Cognitive AI:

—Understanding, Reasoning, and Decision

Jie Tang
Computer Science
Tsinghua University

The slides can be downloaded at http://keg.cs.tsinghua.edu.cn/jietang

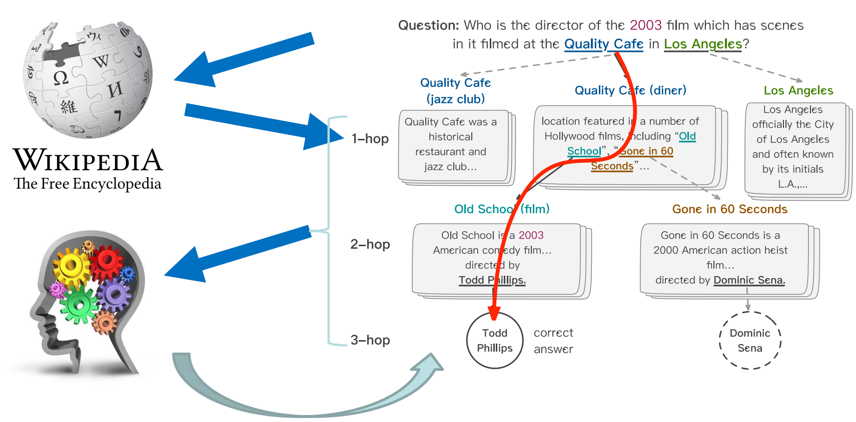
---

- Let us start with an example…

Reasoning

Question: Who is the director of the 2003 film which has scenes in it filmed at the Quality Cafe in Los Angeles?

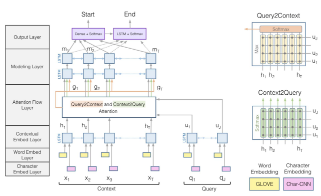
Cognitive Graph



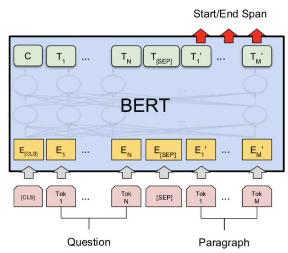
BIDAF, BERT, XLNet

- Modeling the whole document by pre-training
- However, lack of explainability

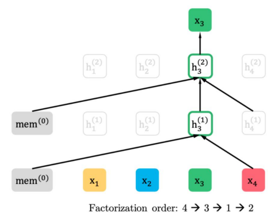
BiDAF



BERT



XLNet

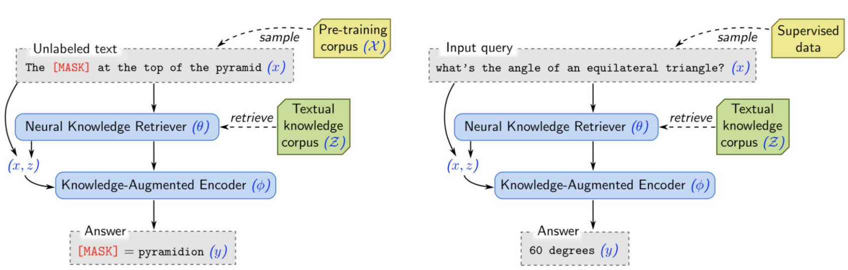


REALM: Retrieval-Augmented LM

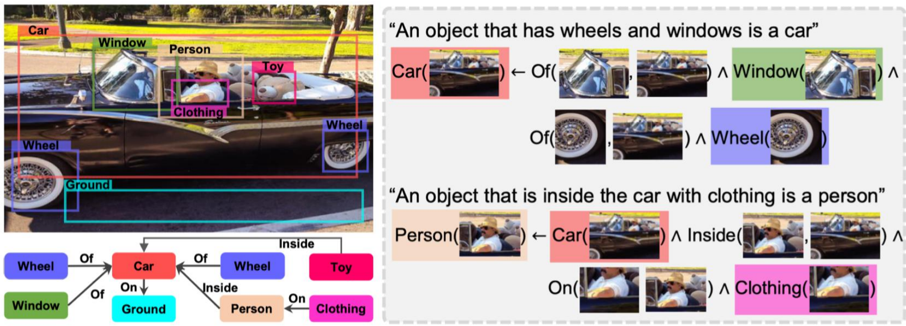
$$\text{BERT} : p(y|x)$$

$$\text{REALM} : p(y|x) = \sum_{z \in \mathcal{Z}} p(y|z,x)p(z|x)$$

- where Z is the supporting set.



Inductive Logic Programming



They apply inductive logic learning on scene graphs generated by deep learning, to extract explainable rules of predicting the object class labels.

1. Yuan Yang and Le Song. Learn to Explain Efficiently via Neural Logic Inductive Learning. ICLR. 2020.

Challenge: only System 1 DL

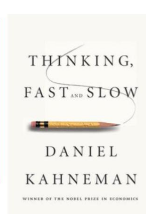
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, non-linguistic, habitual
- Current DL



System 2

- Slow, logical, sequential, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Future DL

Mila

3

1. From Bengio's NIPS'2019 Keynote

System 2 DL: Cognitive Graph

Question: Who is the director of the 2003 film which has scenes in it filmed at the Quality Cafe in Los Angeles?

Quality Café

The Quality Café is a now-defunct diner in Los Angeles, California. The restaurant has appeared as a location featured in a number of Hollywood films, including Old School, Gone in 60 Seconds, …

Los Angeles

Los Angeles is the most populous city in California, the second most populous city in the United States, after New York City, and the third most populous city in North America.

Alessandro Moschitti

Alessandro Moschitti is a professor of the CS Department of the University of Trento, Italy. He is currently a Principal Research Scientist of the Qatar Computing Research Institute (QCRI)

WIKIPEDIA
The Free Encyclopedia

Old School

Old School is a 2003 American comedy film released by Dream Works Pictures and The Montecito Picture Company and directed by Todd Phillips.

