SIC Saarland Informatics Campus

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

# Question Answering Systems

## Conversational Question Answering

**Rishiraj Saha Roy**
Max Planck Institute for Informatics, Germany

11

Saarland University, Summer Semester 2020

# Question of the day

How can we design question answering systems that can handle conversations?

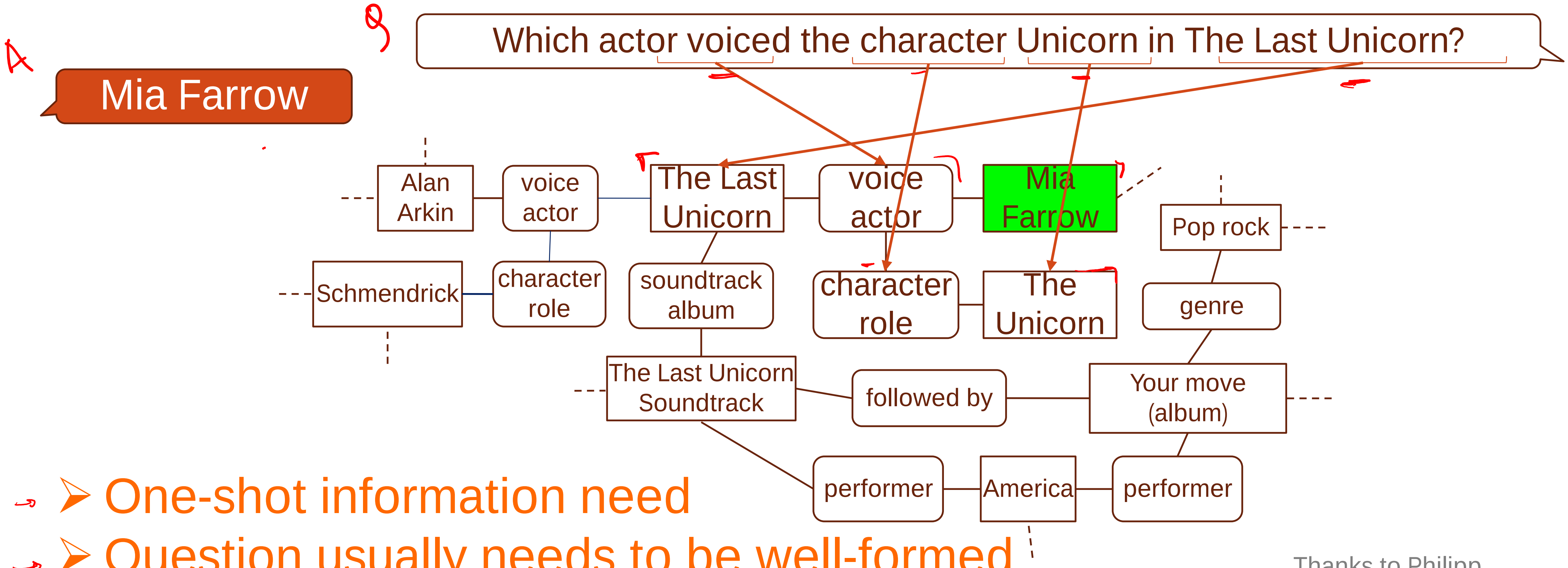
# You'll find this covered in

- Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion
  - Christmann et al.
  - CIKM 2019 3-7 Nov 19
  - https://openreview.net/pdf?id=S1CChZ-CZ
- Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base
  - Shen et al.
  - EMNLP 2019 3-7 Nov 19
  - https://www.aclweb.org/anthology/D19-1248.pdf

# Research paper 1

Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion

# Question answering over KGs




- ➢ One-shot information need
- ➢ Question usually needs to be well-formed

[Abujabal et al. (2018), Diefenbach et al. (2019), Huang et al. (2019)]

Thanks to Philipp Christmann for the slides simulate human-human interaction

# Conversational KG-QA

System | | User | turn
--- | --- | --- | ---
| Q1 | Which actor voiced the character Unicorn in The Last Unicorn? | t1
Mia Farrow | Q2 | And Alan Arkin was behind . . .? | t2
Schmendrick | Q3 | So who performed songs? | t3
America | Q4 | Genre of this band? | t4
Folk rock | Q5 | By the way, who directed the movie? | t5
Jules Bass | | |

{Q} {A} context implicit

- Information needs rarely one-shot
- Natural mode of interaction

remembers + understands

multi-turn

- sequential QA

# Conversational KG-QA

Complete

$2^0$

T1 R1 R2 E2 E1

Which actor voiced the character Unicorn in The Last Unicorn?

Incomplete

And Alan Arkin was behind . . .? → ellipsis

So who performed songs?

Genre of this band?

By the way, who directed the movie?

# Conversational KG-QA

well-formulated :
Which character...

NL methods X

- Ad hoc

And Alan Arkin was behind . . .?

So who performed songs?

Genre of this band?

By the way, who directed the movie?

# Conversational KG-QA

- Ad hoc
- Ungrammatical

And Alan Arkin was behind . . .?

~ and the songs were by... ?

So who performed the songs?

Genre of this band?

By the way, who directed the movie?

No, I meant...

# Conversational KG-QA

- Ad hoc
- Ungrammatical
- Information left out

context implicit

And Alan Arkin was behind . . .?

So who performed songs?

Genre of this band?

By the way, who directed the movie?

Why bother??

(question completion (diff. task))

rewrite

HARD (esp for multi-turn)

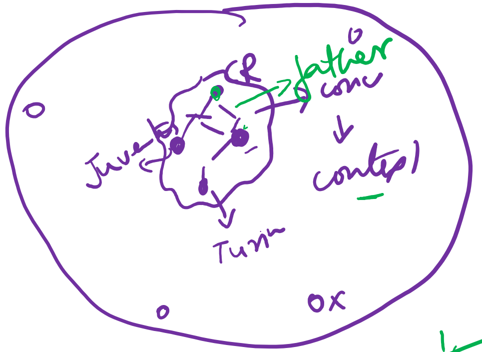
# Desiderata and contributions

→ what is desired/ideal

approach +

- Obs → Large topic jumps in conversations are rare
  - effect → Conversations establish localized context in KG

opportunities from KG-side ?



director Alan Arkin

- Harness underlying KG-connectivity ⇒ as opposed to question completion

who's this father ?

  - Expand context with relevant entities and predicates in neighborhood ⇝ cxt
- 1 CONVEX: CONVersational KG-QA using judicious context EXpansion
  - Completely unsupervised!

$QA_1$ + CONVEX

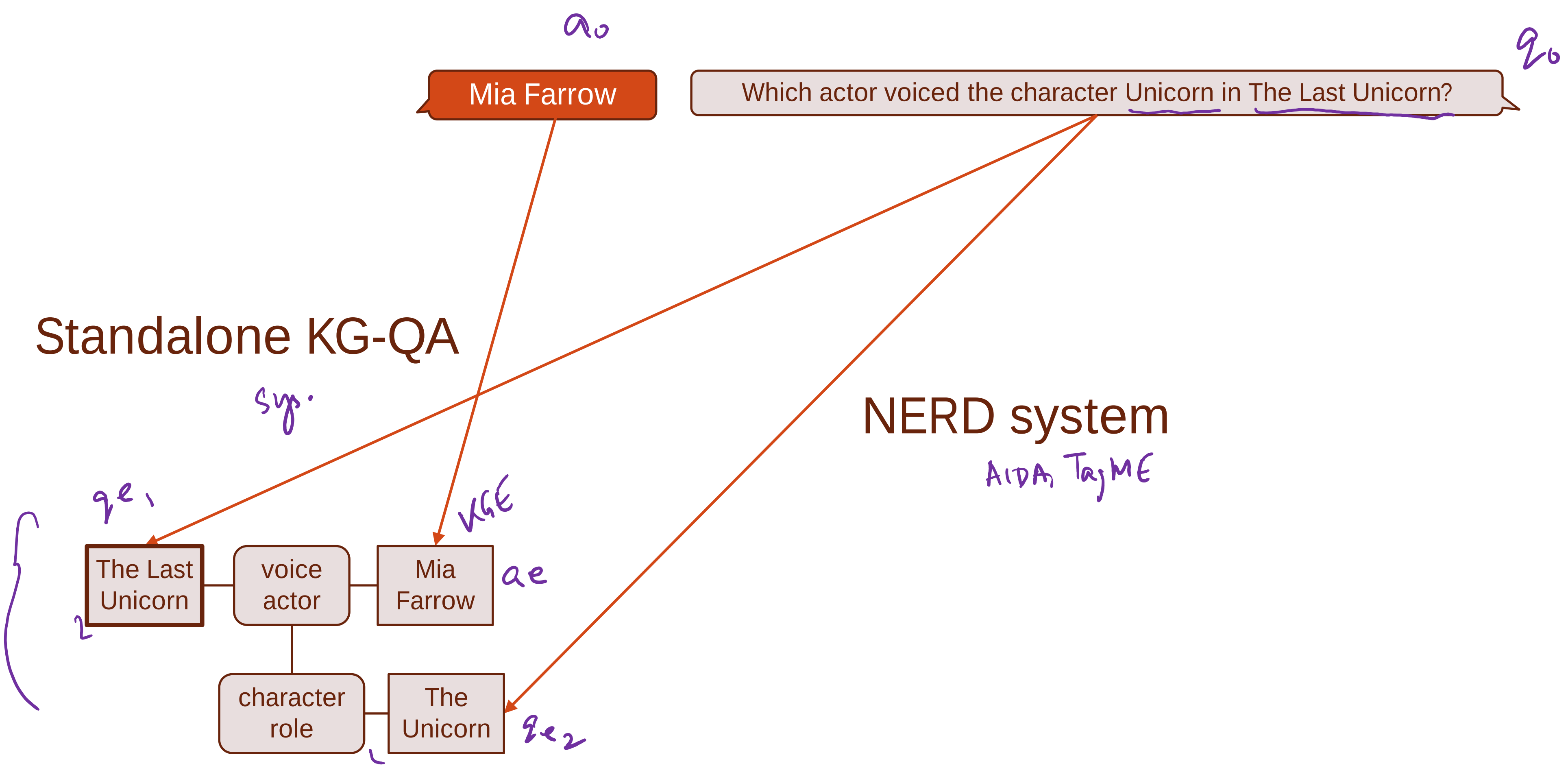
$QA_2$ + CONVEX

...

