

# Beyond NED: Fast and Effective Search Space Reduction for Complex Question Answering over Knowledge Bases

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum  
Max Planck Institute for Informatics, Germany



## CLOCQ can be used for efficient retrieval of relevant KB-facts for any query

### SEARCH SPACE REDUCTION: OVERLOOKED PROBLEM IN KB-QA

- ★ Curated knowledge bases (KB) have billions of facts, with millions of entities and thousands of predicates
- ★ Question answering (QA) system is typically a neural model that requires input to be sufficiently small
- ★ Therefore, QA system not directly applicable to full KB
- ★ Typical KB-QA pipeline first reduces the search space to a few thousands of facts, that are relevant to the question and likely to contain the answer(s)

### STANDARD NED: ONE SIZE FITS ALL

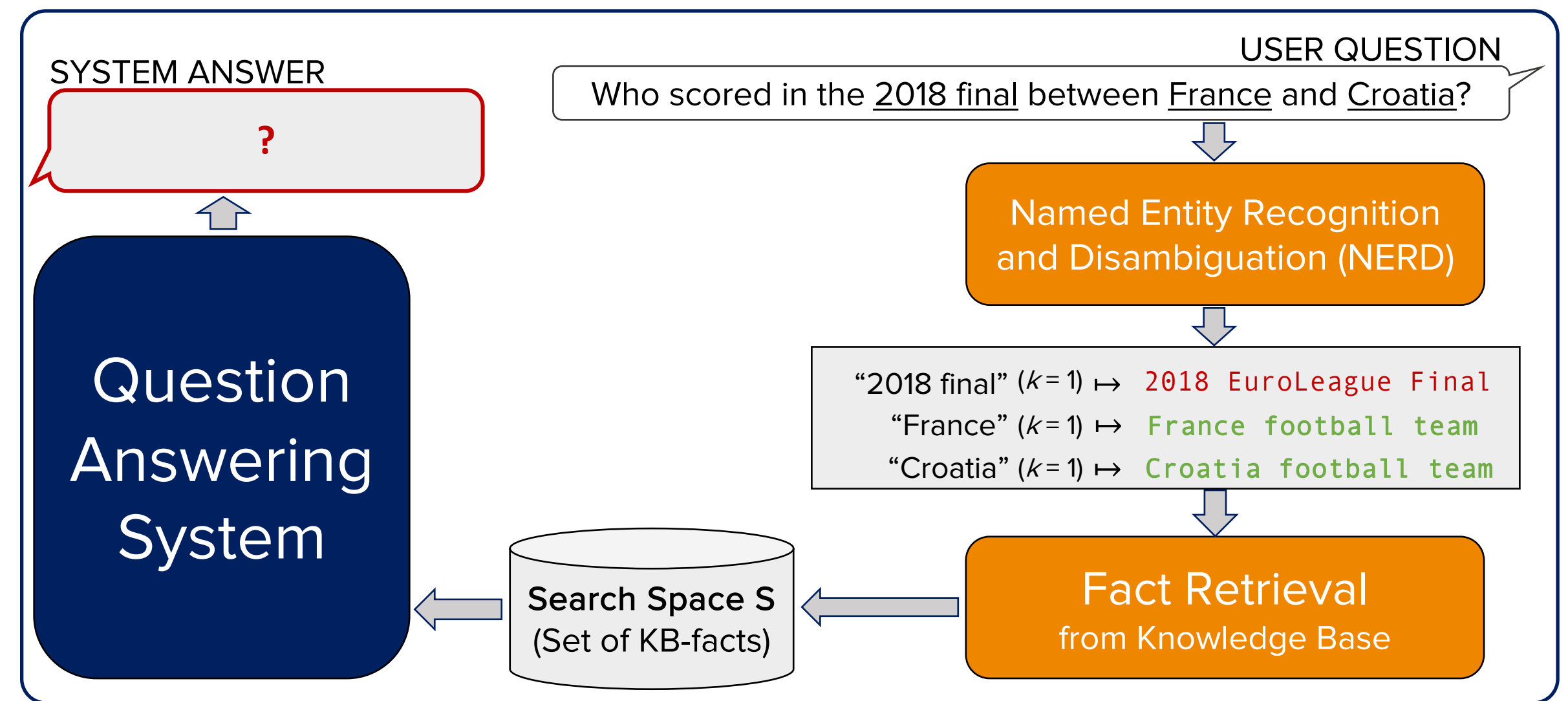
- ★ Top-1 NED too restrictive for QA (single error can cause failure of full QA pipeline)
- ★ Top- $k$  NED can easily include noise, blowing up the search space
- ★ Important cues within the question (e.g. scored in example on the right) are not considered, given that most NED methods disambiguate entity mentions only

