Community Question Answering Entity Linking via Leveraging Auxiliary Data

Yuhan Li, Wei Shen*, Jianbo Gao, Yadong Wang

Nankai University, Tianjin, China

IJCAI 2022, Vienna, Austria

Bridging CQA with KBs

- Community Question Answering (CQA)
  - Quora, Zhihu, Stack Overflow, …
  - Posting questions
  - Seeking answers from other users
- Knowledge Bases (KBs)
  - Wikipedia, Wikidata, Freebase, …
  - Composed of entities and relations
  - Entity with unique identifier

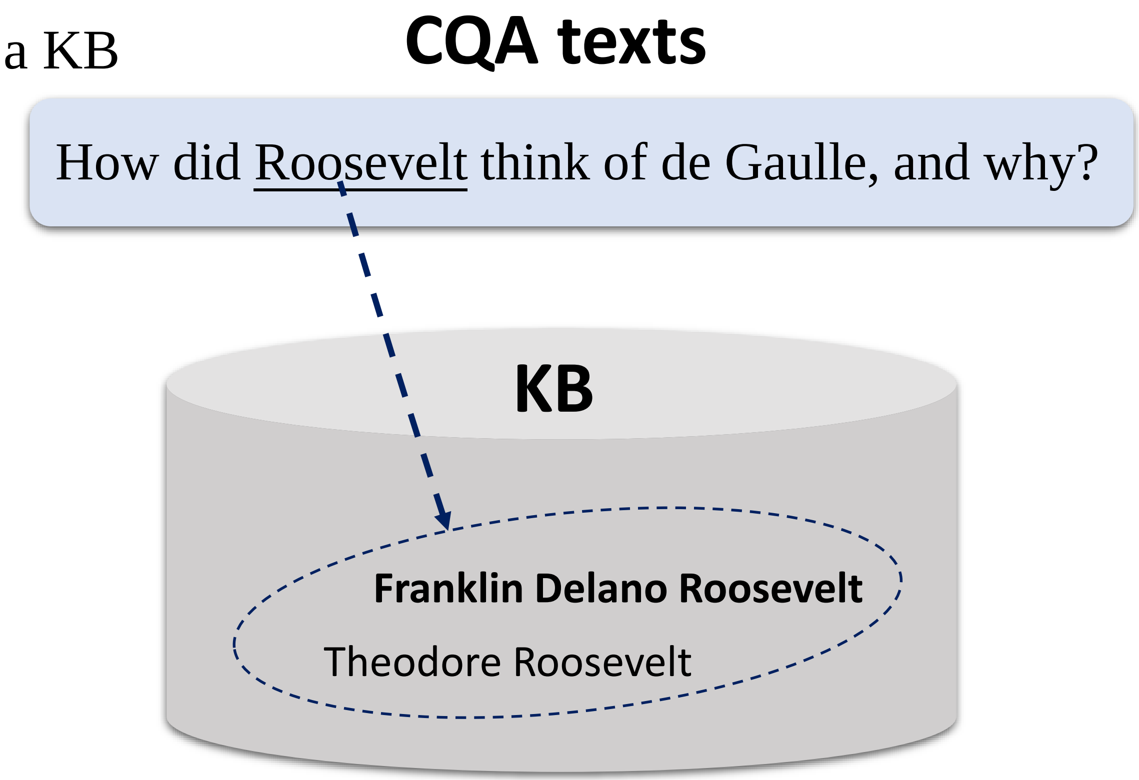
Quora 知乎 www.zhihu.com stackoverflow

Entity Linking

Wikipedia WIKIDATA freebase

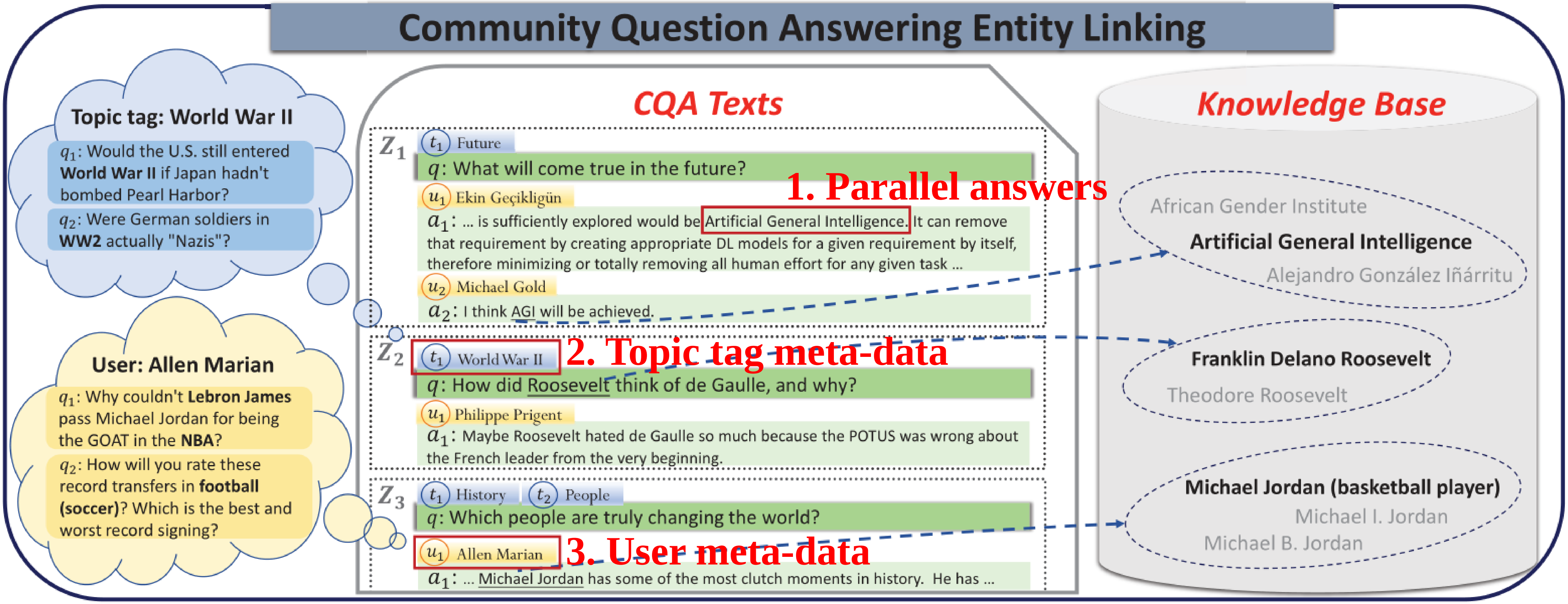
New task: CQA Entity Linking

- Definition
  - Linking textual entity mentions detected from CQA texts with their corresponding named entities in a KB
- CQA texts pose special Challenges
  - Concise and short
  - Informal
- Informative auxiliary data
  - Parallel answers
  - Two types of meta-data
    - ✓ Topic tags ACL'15 [1]
    - ✓ Users WWW'19 [2]



[1] Learning continuous word embedding with metadata for question retrieval in community question answering. Zhou et al. ACL'15.
[2] What we vote for? answer selection from user expertise view in community question answering. Lyu et al. WWW'19.

