Knowledge Base Refinement and Enhancement

Lei Chen
The Hong Kong University of Science and Technology
Outline

• **Background**
  • Overview of Current Challenges
    • Knowledge Base Construction
    • Knowledge Base Refinement
      • Completion (incompleteness)
      • Error detection (incorrectness)
  • Our Works
  • Conclusions
Background: What is a knowledge base?

- A knowledge base stores a collection of **facts** in the form of `<subject, relation, object>` triple

<table>
<thead>
<tr>
<th>Subject</th>
<th>Relation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Lennon</td>
<td>member Of</td>
<td>Beatles</td>
</tr>
<tr>
<td>John Lennon</td>
<td>born In</td>
<td>Liverpool</td>
</tr>
<tr>
<td>Beatles</td>
<td>founded In</td>
<td>Liverpool</td>
</tr>
</tbody>
</table>

**real-world entity**

**real-world entity**

or literal

John Lennon, the singer

born in

Liverpool, the city

member of

Beatles, the band
Background: Why knowledge bases?

- Entity searching, Question Answering, Recommendation...

John Lennon, the singer

born in
Liverpool, the city

found in
Beatles, the band

member of
Liverpool, United Kingdom
Background: Where do knowledge bases come from?

- Structured Text
  - Cyc, Freebase, Wikidata
- Semi-structured Data
  - DBpedia, Yago
- Unstructured Text
  - Nell, Knowledge Vault, Probase
Current Problems in Knowledge Bases

• **Incompleteness**
  • Most KBs can hardly cover all the entities in the real world

• **Incorrectness**
  • There may be errors during KB construction

• **Domain-specific information**
  • It is hard to build a complete and sound KB in certain domains
  • It is hard to involve subjective information in KBs
Summary: Our Recent Works about Knowledge Bases

- Canonicalization of Open KB triples
- Knowledge Fusion
- Subjective Knowledge
- Negative Samples Generation in Knowledge Base Reasoning
- Knowledge Base Completion with Crowd

Data sources

- W₁: <John W. Lennon, be born in, Liverpool>
- W₂: <John W. Lennon, be born in, London>
- W₃: <John W. Lennon, be born in, 1941>
- W₄: <John W. Lennon, be born in, 1940>

Facts

- John Lennon, famous singer
- Liverpool, safe city
- Beatles, popular band

Work with

- Paul McCartney, the singer
- John Lennon, member of
- Liverpool, founded in
- Beatles, member of

Subjective Knowledge

- John Lennon, member of Beatles
Our Recent Works about Knowledge Bases

Work: Canonicalization of Open Knowledge Bases

• Our proposed method: [Lin, ICDE 2019]
  • Model the canonicalization problem of noun phrases and relation phrases in the Open KB triples **jointly**
  • Utilize the **side information** obtained from the source text of the triples

< Peter Cook, hosted, “Not Only… But Also” >

**Peter Cook** hosted The TV series “Not Only ... But Also”. **John Lennon**, born in **Liverpool**, one of the members in **Beatles**, had a pleasant cooperation with **Cook** in this show.

• **Candidate Entities in the Source Text**
  - **Type Lists of These Entities**
  - **The Domain Vector of the Source Text**
Our Recent Works about Knowledge Bases

Work: Subjective Knowledge Base Construction

- Our proposed method: [Xin, SIGMOD 2018]

Input: Existing KBs + Crowd
Goal: Enrich KB with subjective information

Our Recent Works about Knowledge Bases

Work: Knowledge Base Reasoning [Zhang, ICDE 2019]
Goal: Generate negative samples during knowledge base reasoning

Positive samples...
Negative samples...

John Lennon, the singer
Paul McCartney, the singer

born in
Liverpool, the city

founded in
Beatles, the band

work with
member of

member of ?

Our Recent Works about Knowledge Bases

Work: Knowledge Base Enhancement [Jiang, ICDE 2018]
Goal: Update the current KBs with new facts with the help of the crowd

John Lennon, the singer, born in Liverpool, the city, founded the band Beatles, the band, member of Paul McCartney, the singer.

Our Recent Works about Knowledge Bases

Work: Knowledge Fusion [Lin. VLD 2018]

Input: Different sources providing conflicting facts
Goal: Discover trustworthy information (i.e., the truths)

Web sites Facts

\[
\begin{align*}
W_1 &: < \text{John W. Lennon, be born in, Liverpool}> & \checkmark \\
W_2 &: < \text{John W. Lennon, be born in, London}> & \times \\
W_3 &: < \text{John W. Lennon, be born in, 1941}> & \times \\
W_4 &: < \text{John W. Lennon, be born in, 1940}> & \checkmark \\
\end{align*}
\]

Conclusions

• Knowledge base construction needs to be improved
  • Entity Extraction and Entity Resolution
  • Relation Extraction
• Knowledge bases need refinement and enhancement (internal methods vs external methods)
  • Incompleteness
    • Hardly can cover all the entities/relations in the real world
    • How to add new knowledge, e.g., subjective information, to the current KB?
• Incorrectness
  • May have errors when constructing the KB
Future Works

• **Possible research directions:**
  • Knowledge Graph Embedding (for Knowledge Reasoning in KB Completion and Error Detection)
  • Knowledge Fusion
  • Open Information Extraction Refinement
  • Knowledge Base Refinement via Human Power
References

• [Zhang, ICDE 2020], Quanming Yao, Wenyuan Dai, Lei Chen. “AutoSF: Searching Scoring Functions for Knowledge Graph Embedding.”, In 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020