Todd Phillips

Todd Phillips is an American director, producer, screenwriter, and actor. He is best known for writing and directing films, including Road Trip (2000), Old School (2003), Starsky & Hutch (2004), and The Hangover Trilogy.

Tsinghua University

Tsinghua University is a major research university in Beijing and dedicated to academic excellence and global development. Tsinghua is perennially ranked as one of the top academic institutions in China, Asia, and worldwide…

Cognitive Science

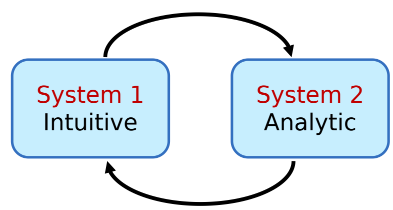
System 1
Intuitive

System 2
Analytic

Dual Process Theory (Cognitive Science)

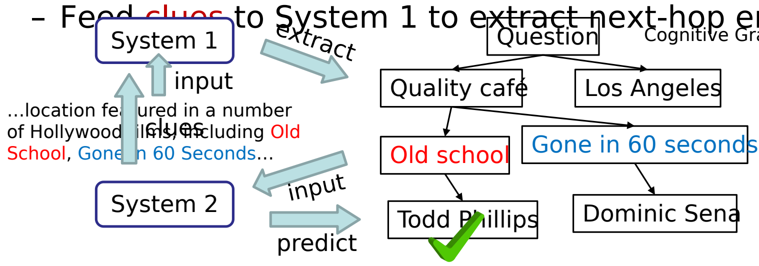
Reasoning w/ Cognitive Graph

- System 1:
  - Knowledge expansion by association in text when reading
- System 2:
  - Decision making w/ all the information



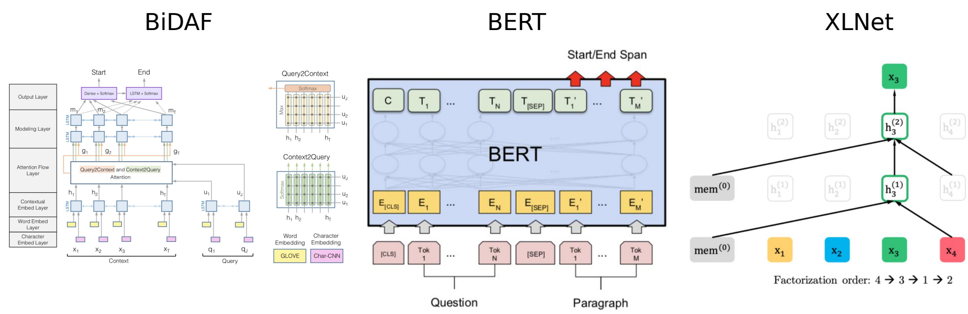
CogQA: Cognitive Graph for QA

- An iterative framework corresponding to dual process theory
- System 1
  - extract entities to build the cognitive graph
  - generate semantic vectors for each node
- System 2
  - Do reasoning based on semantic vectors and graph
  - Feed clues to System 1 to extract next-hop entities



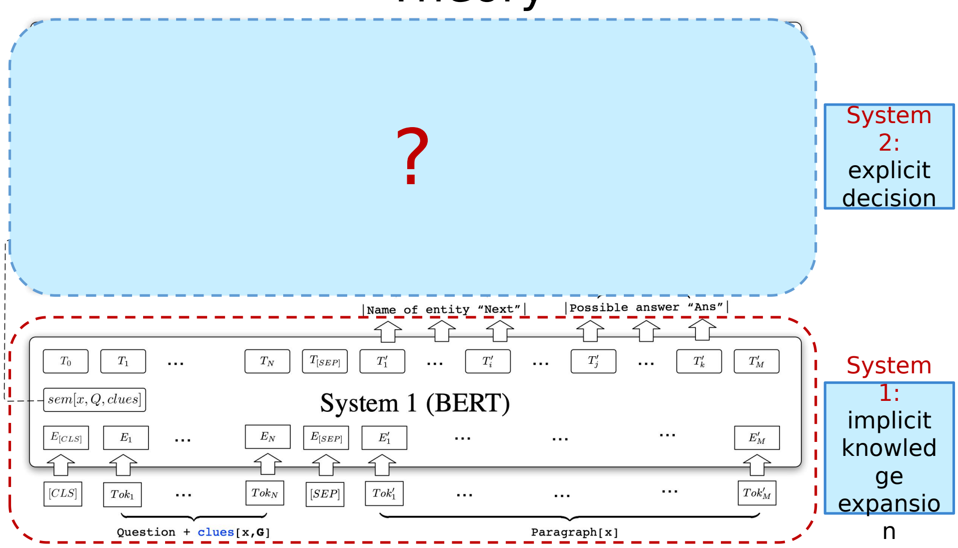
System 1: BIDAF, BERT

- reading comprehension: target at understanding the whole paragraph



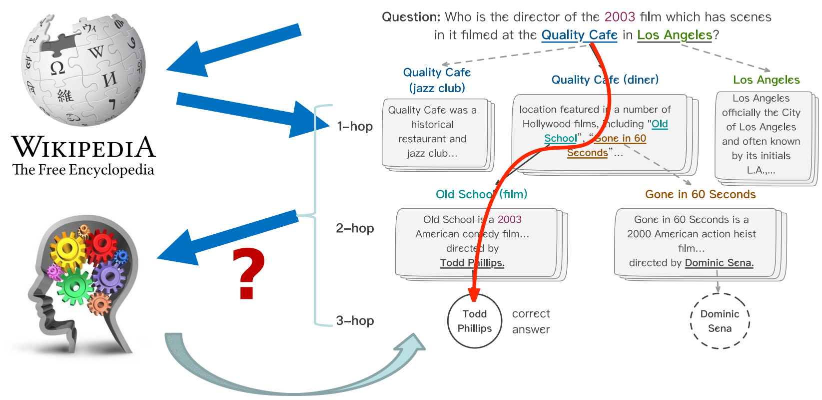
1. Rajpurkar, Pranav, et al. 'SQuAD: 100,000+ Questions for Machine Comprehension of Text.' EMNLP. 2016.

