Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, William W. Cohen

Carnegie Mellon University

'EMNLP 18'

Happy Buzaaba
UNIVERSITY OF TSUKUBA

Presentation Outline

- Introduction
- Representation of Text and Knowledge Base
- Embedding Propagation on graphs
- Results and Discussion

Open Domain Question Answering

- The task of answering factual questions posed in natural language
- Example:
  - Who voiced Meg in Family Guy?
    - Lacey Chabert
  - Which club did Christiano Ronaldo play for in 2011?
    - Real Madrid
- Open-domain: Need to search for the answer in Web-scale data

Structured and Unstructured Knowledge

Text

Cristiano Ronaldo

From Wikipedia, the free encyclopedia

Ahead of the 2009–10 season, Ronaldo joined Real Madrid for a world record transfer fee at the time, of £80 million (€94 million).[105] His co
€11 million per year and contained a €1 billion buy-out clause.[106] At least 80,000 fans attended his presentation at the Santiago Bernabéu
75,000 fans who had welcomed Diego Maradona at Napoli.[107] Since club captain Raúl already wore the number 7, the number Ronaldo wo
received the number 9 shirt,[109] which was presented to him by former Madrid player Alfredo Di Stéfano.[110]

Ronaldo playing against Chelsea in the Premier League during his third season in England

Ronaldo made his debut in La Liga on 29 August 2009, against Deportivo La Coruña, and sco
win.[111] He scored in each of his first four league fixtures with the club, the first Madrid player
club followed with two free kicks in the first group match against Zürich.[113] His strong start to
suffered an ankle injury in October while on international duty, which kept him sidelined for sev
received his first red card in Spain in a match against Almería.[116] Midway through the seasor
d'Or and the FIFA World Player of the Year award, behind Lionel Messi of Barcelona, Madrid's
goals in all competitions, including a hat-trick in a 4–1 win against Mallorca on 5 May 2010, his
season at Real Madrid ended trophyless.[119]

This article is part of a series about
Cristiano Ronaldo

Portuguese professional footballer

International goals · Career achievements · Comparisons to Lionel Messi

Eponyms

Cristiano Ronaldo Campus Futebol · Cristiano Ronaldo International Airport · Galaxy CR7 · Museu CR7

Films

Cristiano Ronaldo: The World at His Feet · *Ronaldo*

V · T · E

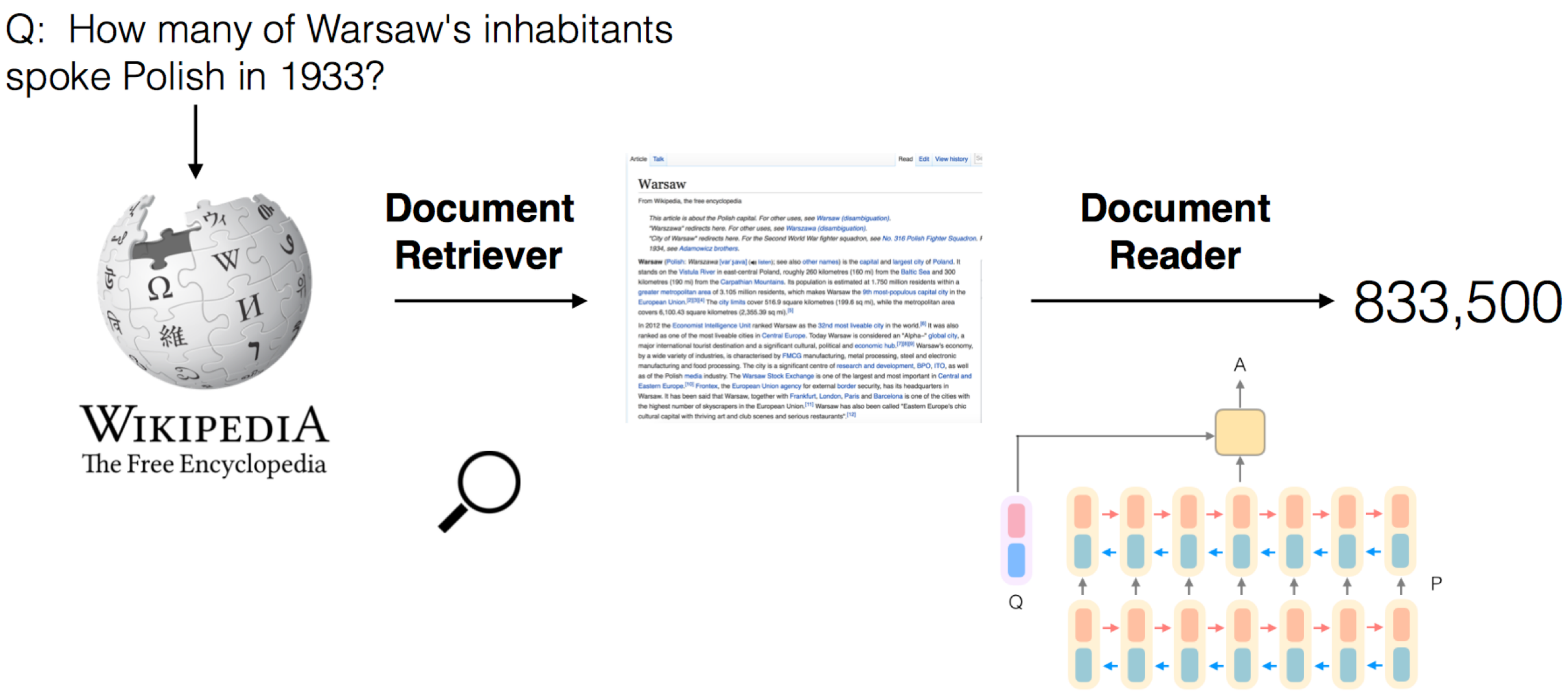
Figure

Real Madrid	2012–13[565]	La Liga	34	34	7	7	—
	2013–14[566]	La Liga	30	31	6	3	—
	2014–15[567]	La Liga	35	48	2	1	—
	2015–16[567]	La Liga	36	35	0	0	—
	2016–17[567]	La Liga	29	25	2	1	—

