



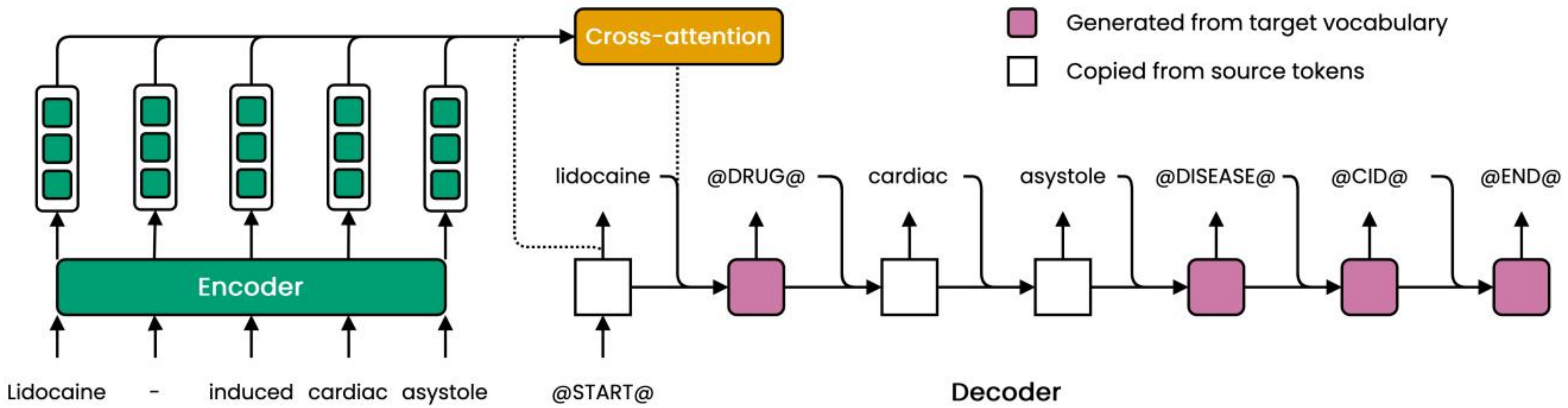
CopyRE



CopyRRL



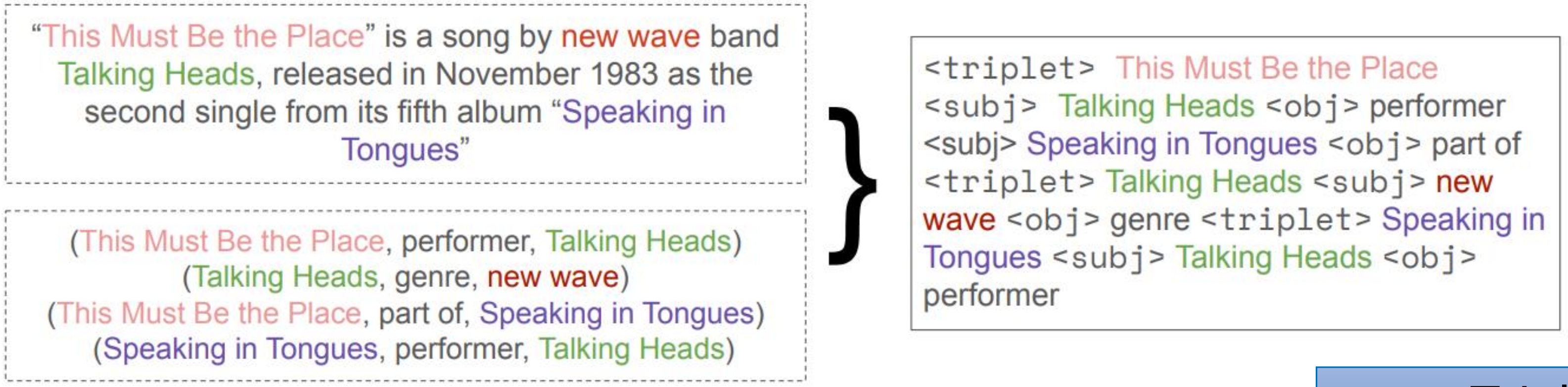
CopyMTL



Seq2rel

Restricted target vocabulary

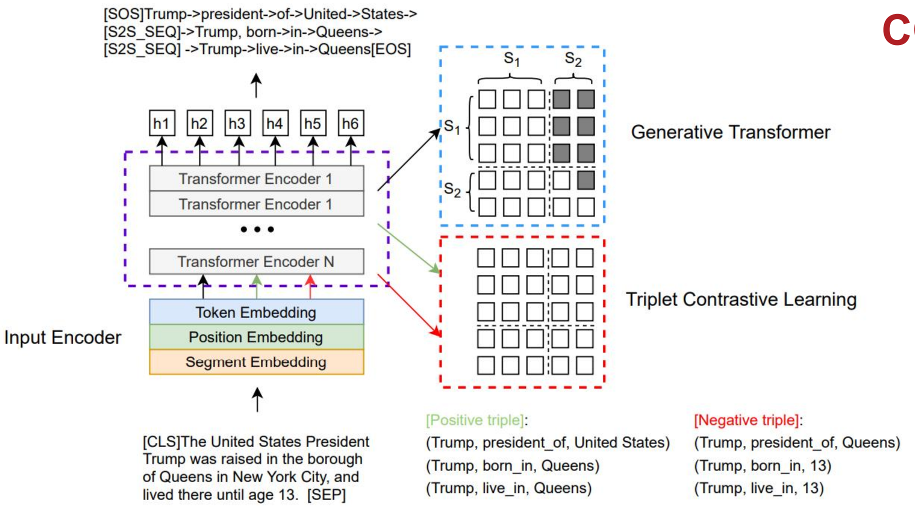
Structure-linearized Sequence



Triplet linearization

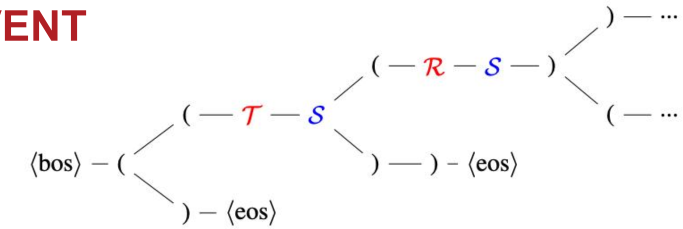
REBEL

CGT

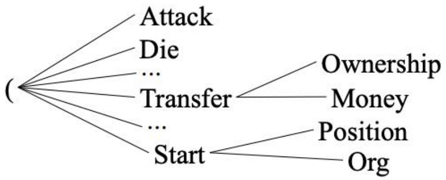


Faithful contrastive learning

TEXT2EVENT



(a) The trie of event structure.



