Complex Query Augmentation for Question Answering Over Knowledge Graphs

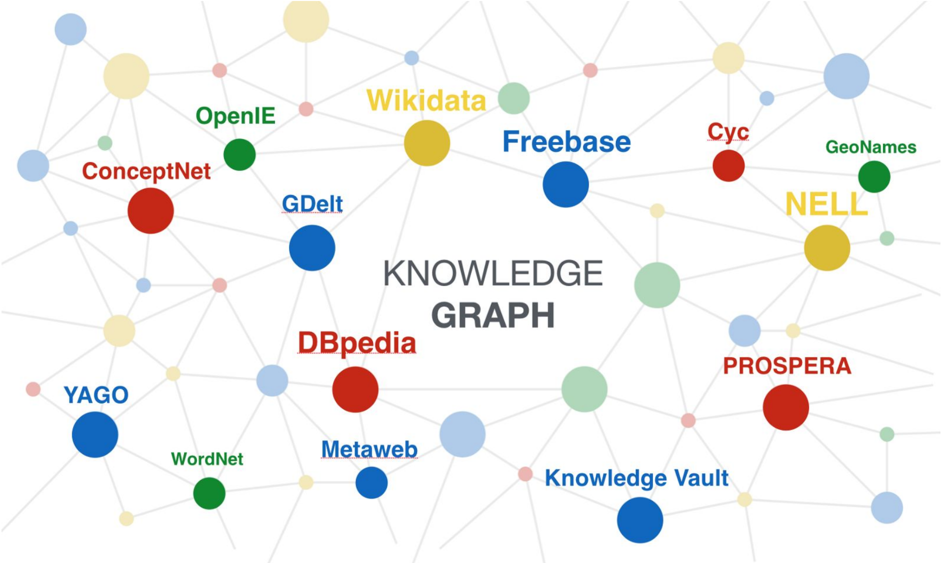
Abdelrahman Abdelkawi**, **Hamid Zafar *,** Maria Melshkova *, Jens Lehmann *

* University of Bonn, Germany

** RWTH Aachen University, Germany

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

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

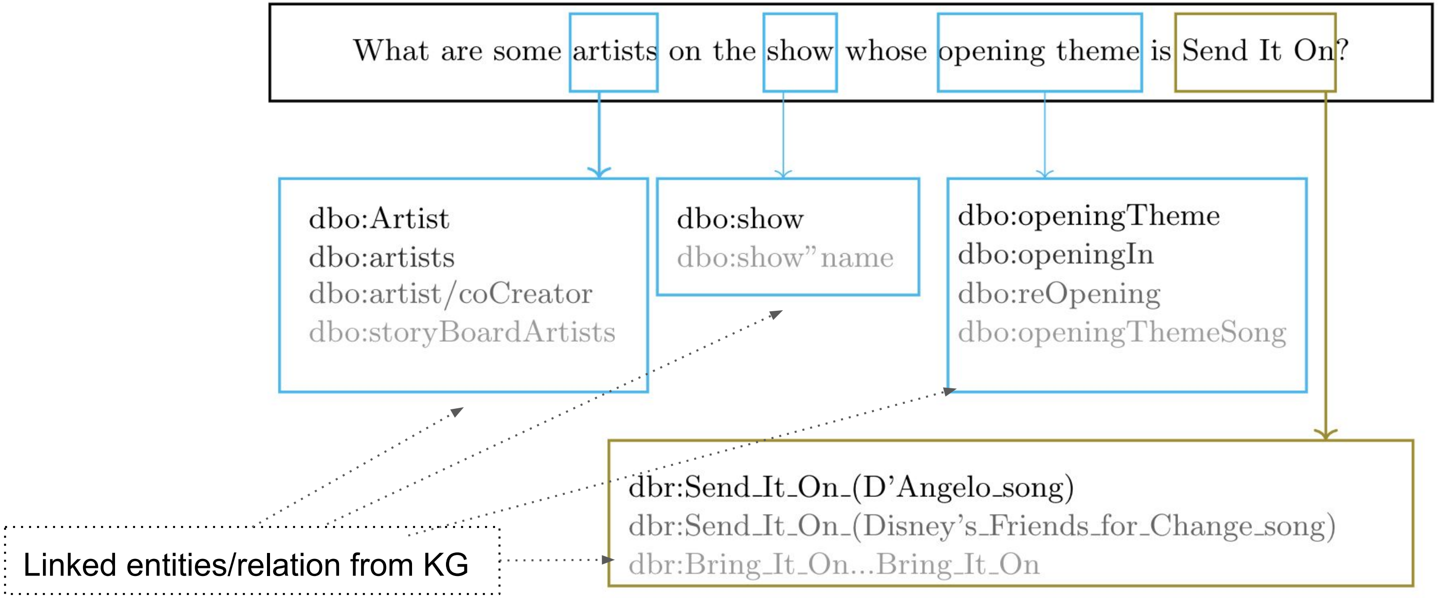
- Cope with large-scale KGs
- Ability to manage noisy input to handle error propagation
- Question type identification
- Support for composite question
- Syntactic ambiguity of the input question