Cognitive Graph: DL + Dual Process Theory

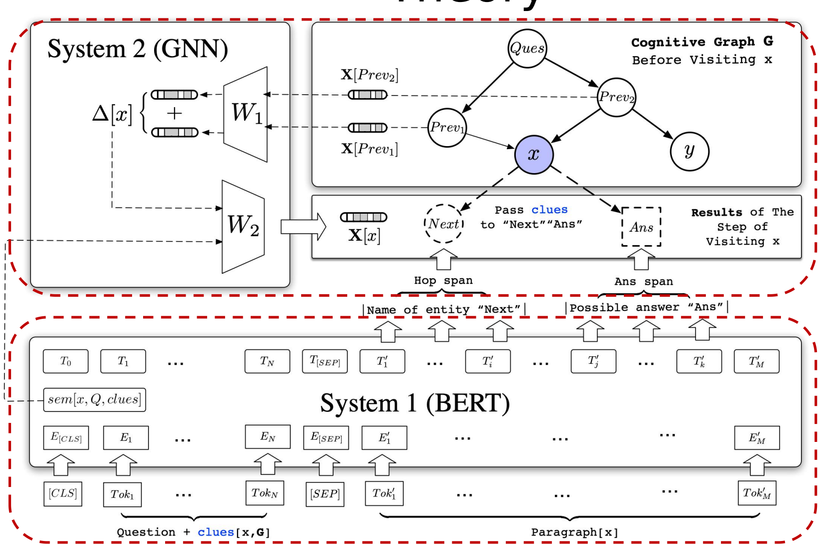


1. M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. ACL'19.

System 2: Reasoning, and Decision



Cognitive Graph: DL + Dual Process Theory

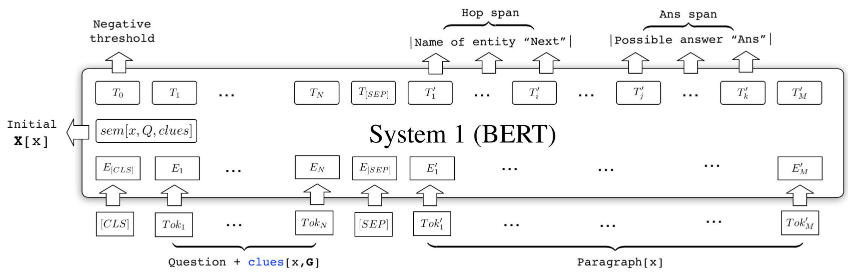


System 2: explicit decision

System 1: implicit knowledge expansion

1. M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. ACL'19.

System 1: the BERT Implementation



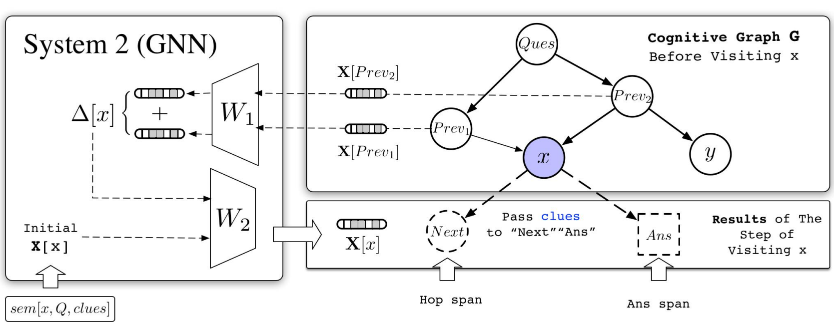
$$\underbrace{[CLS]\ Question\ [SEP]\ clues[x,\mathcal{G}]\ \ [SEP]}_{Sentence\ A}\ \ \underbrace{Para[x]}_{Sentence\ B}$$

$$P_{ans}^{start}[i] = \frac{e^{\mathbf{S}_{ans}\cdot\mathbf{T}_i}}{\sum_j e^{\mathbf{S}_{ans}\cdot\mathbf{T}_j}}$$

$$end_k = \arg\max_{start_k \le j \le start_k + maxL} P_{ans}^{end}[j]$$

- Extract top-k next-hop entities and answer candidates respectively
  - Predict the start and end probabilities of each position
- Generate semantic vectors for entities based on their documents
- Take the 0-th probability as negative threshold
  - Ignore the spans whose start probabilities are small than the negative threshold

System 2: the GNN implementation



At each step, hidden representations **X** for nodes are updated according to the propagation rules:

$$\Delta = \sigma((AD^{-1})^T \sigma(\mathbf{X}W_1))$$

$$\mathbf{X}' = \sigma(\mathbf{X}W_2 + \Delta)$$

Predictor *F* is a two-layer MLP, which predicts the final answer based on hidden representations **X**:

$$answer = \arg\max_{\text{answer node } x} \mathcal{F}(\mathbf{X}[x])$$

Performance

- HotpotQA is a dataset with leaderboard similar to SQuAD
- CogQA ranked 1st from 21, Feb to 15, May (nearly 3 month)