- 2 CONVEX works on top of any KG-QA system to handle conversations

plug-in / reusable module

# Initial context



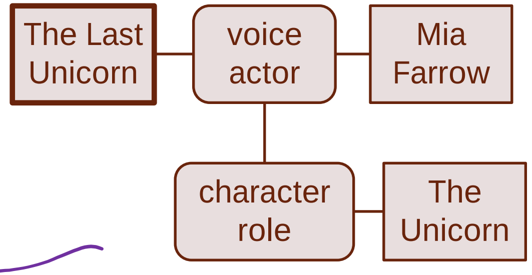
# Initial context graph

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .? $q_1$

$X_0$




ans $a_1$ is not in $cxt_0$

**How to expand the context?**

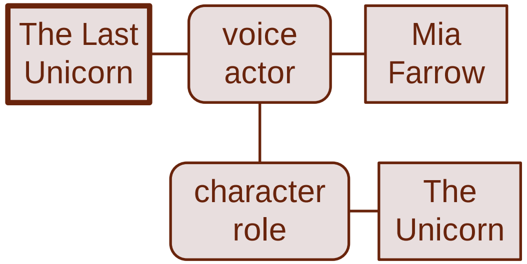

Neighborhood of
Mia Farrow
Neighborhood of
The Last Uni
v. v. large
1-hop / 2hop
Neighborhood of
The Unicorn
flooded with
ans candidates

# Judicious context expansion

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?

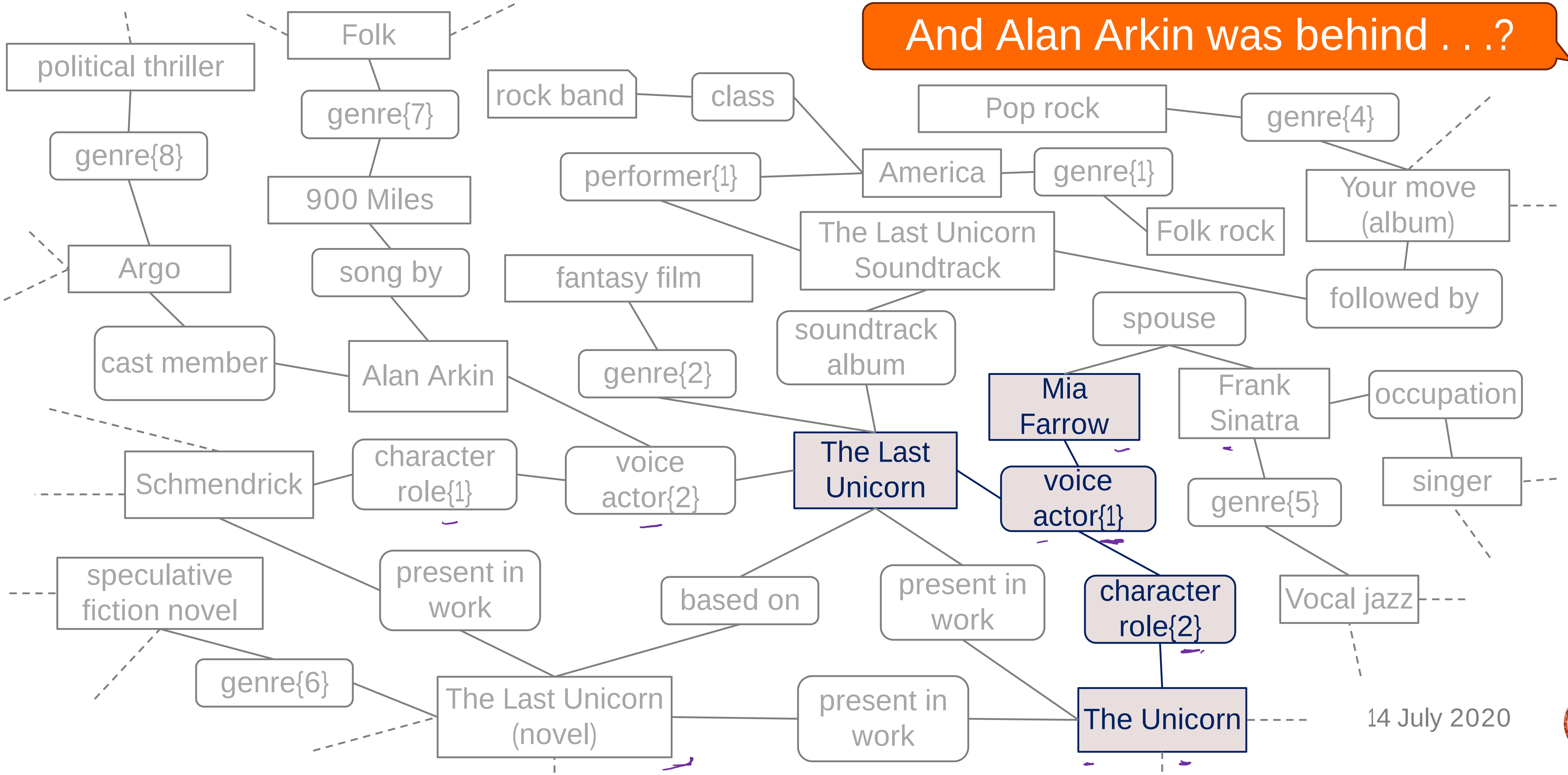



Do not expand with the complete neighborhood!

as of now / term

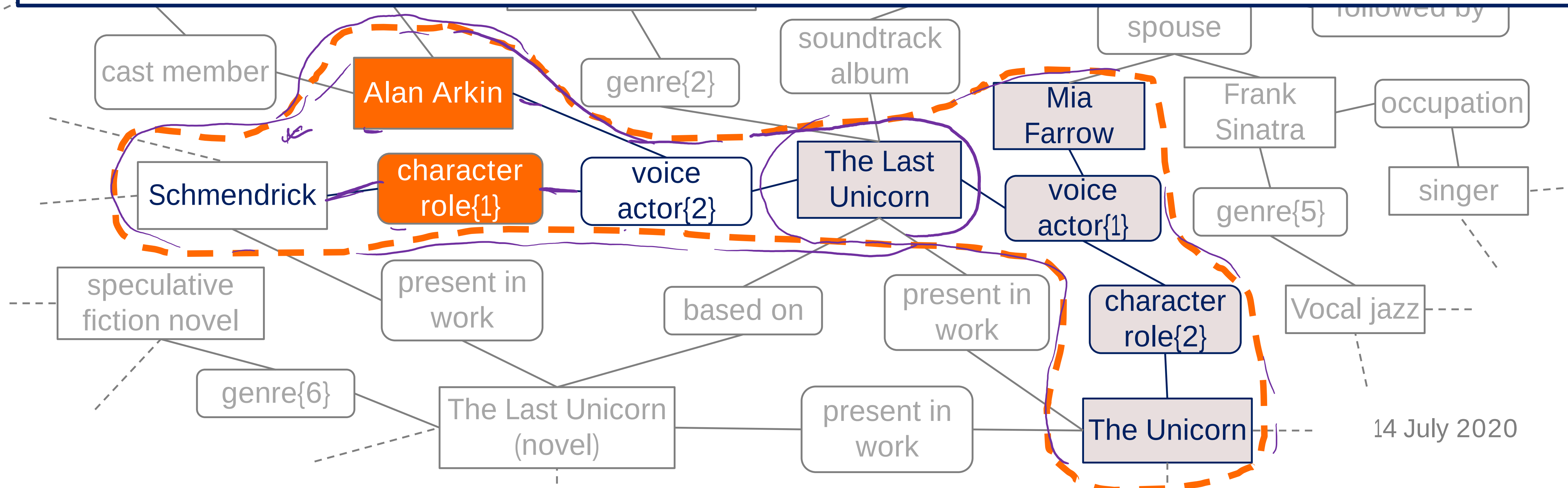
# Exploring context neighborhood

And Alan Arkin was behind . . .?



# Exploring context neighborhood

Determine "Frontier" nodes to describe an expansion border

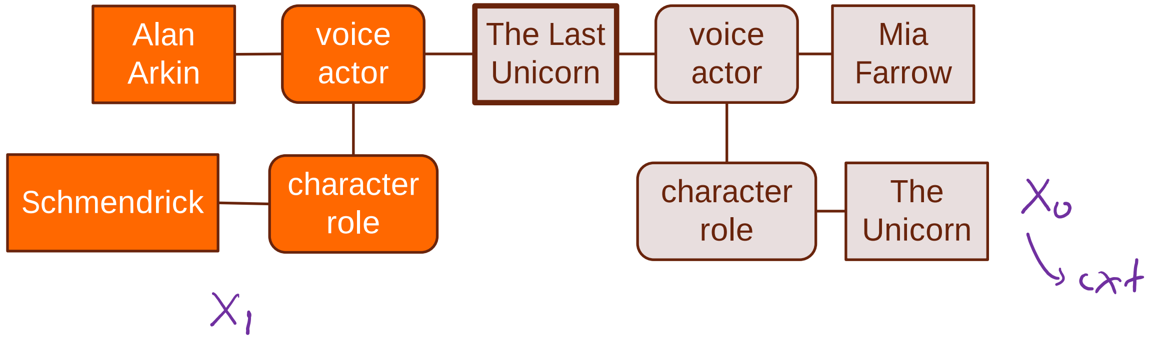


# Context graph

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?




Expand graph accordingly!