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

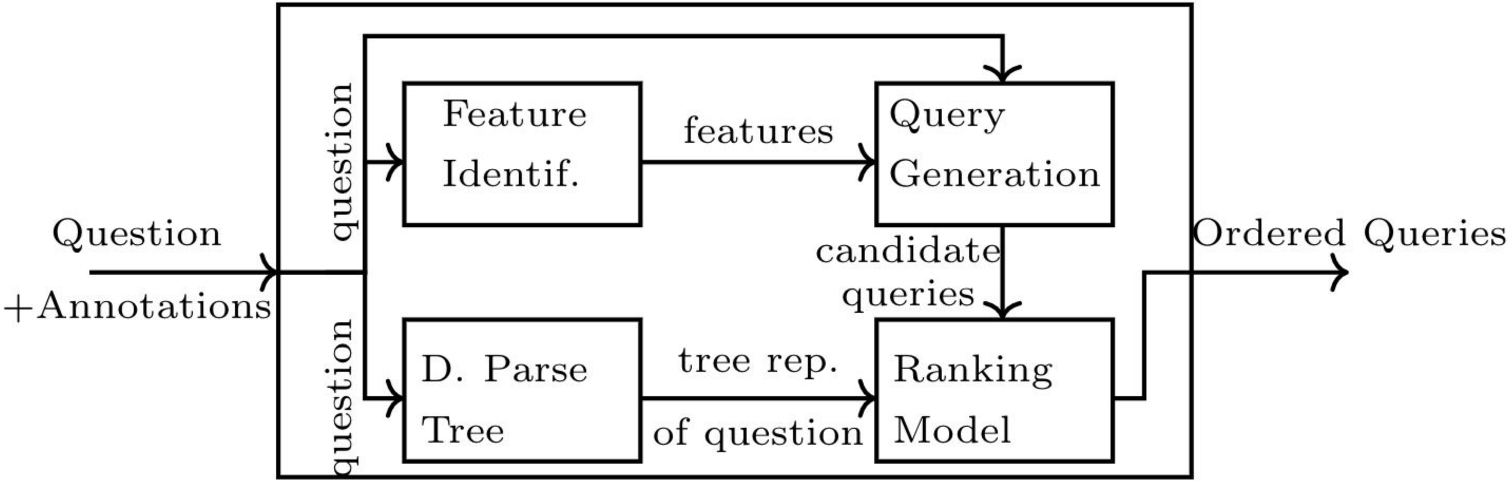
dbo:show
dbo:show”name

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

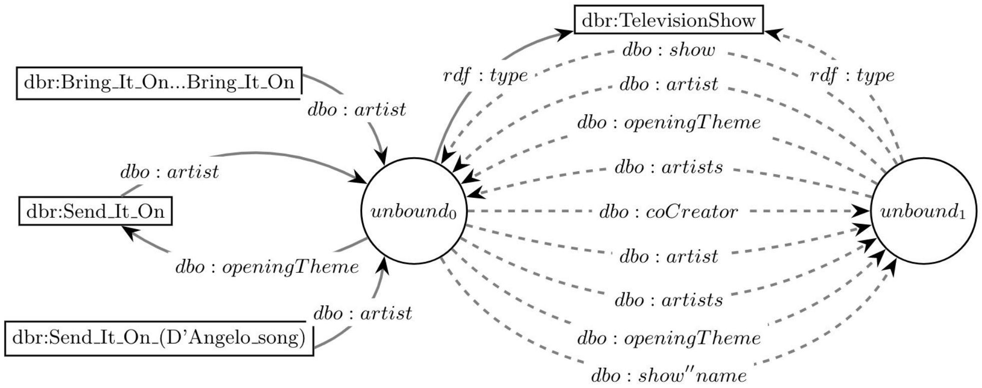
Linked entities/relation from KG

Architecture

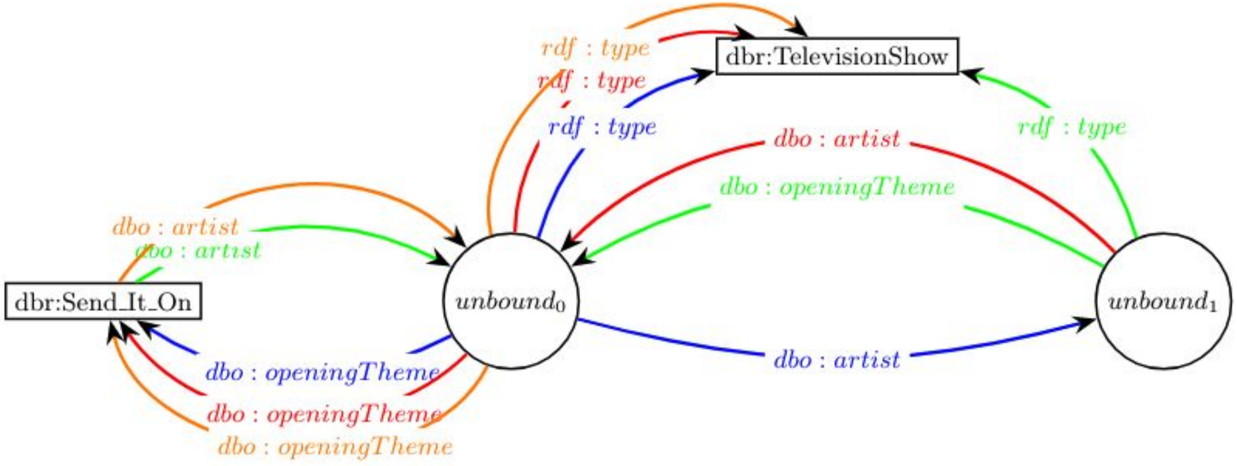


Query Generation- Capturing subgraph

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



Extract valid walks from the subgraph



SQG supports

- **List**
  - Who are the members of the Beatles?
- **Boolean**
  - Is MJ a member of the Beatles?
- **Count**
  - How many members of the Beatles are there?