Table

Question answering with unstructured Knowledge (Text)

- Reading Comprehension:
  - DrQA: Reading wikipedia to answer Open Domain Questions (chen et al, ACL 2017)

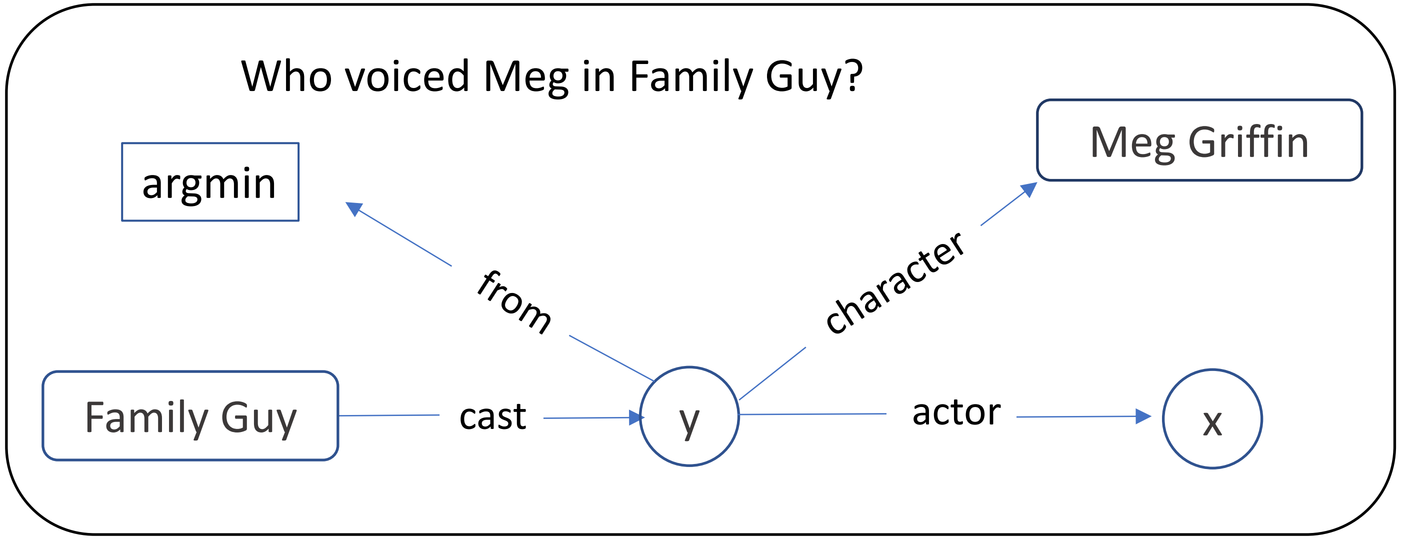


NB: High coverage of answers, but generally difficult to extract the answer in open-domain setting.

Question Answering with Structured Knowledge (KB)

- Knowledge Base
  - (Subject, relation, Object)
  - Example, (Meg_Griffin, voiced_by, Milla_Kunis)

- Semantic Parsing:
  - Example, Neural Symbolic Machine (Liang et al, ACL 2017)



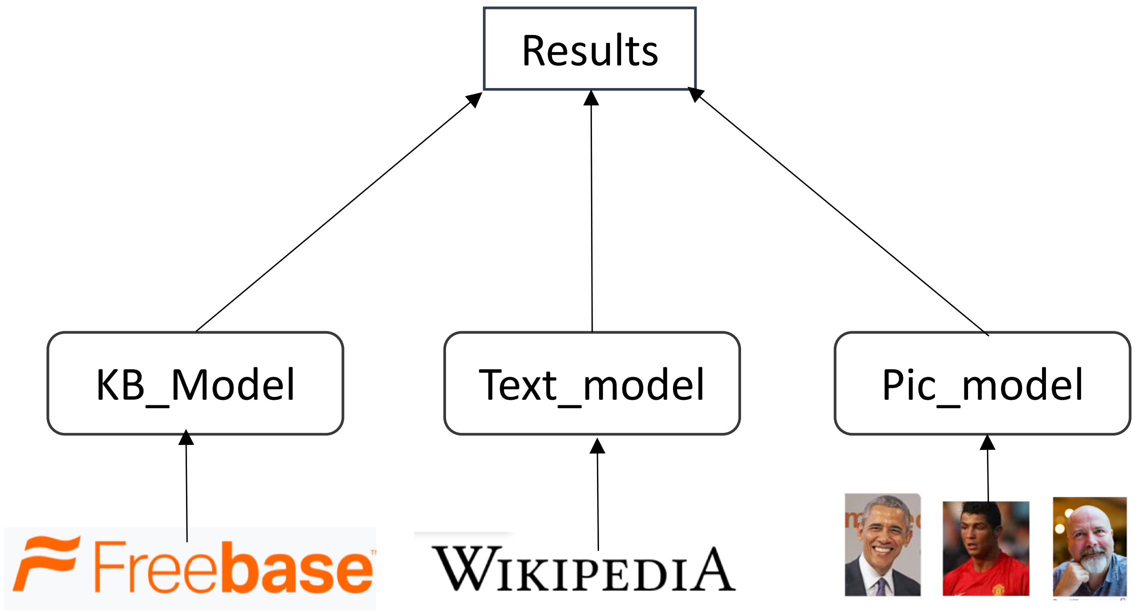
NB: High precision, but knowledge base is always incomplete.