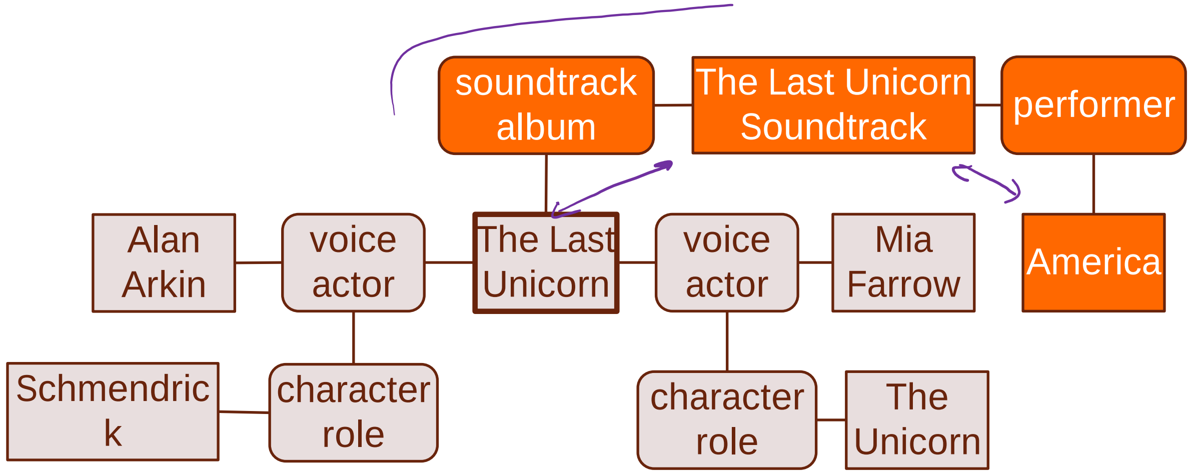
# Context graph

Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn? $q_0$

And Alan Arkin was behind . . .? $q_1$

So who performed songs? $q_2$




Graph expanded with relevant facts only

# Context graph

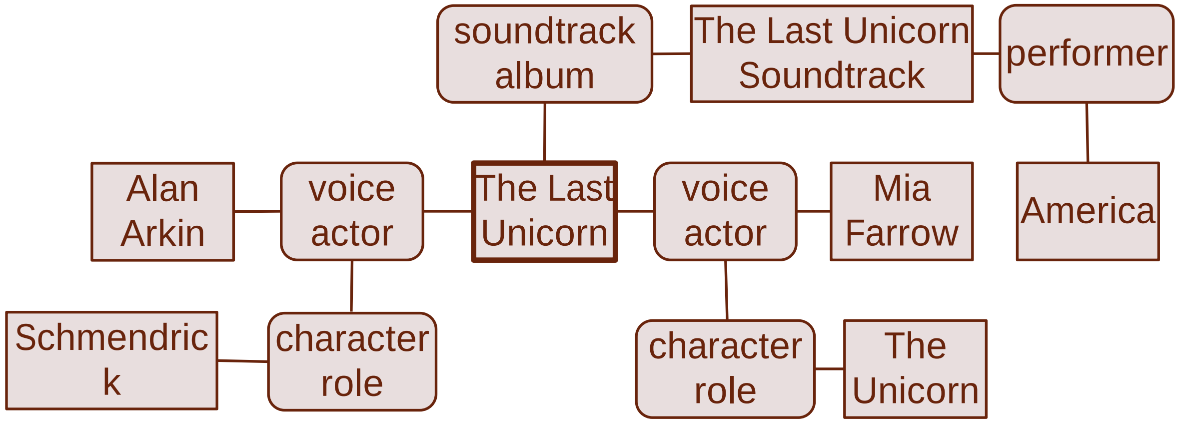
Mia Farrow

Which actor voiced the character Unicorn in The Last Unicorn?

And Alan Arkin was behind . . .?

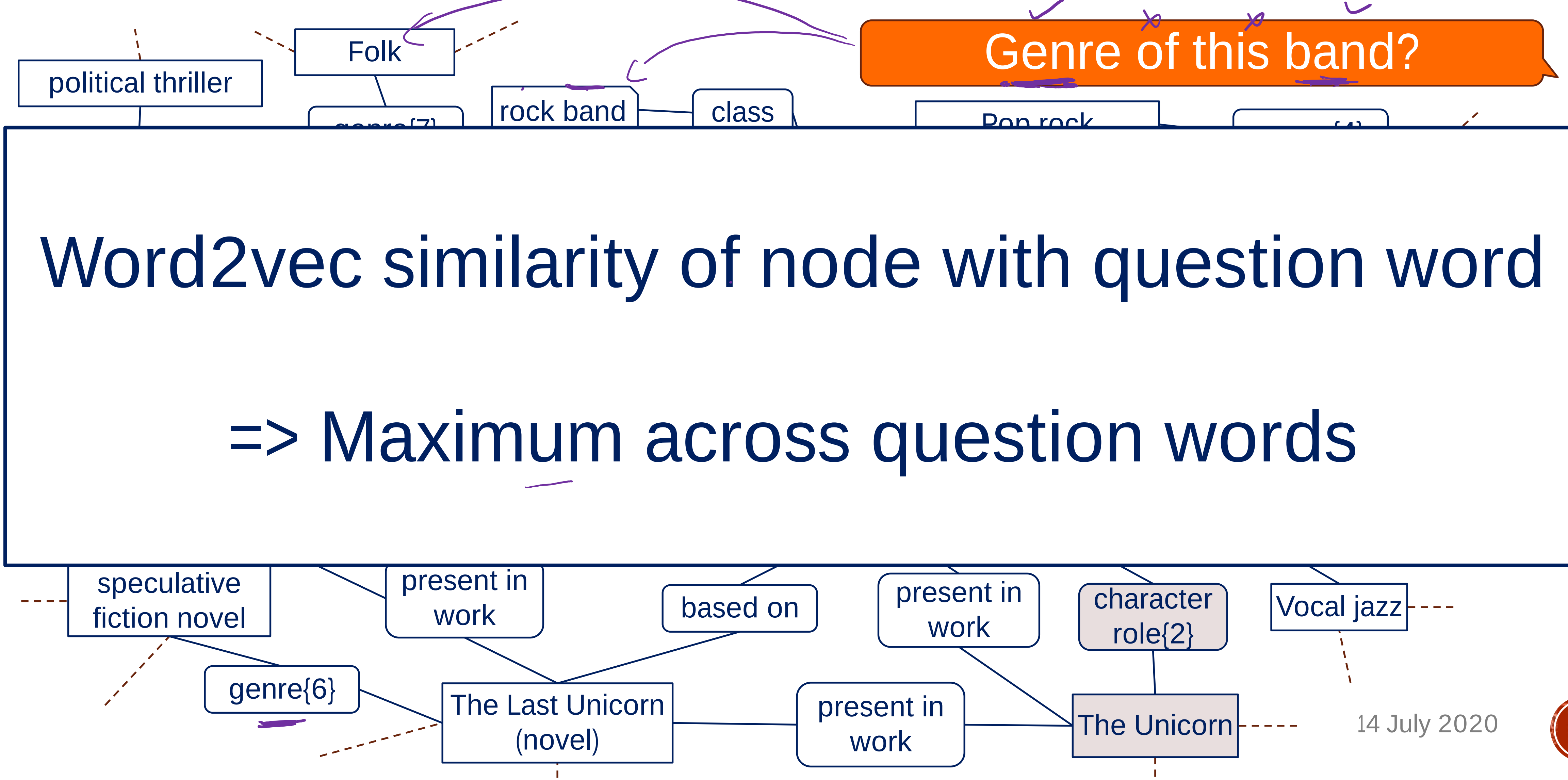
So who performed songs?

Genre of this band?




How to determine Frontier nodes?

