Formal Query Generation for Question Answering over Knowledge Bases

Hamid Zafar *, Giulio Napolitano **, Jens Lehmann *,**

* SDA group, University of Bonn, Germany

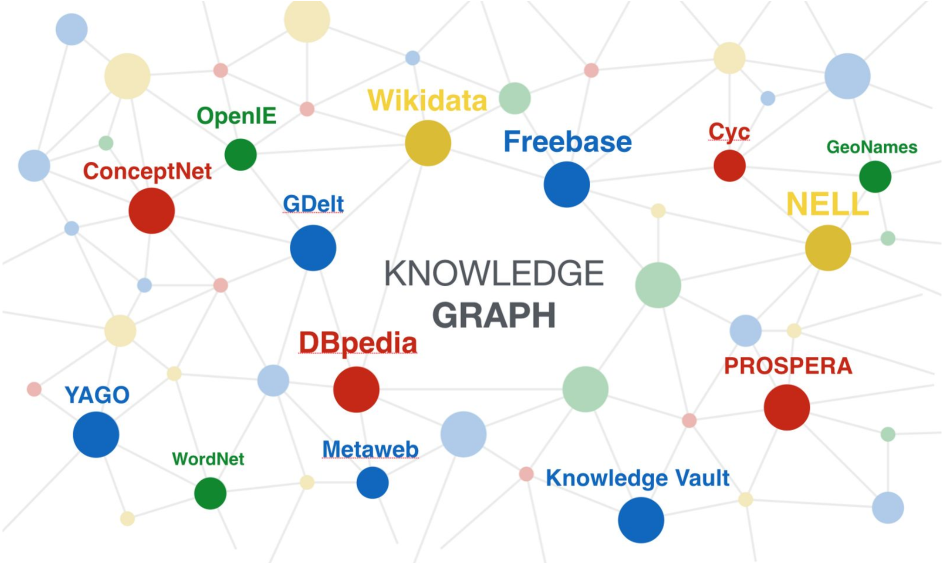
** Fraunhofer IAIS, Germany

Agenda

- Introduction
- SPARQL Query Generator (SQG)
  - Capture subgraph
  - Find valid walks
  - Rank queries
- Empirical results
- Summary

Introduction

Question answering over Knowledge graphs



Introduction

Transform question posed in natural language to a formal language

What are some artists on the show whose opening theme is Send It On?

```
SELECT DISTINCT ?uri WHERE {
?x <http://dbpedia.org/ontology/openingTheme> <http://dbpedia.org/resource/Send_It_On> .
?x <http://dbpedia.org/property/artist> ?uri .
?x <https://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://dbpedia.org/ontology/TelevisionShow>}
```

Common Architectures

End-to-End

- Single process

+ No error propagation
- Limited support for complex questions

Pipeline

- Consists of multiple components including
  - Named Entity Disambiguation
  - Relation Extraction
  - Query Generation (QG)

+ Reusable components
+ Limited focus
- Propagate the error along the pipeline

Pipeline Architecture

Query Generation Component

- The **Query Generation** is a common components in QA systems
- Error analysis from [4] showed the importance of the **Query Generation** and its effect on the overall performance of the QA pipeline