Extended-SQG (Ex-SQG) Objectives

- To provide support for more complex questions, namly:
  - **Sort** Questions
  - **Filter** Questions

Sort Questions

- Question: what is the highest mountain in Australia

- SPARQL: select distinct ?uri where {

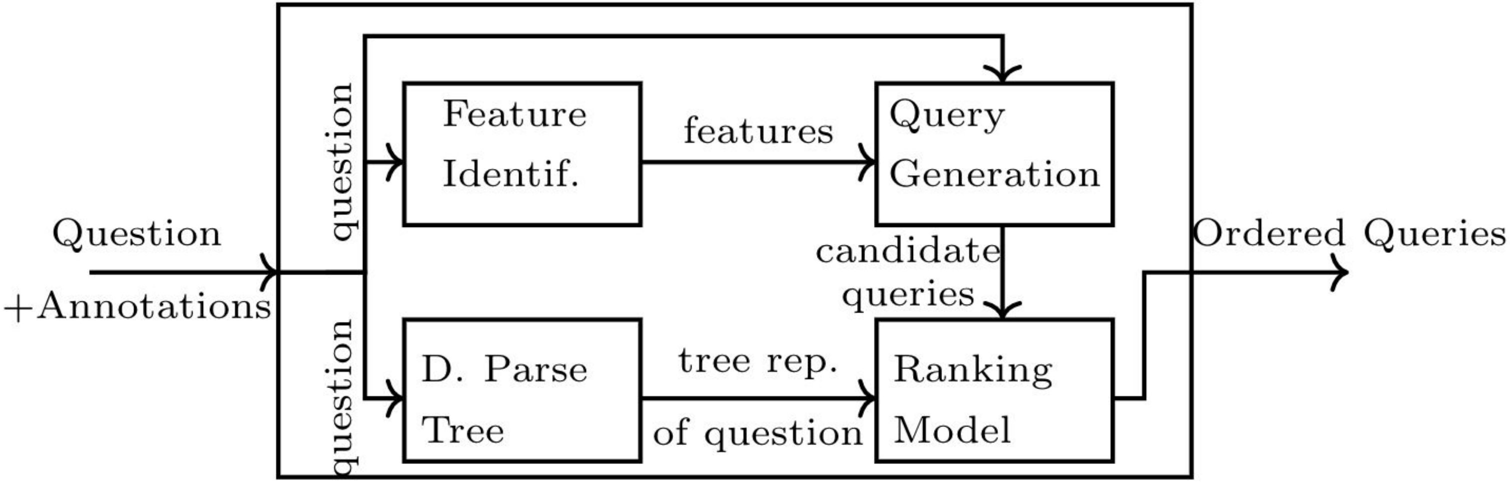
  ```
  ?uri dbo:locatedInArea dbr:Australia .
  ?uri rdf:type dbo:Mountain .
  ?uri dbo:elevation ?elevation }
  order by desc (?elevation)  limit 1
  ```

Filter Questions

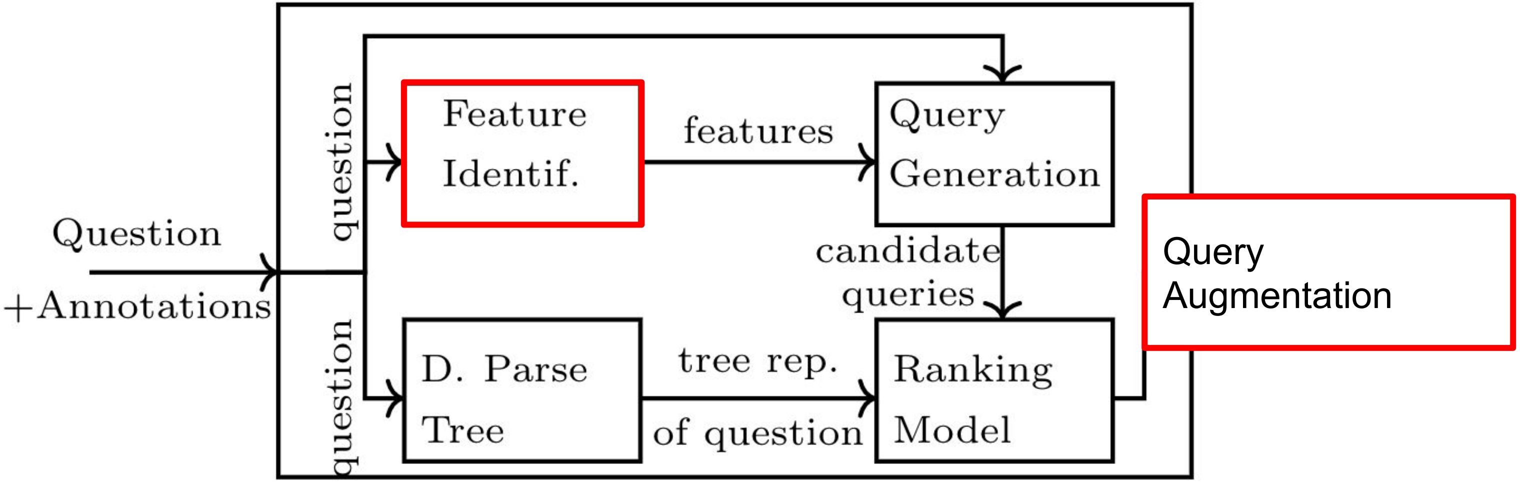
- Question: is lake baikal bigger than the great bear lake?