(b) The trie of event type $\mathcal{T}$.

Prefix tree constraint decoding

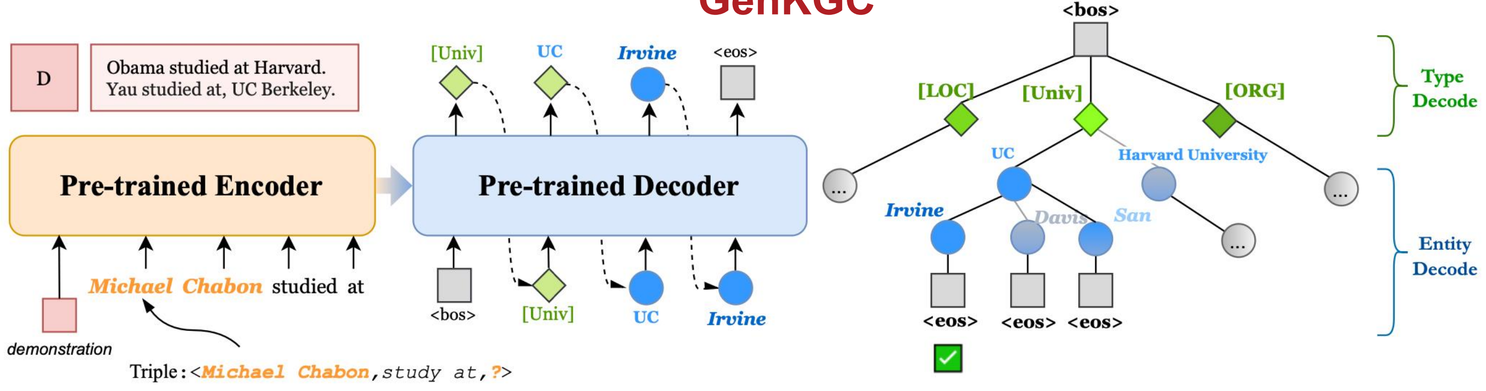
Structure-linearized Sequence

Nested-seq

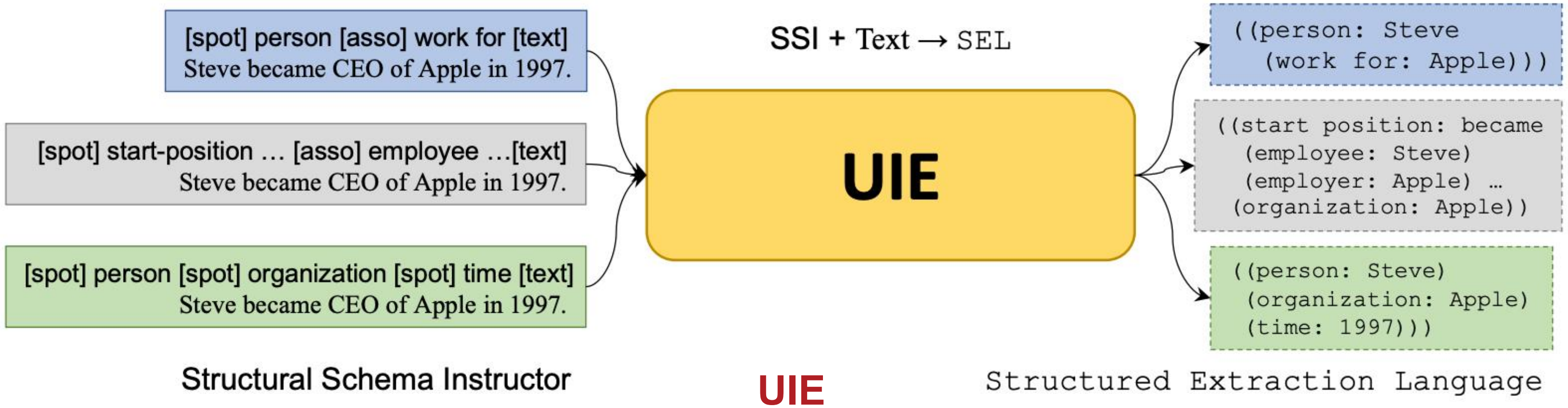
in	O
the	B-ORG
US	I-ORG|U-GPE
Federal	I-ORG
District	I-ORG|U-GPE
Court	I-ORG
of	I-ORG
New	I-ORG|B-GPE
Mexico	L-ORG|L-GPE
.	O