# Relevance to the question

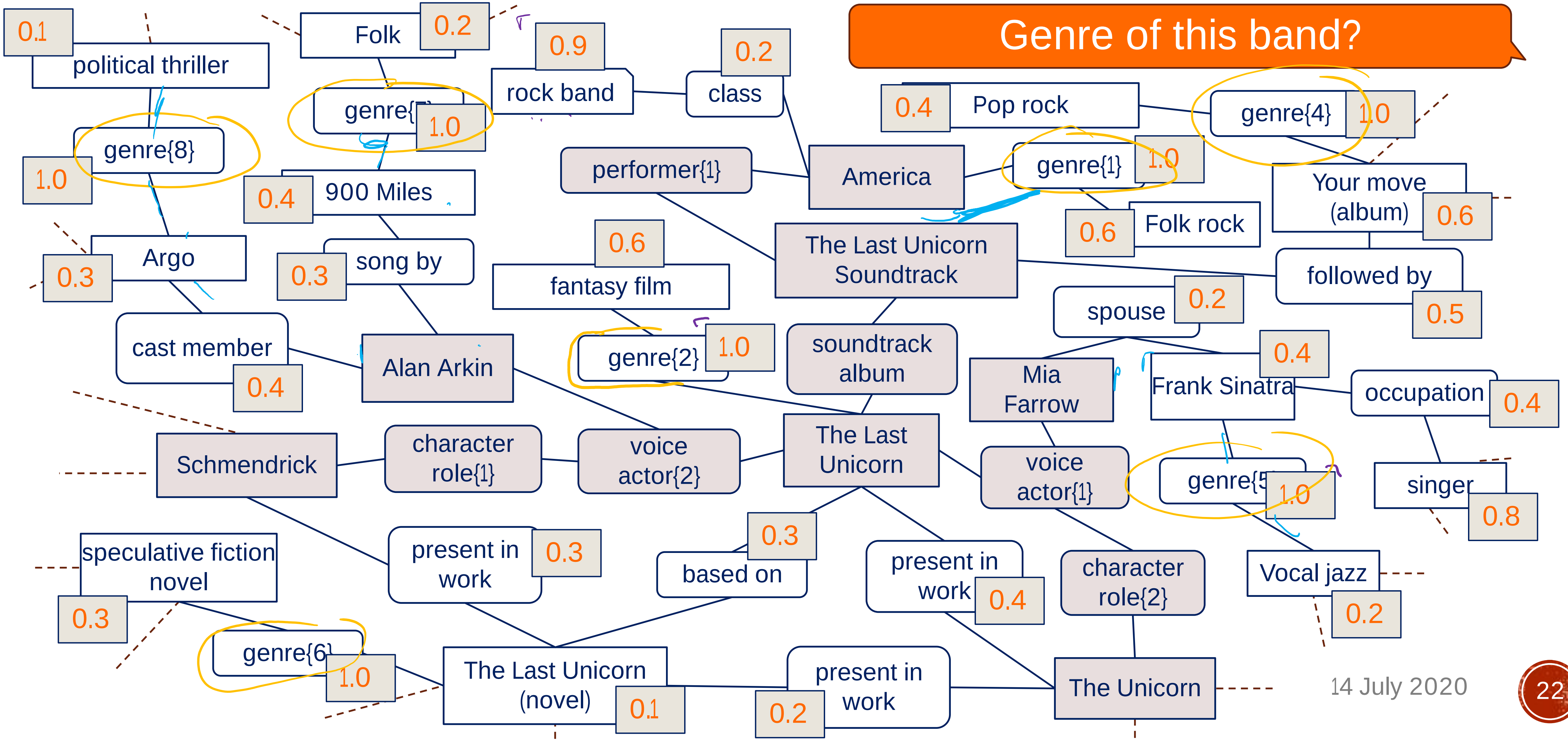



Word2vec similarity of node with question word

=> Maximum across question words

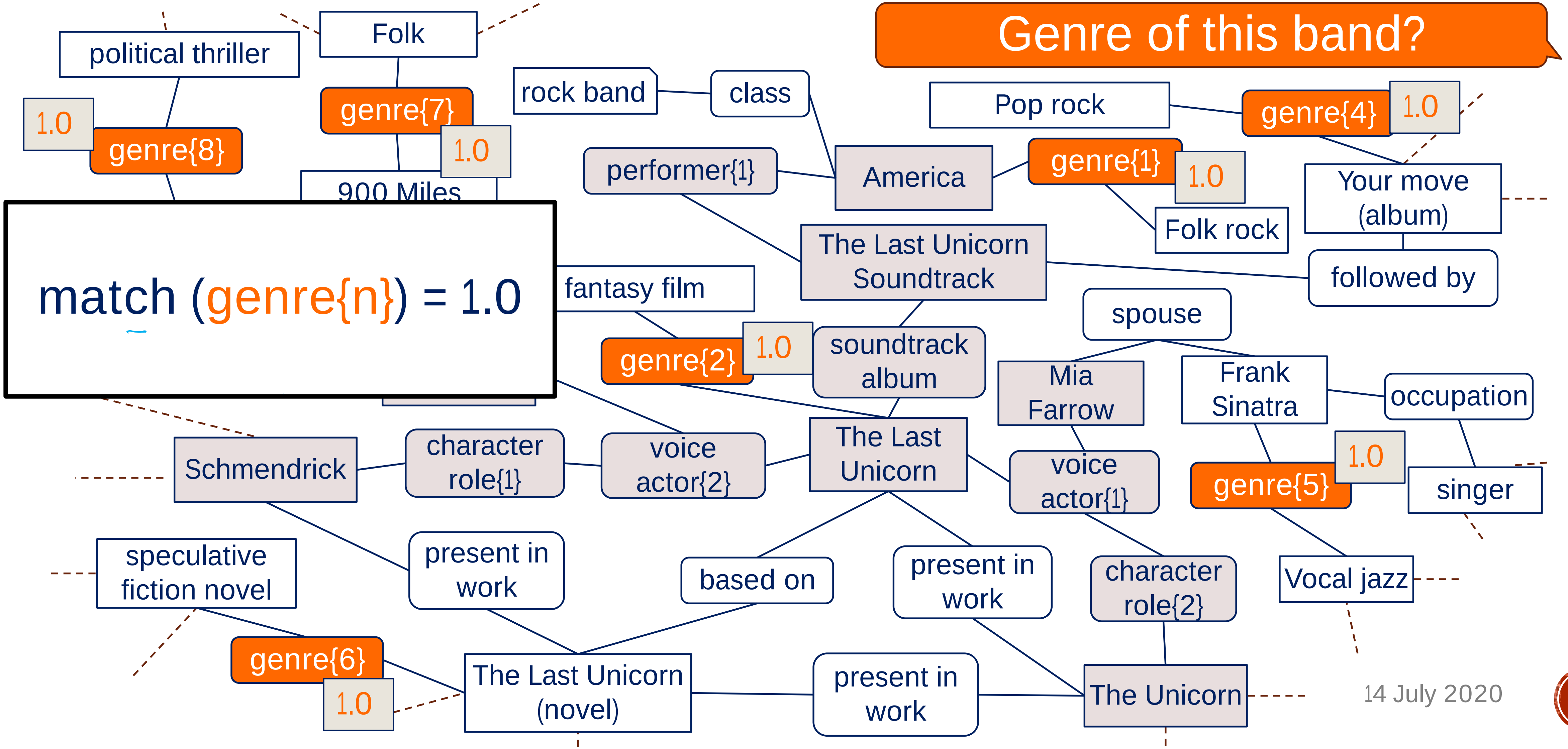
# Relevance to the question

incomplete cxt

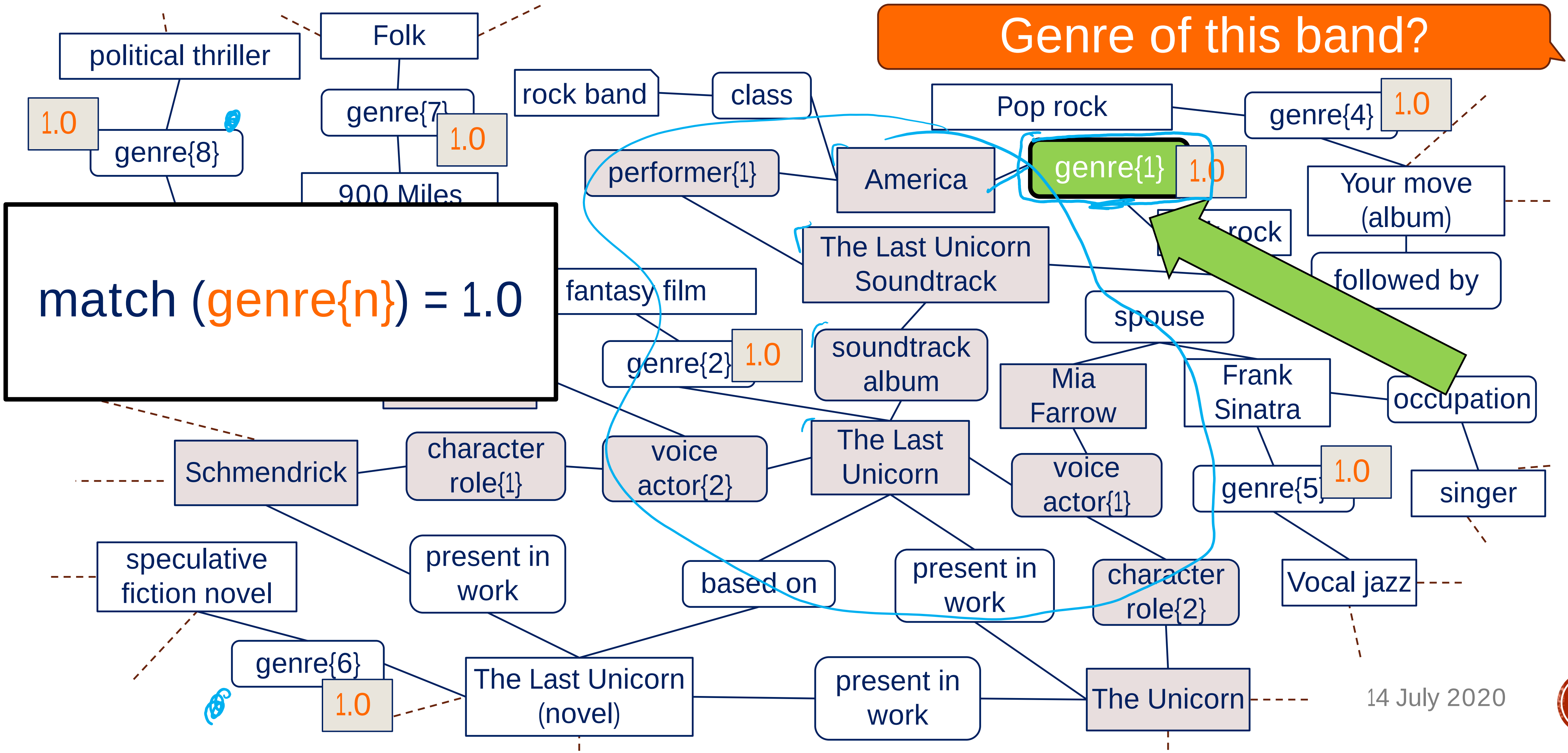


2

# Relevance to the context

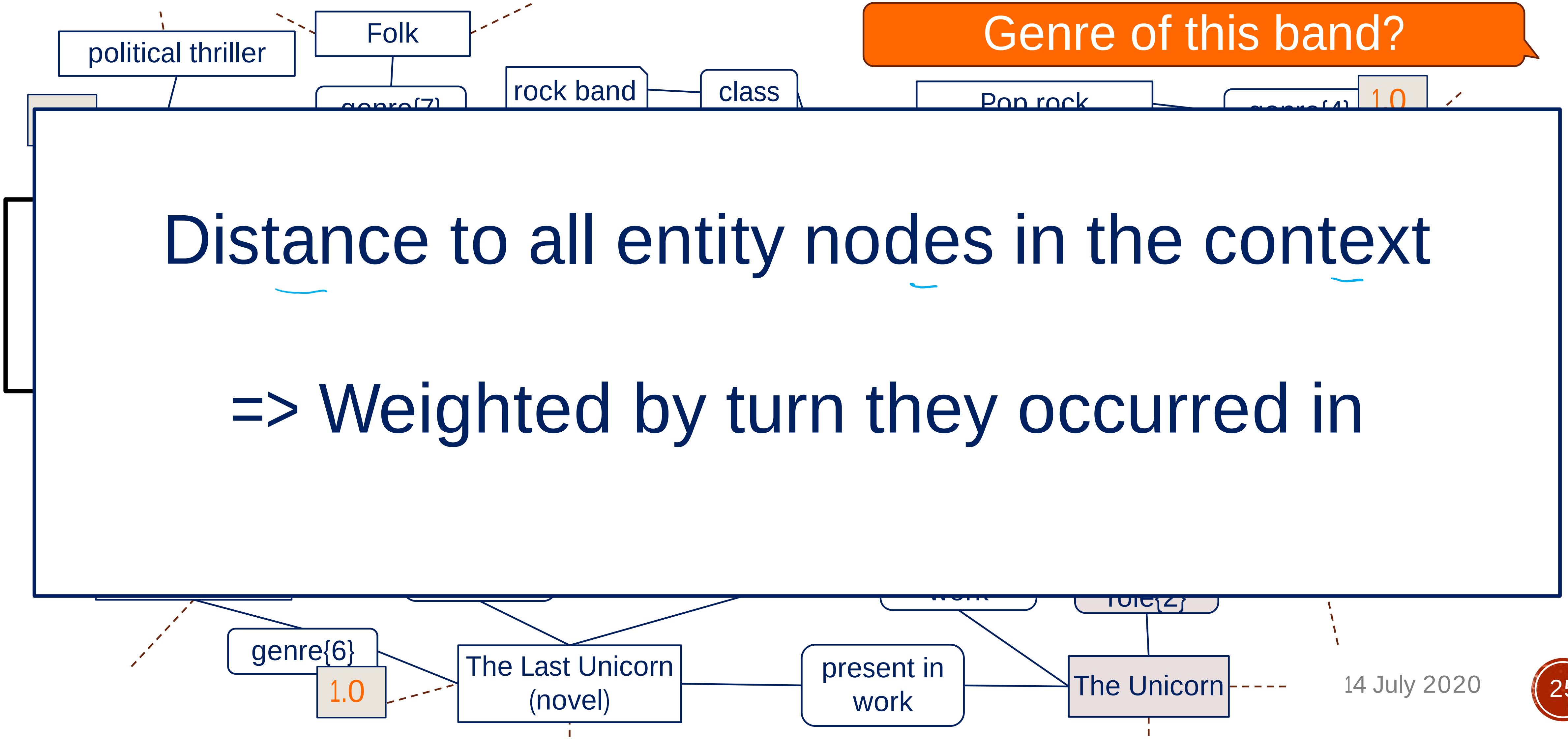


# Relevance to the context




match (genre{n}) = 1.0

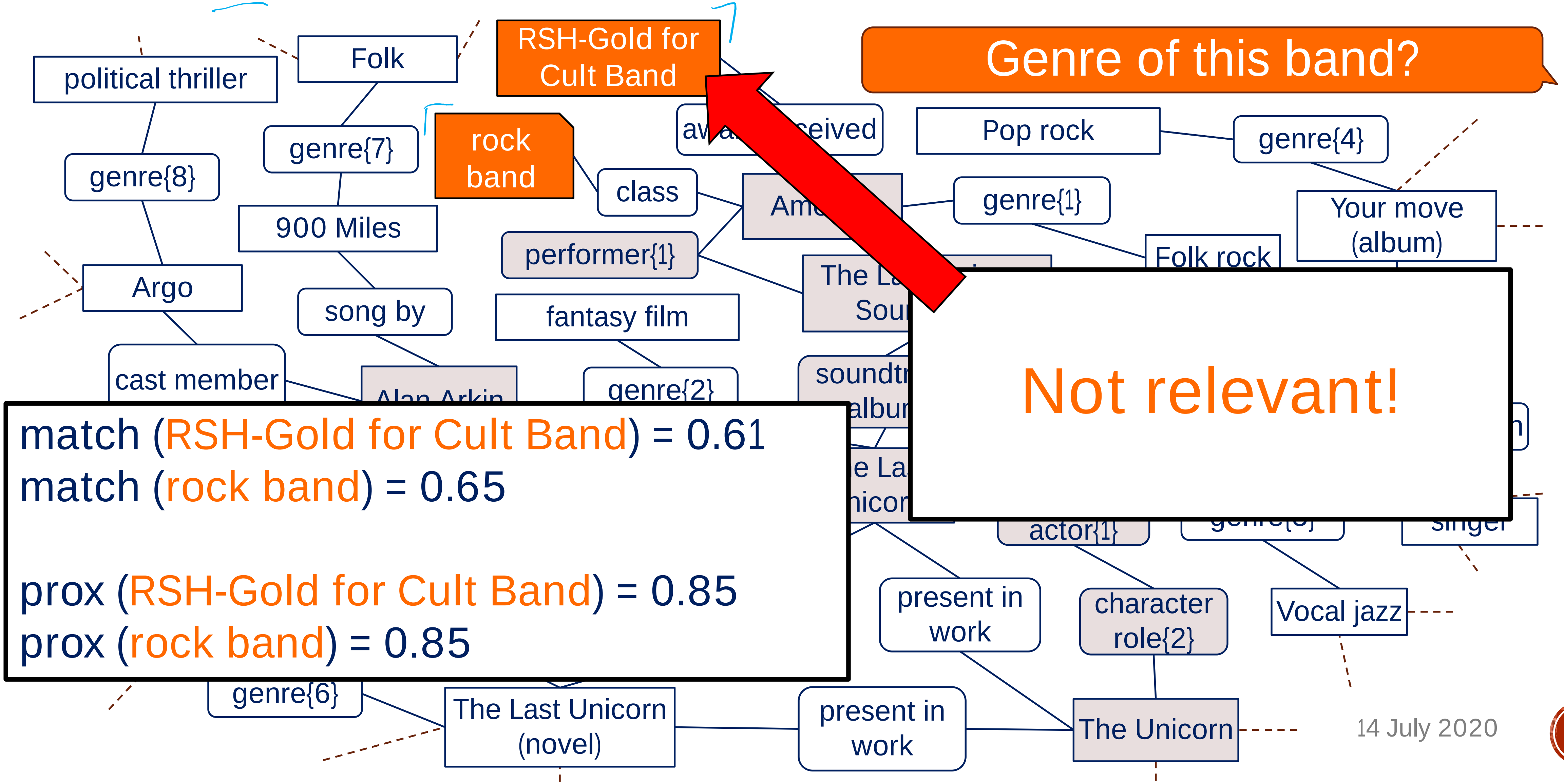
# Relevance to the context




Distance to all entity nodes in the context

=> Weighted by turn they occurred in

# KG priors




match (RSH-Gold for Cult Band) = 0.61
match (rock band) = 0.65

prox (RSH-Gold for Cult Band) = 0.85
prox (rock band) = 0.85

# KG priors

RSH-Gold for

Prioritize the more frequent/prominent entities and predicates

=> Normalize the value with maximum frequency

prox (rock band) = 0.85

genre{6}

The Last Unicorn (novel)

present in work

The Unicorn

# Frontier score

① **Matching similarity**