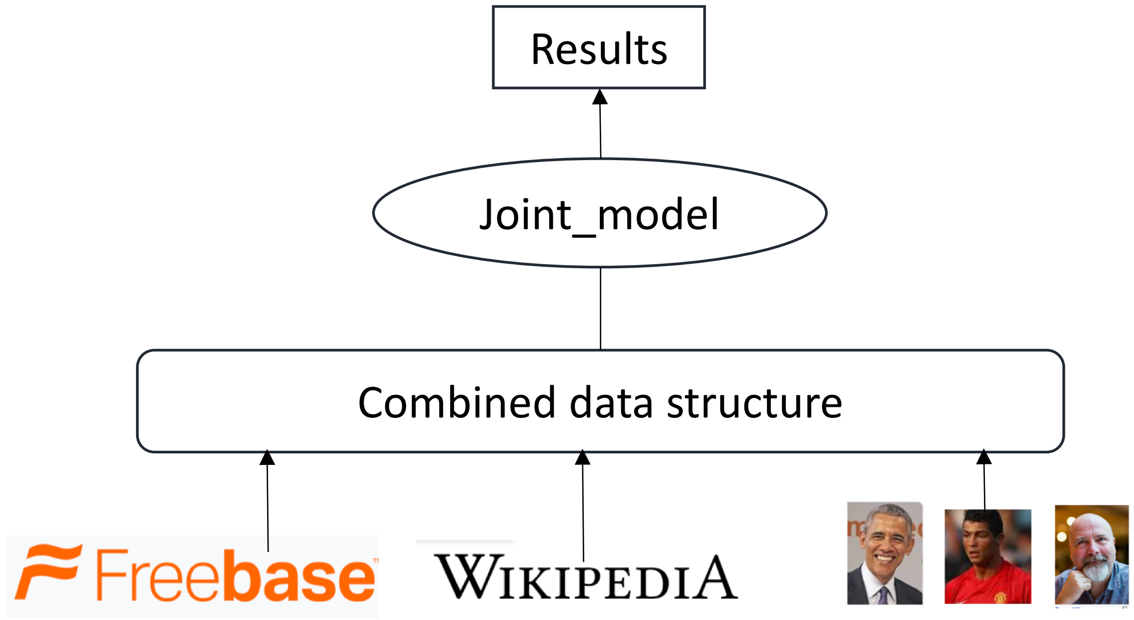
Question

Possibility of jointly using Structured and Unstructured Knowledge in QA

Question Answering with Structured and Unstructured Knowledge

- Option 1 Late Fusion
  - Train QA system on each Knowledge Source separately, and then ensemble the predictions
- Option 2: Early Fusion
  - Combine all knowledge sources into a single data structure





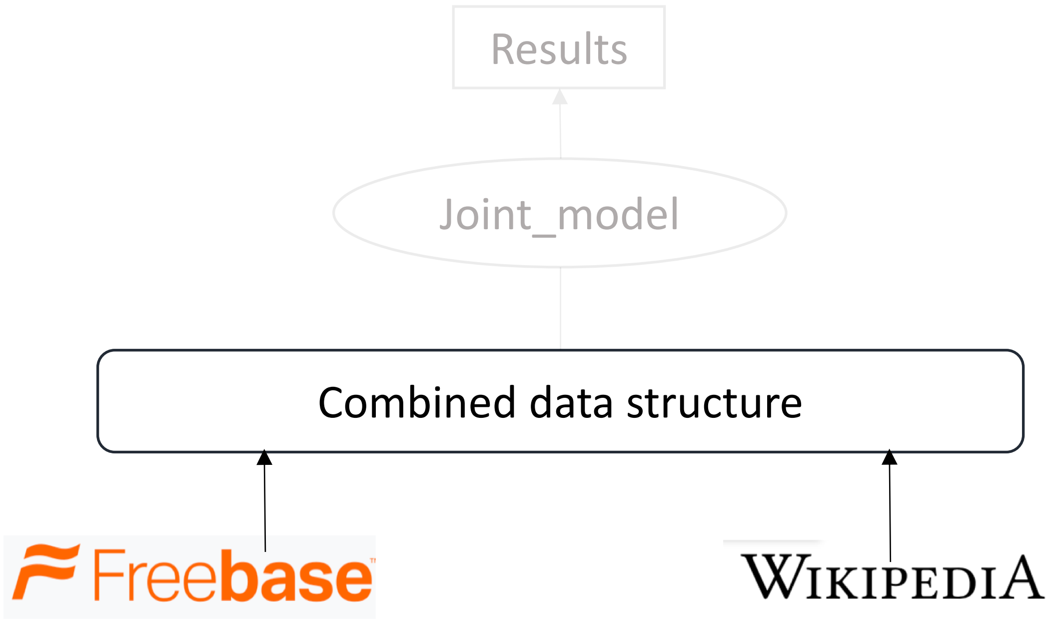
Proposed Model

❖ **GRAFT-Nets**

- Can we use multiple knowledge sources for question answering
  - Text and incomplete KB
- Is early fusion better than late fusion