Per-token tag encoding

GenKGC



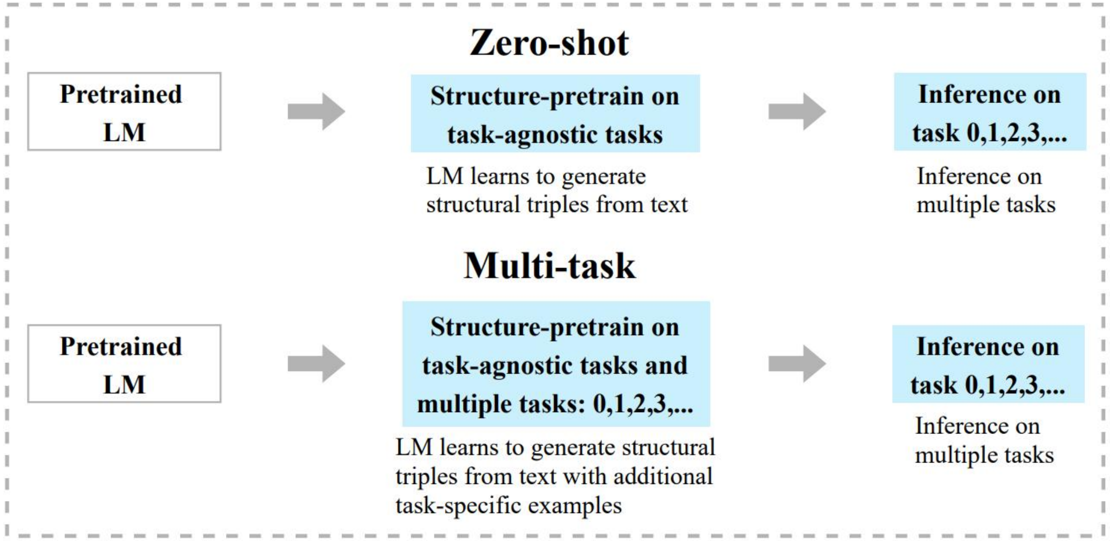
Entity-aware hierarchical decoding



UIE

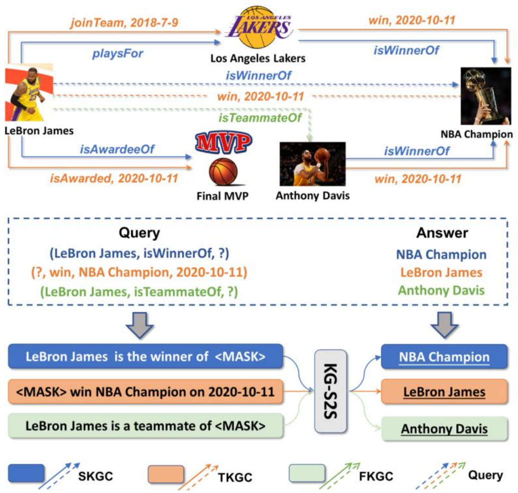
Unified structure generation

Structure-linearized Sequence

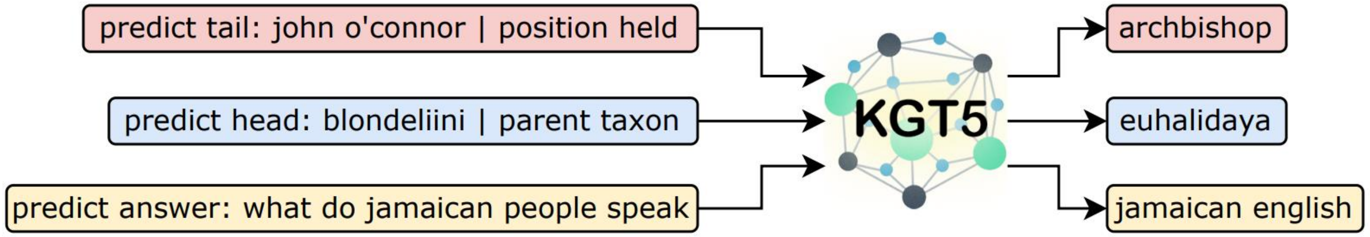


DEEPSTRUCT

Reformulating triple prediction



KG-S2S

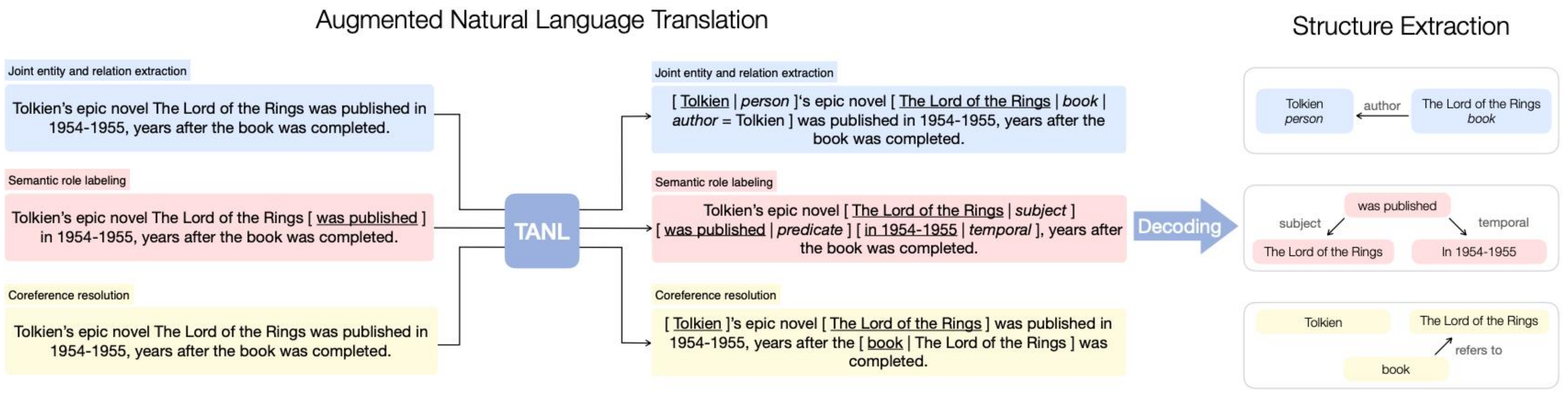


KGT5

Query Verbalization

求是创新

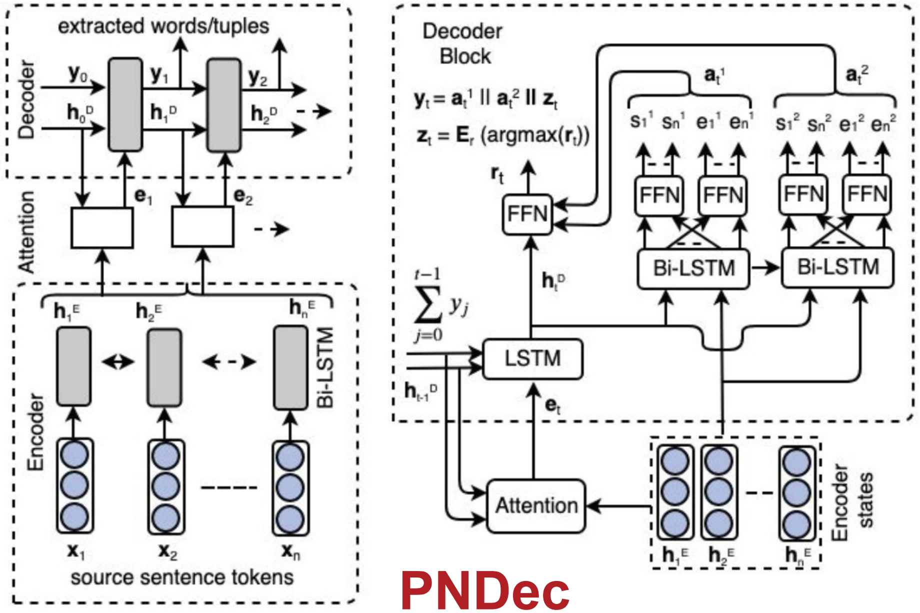
Label-based Sequence



TANL

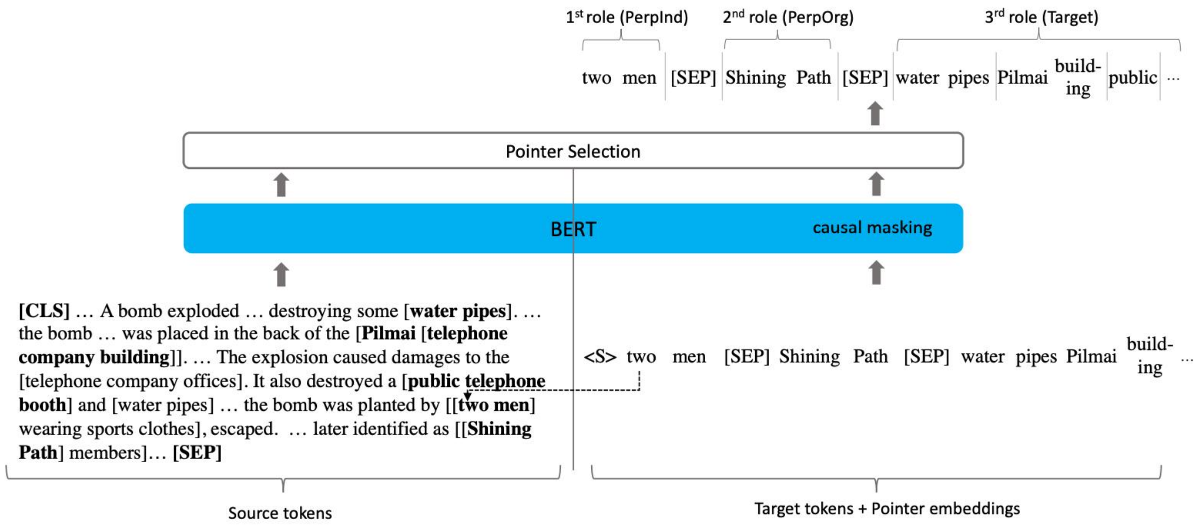
Augmented natural language

Indice-based Sequence



PNDec

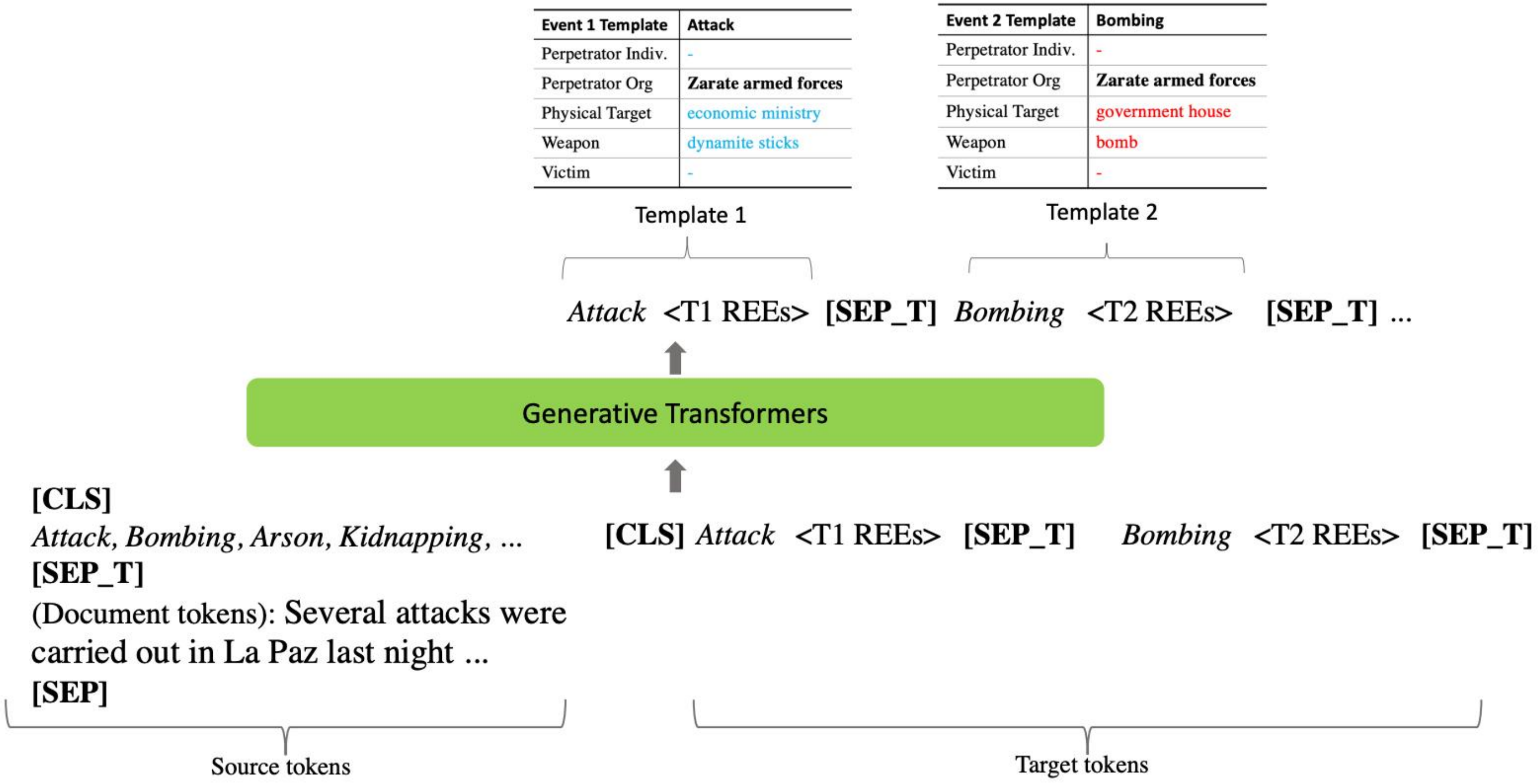
Pointer mechanism



GRIT

Pointer selection

Blank-based Sequence

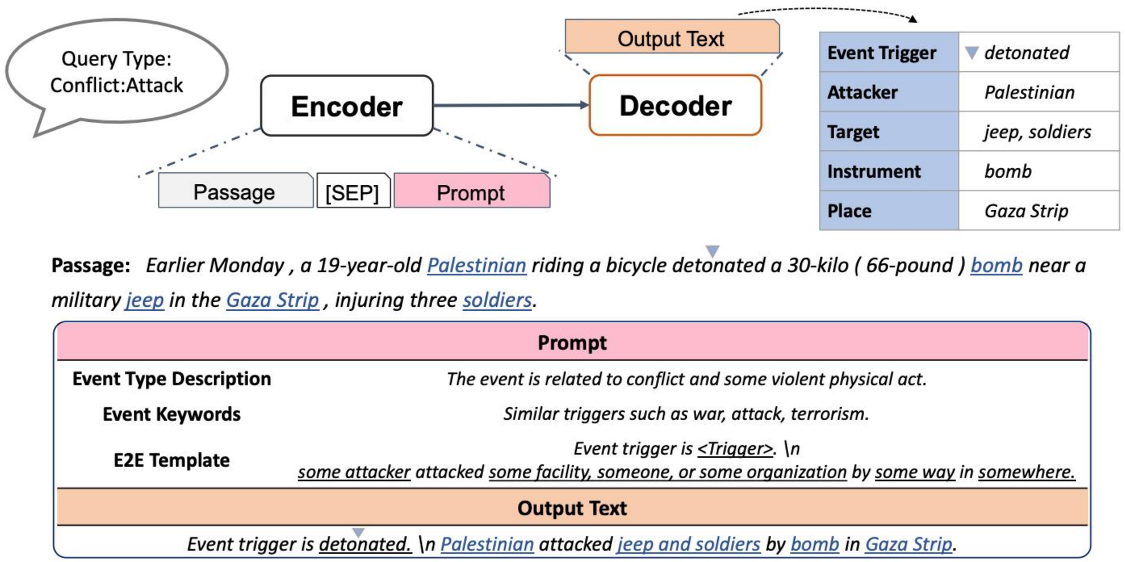


GTT

Template filling as generation

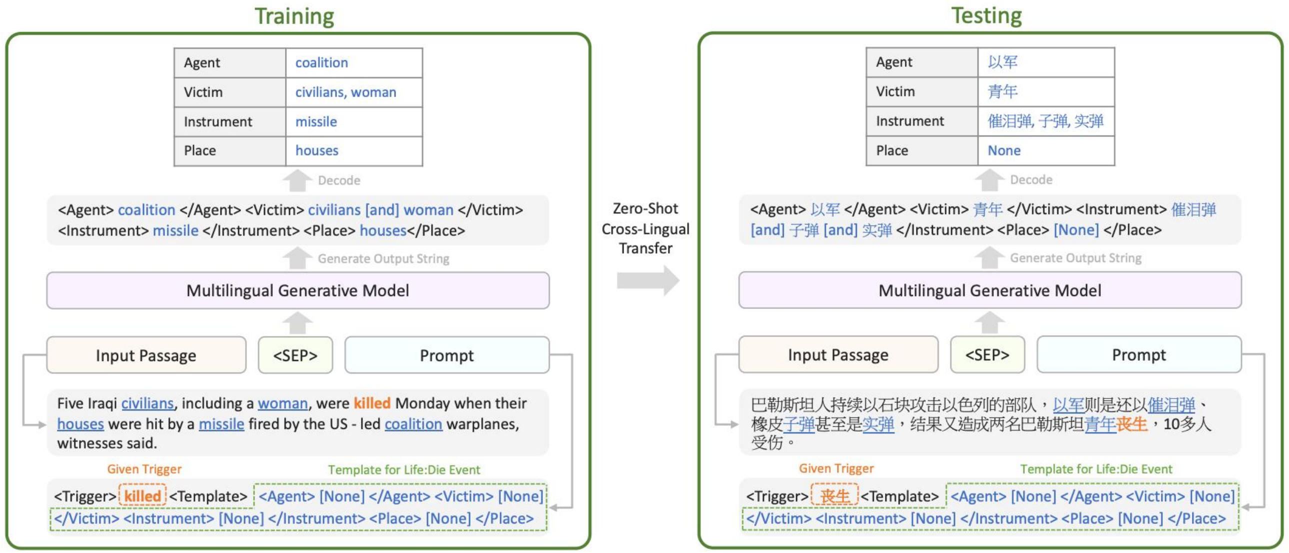
求是创新

Blank-based Sequence



DEGREE