- SPARQL: ASK WHERE {
  dbr:Lake_Baikal dbo:areaTotal ?a1 .
  dbr:Great_Bear_Lake dbo:areaTotal ?a2 .
  FILTER (?a1 > ?a2) }

SQG Architecture



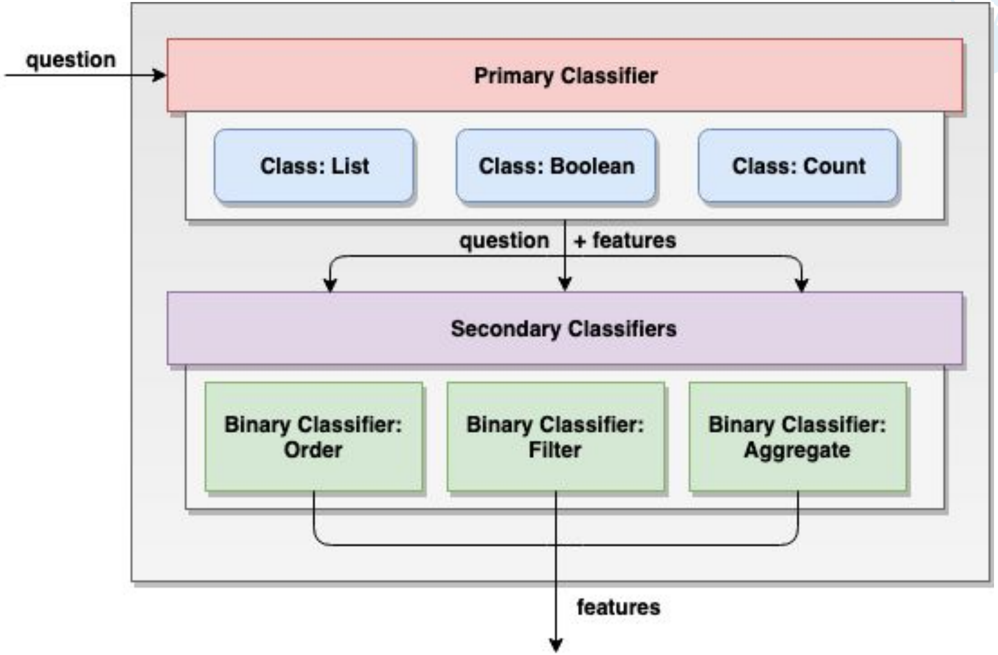
Ex-SQG Architecture



Ex-SQG Architecture

- Feature Identification
  - Question Classifier
- Query Augmentation
  - Keyword Extraction
  - KB Ontology Selection
  - Type-Specific Parameters