Requirements of Query Generation

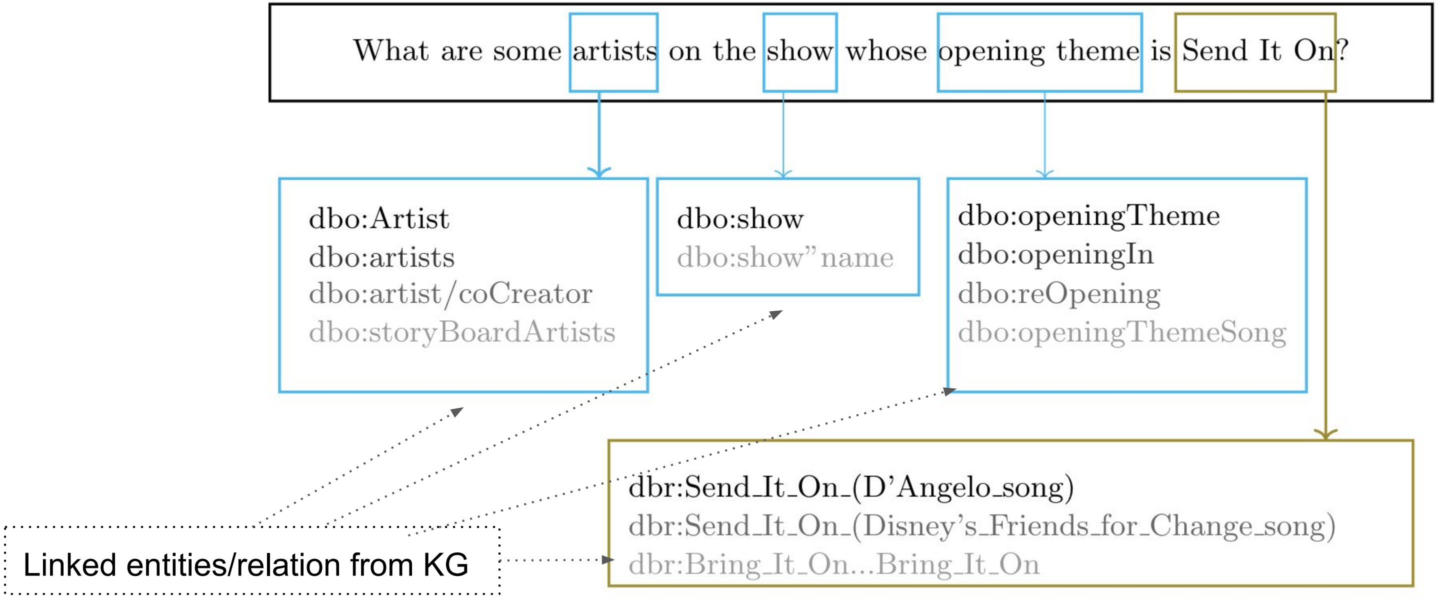
- Cope with large-scale KGs
- Ability to manage noisy input to handle error propagation
- Question type identification
- Support for composite question
- Syntactic ambiguity of the input question

SPARQL Query Generation (SQG)

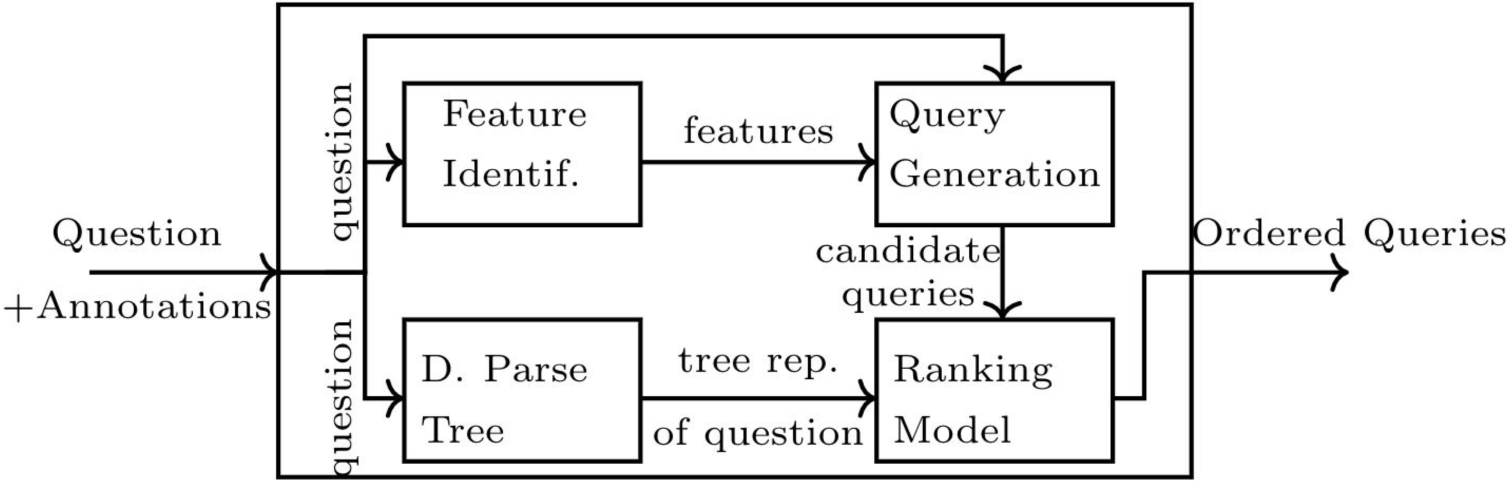
- **Hypothesis:** The formal interpretation of the question is a walk in the KG which contains the target entities and relations of the input questions plus the answer node.