### APPROACH: GOING BEYOND NED

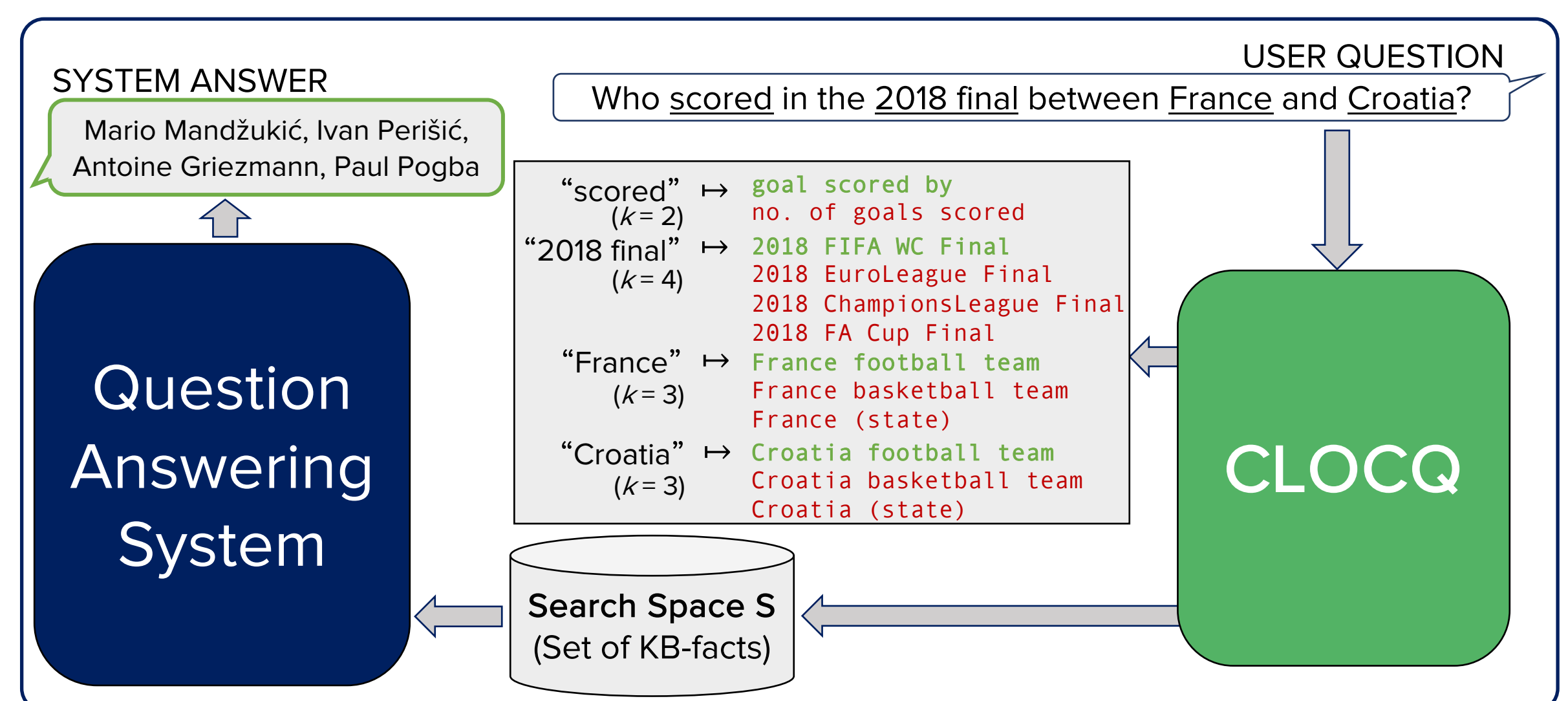
- ★ Adaptive top- $k$ : choose appropriate  $k$ , for each question word individually
- ★ Disambiguate all KB-items (entities, types, concepts, predicates)
- ★ Consider proximity of disambiguated KB-items within graph underlying the KB
- ★ Establish a novel fact-centric KB-index for more efficient access to the KB

