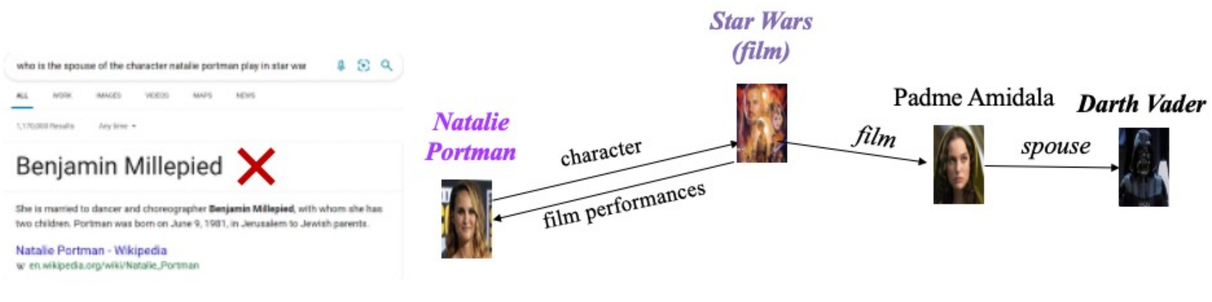
TIARA: Multi-grained Retrieval for Robust Question Answering over Large Knowledge Base

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F. Karlsson, Tingting Ma, Yuzhong Qu, Chin-Yew Lin

南京大學 NANJING UNIVERSITY · Microsoft · 南開大學 Nankai University · 哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

Motivation

who is *the husband of* the character natalie portman play in star wars?



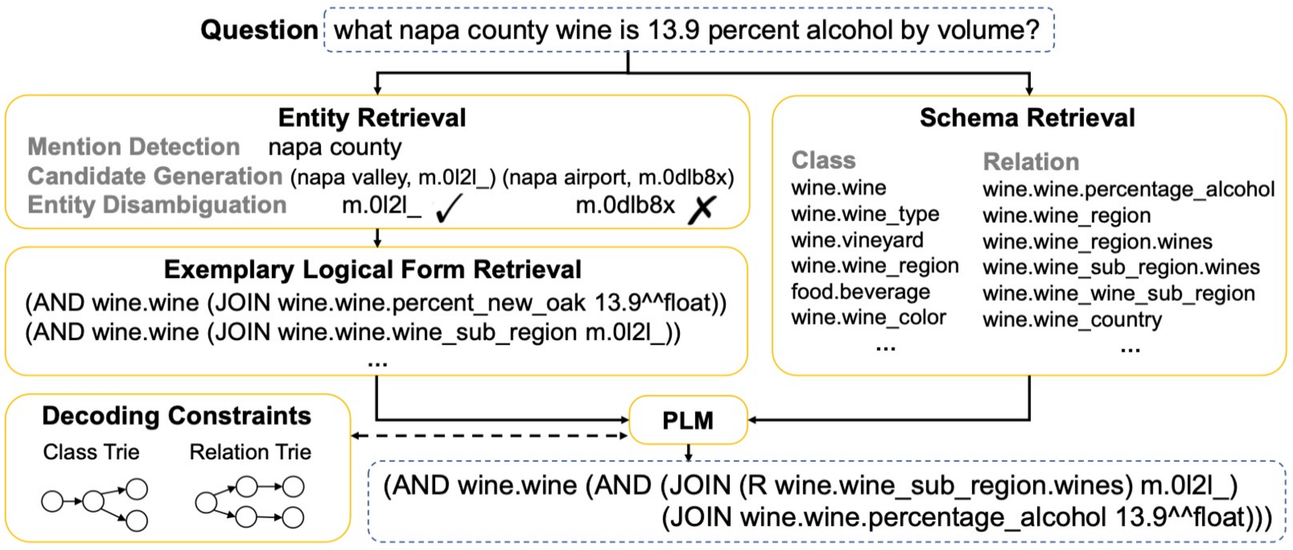
Search engine based on text vs. Reasoning over knowledge base

KBQA unique characteristics:
- Strong interpretability
- Abundant curated data
- Multi-hop and numerical reasoning

Semantic parsing-based KBQA: converts natural language questions into executable logical forms (e.g., s-expression, SPARQL)

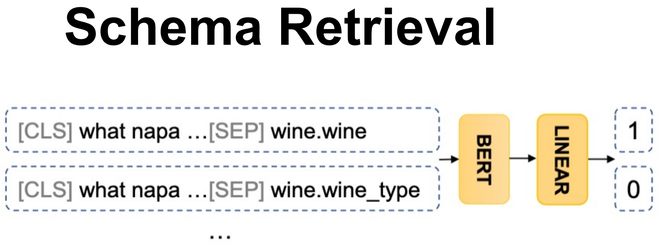
KBQA challenges:
- Question understanding (e.g., implicit relations & diverse functions)
- Large search space (e.g., Freebase has millions of entities)
- Robustness (e.g., compositional and zero-shot generalization)