- **Inputs:** Question along with the linked entities and relations

Inputs


What are some artists on the show whose opening theme is Send It On?
dbo:Artist
dbo:artists
dbo:artist/coCreator
dbo:storyBoardArtists
dbo:show
dbo:show"name
dbo:openingTheme
dbo:openingIn
dbo:reOpening
dbo:openingThemeSong
dbr:Send_It_On_(D'Angelo_song)
dbr:Send_It_On_(Disney's_Friends_for_Change_song)
dbr:Bring_It_On...Bring_It_On
Linked entities/relation from KG

Architecture

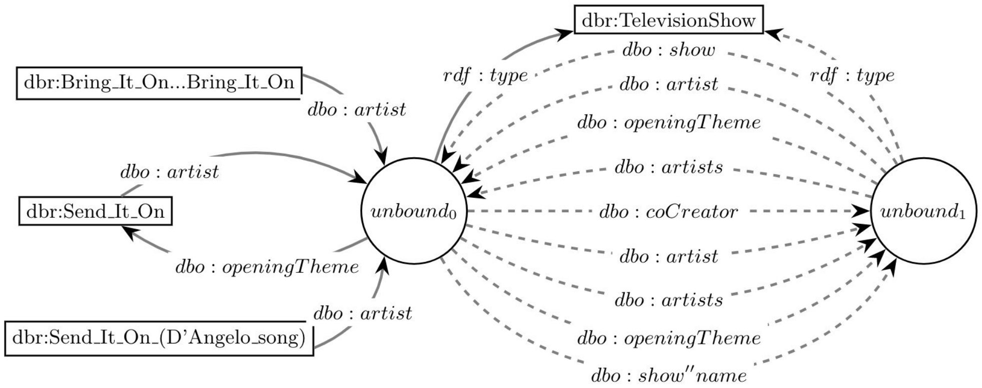


Feature identification

- SVM model on tf-idf representation of input questions
  - Establish the type of the question (e.g. boolean, count or list)
    - Affects the query formation process
  - Hidden relation identification (e.g. what is the birthplace of X and Y)