### PIPELINE: SEARCH SPACE REDUCTION FOR COMPLEX KB-QA

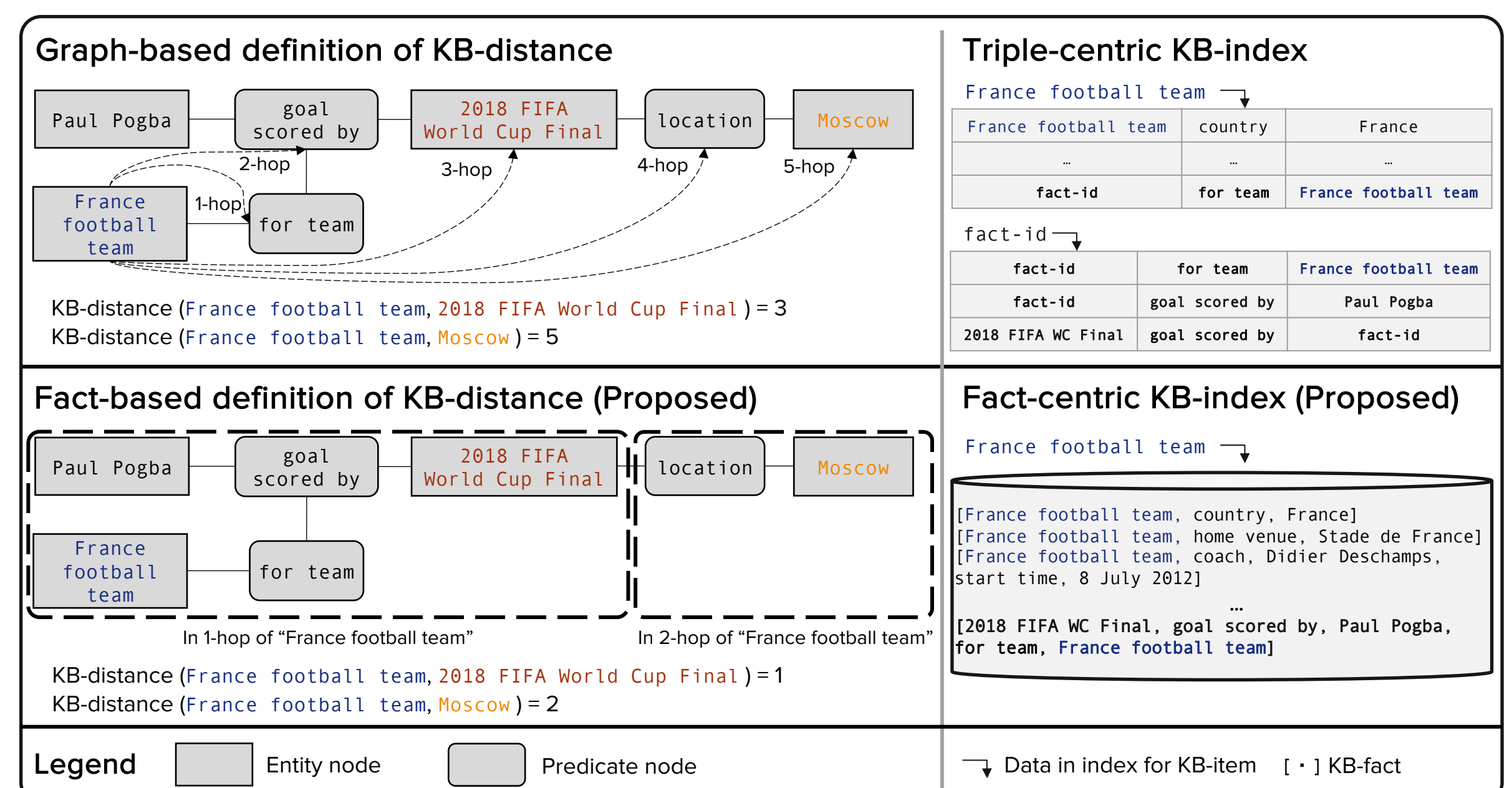
1. KB-item retrieval: Retrieve  $d$  KB-item candidates per question word
2. Auto- $k$ : Choose  $k$  automatically for each question word based on its ambiguity
3. Candidate scoring: Identify top- $k$  candidates using signals for pair-wise KB-proximity, pair-wise semantic coherence, question relevance, lexical matching
4. Search space retrieval: Retrieve facts for disambiguated KB-items from knowledge base using our fact-centric KB-index; consider only salient facts, pruning potentially noisier facts based on a parameter  $p$  (e.g. facts with US as object or qualifier-object)