*match (candidate c)*

② **Context relevance**

*prox (candidate c)*

③ **KG priors**

*prior (candidate c)*

$$frontier\_score(candidate\ c) = h_1 \cdot match(c) + h_2 \cdot prox(c) + h_3 \cdot prior(c)$$

With hyperparameters $h_1$, $h_2$, $h_3$

↳ dev set

# Frontier nodes

1-hop nbd of $x_t$

## Matching similarity

question ranked list

| *Candidate* | *Match* |
|---|---|
| genre{1} | 1.00 |
| genre{2} | 1.00 |
| … | … |
| folk rock band | 0.89 |
| RSH-Gold for Cult Band | 0.87 |
| fantasy film | 0.36 |
| … | … |

## Context relevance

| *Candidate* | *Prox* |
|---|---|
| genre{1} | 0.91 |
| folk rock band | 0.86 |
| RSH-Gold for Cult Band | 0.86 |
| … | … |
| genre{2} | 0.34 |
| fantasy film | 0.36 |
| … | … |

## KG priors

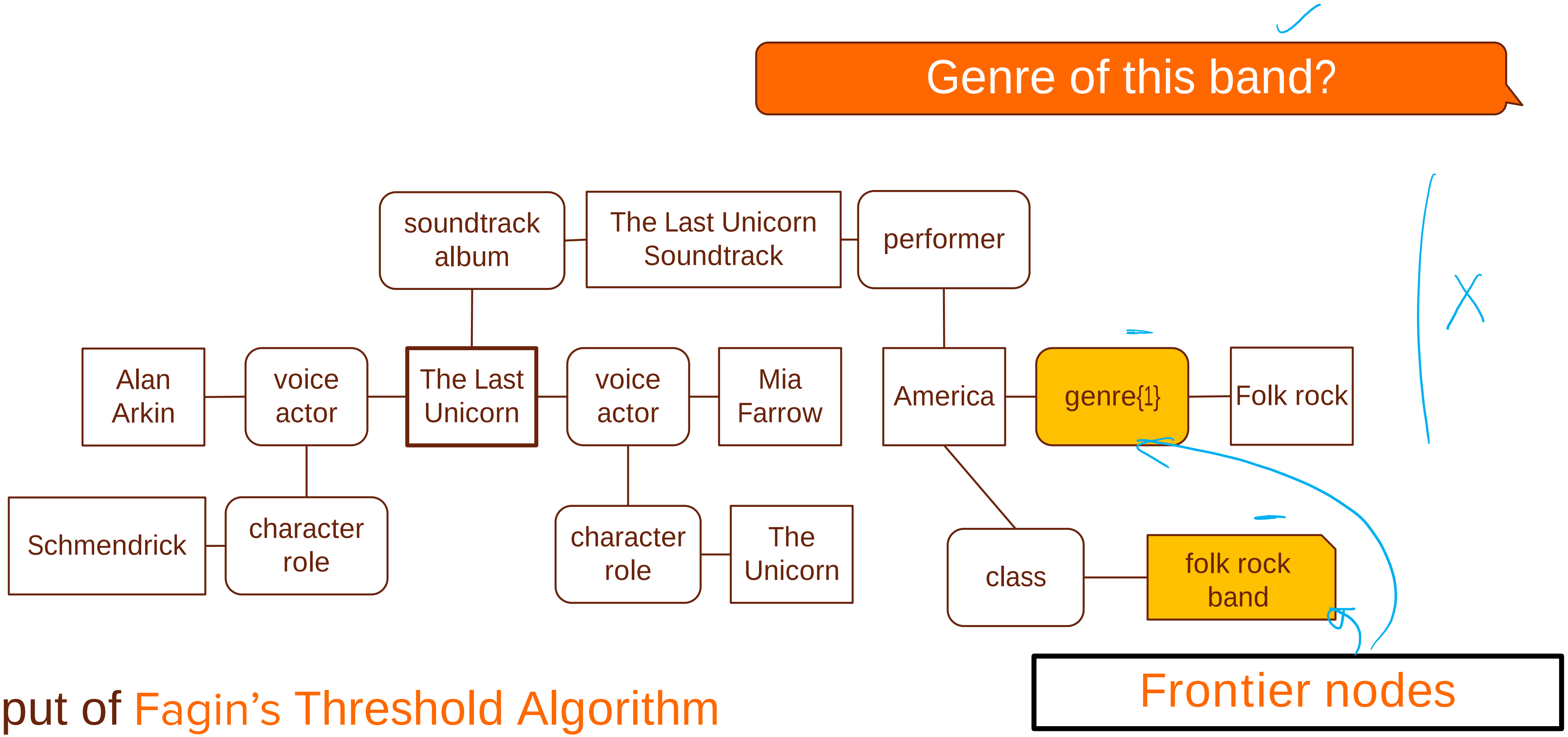
| *Candidate* | *KG priors* |
|---|---|
| … | … |
| genre{1} | 0.56 |
| genre{2} | 0.56 |
| … | … |
| folk rock band | 0.34 |
| … | … |
| RSH-Gold for Cult Band | 0.01 |

rank aggregation (IR)