Presentation Outline

- Introduction
- Representation of Text and Knowledge Base
- Embedding Propagation on graphs
- Results and Discussion

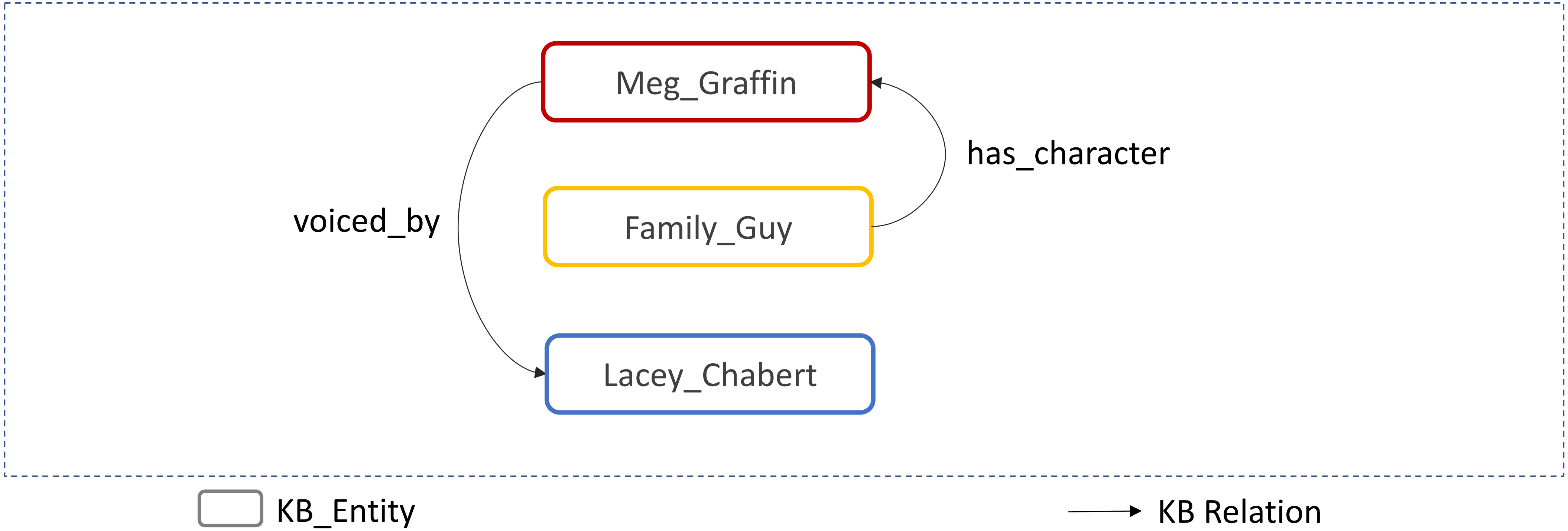


Representation of Text and Knowledge Base

❖ KB Facts:

(Meg_Graffin, Voiced_by, Lacey_Chabert)

(Family_Guy, has_character, Meg_Graffin)



Representation of Text and Knowledge Base

- KB Facts $\mathcal{K}$:
  - (Meg_Graffin, Voiced_by, Lacey_Chabert)
  - (Family_Guy, has_character, Meg_Graffin)
- Text $\mathcal{D}$:

  "Meg was originally voiced by Lacey Chabert during the first season ...."

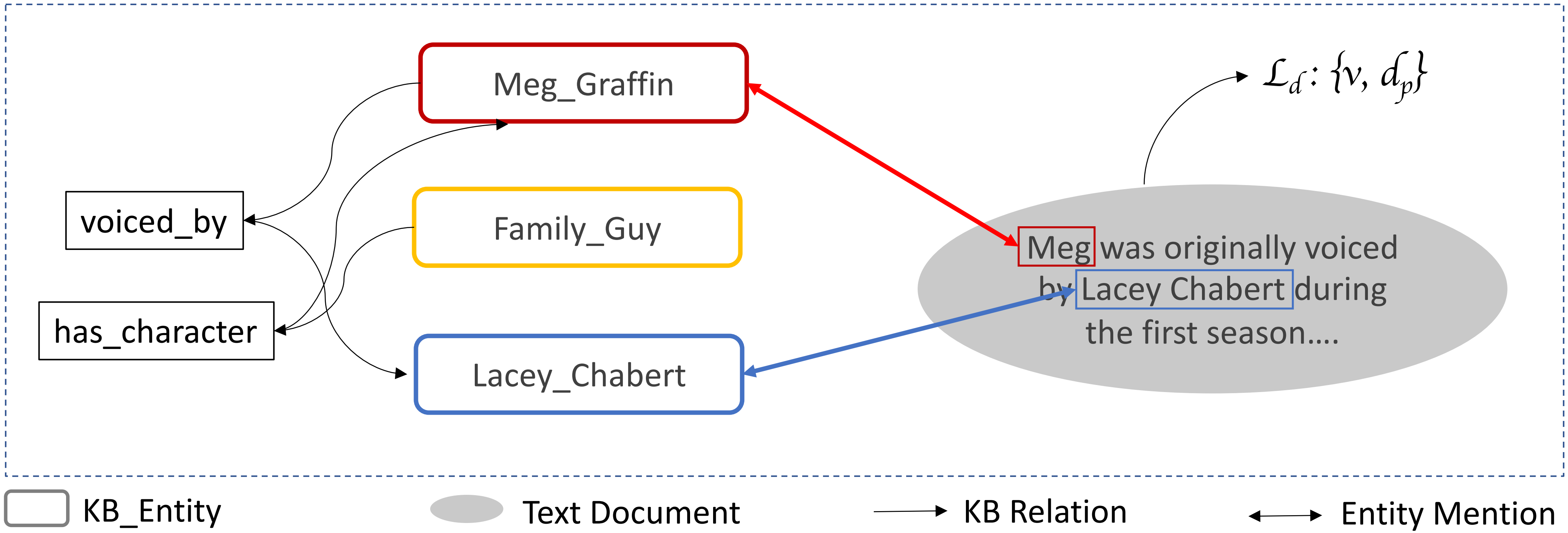
$\mathcal{K} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$