Methods



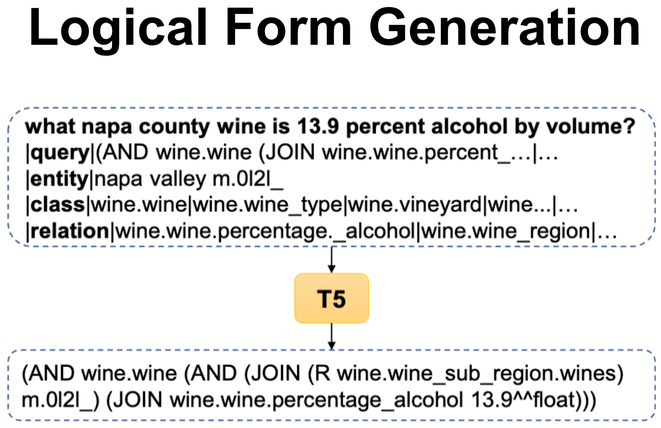
- Exemplary logical form retriever for **KB structures**
- Decouple the entity linker and schema retriever for **semantic supplement**
- Contexts for PLM: entities + top-5 LFs + top-10 classes/relations

Schema Retrieval



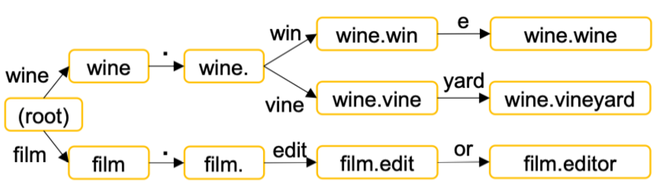
Schema retrieval learns if a question and a schema item are a match or not.

Logical Form Generation



Given a set of retrieved contexts, including entities, exemplary logical forms, classes, and relations, T5 generates the target logical form.

Constrained Decoding



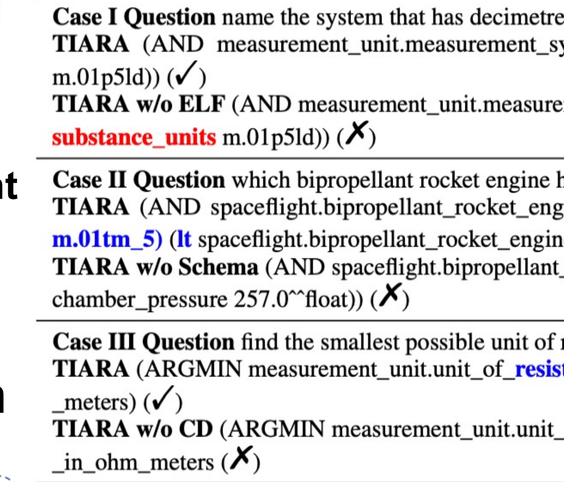
An example of a trie (prefix tree) that stores KB classes. Each edge represents a token that the PLM can select.