Query Generation- Capturing subgraph

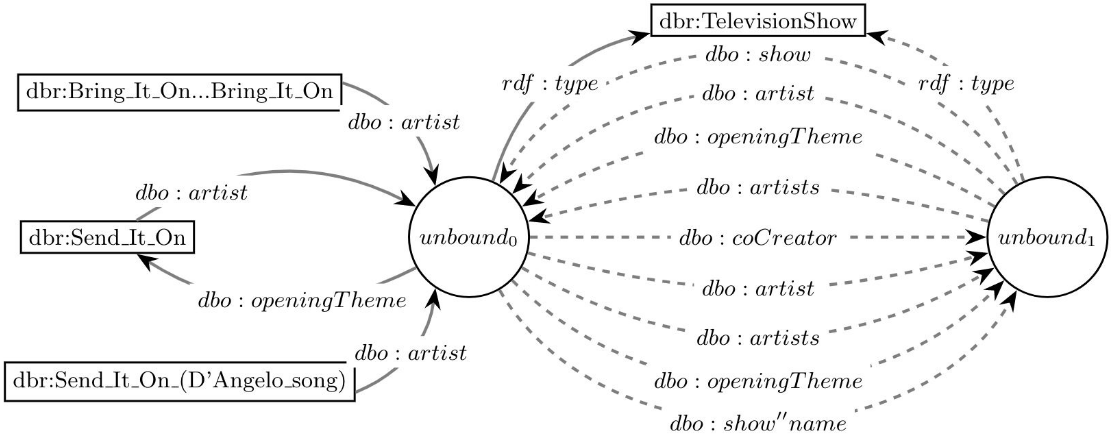
- Capture the connected subgraph which contains the linked entities/relation and arbitrary unbounded nodes.
- Limited to one and two hop distance



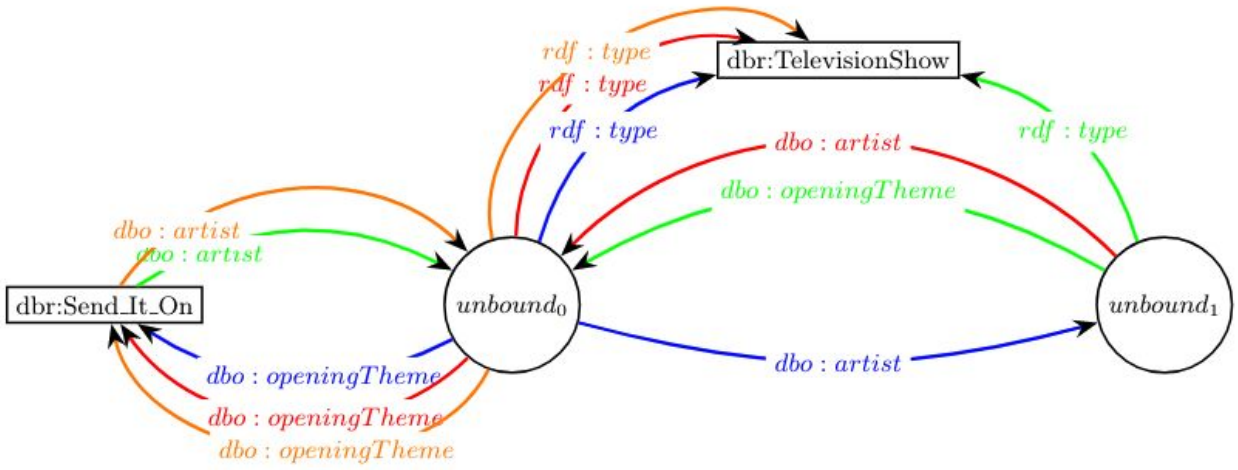
Valid walks

- Walk: A walk in a knowledge graph is a sequence of edges along the nodes they connect.
- Valid Walk: A walk is valid w.r.t a set of entities and relations, if and only if it contains all of them.