Question Classifier



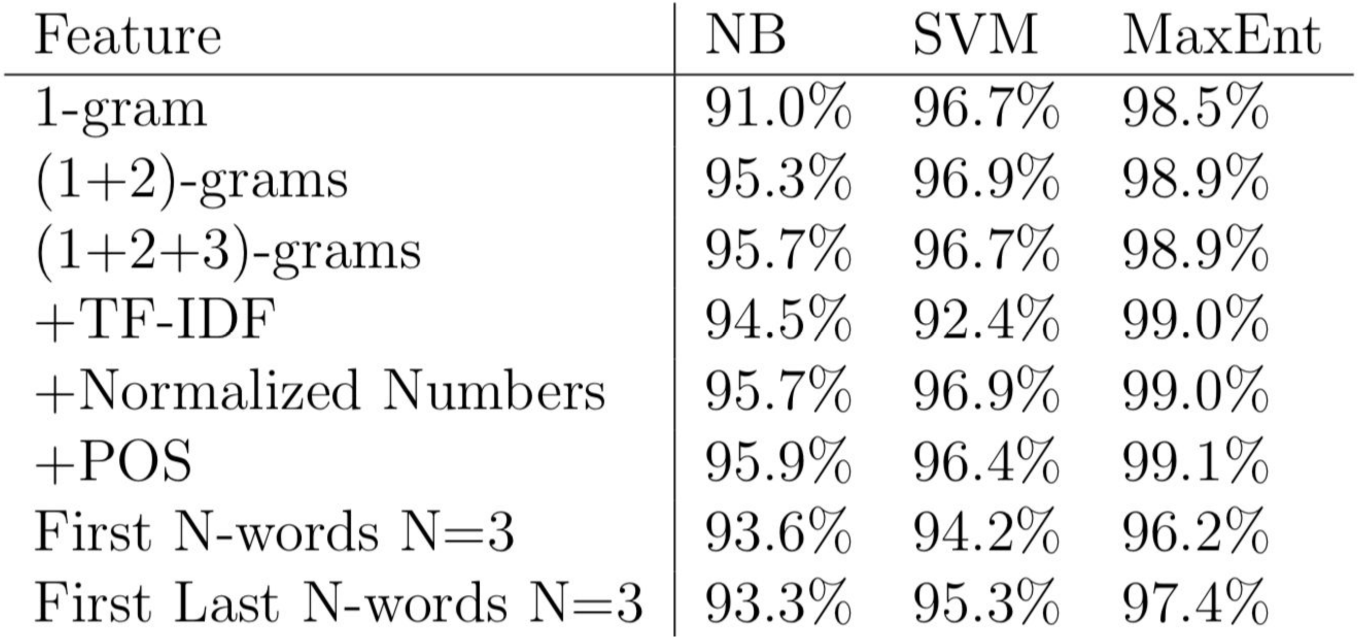
Question Classifier

Feature Engineering and Model selection

- Models: NB, SVM, MaxEnt Classifiers
- Features: Unigrams, Bigrams, Trigrams, TF-IDF, POS Tags

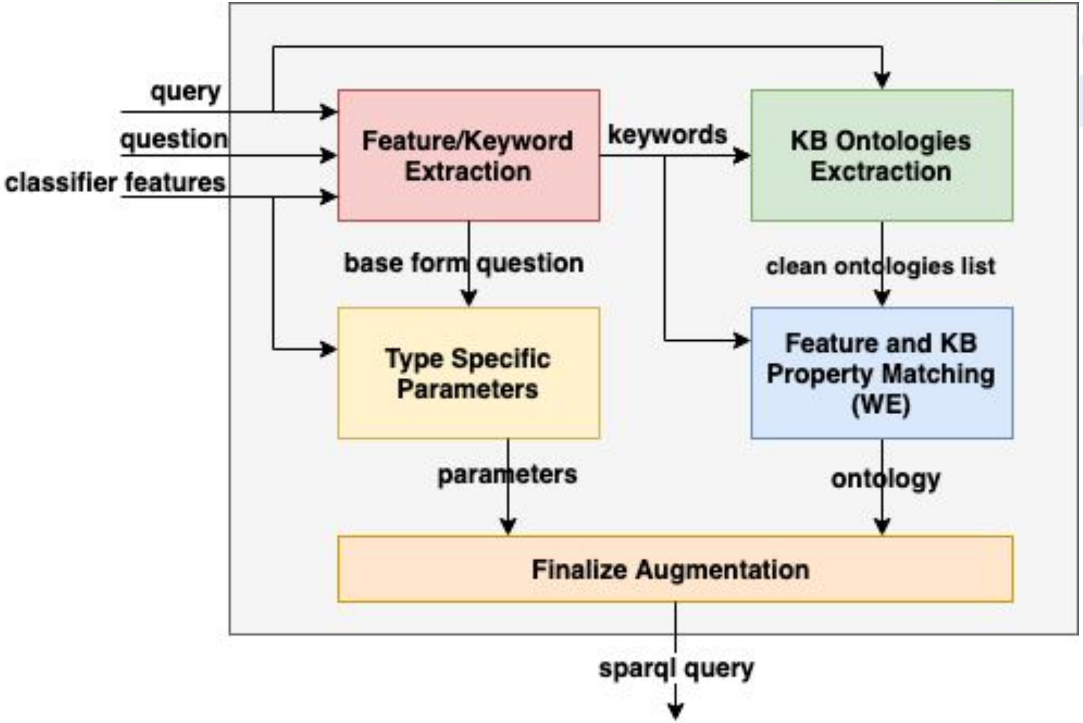
Question Classifier

Feature Engineering and Model selection



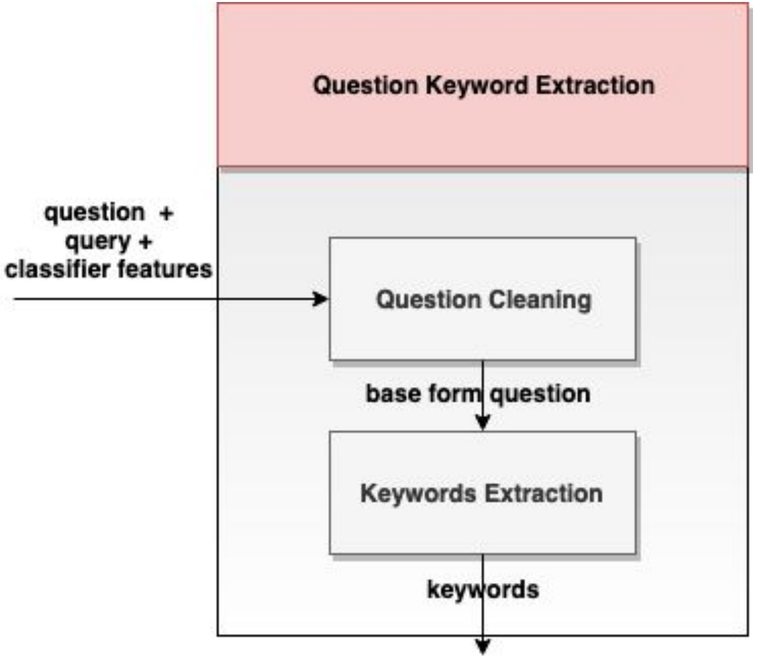
Feature	NB	SVM	MaxEnt
1-gram	91.0%	96.7%	98.5%
(1+2)-grams	95.3%	96.9%	98.9%
(1+2+3)-grams	95.7%	96.7%	98.9%
+TF-IDF	94.5%	92.4%	99.0%
+Normalized Numbers	95.7%	96.9%	99.0%
+POS	95.9%	96.4%	99.1%
First N-words N=3	93.6%	94.2%	96.2%
First Last N-words N=3	93.3%	95.3%	97.4%