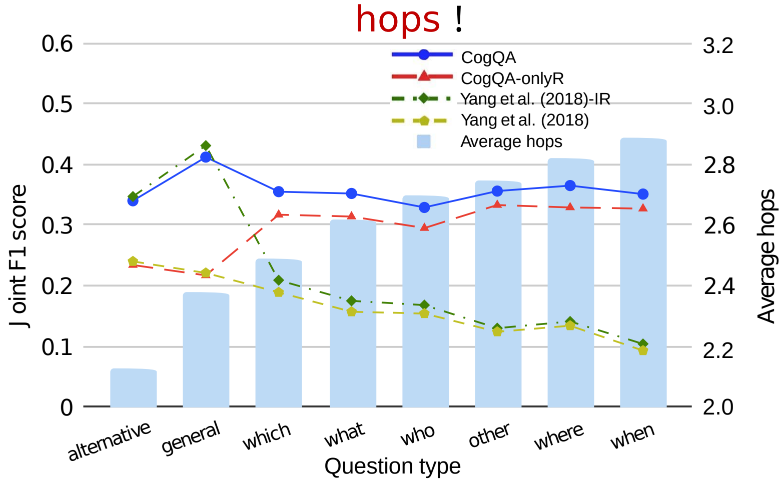
	Model	Ans EM	Ans F_1	Ans Prec	Ans Recall	Sup EM	Sup F_1	Sup Prec	Sup Recall	Joint EM	Joint F_1	Joint Prec	Joint Recall
Dev	Yang et al. (2018)	23.9	32.9	34.9	33.9	5.1	40.9	47.2	40.8	2.5	17.2	20.4	17.8
	Yang et al. (2018)-IR	24.6	34.0	35.7	34.8	10.9	49.3	52.5	52.1	5.2	21.1	22.7	23.2
	BERT	22.7	31.6	33.4	31.9	6.5	42.4	54.6	38.7	3.1	17.8	24.3	16.2
	CogQA-sys1	33.6	45.0	47.6	45.4	**23.7**	58.3	**67.3**	56.2	12.3	32.5	39.0	31.8
	CogQA-onlyR	34.6	46.2	48.8	46.7	14.7	48.2	56.4	47.7	8.3	29.9	36.2	30.1
	CogQA-onlyQ	30.7	40.4	42.9	40.7	23.4	49.9	56.5	48.5	**12.4**	30.1	35.2	29.9
	CogQA	**37.6**	**49.4**	**52.2**	**49.9**	23.1	**58.5**	64.3	**59.7**	12.2	**35.3**	**40.3**	**36.5**
Test	Yang et al. (2018)	24.0	32.9	-	-	3.86	37.7	-	-	1.9	16.2	-	-
	QFE	28.7	38.1	-	-	14.2	44.4	-	-	8.7	23.1	-	-
	DecompRC	30.0	40.7	-	-	N/A	N/A	-	-	N/A	N/A	-	-
	MultiQA	30.7	40.2	-	-	N/A	N/A	-	-	N/A	N/A	-	-
	GRN	27.3	36.5	-	-	12.2	48.8	-	-	7.4	23.6	-	-
	CogQA	**37.1**	**48.9**	-	-	**22.8**	**57.7**	-	-	**12.4**	**34.9**	-	-

Table 1: Results on HotpotQA (fullwiki setting). The test set is not public. The maintainer of HotpotQA only offers EM and F_1 for every submission. N/A means the model cannot find supporting facts.

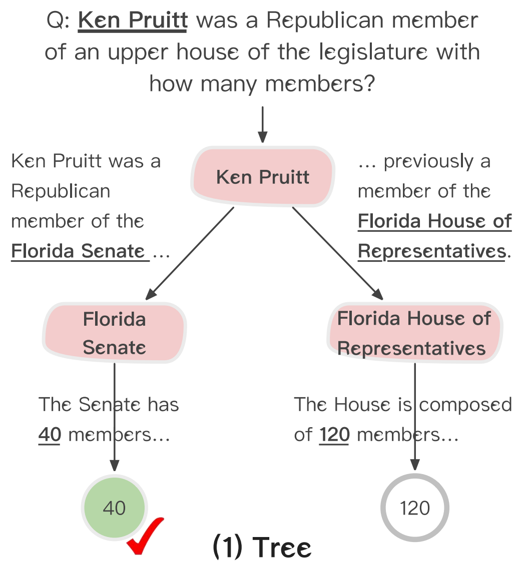
** Code available at https://github.com/THUDM/CogQA

Reasoning Power

CogQA Performs much better on question with more hops !

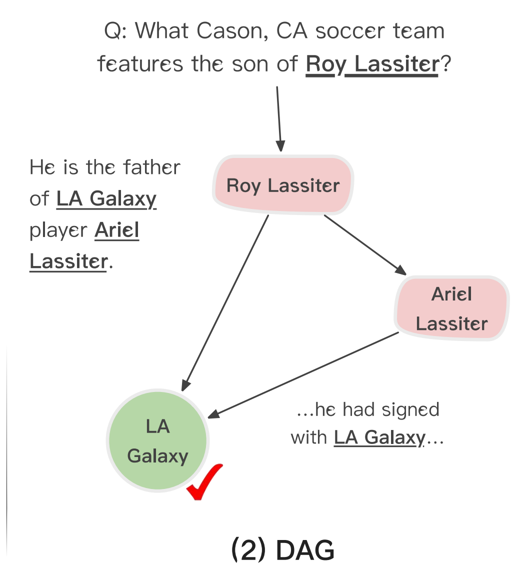


Case Study



- Tree-shape Cognitive Graph
- Users can verify the answer by comparing it with another possible reasoning chain.
- "*Upper House*" in the question is similar to "*Senate*" not "*House of Representative*"

Case Study



- DAG-shape Cognitive Graph
- Multiple supporting facts provides richer information, increasing the credibility of the answer.