Prompt semantic guidance



X-GEAR

Language-agnostic template

Theoretical Insight

浙江大學 ZHEJIANG UNIVERSITY

Taxonomy	Generative Strategy	Representative Model	Evaluation Scope			
			SU↑	SS↓	AS↓	TS↓
Copy-based (§ 3.1)	Directly copy entity	CopyRE (Zeng et al., 2018)	L	L	M	L
	Restricted target vocabulary	Seq2rel (Giorgi et al., 2022)	L	L	H	L
Structure-based (§ 3.2)	Per-token tag encoding	Nested-seq (Straková et al., 2019)	L	L	H	L
	Faithful contrastive learning	CGT (Zhang et al., 2021b)	M	M	H	L
	Prefix tree constraint decoding	TEXT2EVENT (Lu et al., 2021)	M	M	H	L
	Triplet linearization	REBEL (Cabot and Navigli, 2021)	M	H	M	L
	Entity-aware hierarchical decoding	GenKGC (Xie et al., 2022)	M	L	M	L
	Unified structure generation	UIE (Lu et al., 2022)	M	H	H	L
	Reformulating triple prediction	DEEPSTRUCT (Wang et al., 2022)	M	H	H	L
	Query Verbalization	KGT5 (Saxena et al., 2022)	M	H	M	L
Label-based (§ 3.3)	Augmented natural language	TANL (Paolini et al., 2021)	M	H	H	L
Indice-based (§ 3.4)	Pointer mechanism	PNDec (Nayak and Ng, 2020)	L	L	M	L
	Pointer selection	GRIT (Du et al., 2021a)	M	L	M	L
Blank-based (§ 3.5)	Template filling as generation	GTT (Du et al., 2021b)	H	H	H	H
	Prompt semantic guidance	DEGREE (Hsu et al., 2022)	H	H	H	H
	Language-agnostic template	X-GEAR (Huang et al., 2022)	H	M	H	H

Comparison of generation methods from different evaluation scopes. "SU" indicates semantic utilization, "SS" indicates search space, "AS" indicates application scope, and "TS" indicates template cost. We divide the degree into three grades: L (low), M (middle), and H (high), and the ↑ indicates that the higher grade performance is better while the ↓ is the opposite.