DB community

Fagin's Threshold Algorithm to retrieve $\boxed{\text{top-}k}$ ranked nodes according to frontier score

# Frontier nodes

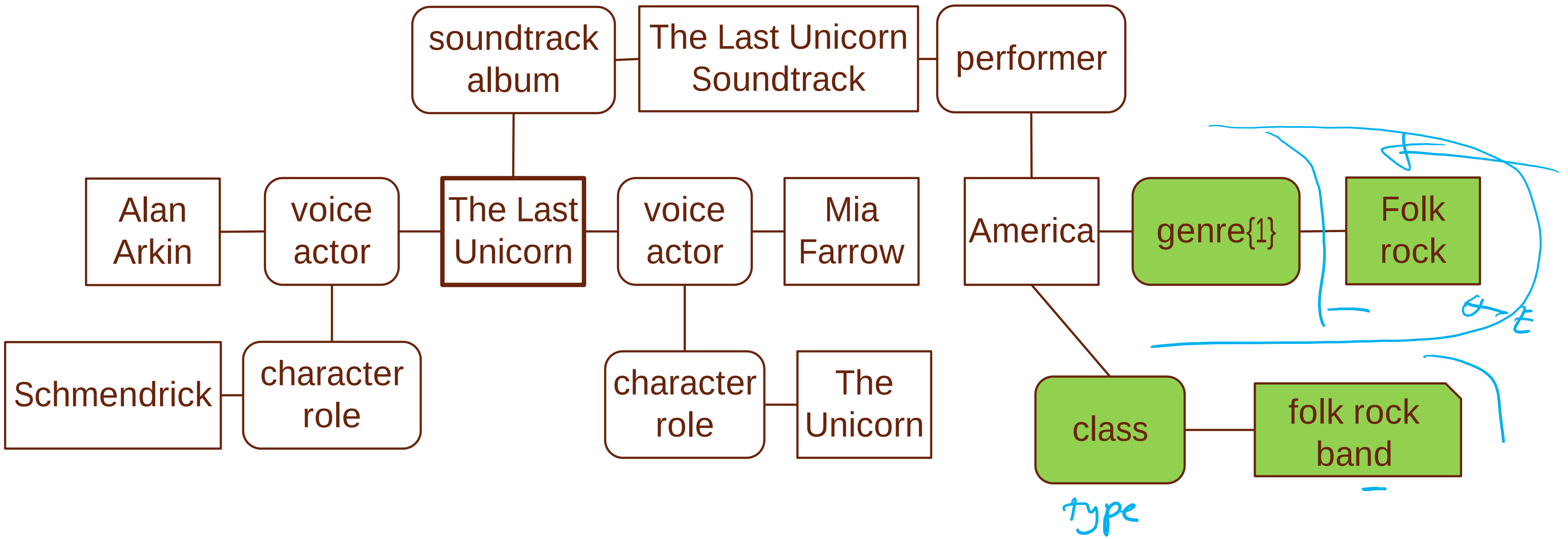



- Output of Fagin's Threshold Algorithm

  ⇒ Top-ranked candidates according to Frontier score

# Frontier nodes

Genre of this band?




$X_t^+$

# Answer to the question

Genre of this band?

- Distance to Frontier nodes
  - Weighted by the frontier score
  - distance_F

=> Explicit part

- Distance to all nodes in context graph $X_t$
  - Weighted by the turn they occurred in
  - distance_X

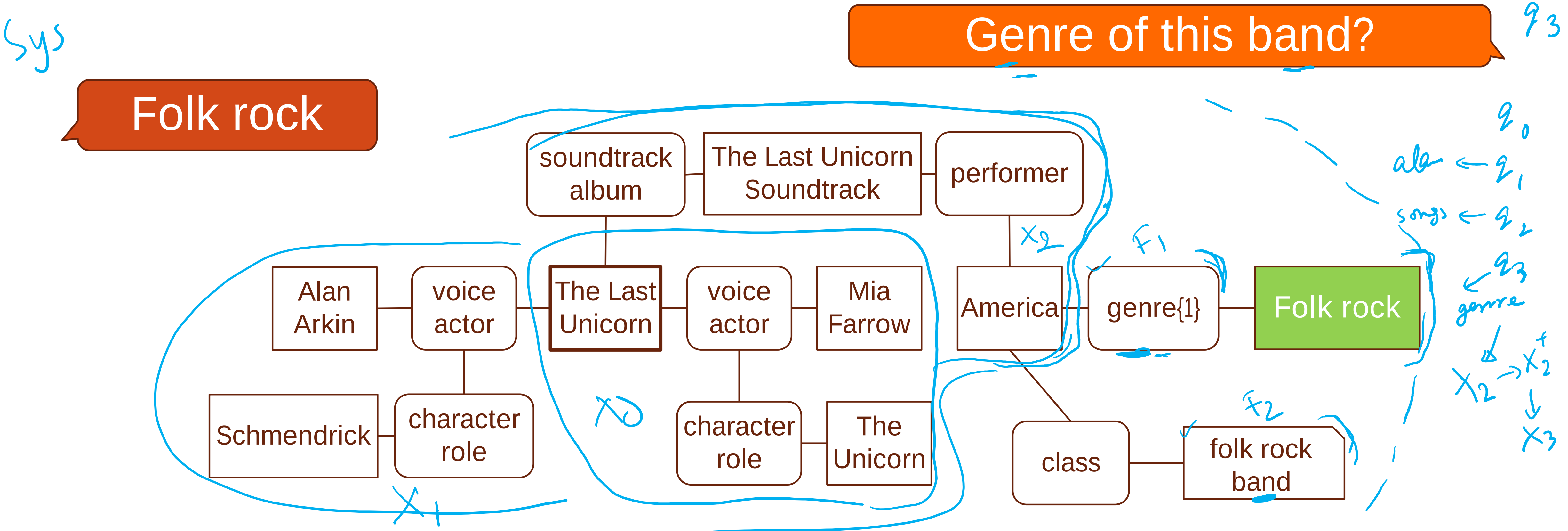
=> Implicit part

$X_t^+$

$$answer\_score(candidate\ c) = h_4 \cdot distance\_F + h_5 \cdot distance\_X$$

linear combination

# Answer detection




- Top-ranked node according to *answer_score*

# Answering steps

- Define expansion border
- Determine most relevant nodes in neighborhood of context
  $X_t$
  - Frontier nodes

  $X_t^+$
- Expand context according to frontier nodes
- Detect answer in expanded graph $X_t^+$

# Experimental dataset: ConvQuestions

CSQA
↓
not realistic

- 11,200 distinct conversations

turns

- 5 utterances per conversation

incomplete

- Initial question + 4 follow-up questions

1 2 3 4 5

- Domains: Books, Movies, Music, TV Series, Soccer
- Gathered via crowdsourcing AMT

# Experimental dataset: ConvQuestions

- Realistic benchmark
  - Questions created by humans from Amazon Mechanical Turk
  - In topic of their choice → informed
    → entity
- Natural flow of conversations
  - Conversations were not interleaved
  - Order of utterances was not permuted
    → sequence dependent

| Books | Movies | Soccer | Music | TV series |
|---|---|---|---|---|
| When was the first book of the book series The Dwarves published? | Who played the joker in The Dark Knight? | Which European team did Diego Costa represent in the year 2018? | Led Zeppelin had how many band members? | Who is the actor of James Gordon in Gotham? |
| *2003* | *Heath Ledger* | *Atletico Madrid* | *4* | *Ben McKenzie* |
| What is the name of the second book? | When did he die? | Did they win the Super Cup the previous year? | Which was released first: Houses of the Holy or Physical Graffiti? | What about Bullock? |
| *The War of the Dwarves* | *22 January 2008* | *No* | *Houses of the Holy* | *Donal Logue* |
| Who is the author? | Batman actor? | Which club was the winner? | Is the rain song and immigrant song there? | Creator? |
| *Markus Heitz* | *Christian Bale* | *Real Madrid C.F.* | *No* y/N | *Bruno Heller* |
| In which city was he born? | Director? | Which English club did Costa play for before returning to Atletico Madrid? | Who wrote those songs? | Married to in 2017? |
| *Homburg* | *Christopher Nolan* | *Chelsea F.C.* | *Jimmy Page* | *Miranda Cowley* |
| When was he born? | Sequel name? | Which stadium is this club's home ground? | Name of his previous band? | Wedding date first wife? |
| *10 October 1971* | *The Dark Knight Rises* | *Stamford Bridge* | *The Yardbirds* | *19 June 1993* |

https://convex.mpi-inf.mpg.de/

# Research paper 2

Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base multi-task semantic parsing

# MaSP Outline

- Task Definition
- ~~Existing Methodology~~
- Issues
- Proposed Approaches: Multi-task Learning

NED

QA

$q_3$: who played alfred?

Thanks to Tao Shen for the slides

# Task definition

- Targeting Knowledge-based Question Answering (KB-QA)
  - The backend *Knowledge Based* (KB) is large-scale, e.g., several million entities
  - The QA is conversational, i.e., *Co-reference* or *Ellipsis* might occur

Large-scale KB:

Data Format:
(subject, predicate, object)

Common KB:
Wikidata: 57M data items
Freebase: 1.9B data items
...