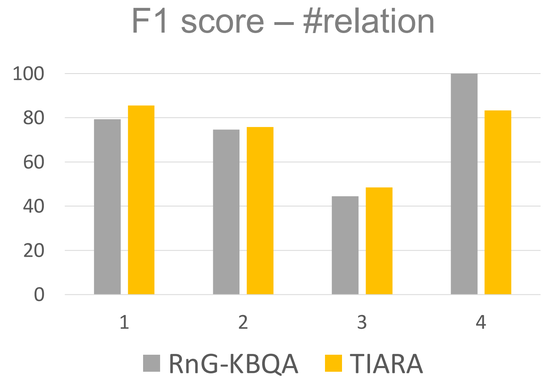
Experiment

Evaluation on two important benchmark GrailQA & WebQuestionsSP

Method	Overall EM	Overall F1	I.I.D. EM	I.I.D. F1	Compositional EM	Compositional F1	Zero-shot EM	Zero-shot F1
GloVe + TRANSDUCTION (Gu et al., 2021)	17.6	18.4	50.5	51.6	16.4	18.5	3.0	3.1
QGG (Lan and Jiang, 2020)	-	36.7	-	40.5	-	33.0	-	36.6
BERT + TRANSDUCTION (Gu et al., 2021)	33.3	36.8	51.8	53.9	31.0	36.0	25.7	29.3
GloVe + RANKING (Gu et al., 2021)	39.5	45.1	62.2	67.3	40.0	47.8	28.9	33.8
BERT + RANKING (Gu et al., 2021)	50.6	58.0	59.9	67.0	45.5	53.9	48.6	55.7
ReTraCk (Chen et al., 2021)	58.1	65.3	84.4	87.5	61.5	70.9	44.6	52.5
S^2QL (Zan et al., 2022)	57.5	66.2	65.1	72.9	54.7	64.7	55.1	63.6
ArcaneQA (Gu and Su, 2022)	63.8	73.7	85.6	88.9	65.8	75.3	52.9	66.0
RnG-KBQA (Ye et al., 2021)	68.8	74.4	86.2	89.0	63.8	71.2	63.0	69.2
TIARA (Ours)	**73.0**	**78.5**	**87.8**	**90.6**	**69.2**	**76.5**	**68.0**	**73.9**

- **SOTA** on both GrailQA & WebQSP (May 31, 2022).
- Performance improved on **all three generalization** levels.
- F1 higher than methods even with oracle entities.





Method	F1	Hits@1
IR-based methods		
EmbedKGQA* (Saxena et al., 2020)	-	66.6
GRAFT-Net (Sun et al., 2018)	62.8	67.8
PullNet (Sun et al., 2019)	-	68.1
TransferNet♡ (Shi et al., 2021)	-	71.4
Relation Learning♡♣ (Yan et al., 2021)	64.5	72.9
NSM*♡ (He et al., 2021)	67.4	74.3
Subgraph Retrieval* (Zhang et al., 2022)	74.5	**83.2**
SP-based (feature-based ranking) methods		
TextRay♡ (Bhutani et al., 2019)	60.3	-
Topic Units♡ (Lan et al., 2019)	67.9	-
UHop (Chen et al., 2019)	68.5	-
GrailQA RANKING*♡♣ (Gu et al., 2021)	70.0	-
STAGG♡ (Yih et al., 2016)	71.7	-
QGG♡ (Lan and Jiang, 2020)	74.0	-
SP-based (seq2seq generation) methods		
NSM♡ (Liang et al., 2017)	69.0	-
ReTraCk (Chen et al., 2021)	71.0	71.6
CBR-KBQA (Das et al., 2021)	72.8	-
ArcaneQA (Gu and Su, 2022)	75.6	-
RnG-KBQA (Ye et al., 2021)	75.6	-
Program Transfer*♣ (Cao et al., 2022b)	76.5	74.6
TIARA (Ours)	**76.7**	73.9
w/o Schema	76.4	73.7
w/o ELF	75.0	73.4
w/o ELF & Schema	73.2	71.1
TIARA*	**78.9**	75.2
w/o Schema	78.8	75.0
w/o ELF	76.2	74.5
w/o ELF & Schema	75.4	73.1

∗ denotes using oracle entity linking annotations. ♥ denotes the assumption of a fixed number of hops. ♣ denotes pre-training on an auxiliary task or other KBQA datasets.

Case I Question name the system that has decimetre as a measurement unit.
TIARA (AND measurement_unit.measurement_system (JOIN measurement_unit.measurement_system.**length_units** m.01p5ld)) (✓)
TIARA w/o ELF (AND measurement_unit.measurement_system (JOIN measurement_unit.measurement_system. **substance_units** m.01p5ld)) (✗)

Case II Question which bipropellant rocket engine has a chamber pressure of less than 257.0 and uses an oxidizer of lox?
TIARA (AND spaceflight.bipropellant_rocket_engine (AND **(JOIN spaceflight.bipropellant_rocket_engine.oxidizer m.01tm_5)** (**lt** spaceflight.bipropellant_rocket_engine.chamber_pressure 257.0^^float))) (✓)
TIARA w/o Schema (AND spaceflight.bipropellant_rocket_engine (**JOIN** spaceflight.bipropellant_rocket_engine. chamber_pressure 257.0^^float)) (✗)

Case III Question find the smallest possible unit of resistivity.
TIARA (ARGMIN measurement_unit.unit_of_**resistivity** measurement_unit.unit_of_resistivity.resistivity_in_ohm _meters) (✓)
TIARA w/o CD (ARGMIN measurement_unit.unit_of_**resistance_unit** measurement_unit.unit_of_resistivity.resistivity _in_ohm_meters (✗)

Case study of predicted logical forms by TIARA variants (without exemplary logical form retrieval, schema retrieval or constrained decoding). Errors are **red**, and correct parts are **blue**.

Conclusion

- Multi-grained Retriever is critical for the system robustness

Entity LF KB structure Schema semantic supplements

- Given enough contexts, PLMs can reason with high accuracy

Limitations

- Logical form retrieval is not efficient
- Strong supervision is required, which needs expensive annotations
- Gap between pre-training tasks and semantic parsing over KBs