$\mathcal{V}$: set of entities,

$\mathcal{E}$: edges are triplets (s, r, o) where $r \in \mathcal{R}$ holds between $s \in \mathcal{V}$ and $o \in \mathcal{V}$

$\mathcal{D}$: $\{d_1, d_2, \ldots\ldots, d_{|D|}\}$

d : $(w_1, w_2, \ldots., w_{|di|})$



Representation of Text and Knowledge Base

Given NLQ: q = (w_1, w_2, …., $w_{|q|}$) extract an answer to the question $\{a_{|q|}\}$ from the $\mathcal{G} = (\mathcal{K}, \mathcal{D}, \mathcal{L})$

Steps for answer extraction:

1. Extracting subgraph $\mathcal{G}_q \in \mathcal{G}$ which contains an answer to the question to ensure high recall for answers

2. Use **Graft-net** to learn Node representation in the $\mathcal{G}_q$ conditioned to the question which are then used to classify the node as answer or not.

Extracting Sub-graph from graph

KB Retrieval:

1. Entity linking on question to produce seed entities S_q
2. Run **PPR** (Haveliwala, 2002) around S_q to identify top 50 entities which might be an answer to the question
3. Add the Top entities by **PPR** score and edges between them to the $\mathcal{G}_q$

Text Retrieval:

1. Retrieve top 5 most relevant Wikipedia articles using weighted bag of words model from (Chen et al 2017)
2. Populate **Lucene index** with sentences from the article and retrieve the top ranking documents based on words in the question.
3. Add the retrieved documents with entities linked to them to the $\mathcal{G}_q$

Sub-graph

Final Sub-graph becomes: $\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q, \mathcal{R}^+)$. Where;

$\mathcal{V}_q$ **:** All retrieved entities and Documents

$$\mathcal{V}_q = \{v_1, v_2, \ldots\ldots\ldots, v_E\} \cup \{d_1, d_2, \ldots\ldots\ldots, d_D\}$$

$\mathcal{E}_q$: All relations from the KB and entity links between documents and entities

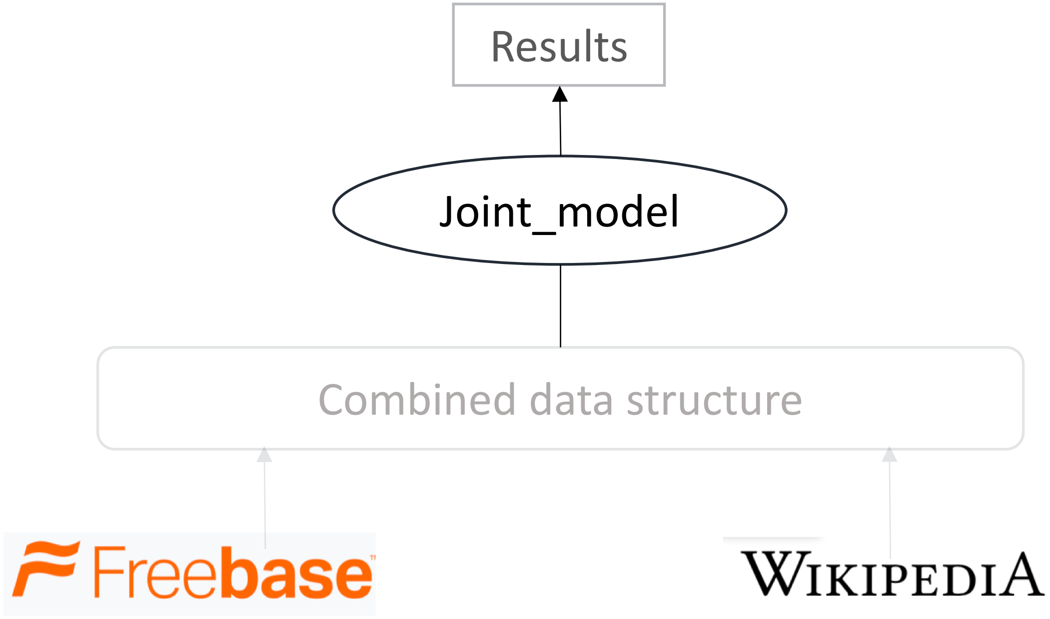
$$\mathcal{E}_q = \{(s, r, o) \in \mathcal{E} : s, o \in \mathcal{V}_q, r \in \mathcal{R}\} \cup \{(v, d_p, rL) : (v, d_p) \in \mathcal{L}_d, d \in \mathcal{V}_q\}$$