Case Study

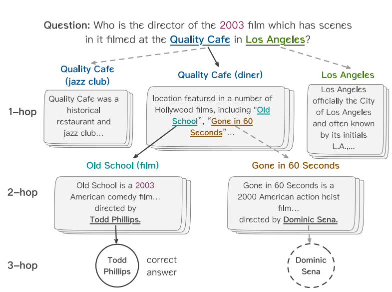
(3) Cyclic Graph

- CogQA gives the answer "Marijus Adomaitis" while the ground truth is "Ten Walls".
- By examining, Ten Walls is just the stage name of Marijus Adomaitis!
- Without cognitive graphs, black-box models cannot achieve it

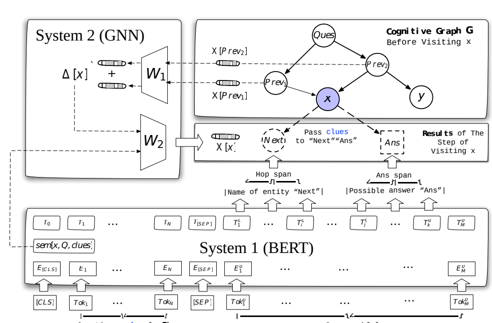
Summary

- Iterative Framework --> Myopic Retrieval
- Cognitive Graph --> Explainability
- Dual process theory --> System 2 Reasoning

Cognitive graph



CogQA framework



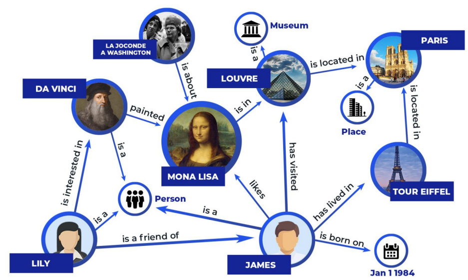
Applications:
—Knowledge Graph and Search

Cognitive Graph for Multi-hop Knowledge Graph Reasoning

Introduction

- Knowledge graphs (KGs), have been built in the last decades and nourished a wide range of downstream tasks.

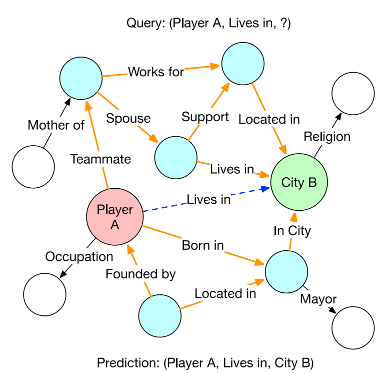
- A lot of important facts are missing in existing KGs, making it essential to enhance the ability to infer new facts.



Multi-hop KG Reasoning

- It is also important to provide explainable reasoning processes for prediction, so that humans can verify the inferred facts.

- The reasoning process often involves multiple entities and relations.

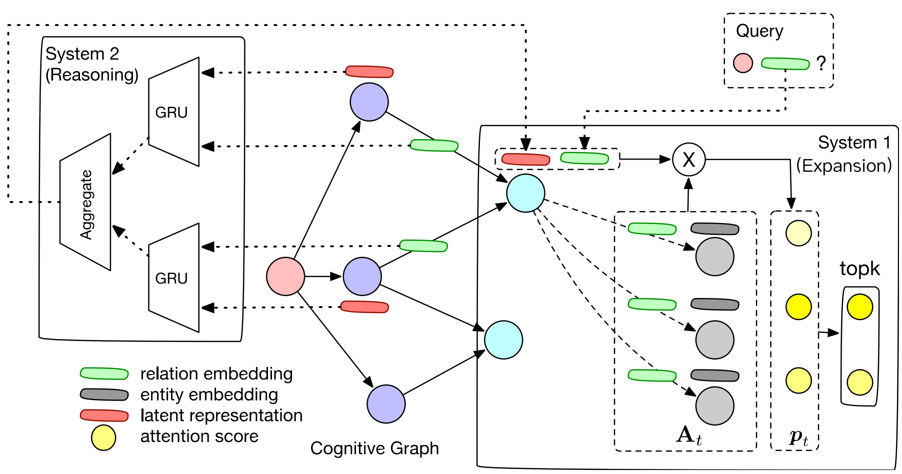


An example of multi-hop KG reasoning

Dual-process Theory

- The dual-process theory of cognitive science states that, the reasoning system of human beings consists of two distinct processes:
  - System 1: retrieve relevant information intuitively
  - System 2: reason over the collected information via a controllable, sequential, and logical reasoning process
- System 1 updates the working memory with the retrieved content, while System 2 operates on the content of the working memory.
- With inspiration from dual-process theory, we propose a new framework for multi-hop reasoning that can cope with more complex reasoning scenarios in the form of subgraphs.

Overview



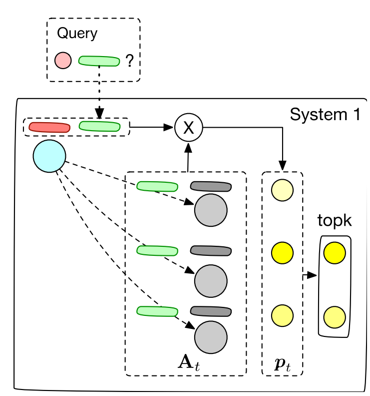
Our framework iteratively coordinates an expansion module and a reasoning module to set up a cognitive graph.

Cognitive Graph

- The cognitive graph is a subgraph of the original KG that contains entities and edges selected from G as relevant evidence and latent representations for its entities as the reasoning results.
- In the beginning, the cognitive graph only contains the initially given head entity. At each step, The entity and edge sets are expanded based on current involved nodes by System 1, and then the representations for the expanded nodes are updated by System 2

Expansion Module

- The expansion model constructs an attention flow on the KG to represent the current focus.
- At each step, the new attention distribution is computed from the neighborhood of the old attention. The top-k nodes in the attention distribution is added to the cognitive graph



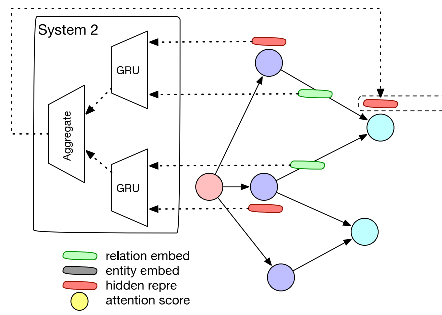
Reasoning Module

- The reasoning module updates the hidden representations of newly expanded nodes by neighborhood aggregation

$$\mathbf{X}[e] = \boldsymbol{U}(e, \boldsymbol{m}_e)$$
$$\boldsymbol{m}_e = \sum_{(e_k, r_k) \in E_e} \boldsymbol{M}(e_k, r_k, e).$$

- Inspired by the success of RNN-based models in path-based methods, we use GRU as the message function

$$\boldsymbol{M}(e_k, r_k, e) = \mathrm{GRU}(\mathbf{X}[e_k], \boldsymbol{v}_{r_k} \oplus \boldsymbol{v}_e)$$



Experiments

We evaluate CogKR on two public benchmarks for knowledge graph completion, against both embedding-based and path-based state-of-the-arts.