Query Augmentation



Query Augmentation

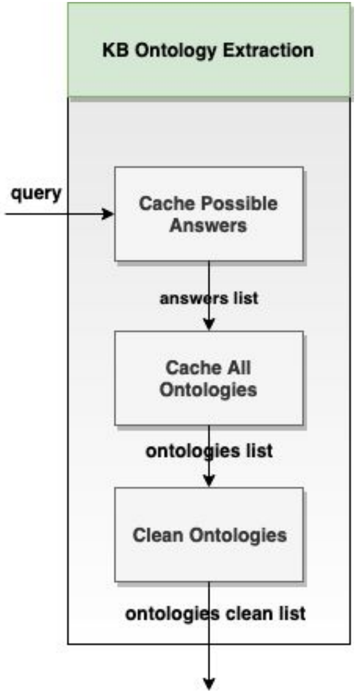
Keyword Extraction



Query Augmentation
Keyword Extraction

- Who is the second youngest football player in the Liga Futsal?
- Clean Question: second youngest football player liga futsal
- Base-form: second youngest football player
- Keyword: youngest player

Query Augmentation
KB Ontology Extraction



Query Augmentation
KB Ontology Extraction

- who is the second youngest football player in the liga futsal?
- SQG Query: select distinct ?player where {
  ?t dbo:league dbr:Liga_Futsal
  ?player dbo:team ?t . }
- Answer: List of players in the Liga Futsal
- One-hop Clean Ontologies: [“height”, “formationDate”,

  “squadNumber”, “birthDate”, “numberOfGoals”, “capacity”]

Query Augmentation

KB Ontology Extraction-Word Embeddings

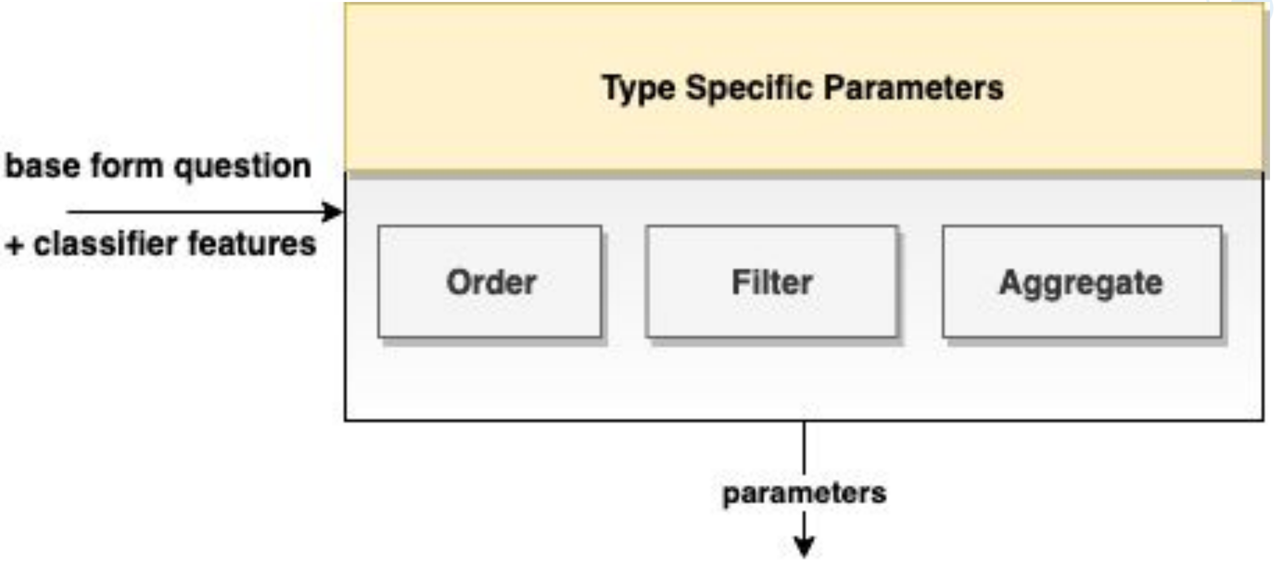
- To leverage the semantic similarity to opt the correct item
- Distance Measure: Cosine Distance
- Compound Vector Representation
  - Addition
  - Mean
  - Word Mover Distance (WMD)

Query Augmentation
KB Ontology Extraction-Word Embeddings

- who is the second youngest football player in the liga futsal?
- Keyword: youngest player
- One-hop Clean Ontologies: [“height”, “formationDate”, “squadNumber”, “birthDate”, “numberOfGoals”, “capacity”]
- select distinct ?play where {
  ?t dbo:league dbr:Liga_Futsal .
  ?play dbo:team ?t.
  ?play dbo:birthDate ?date}