Standard KB-QA pipeline, leveraging facts of disambiguated entities to prune the search space.



KB-QA pipeline with CLOCQ: All KB-items are disambiguated, with an adaptive  $k$  per question word.



CLOCQ makes use of a fact-centric KB-index, which stores n-ary facts in a way that allows for efficient retrieval.

## CLOCQ enhances answer presence in search space for QA systems

### ConvQuestions

Question: "How is the main soccer club of the german city Düsseldorf called?"

#### CLOCQ

"main"  $\rightarrow$  {Frankfurt am Main, Main(river), Offenbach am Main};  
"soccer"  $\rightarrow$  {football, football team};  
"club"  $\rightarrow$  {Nightclub, Torina F.C.};  
"german"  $\rightarrow$  {German, German Empire};  
"city"  $\rightarrow$  {Manchester City F.C., Birmingham City F.C., Stoke City F.C., Cardiff City F.C.};  
"Düsseldorf"  $\rightarrow$  {Düsseldorf, Fortuna Düsseldorf};

#### TagME

"main"  $\rightarrow$  {Main (river)};  
"soccer"  $\rightarrow$  {football};  
"club"  $\rightarrow$  {sports club};  
"german"  $\rightarrow$  {Germany};  
"city"  $\rightarrow$  {City of London};

#### ELQ

"german"  $\rightarrow$  {Germany};  
"soccer"  $\rightarrow$  {football};

Question: "Who was the screenwriter for Crazy Rich Asians?"

#### CLOCQ

"screenwriter"  $\rightarrow$  {screenwriter};  
"Crazy Rich Asians"  $\rightarrow$  {Crazy Rich Asians (film)};

#### TagME

"Crazy Rich Asians"  $\rightarrow$  {Crazy Rich Asians (book)};

#### ELQ

"Crazy Rich Asians"  $\rightarrow$  {Crazy Rich Asians (book)};

### LC-QuAD 2.0

Question: "Who is the composer of All We Know?"