Coreference: he/she

**USER :** Can you tell me which cities border Verderio Inferiore ?
**SYSTEM :** Cornate d'Adda, Bernareggio, Robbiate

**USER :** And which cities flank that one ?
**SYSTEM :** Did you mean Robbiate ?

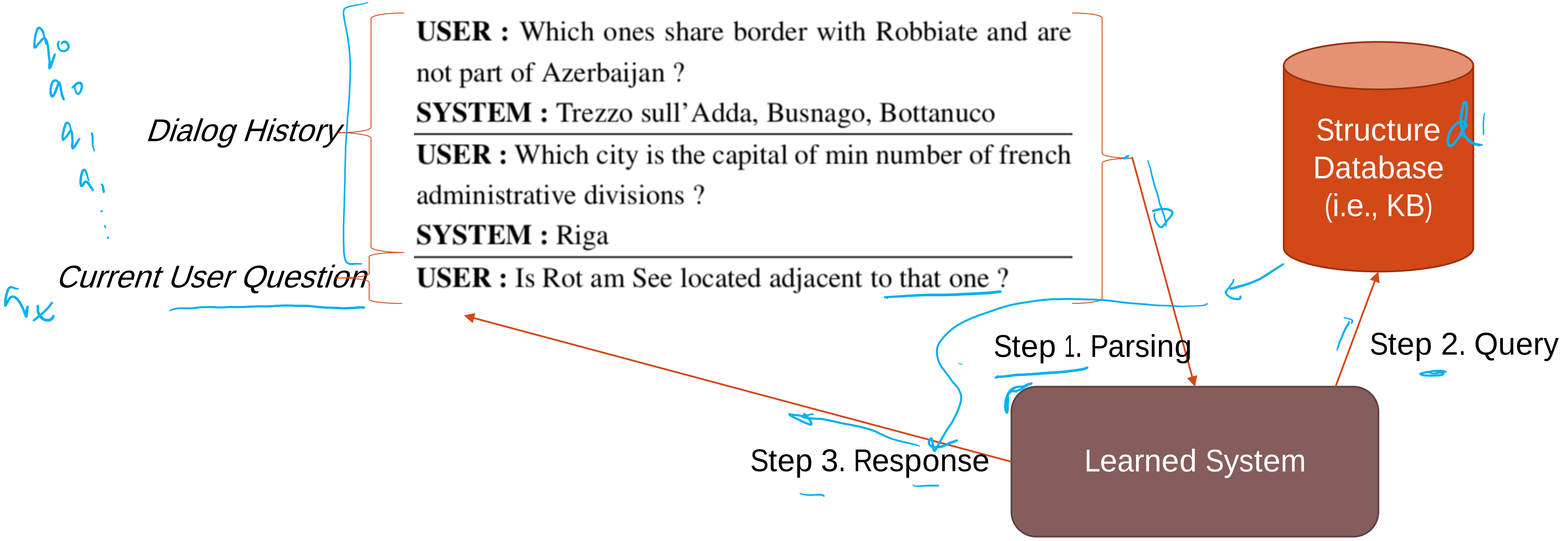
Ellipsis:

**USER :** How many countries are diplomatically related to Italy ?
**SYSTEM :** 74

countries

**USER :** And how many of them also with Australia ?
**SYSTEM :** 56

# Solution pipeline



# Semantic parsing-based KB-QA

hand-coded → lead to logical form  sequence

An Example of Grammar Definition:

formal CSG action CFG PCFG

| Alias | Operator | Comments |
|---|---|---|
| A1/2/3 | $start \rightarrow set/num/bool$ | |
| A4 | $set \rightarrow \text{find}(set, p)$ | set of entities with a predicate $p$ edge to entity $e$ |
| A5 | $num \rightarrow \text{count}(set)$ | number of distinct elements in the input $set$ |
| A6 | $bool \rightarrow \text{in}(e, set)$ | whether the entity $e$ in $set$ or not |
| A7 | $set \rightarrow \text{union}(set_1, set_2)$ | $set_1 \cup set_2$ |
| A8 | $set \rightarrow \text{inter}(set_1, set_2)$ | $set_1 \cap set_2$ |
| A9 | $set \rightarrow \text{diff}(set_1, set_2)$ | $set_1$ - $set_2$ |
| A10 | $set \rightarrow \text{large}(set, p, num)$ | subset of set linking to more than $num$ entities with predicate $p$ |
| A11 | $set \rightarrow \text{less}(set, p, num)$ | subset of set linking to less than $num$ entities with predicate $p$ |
| A12 | $set \rightarrow \text{equal}(set, p, num)$ | subset of set linking to $num$ entities with predicate $p$ |
| A13 | $set \rightarrow \text{argmax}(set, p)$ | subset of set linking to most entities with predicate $p$ |
| A14 | $set \rightarrow \text{argmin}(set, p)$ | subset of set linking to least entities with predicate $p$ |
| A15 | $set \rightarrow \text{filter}(tp, set)$ | subset where entity $e$ in set and belong to entity type $tp$ |
| A16 | $num \rightarrow u\_num$ | transform number in utterance $u\_num$ to intermediate number $num$ |
| A17 | $set \rightarrow \text{set}(e)$ | |
| A18/19/20 | $e/p/tp/u\_num \rightarrow \text{constant}$ | instantiation for $e$, $p$, $tp$, $u\_num$ from parsing results of the question |

◦ Non-terminals
◦ Terminals

rule

subs

$$set \rightarrow \text{find}(set, p)$$

Semantic Category Name

Arguments w/ specified semantic category

| | |
|---|---|
| Entry Semantic Category | e, p, tp, u_num |
| Intermediate Semantic Category | start, set, num, bool |

# Logical form/program

learning program induction → sequence of steps

1 0
2 0
3 0
4 0
5 0

Natural Language

*Question: who founded Apple Inc.*

Tree-Structured

generative model/ process

query

DFS

start A1

filter(set, tp) A15

A4 find(set, p)   tp: person A20

A17 set(e)   p: Founded by A19

e: Apple Inc.

A18 terminal

simplify with aliases

A1

A15

A4   tp: person

A17   p: Founded By

e: Apple Inc.

tree

flatten

Transform by DFS

Sequence-formatted

A1 A15 A4 A17 e:AppleInc. p:FoundedBy tp:person

seq2seq model for encoding

All transformation operation is reversible with the guidance from grammars.

# Issues

NERD before QA

- The errors in upstream tasks (e.g., entity detection & linking) lead to error propagation for downstream subtasks (e.g. , logical form generation)
- The subtasks are learned separately and thus cannot share supervisions
- Only low-level features of entities (e.g., mean-pooling over embeddings of composing words) are used regardless of any context information over the entities.

NER prive & persis

neural NED

# Multi-task learning over subtasks

MaSP

- Highlights
  - Pointer-equipped semantic parsing model
    - Pointer networks are used to point toward entity mention and number in question
    - The pointer-based model facilitates multi-task learning with upstream sequence labeling subtask, i.e., entity detection
    - The pointer-based model explicitly takes into account the context of entity mentions
  - Type-aware entity detection method
    - A joint prediction space combining entity detection and entity type is employed
    - The predicted type is then used to filter entity linking results during inference phase

Harry movie

jt setup

which actor

Nolan → dir

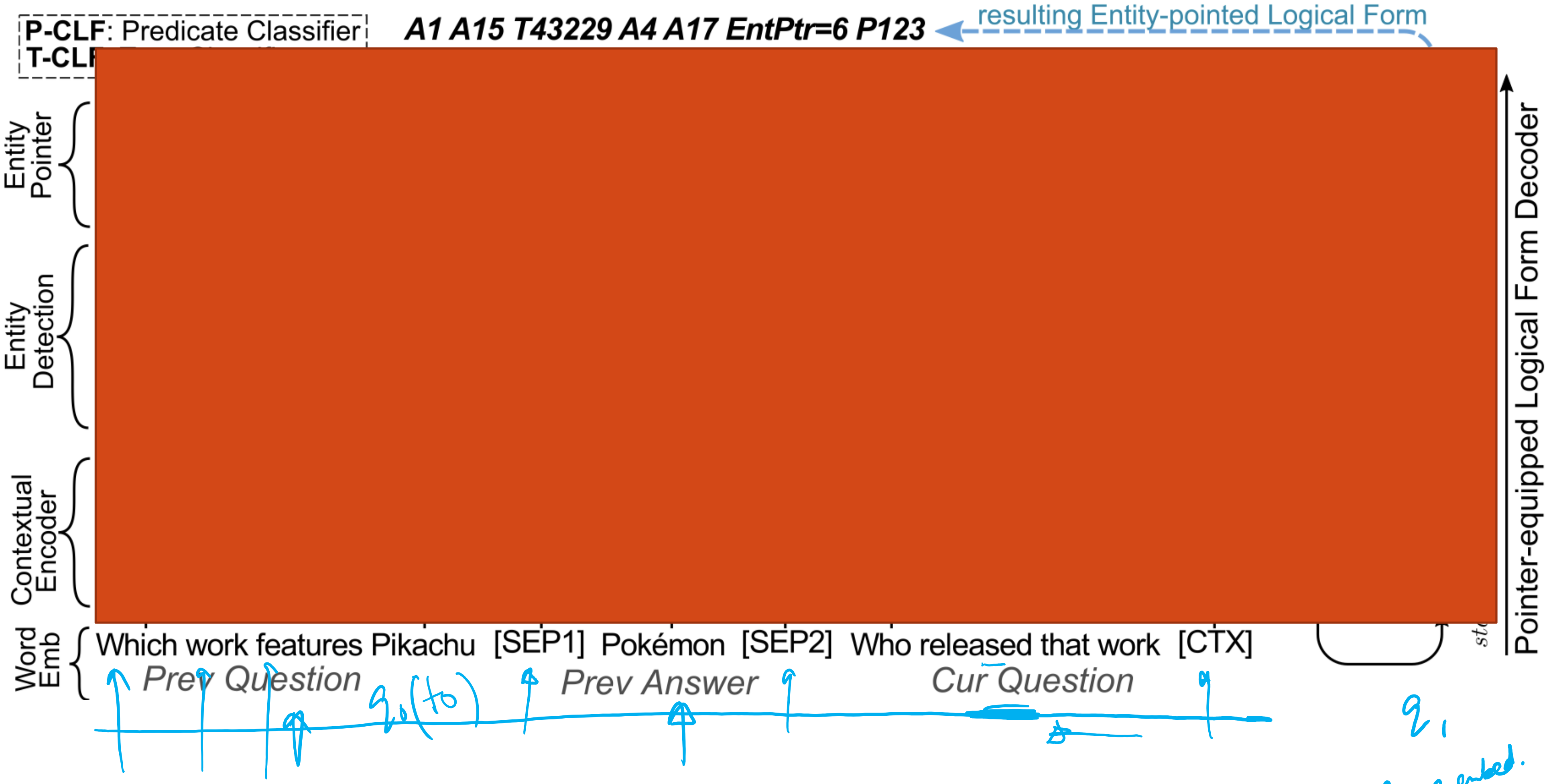
# Multi-task learning over subtasks

- Highlights
  - Pointer-equipped semantic parsing model
    - Pointer networks are used to point toward entity mention and number in question
    - The pointer-based model facilitates multi-task learning with upstream sequence labeling subtask, i.e., entity detection
    - The pointer-based model explicitly takes into account the context of entity mentions
  - Type-aware entity detection method
    - A joint prediction space combining entity detection and entity type is employed
    - The predicted type is then used to filter entity linking results during inference phase