求是創新

Theoretical Insight

浙江大学 ZHEJIANG UNIVERSITY

NLG are normally modeled by parameterized probabilistic models：

$$p_{gen}(y \mid x) = \prod_{i=1}^{\text{len}(y)} p_{gen}(y_i \mid y_{<i}, x)$$

$$y = \arg\max_{y \in \mathcal{T}} \log P_\theta(D(y) \mid x)$$

Generating architecture

Non-autoregressive parallel decoding architecture

It is necessary to develop sophisticated, efficient decoding strategies for generative KGC

Generate Objectives

Representation space modeling

Balance parametric optimization as well as semantic utilization

Generation Mechanism

Learning with constraints

Vanilla decoding solutions such as beam search or greedy

Knowledge-guided (or schema-guided) decoding

Empirical Analysis

Type	Models	NYT		
		P	R	F
Discrimination	CasRel (Wei et al., 2020)	89.7	89.5	89.6
	TPLinker (Wang et al., 2020)	91.4	92.6	92.0
	OneRel (Shang et al., 2022)	92.8	92.9	92.8
Copy-based	CopyRE (Zeng et al., 2018)	61.0	56.6	58.7
	CopyRRL (Zeng et al., 2019)	77.9	67.2	72.1
	CopyMTL (Zeng et al., 2020)	75.7	68.7	72.0
Structure-based	CGT (Ye et al., 2021)	94.7	84.2	89.1
	REBEL (Cabot and Navigli, 2021)	91.5	92.0	91.8
	UIE (Lu et al., 2022)	-	-	93.5
	DEEPSTRUCT (Wang et al., 2022)	-	-	93.9
Label-based	TANL (Paolini et al., 2021)	-	-	90.8
Indice-based	PNDec (Nayak and Ng, 2020)	89.3	78.8	83.8
Others*	SPN (Sui et al., 2020)	93.3	91.7	92.5
	Seq2UMTree (Zhang et al., 2020b)	79.1	75.1	77.1

Table 2: Main results of NYT dataset. The top section refers to the discrimination models, and the bottom section indicates generation models. "*" refers to the non-autoregressive models.

Empirical Analysis

浙江大學 ZHEJIANG UNIVERSITY

Type	Models	Trigger		Argument	
		Id	Cl	Id	Cl
Discrimination	JMEE (Liu et al., 2018)	75.9	73.7	68.4	60.3
	DYGIE++ (Wadden et al., 2019)	-	69.7	53.0	48.8
	OneIE (Lin et al., 2020)	78.6	75.2	60.7	58.6
	QAEE (Du and Cardie, 2020)	75.8	72.4	55.3	53.3
	MQAEE (Li et al., 2020a)	74.5	71.7	55.2	53.4
	RCEE (Liu et al., 2020)	-	74.9	-	63.6
Structure-based	TEXT2EVENT (Lu et al., 2021)	-	71.9	-	53.8
	UIE (Lu et al., 2022)	-	73.4	-	54.8
	DEEPSTRUCT (Wang et al., 2022)	73.5	69.8	59.4	56.2
Label-based	TANL (Paolini et al., 2021)	72.9	68.4	50.1	47.6
Blank-based	BART-Gen (Du et al., 2021b)	74.4	71.1	55.2	53.7
	DEGREE (Hsu et al., 2022)	-	73.3	-	55.8
	GTEE (Liu et al., 2022)	-	72.6	-	55.8
	PAIE (Ma et al., 2022)	-	-	75.7*	72.7*

Table 3: F1 results (%) of ACE-2005. The top section refers to the discrimination models, and the bottom section indicates the generation models. Id is Identification, and Cl is Classification. "*" refers to experiments only in argument extraction tasks with the golden trigger.

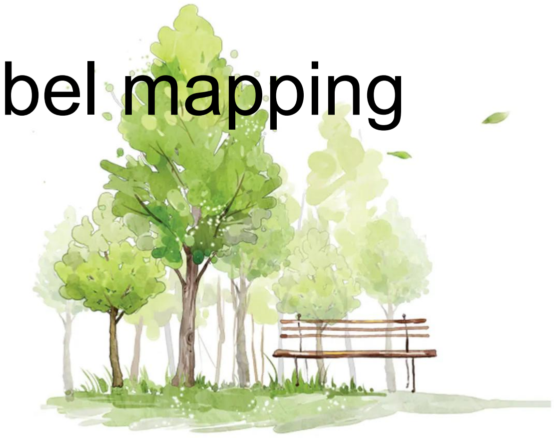
Empirical Analysis

- We can observe that:
  - Structure-based and label-based methods can better utilize label semantics and structural knowledge than other generation methods.
  - Structure-based methods develop a universal architecture to solve different tasks.
  - Structure-based methods can be pre-trained in multiple downstream tasks or cross-language tasks.

Empirical Analysis

浙江大學
ZHEJIANG UNIVERSITY

- We can observe that:
  - ♦ Owing to the complete template design of the Blank-based approach, PLMs can understand complex task knowledge, structural knowledge and label semantics.
  - ♦ In contexts with nested labels in NER and overlapping triples in RE , generative method implicitly models the structure between labels, which requests complex multi-label mapping for traditional discriminative models.