Extract valid walks from the subgraph



Extract valid walks from the subgraph

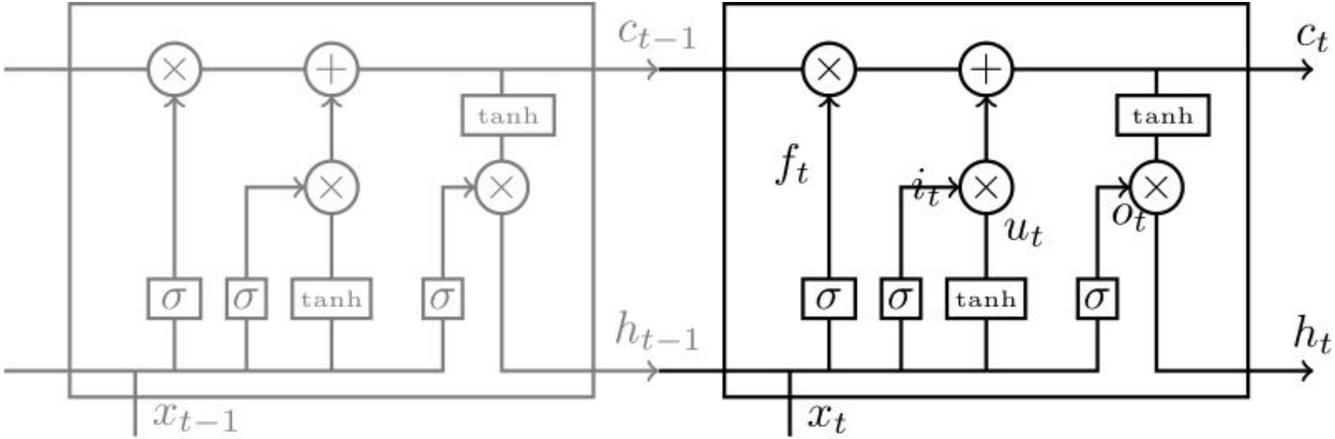


Ranking Model

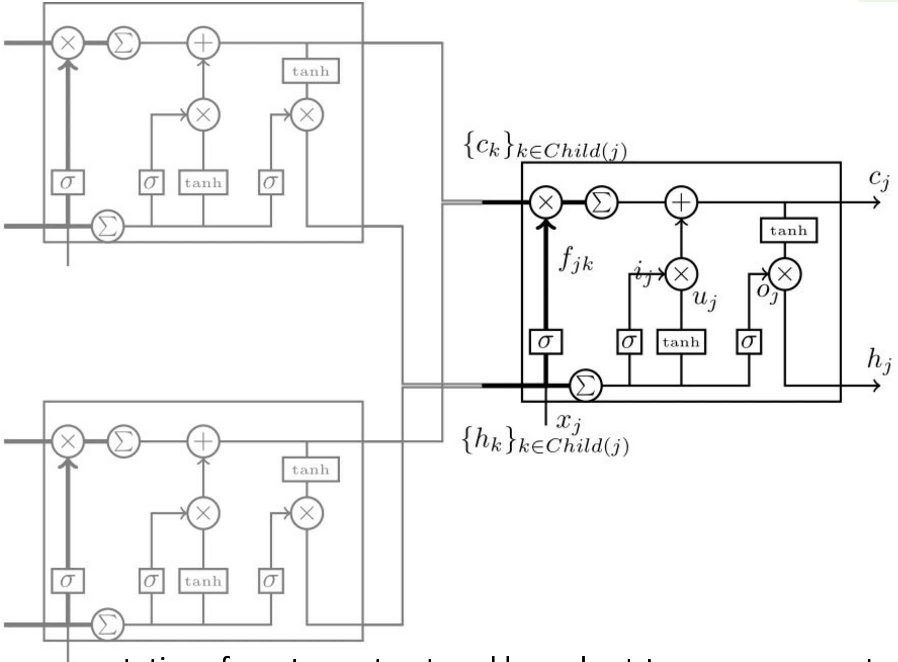
- **Goal:** Rank the valid walks w.r.t. the semantic of the input question

- **Hypothesis:** the structure of the walks is a distinctive feature to distinguish the similarity between the candidate walks and the input question