$\mathcal{R}^+$: Set of all edge types in the subgraph

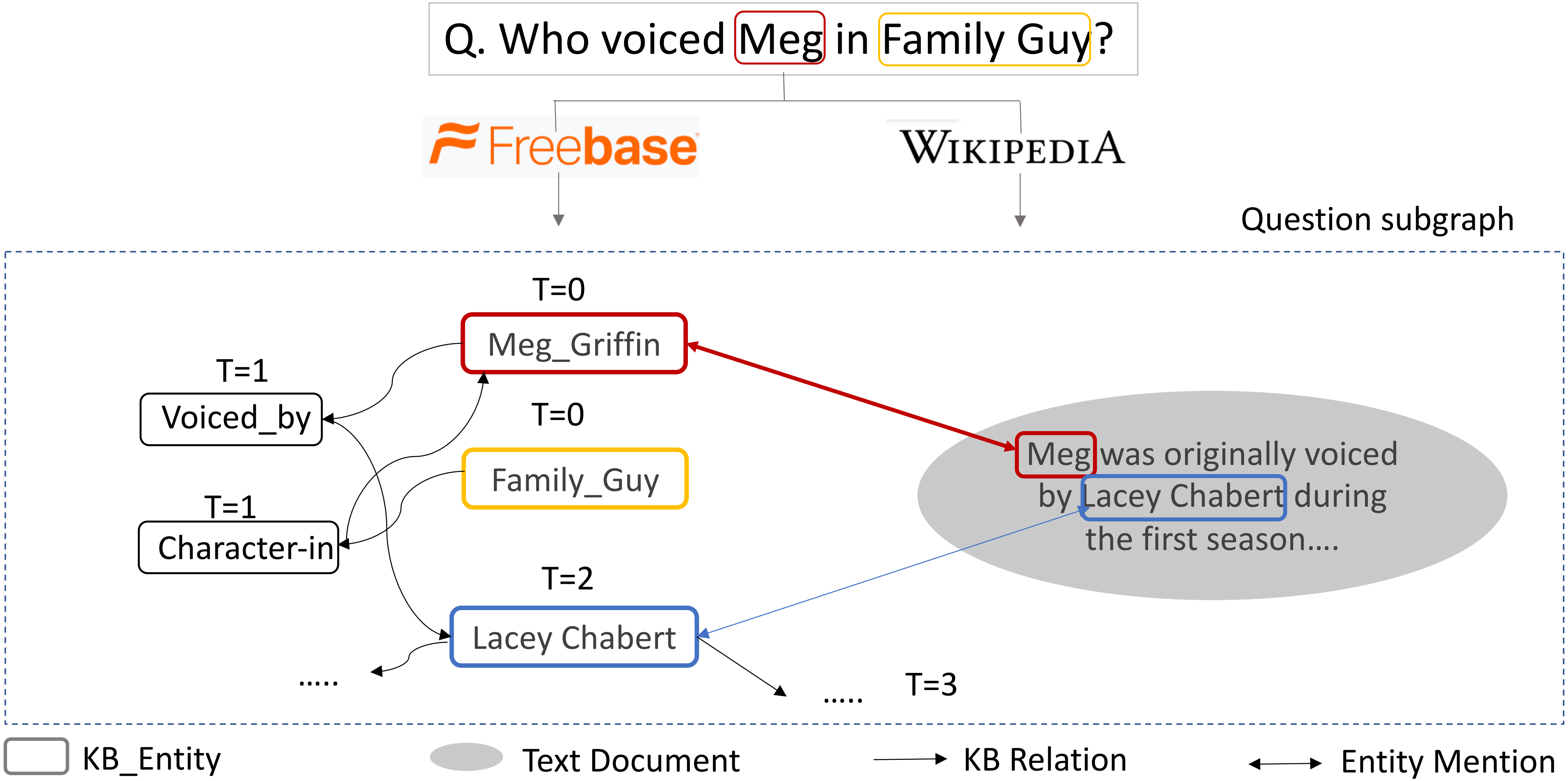
$$\mathcal{R}^+ = \mathcal{R} \cup \{rL\}$$