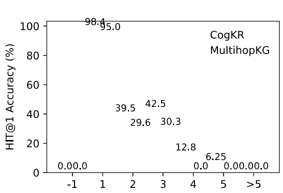
Model	FB15K-237				WN18RR			
	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR
DistMult [♠]	27.5	41.7	56.8	37.0	41.0	44.1	47.5	43.3
ComplEx [♠]	30.3	43.4	57.2	39.4	38.2	43.3	48.0	41.5
R-GCN	19.9	32.0	47.5	29.1	7.6	11.8	18.4	11.1
ConvE	31.3	45.7	60.0	41.0	40.3	45.2	51.9	43.8
RotatE	33.1	48.2	**63.7**	43.4	44.6	52.1	60.3	49.9
TuckER	33.6	46.7	60.7	42.7	45.5	50.2	53.5	48.5
NeuralLP [♠]	18.2	27.2	-	24.9	37.2	43.4	-	43.5
MINERVA[♡]	21.7	32.9	45.6	29.3	41.3	45.6	51.3	44.8
MultihopKG[♡]	32.9	-	54.4	39.3	43.7	-	54.2	47.2
M-Walk[♡]	16.5	24.3	-	23.2	41.4	44.5	-	43.7
DIVINE[♡]	22.3	33.1	-	29.6	-	-	-	-
CogKR	**36.9**	**49.2**	61.4	**44.9**	**48.4**	**54.3**	**60.6**	**52.3**

One-shot KG Reasoning

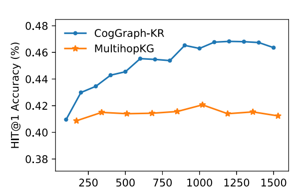
To further validate the effectiveness of CogKR, we evaluate the model on a more challenging task proposed recently: one-shot link prediction on KGs. In this task, we need to perform link prediction for relation types with only one training fact per relation.

	NELL-One				Wiki-One			
Model	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR
TransE	4.4	14.9	29.6	11.1	2.5	4.3	5.2	3.5
ComplEx	9.4	19.4	23.9	14.1	4.0	9.2	12.1	6.9
DistMult	12.3	23.1	26.9	16.3	1.9	7.0	10.1	4.8
MINERVA	16.0	26.1	29.0	20.6	21.4	25.6	26.6	23.3
GMatching	13.3	22.6	29.6	18.3	17.0	**27.3**	**33.5**	22.6
CogKR	**23.4**	**34.1**	**36.9**	**28.4**	**23.4**	26.5	27.0	**24.7**

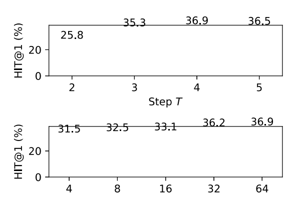
Quantitative Analysis



Multi-hop reasoning performance



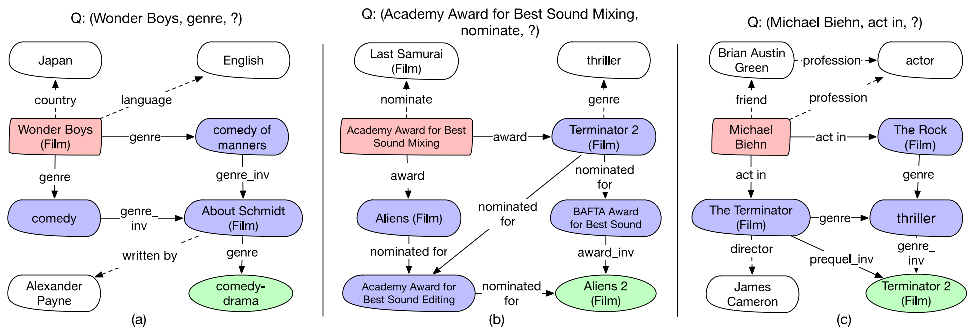
Convergence speed



Hyperparameter analysis

Case Study

We conduct a case study to show examples of cognitive graphs built by CogKR. We can observe from the examples: 1) the expressiveness of cognitive graphs is more powerful than individual paths. 2) Cognitive graphs can provide graphical explanations for prediction of CogKR.



Conclusion

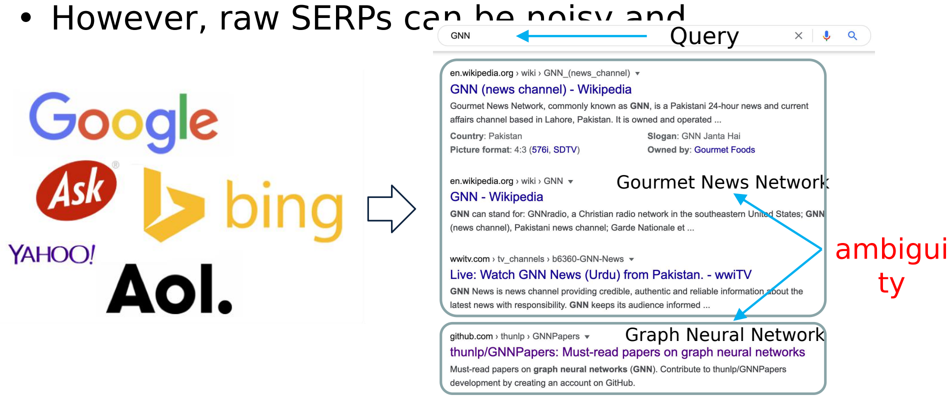
- We present CogKR, a novel framework to tackle the multi-hop KG reasoning problem.
- We organize the reasoning process with a cognitive graph, achieving more powerful reasoning ability than previous path-based methods and end-to-end training following gradient methods.
- Experimental results on both knowledge graph completion and one-shot link prediction benchmarks demonstrate the superiority of our framework.