CQAEL via Leveraging Auxiliary data



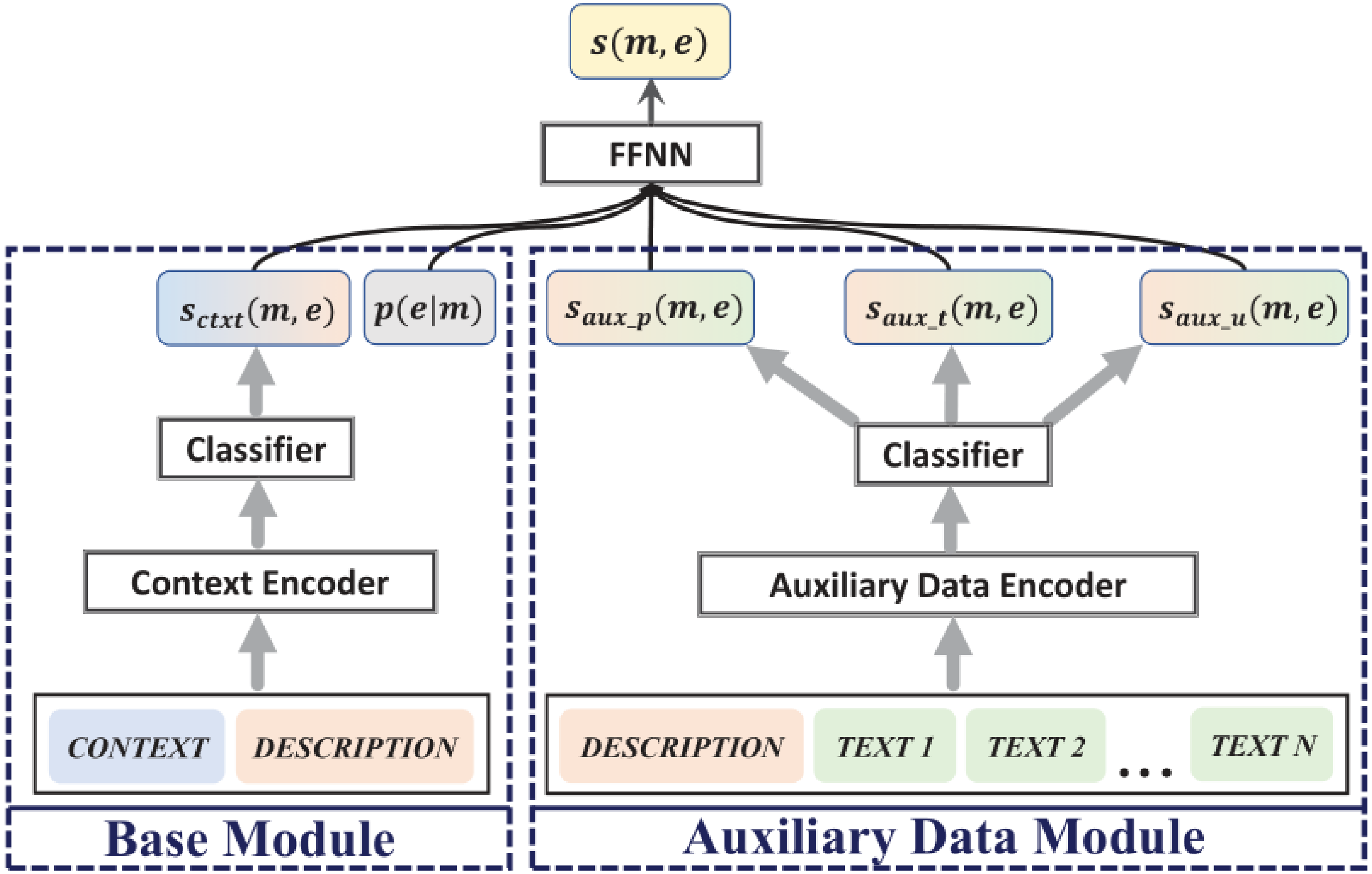
New dataset: QuoraEL

- CQA platform & Knowledge Base
  - Quora
  - Wikipedia
- Two-stage annotation
  - First: Stanford CoreNLP package
  - Second: Human annotators
- Statistics

# Total CQA texts	504
# Total entity mentions	8030
# Total answers	2192
# Total topic tags	1165
# Average entity mentions per CQA text	15.93
# Average answers per CQA text	4.35
# Average topic tags per CQA text	2.31
# Max questions per topic tag	10
# Max questions per user	20

The proposed framework

- Base module
  - Mention context & Entity description
  - XLNet [1]
- Auxiliary data module
  - Entity description & Different kinds of auxiliary data
  - Longformer [2]
- Combination
  - FFNN



[1] Xlnet: Generalized autoregressive pretraining for language understanding. Yang et al. NeurIPS'19.
[2] Longformer: The long-document transformer. Beltagy et al. arXiv'20.

Experiments

Models	Base Setting	Aux Setting
Deep-ED (2017)	82.56	82.97
Ment-Norm (2018)	82.99	83.19
Zeshel (2019)	88.72	88.91
REL (2020)	80.49	81.05
FGS2EE (2020)	82.59	83.07
BLINK (2020)	87.97	87.92
GENRE (2021)	86.26	87.06
Base Module (ours)	**89.37**	-
Full Module (ours)	-	**92.02**

Table 1: Effectiveness performance.

Models	Accuracy (%)	
	Total	△
Base Module	89.37	-
+ *Parallel answers*	91.55	+2.18
+ *User*	91.26	+1.89
+ *Topic*	91.77	+2.40
+ *User, Parallel answers*	91.61	+2.24
+ *Topic, User*	91.89	+2.52
+ *Topic, Parallel answers*	91.76	+2.39
Full Module	92.02	+2.65
Deep-ED [Ganea and Hofmann, 2017]	82.56	-
+ *Our Auxiliary Data Module*	88.16	+5.60
Zeshel [Logeswaran *et al.*, 2019]	88.72	-
+ *Our Auxiliary Data Module*	91.49	+2.77

Table 2: Ablation performance.

Conclusion

- A new task
- A new dataset
- A novel framework
  - Leveraging auxiliary data effectively
- A thorough experimental study
  - Outperform state-of-the-art baselines

Thanks for your watching!

More questions please email **yuhanli@mail.nankai.edu.cn**