Future Directions

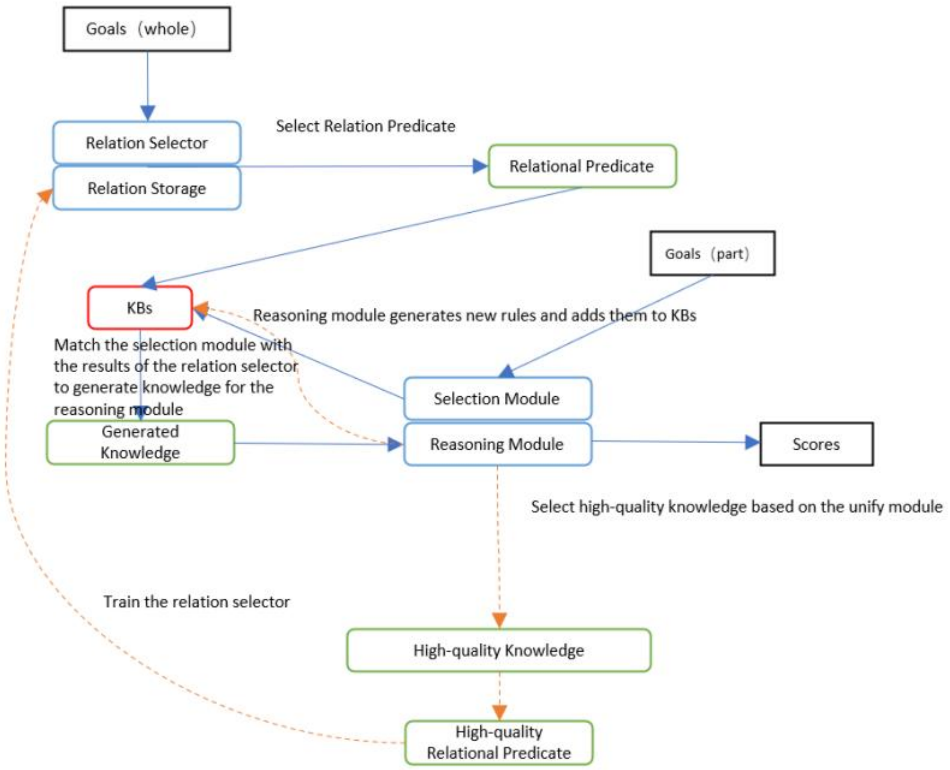
- **Generation Architecture**
  - Most of the recent generative KGC frameworks face serious homogenization with Transformer.
- **Generation Quality**
  - Considering the target reliability of generation methods, more sophisticated strategies can be leveraged to control the quality of generative KGC.
- **Training Efficiency**
  - In practical applications, it is essential to reduce data annotation and training costs.

Future Directions

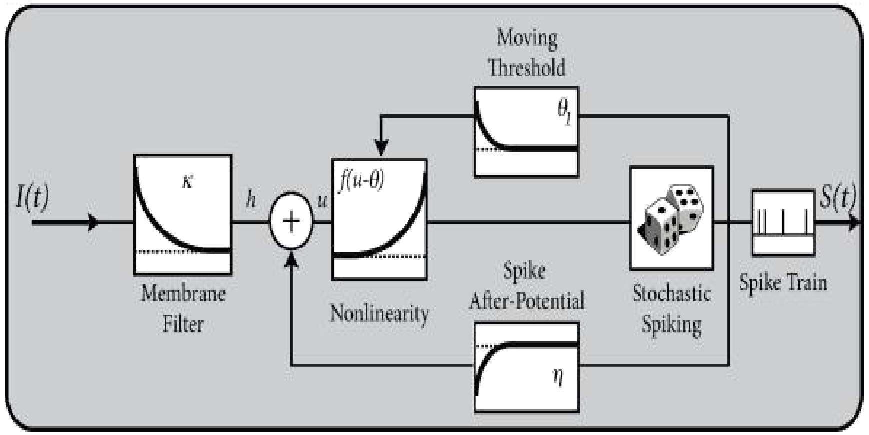
- **Universal Deployment**
  - Inspired by the T5, which transforms all NLP tasks into Text-to-Text tasks, generation models can be generalized to the multi-task and multi-modal domain.
- **Inference Speed**
  - Generative tasks are still limited by auto-regressive decoders.
  - The autoregressive decoder generates each token based on previously generated tokens during inference.
  - Auto-regressive process is not parallelizable.