Query Augmentation

Type-Specific Parameters



Query Augmentation
Type-Specific Parameters

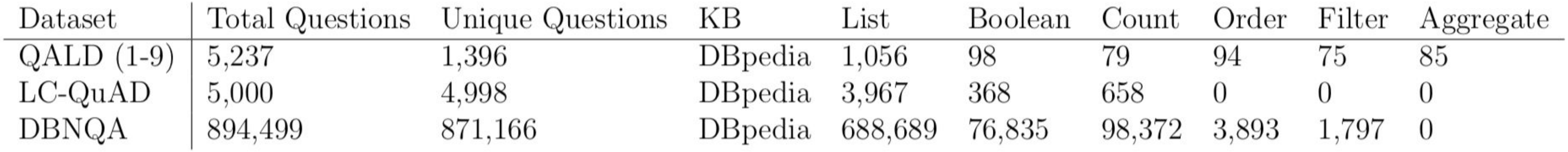
- Ordinal Questions
  - Offset: Ordinal Detection
  - Direction of Sort: Direction Classifier
  - Limit: Using POS Tags
- Fitler Questions
  - Comparison Operator: Operator Classifier

Query Augmentation
Type-Specific Parameters

- who is the second youngest football player in the liga futsal?
-
```
select distinct ?player where {
   ?t dbo:league dbr:Liga_Futsal.
   ?player dbo:team ?t .
   ?player dbo:birthDate ?date}
order by desc(?date)  offset 1 limit 1
```

Experimental Results

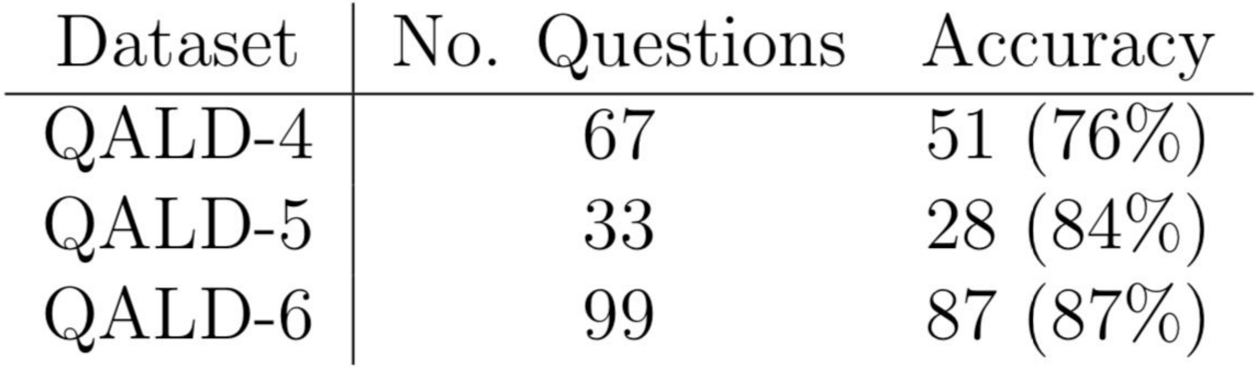
Datasets



Dataset	Total Questions	Unique Questions	KB	List	Boolean	Count	Order	Filter	Aggregate
QALD (1-9)	5,237	1,396	DBpedia	1,056	98	79	94	75	85
LC-QuAD	5,000	4,998	DBpedia	3,967	368	658	0	0	0
DBNQA	894,499	871,166	DBpedia	688,689	76,835	98,372	3,893	1,797	0

Experimental Results

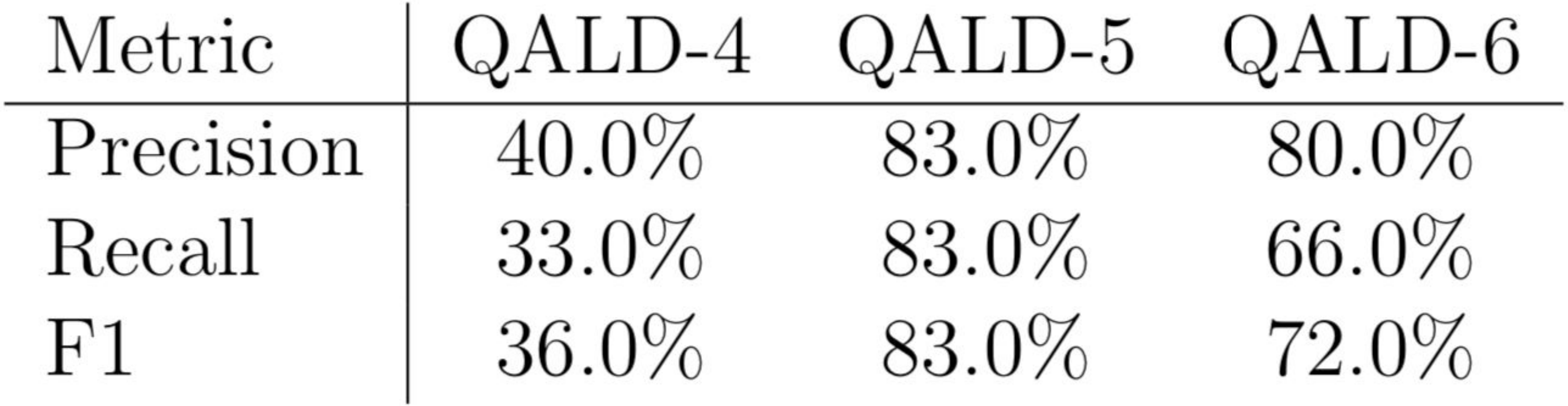
Question Classifier Performance



Dataset	No. Questions	Accuracy
QALD-4	67	51 (76%)
QALD-5	33	28 (84%)
QALD-6	99	87 (87%)