LSTM

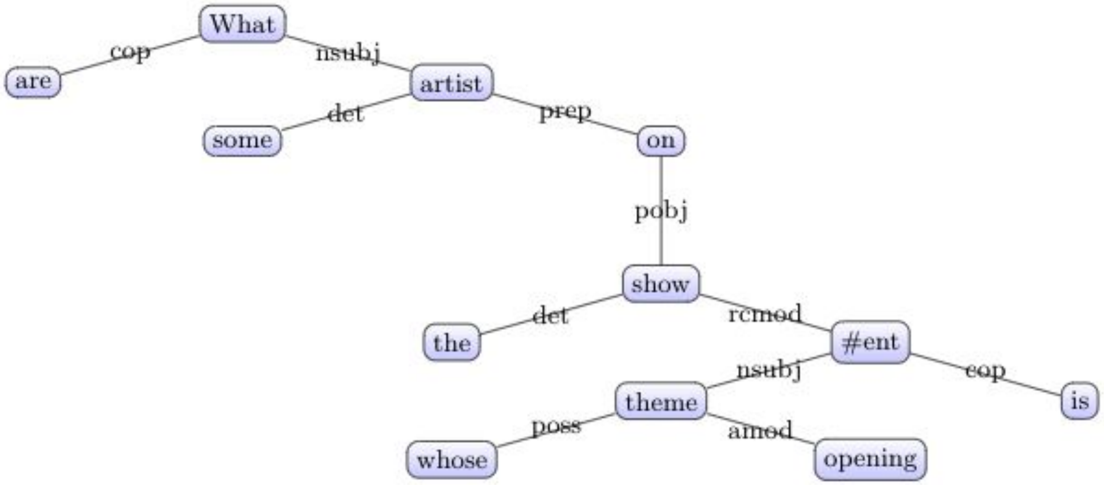


Tree-LSTM

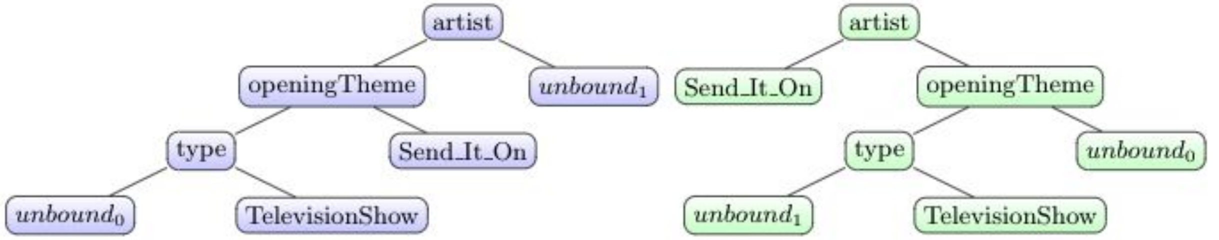


[1] Tai et al."Improved semantic representations from tree- structured long short-term memory networks"

Dependency Parsing Tree

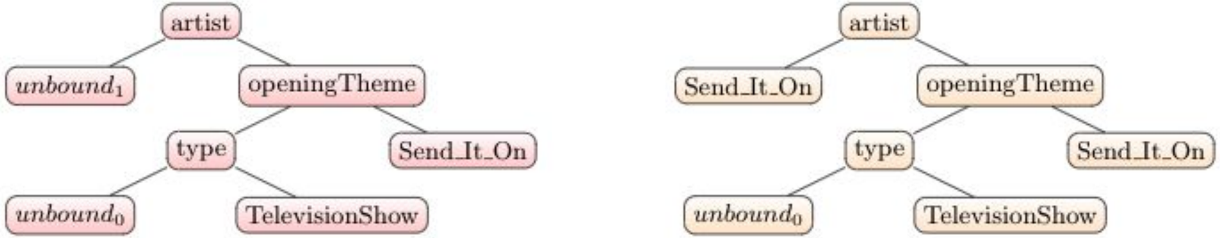


Tree-Rep. of Candidate Queries



(a) What are some artists on the show whose opening theme is Send It On?