Future Directions

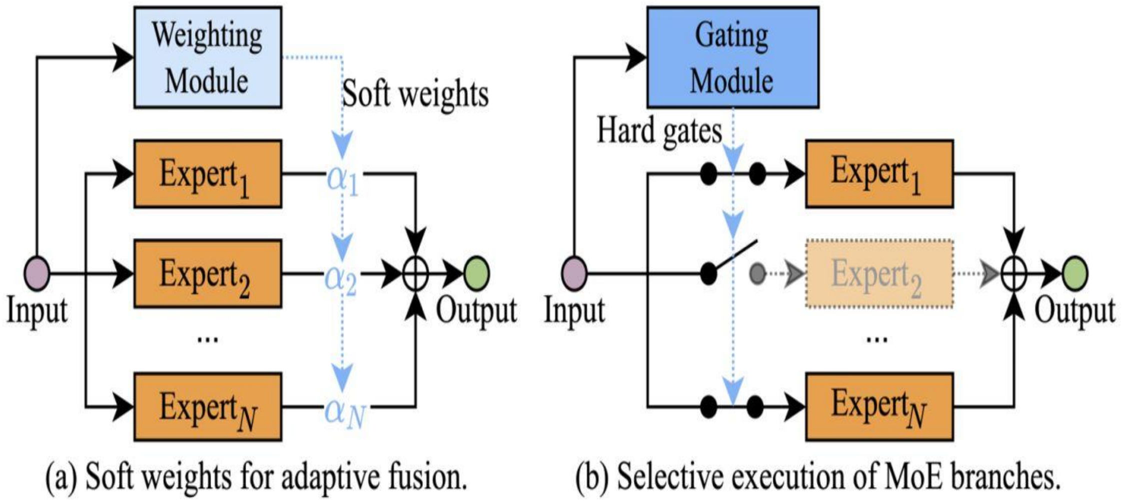
➢ Generation Architecture



Neural-Symbolic Transformer



Spiking Neural Network



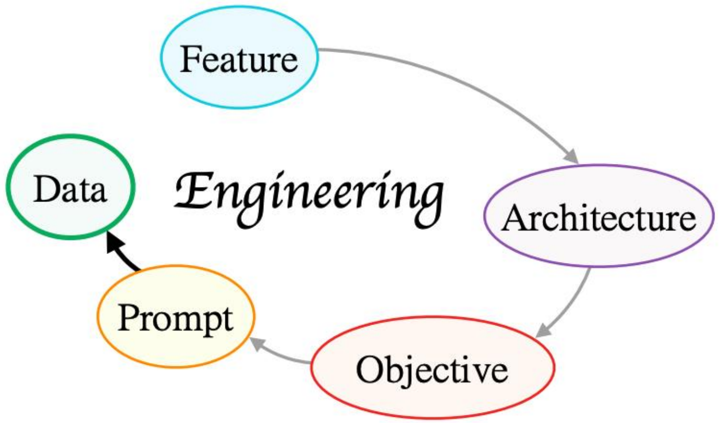
Dynamic Neural Network

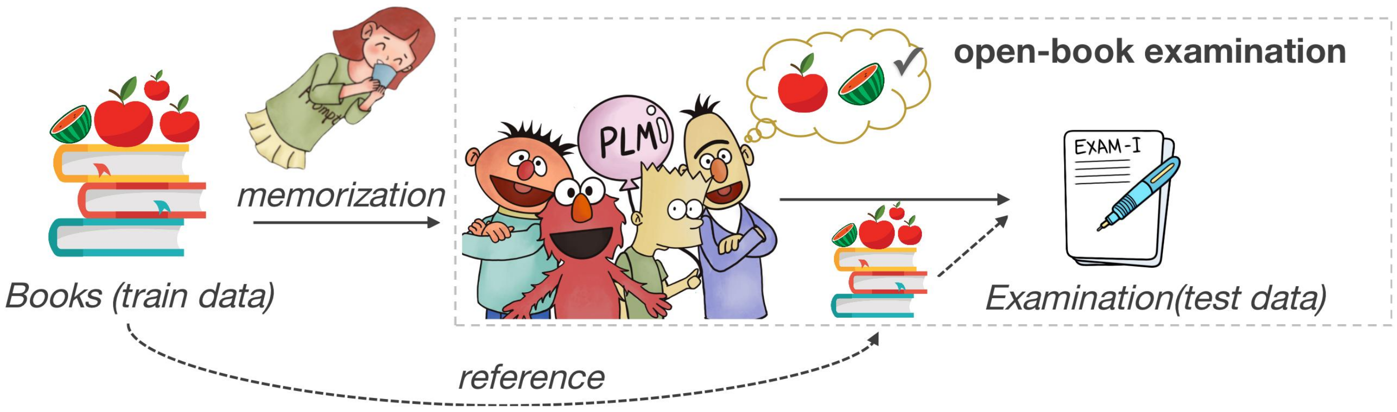
ZHEJIANG UNIVERSITY

- Generation Quality

Generation Strategies	References
Control code construction	CTRL(Keskar et al., 2019) Gsum(Dou et al., 2021)
Decoding strategy such as introducing external feedback and generative discriminator	L2W(Holtzman et al., 2018) Gedi(Krause et al., 2021)
Loss function design	Cocon(Chan et al., 2021)
Prompt design	Prefix-turing(Qian et al., 2020)
Retrieval augmentation	REALM(Kelvin et al., 2022)
Write-then-Edit strategy	PPLM(Dathathri et al., 2020)
Diffusion process	Diffuseq(Gong et al., 2022)

- Training Efficiency
  - Freeze most of the generation model parameters
  - Leverage prompt learning
  - Knowledge decoupling intervention training models



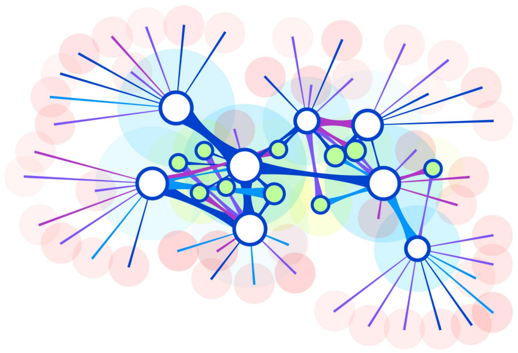


Future Directions

- Universal Deployment
  - ♦ Align structure and expand to joint multi-task training
  - ♦ Cross-language to compensate for the lack of linguistic corpus resources
  - ♦ Cross-modal enhancement of interactions between perceptual elements
  - ♦ Integrating symbolic knowledge to guide KGC with cognitive knowledge

Future Directions

- Inference Speed
  - Utilizes the transformer-based non-autoregressive decoder
  - Formulates end-to-end knowledge base population as a direct set generation problem
  - Unordered multi-tree decoding strategy to avoid aggregation of information at prediction
  - Semi-autoregressive decoding
  - Pathways can dynamically assign competencies to different parts of the neural network



Future Work

GENKGC
GENERATIVE KNOWLEDGE GRAPH CONSTRUCTION

[🔔 News! 🔔] We have released a new survey paper:"Generative Knowledge Graph Construction: A Review" based on this repository, with a perspective of existing Generative Knowledge Graph Construction! We are looking forward to any comments or discussions on this topic :)

🕵️ Introduction

Generative Knowledge Graph Construction (KGC) refers to those methods that leverage the sequence-to-sequence framework for building knowledge graphs, which is flexible and can be adapted to widespread tasks. In this study, we summarize the recent compelling progress in generative knowledge graph construction. We present the advantages and weaknesses of each paradigm in terms of different generation targets and provide theoretical insight and empirical analysis. Based on the review, we suggest promising research directions for the future. Our contributions are threefold: (1) We present a detailed, complete taxonomy for the generative KGC methods; (2) We provide a theoretical and empirical analysis of the generative KGC methods; (3) We propose several research directions that can be developed in the future. For more resources about knowledge graph construction, please check our paper tookit DeepKE and PromptKG.

Public Repository: https://github.com/zjunlp/Generative_KG_Construction_Papers

浙江大學
ZHEJIANG UNIVERSITY

Thank You!

zhangningyu@zju.edu.cn

ACCEPT MY ENDLESS GRATITUDE

DeepKE

PromptKG