Presentation Outline

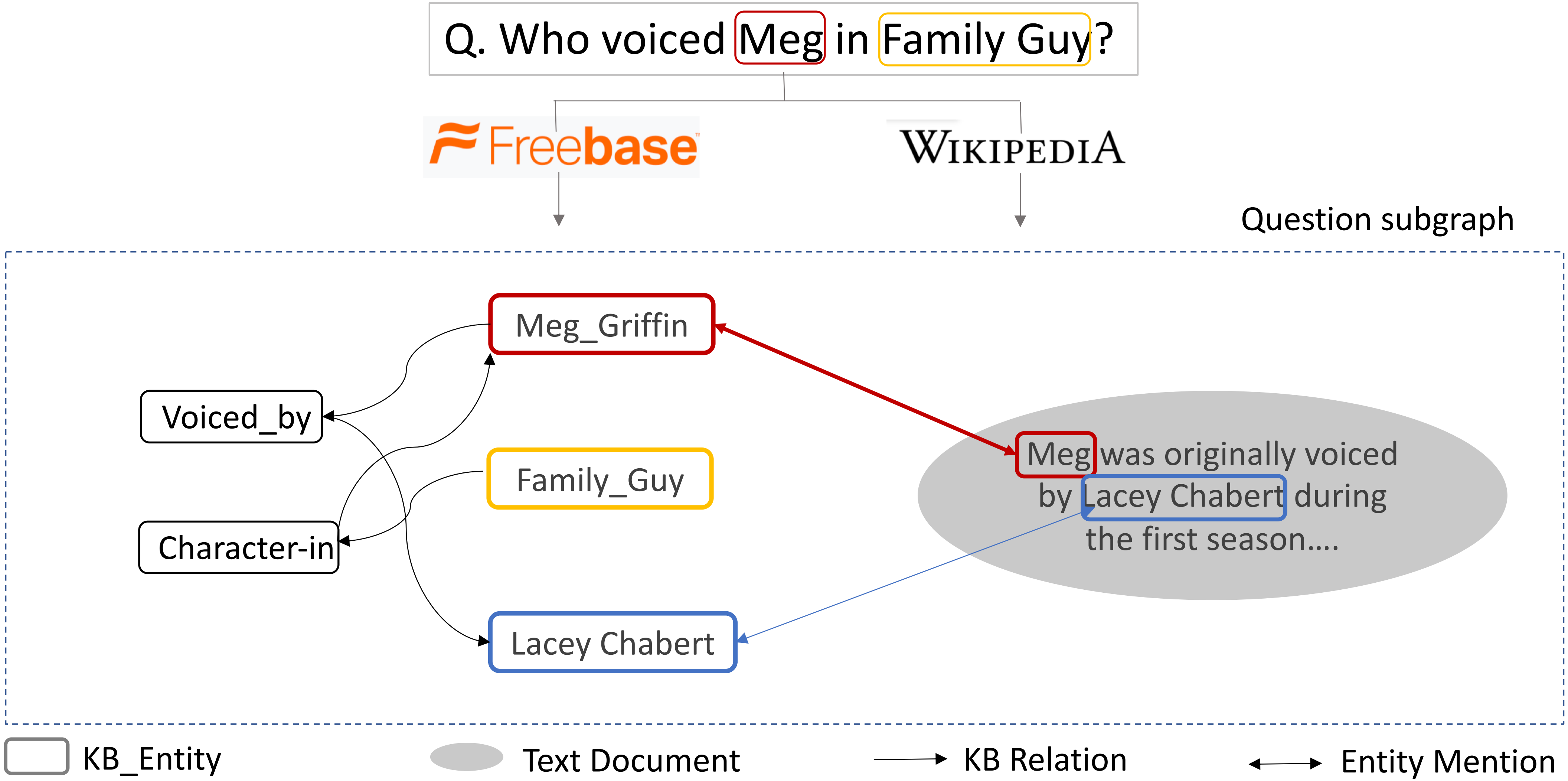
- Introduction
- Representation of Text and Knowledge Base
- Embedding Propagation on graphs
- Results and Discussion



Embedding Propagation



Fact Dropout



Presentation Outline

- Introduction
- Representation of Text and Knowledge Base
- Embedding Propagation on graphs
- Results and Discussion