#### CLOCQ

"composer"  $\rightarrow$  {composer, film score composer};  
"All We Know"  $\rightarrow$  {All We Know (Paramore), For All We Know (album), All We Know (Chainsmokers), For All We Know (Carpenters), For All We Know (1934 song)};

#### TagME

"Who"  $\rightarrow$  {The Who};  
"composer"  $\rightarrow$  {composer};  
"All We Know"  $\rightarrow$  {For All We Know (Carpenters)};

#### ELQ

"All We Know"  $\rightarrow$  {All We Know (Chainsmokers)};

Question: "Who is the son of the brother of Queenie Padilla?"

#### CLOCQ

"son"  $\rightarrow$  {Son en Breugel, nephew, Mae Hong Son, Porto de Son};  
"brother"  $\rightarrow$  {sibling};  
"Queenie Padilla"  $\rightarrow$  {Queenie Padilla};

#### TagME

"Who"  $\rightarrow$  {World Health Organization};  
"brother"  $\rightarrow$  {Brother};

#### ELQ

"Padilla"  $\rightarrow$  {Zsa Zsa Padilla};

Anecdotal examples for which only CLOCQ had an answer in the search space.

### RESULTS ON LC-QuAD 2.0

CLOCQ 82.6 ★

CLOCQ ( $k=1$ ) 80.0

TagME+HDT 76.8

ELQ+HDT 76.7

EARL+HDT 65.9

AIDA+HDT 60.5

REL+HDT 55.8

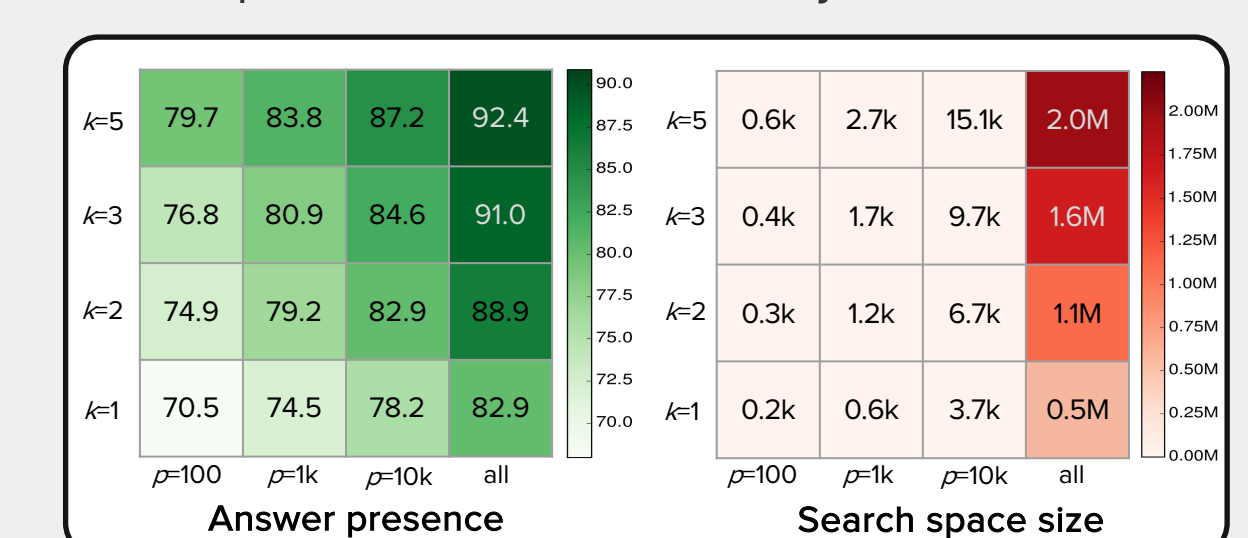
Answer presence after search space reduction on LC-QuAD 2.0 dataset.

CLOCQ+GRAFT-Net 19.7 ★

TagME+HDT+GRAFT-Net 17.1

ELQ+HDT+GRAFT-Net 16.8

Precision at 1 for QA task when feeding outputs of search space reduction into QA system GRAFT-Net.



Trade-off between answer presence and search space size.