CogWS: Improving Search Engine Result Pages with Reasoning

Tsinghua University

Introduction

- Search engines play a fundamental role for users to browse the Web as the number of Web pages increase.
- However, raw SERPs can be noisy and

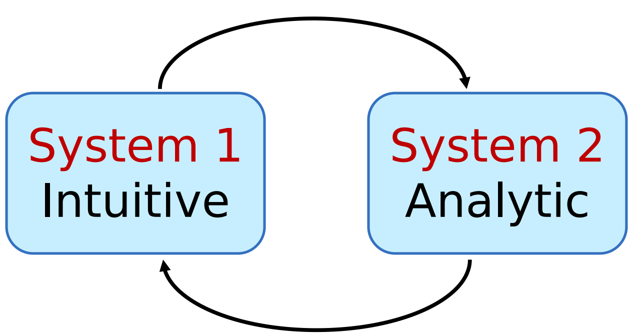


Motivation

A Coarse-to-Fine Process

- How do users refine search results?
  - Go over Web page snippets to understand the overall contents
  - Identify related pages through reasoning over Web page content and *prior knowledge.*
  - Extract informative clues about the intended entity and search for more related pages.

Connecting to Cognitive Science

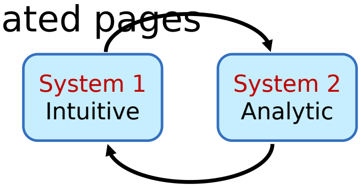


Dual Process Theory

1. Jonathan St BT Evans. 1984. Heuristic and analytic processes in reasoning. British Journal of Psychology, 75(4):451–468.

Connecting to Cognitive Science

- System 1:
  - Matching semantic representations of web pages to entities
- System 2:
  - Reasoning using semantic representations and relationships between web pages
  - Extracting clues from related pages



Our Solution: **Cog**nitive **W**eb **S**earch (CogWS)

Based on dual process theory, we propose a novel framework named CogWS:
- Utilizing system 1 for semantic understanding of each Web page
- Utilizing system 2 for relational reasoning and clues extraction
- This process is **iterative** in order to be complete.

Algorithm 1 Overall framework of CogWS

Input: name of the target entity v, prior knowledge of the target entity k, System 1 model S_1, System 2 model S_2, the total number of iterations N.
Output: the set of relevant Web pages R and collected clues K.

Initialize Q as a queue containing an empty word
$i \leftarrow 0$
$K \leftarrow \emptyset$
$R \leftarrow \emptyset$
while $i < N$ **do**
  $i \leftarrow i + 1$
  $clue \leftarrow Q$.pop()
  $key \leftarrow v \oplus clue$
  $C \leftarrow$ GoogleSearch(key)
  Construct cognitive graph $\mathcal{G}$ based on C
  Generate hidden representation $\mathbf{X}$ from S_1
  Update $\mathbf{X}$ using S_2
  Get relevant pages R_i from C based on $\mathbf{X}$
  $R \leftarrow R \cup R_i$
  Extract $clues_i$ from pages in R_i using S_2
  for $clue \in clues_i$ **do**
    if $clue \notin K$ **then**
      Q.append($clue$)
    end if
  end for
  $K \leftarrow K \cup clues_i$
end while
return R, K

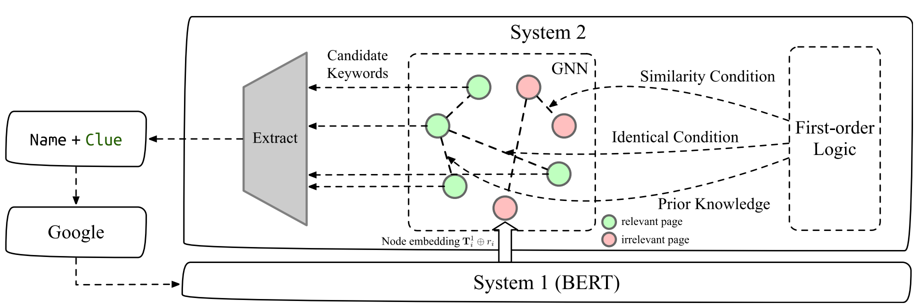
Two Systems

- System 1:
  - Extracting semantic representations using BERT with input $\texttt{[CLS]} \oplus t_i \oplus l_i \oplus s_i \oplus \texttt{[SEP]} \oplus v \oplus k$
- System 2:
  - Cognitive graph is constructed under first-order logic $\mathrm{sim}(S_i, S_j) \to \mathrm{rel}(p_i, p_j)$
    - Similarity condition $(\exists l)(l \in L(p_i)) \wedge \mathrm{equal}(l, l_j) \to \mathrm{rel}(p_i, p_j)$
    - Identical condition $(\exists u)(u \in I(p_i)) \wedge (u \in I(p_j)) \to \mathrm{rel}(p_i, p_j)$
    - Prior knowledge.
  - GNN for relational reasoning and classification
  - TF-IDF for clue extraction

$$O = \{D\} \bigcup \{T(p_i) \,|\, P(p_i \in R \mid \Theta, W, b) \leq \theta, p_i \in C\}$$

$$\mathrm{clues}_i = \{w \mid w \in D, \ \text{TF-IDF}(w, O) > \tau\}$$

Overall framework



Experiments

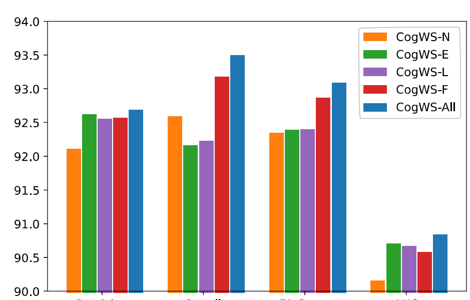
- Scholar Dataset
  - 2214 scholars randomly selected with their name and affiliation
  - 39413 web pages searched by scholar's name
  - Manually label whether the page is *relevant* to the given scholar.