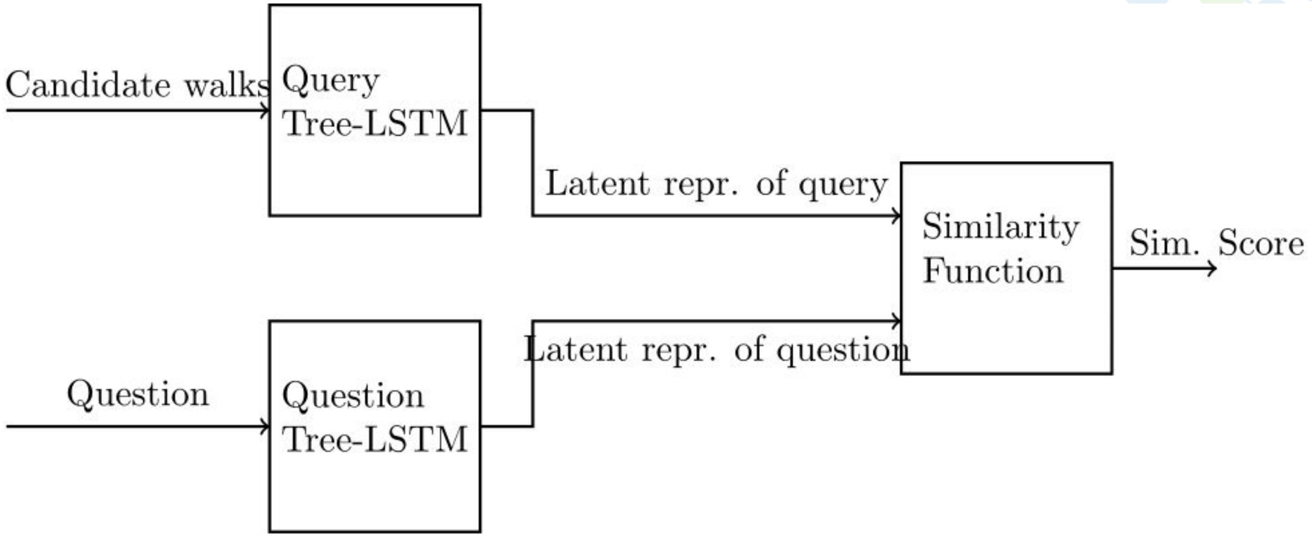
(b) What TV shows with Send It On as their opening theme are the artists of Send it On?



(c) TV shows with Send It On as their opening theme are the artists of what?

(d) Which TV shows has a opening them which is among the artists of Send it On?

Tree-LSTM as Ranking model



Evaluation Setup

- Dataset: LC-QuAD
  - 5000 Q/A pairs with different complexity and types of questions
- Baseline QA systems:
  - Sina Shekarpour et al. ”SINA:semantic interpretation of user queries for question answering on interlinked data. Web Semant”
  - NLIWOD https://github.com/dice-group/NLIWOD
- Baseline for the Ranking Model:
  - LSTM

Evaluation- Scenarios

- Top-1 correct: Questions annotated w. correct entities/relations
- Top-5 EARL+correct: Questions annotated w. list of candidate entities/relations (correct ones forcefully injected if not exists)
- Top-5 EARL: Questions annotated w. list of candidate entities/relations

- EARL: an entity/relation linking tool, Dubey et a.l. "EARL: joint entity and relation linking for question answering over knowledge graphs"

Evaluation- Ranking model

Considering the tree-representation significantly improves the performance.

Scenario	LSTM F1-measure	Tree-LSTM F1-measure
Top-1 correct	0.54	0.75
Top-1 EARL+correct	0.41	0.84
Top-1 EARL	0.32	0.74

Better generalization

- Micro F1-measure

Evaluation- vs. Baselines

Approach	Precision	Recall	F1-measure
Sina*	0.23	0.25	0.24
NLIWOD*	0.48	0.49	0.48
SQG	0.76	0.74	0.75

* Sina and NLIWOD results are taken from Singh et al.”Why reinvent the wheel–lets build question answering systems together”

- on a subset of LC-QuAD containing 3,200 questions

Summary

- Reusable and Scalable approach
- Managing noisy annotations
- Exploit structural similarity of input question and candidate queries

Thanks you for you attention.

Questions?

Code is available at:
https://github.com/AskNowQA/SQG