Experimental Results

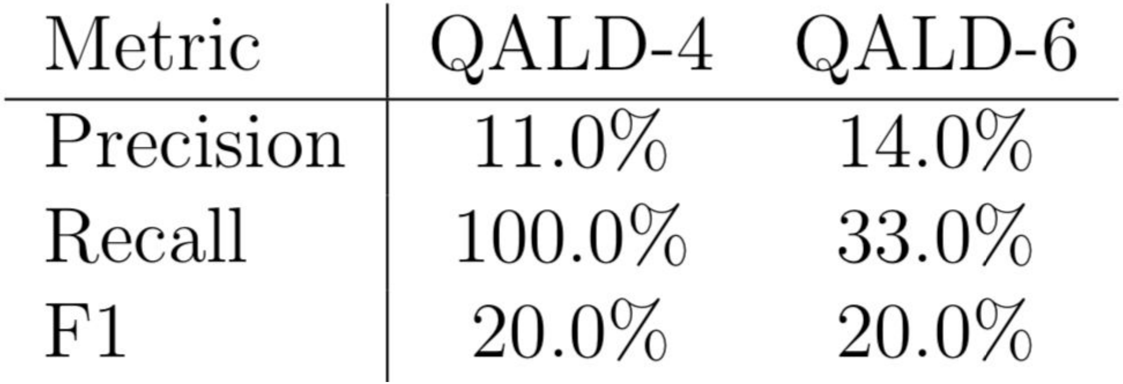
Ordinal Questions Pipeline Performance



Metric	QALD-4	QALD-5	QALD-6
Precision	40.0%	83.0%	80.0%
Recall	33.0%	83.0%	66.0%
F1	36.0%	83.0%	72.0%

Experimental Results

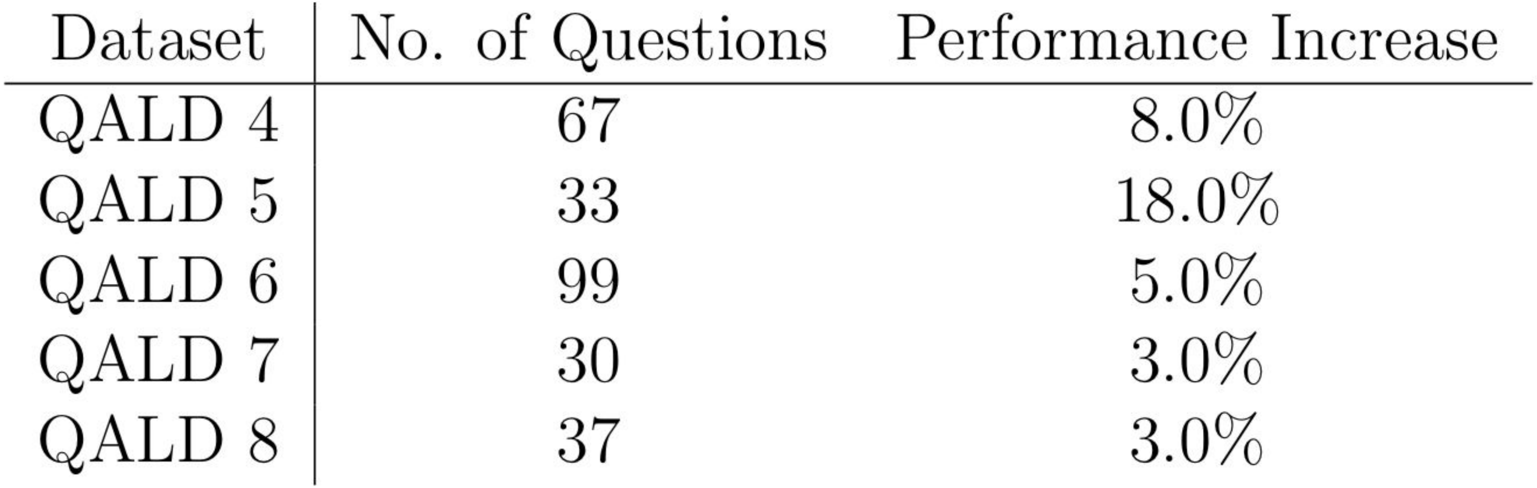
Filter Questions Pipeline Performance



Metric	QALD-4	QALD-6
Precision	11.0%	14.0%
Recall	100.0%	33.0%
F1	20.0%	20.0%

Experimental Results

End-to-end Performance



Dataset	No. of Questions	Performance Increase
QALD 4	67	8.0%
QALD 5	33	18.0%
QALD 6	99	5.0%
QALD 7	30	3.0%
QALD 8	37	3.0%

Summary

- Filling a gap by supporting ordinal and filter questions in KG-QA

Thanks you for you attention.

Questions?

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