Results

- Web page classification task

Methods	Prec.	Rec.	F1	AUC
Rule	84.77	32.35	46.83	–
SVM	89.29	90.66	89.97	85.79
DUET	88.85	92.85	90.80	86.20
MatchLSTM	88.53	92.72	90.58	86.34
ConvKNRM	86.87	89.85	88.33	81.82
DRMMTKS	81.92	92.00	86.67	74.99
ARCII	90.09	89.32	89.70	86.08
CogWS-sys1	92.11	92.59	92.35	90.16
CogWS	**92.69**	**93.50**	**93.09**	**90.84**

Performance comparison

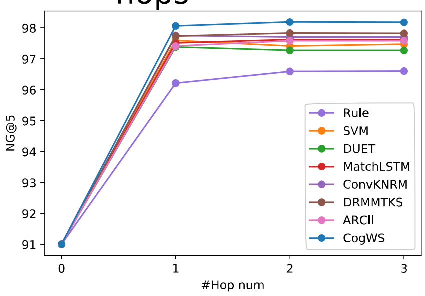


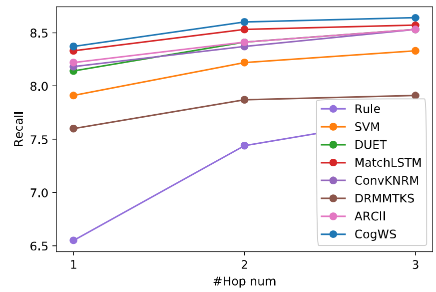
Edge feature contribution analysis

Reasoning do help!

Results

- Effectiveness of iterative retrieval
  - Divide the dataset in half, have the model view only the first half. Examining how the second half of the pages goes to the first half after multiple hops

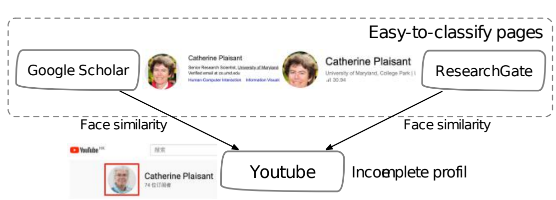




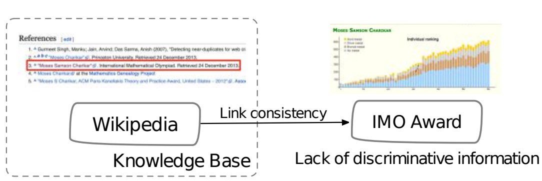
We get more relevant pages!

Case Study

- Reasoning for web page classification



(a) Reasoning based on prior knowledge



(b) Reasoning based on identical condition

Related Publications

For more, check http://keg.cs.tsinghua.edu.cn/jietang

- Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, and Jie Tang. Graph Random Neural Networks for Semi-Supervised Learning on Graphs. **NeurIPS'20**.
- Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. CogLTX: Applying BERT to Long Texts. **NeurIPS'20**.
- Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, and Jie Tang. Concentration Bounds for Co-occurrence Matrices of Markov Chains. **NeurIPS'20**.
- Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised Learning: Generative or Contrastive. https://arxiv.org/pdf/2006.08218.pdf
- Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. GCC: Graph Contrastive Coding for Structural Graph Representation Pre-Training. **KDD'20**.
- Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, and Jie Tang. Understanding Negative Sampling in Graph Representation Learning. **KDD'20**.
- Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable Multi-Interest Framework for Recommendation. **KDD'20**.
- Yuxiao Dong, Ziniu Hu, Kuansan Wang, Yizhou Sun and Jie Tang. Heterogeneous Network Representation Learning. **IJCAI'20**.
- Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. **ACL'19**.
- Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. ProNE: Fast and Scalable Network Representation Learning. **IJCAI'19**.
- Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou and Jie Tang. Representation Learning for Attributed Multiplex Heterogeneous Network. **KDD'19**.
- Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, and Kuansan Wang. OAG: Toward Linking Large-scale Heterogeneous Entity Graphs. **KDD'19**.
- Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou and Jie Tang. Towards Knowledge-Based Personalized Product Description Generation in E-commerce. **KDD'19**.
- Yifeng Zhao, Xiangwei Wang, Hongxia Yang, Le Song, and Jie Tang. Large Scale Evolving Graphs with Burst Detection. **IJCAI'19**.
- Yu Han, Jie Tang, and Qian Chen. Network Embedding under Partial Monitoring for Evolving Networks. **IJCAI'19**.
- Yifeng Zhao, Xiangwei Wang, Hongxia Yang, Le Song, and Jie Tang. Large Scale Evolving Graphs with Burst Detection. **IJCAI'19**.
- Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Chi Wang, Kuansan Wang, and Jie Tang. NetSMF: Large-Scale Network Embedding as Sparse Matrix Factorization. **WWW'19**.
- Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. DeepInf: Modeling Influence Locality in Large Social Networks. **KDD'18**.
- Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec. **WSDM'18**.

Thank you !

Collaborators:

Jie Zhang, Ming Ding, Jiezhong Qiu, Qibin Chen, Yifeng Zhao, Yukuo Cen, Yu Han, Fanjin Zhang, Xu Zou, Yan Wang, et al. (**THU**)

Hongxiao Yang, Chang Zhou, Le Song, Jingren Zhou, et al. (**Alibaba**)

Yuxiao Dong, Chi Wang, Hao Ma, Kuansan Wang (**Microsoft**)

Jie Tang, KEG, Tsinghua U — http://keg.cs.tsinghua.edu.cn/jietang

Download all data & Codes — https://keg.cs.tsinghua.edu.cn/cogdl/ — https://github.com/THUDM