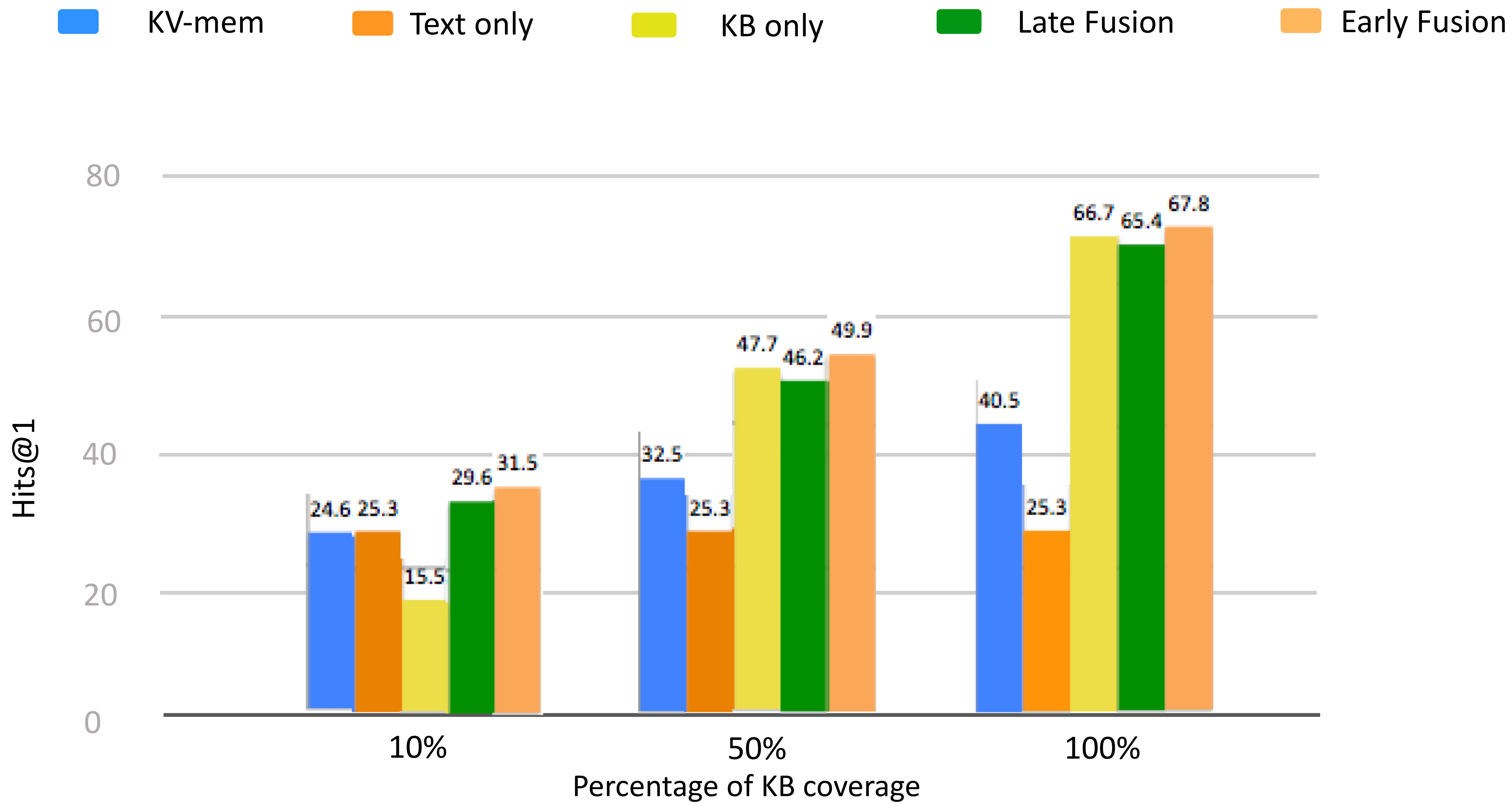
Experiments Datasets

- WikiMovies (Miller et al, 2016)
  - KB: OMDB – 43,233 entities
  - Text: Wikipedia – 79,728 documents
- WebQuestionSP (Yih et al, 2016)
  - KB: Freebase – 528,617
  - Text: Wikipedia – 235,567 documents
- Simulated datasets with incomplete KB:
- Uniformly sample KB: 10%, 50%

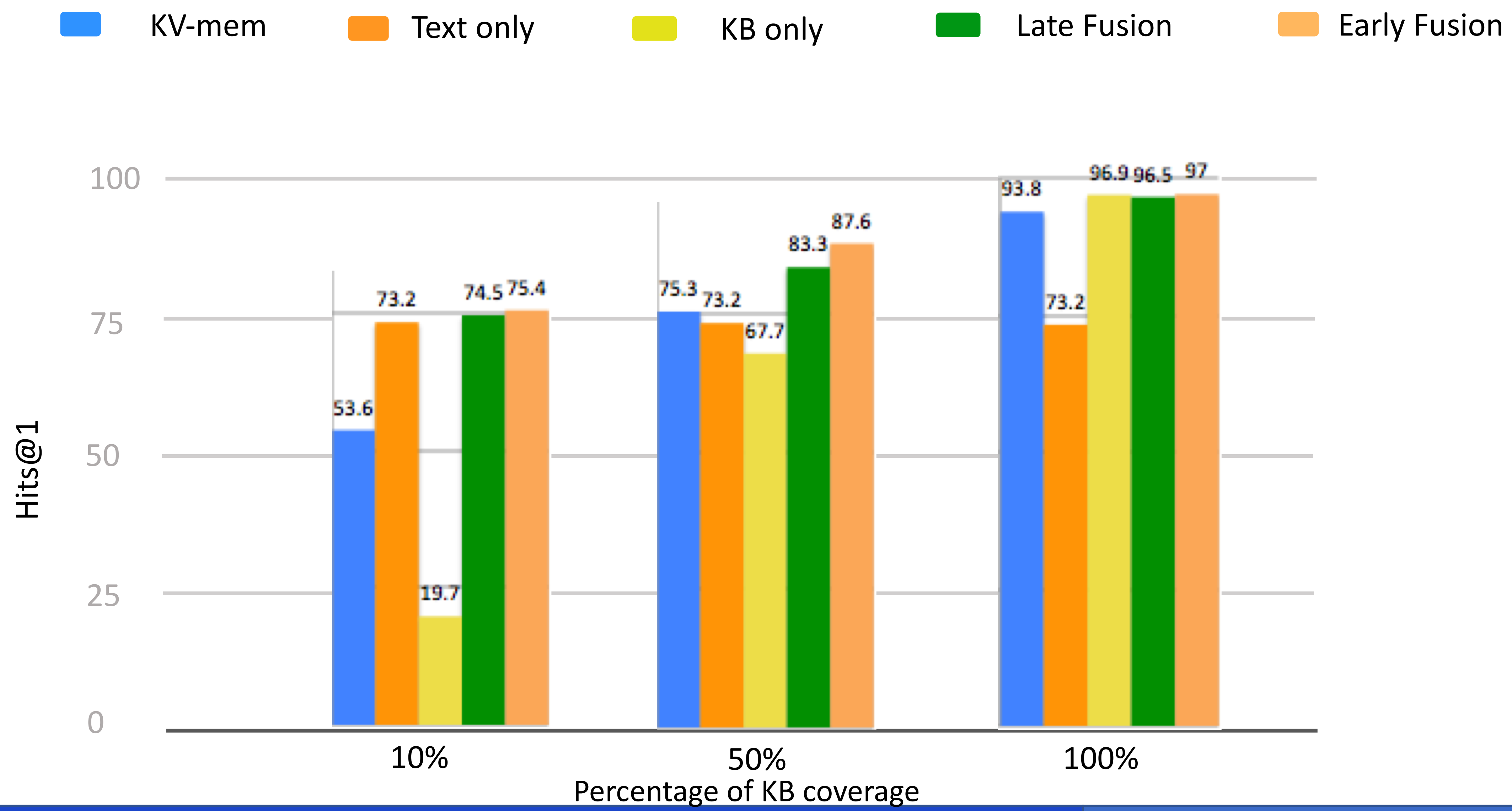
Experimental Questions

- Can we use multiple knowledge Sources (KB and Text) for QA?
- Is early fusion better than late fusion?
- Does Graph structure help?
  - Compare with Key-Value memory Network (Miller et al. 2016)

Results Web-Questions



Results WikiMovie



Conclusion and Future work

Conclusion

- Graft-net can effectively use KB and Text for Question Answering
- Graft-net with early fusion of KB and Text works better than late fusion

Future Work

- Extend Graft-net to pick spans of text as answers rather than only entities
- Improving subgraph retrieval process