# Pointer-equipped Semantic Parsing

1. Word Embedding

$$X = W^{(enc)}U$$
$$= [x_1, x_2, ..., x_n] \in \mathbb{R}^{d_e \times n}$$



# Pointer-equipped Semantic Parsing

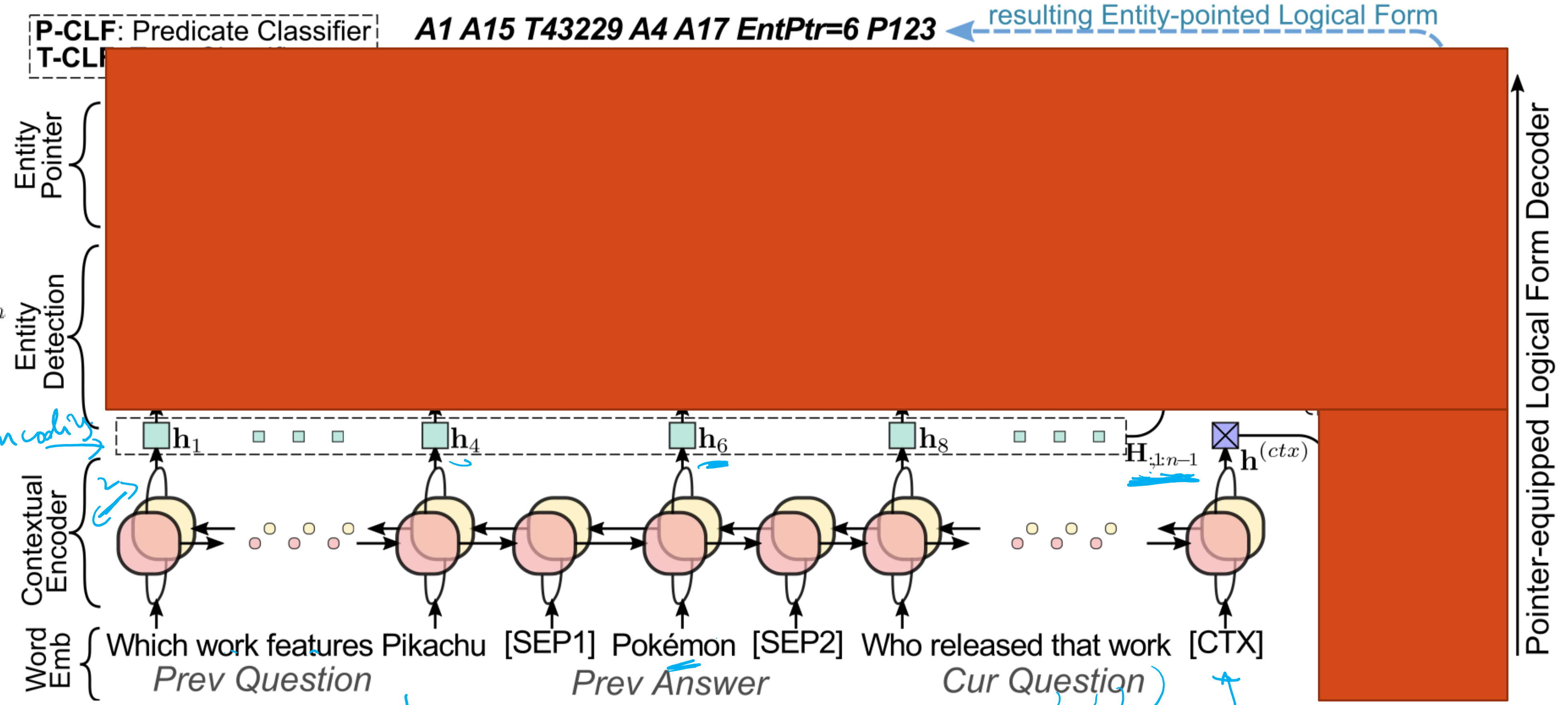
1. Word Embedding

$$\boldsymbol{X} = \boldsymbol{W}^{(enc)}\boldsymbol{U}$$
$$= [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}] \in \mathbb{R}^{d_e \times n}$$

2. Contextual Encoder

encoding

$$\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_n}] \triangleq \boldsymbol{X'} \in \mathbb{R}^{d_e \times n}, \text{where,}$$
$$2\times[\boldsymbol{X'} = \mathrm{FFN}(\mathrm{MultiHead}(\boldsymbol{X'}, \boldsymbol{X'}, \boldsymbol{X'}))],$$
$$\boldsymbol{X'} = \boldsymbol{X} + \boldsymbol{W}^{(pe)},$$



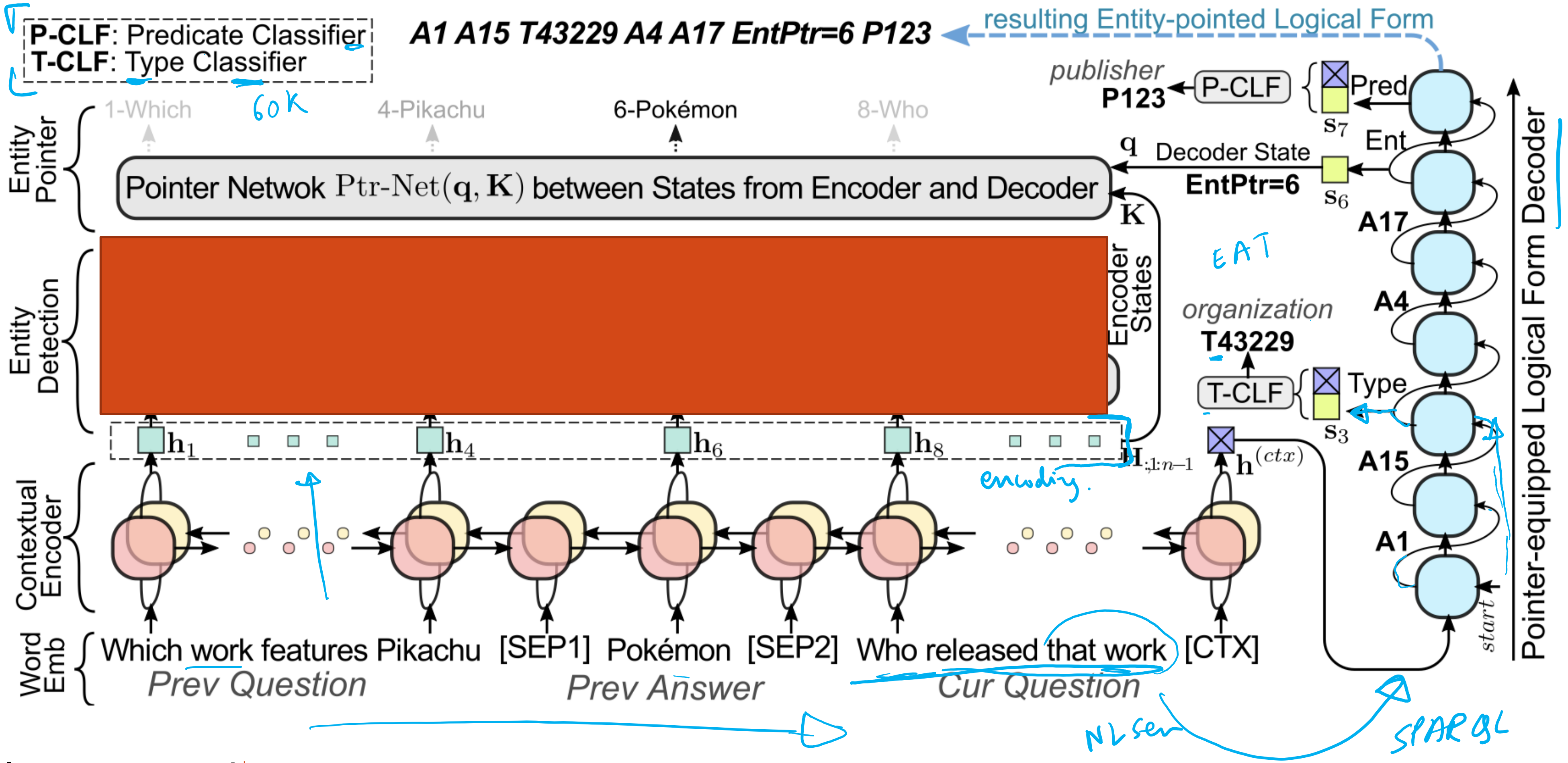

2017 | transformer

ip text sequence model LSTM → GRU RNN

(w2v)

encoding of input text

# Pointer-equipped Semantic Parsing




## 1. Word Embedding

$$X = W^{(enc)}U = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d_e \times n}$$

## 2. Contextual Encoder

$$H = [h_1, \cdots, h_n] \triangleq X' \in \mathbb{R}^{d_e \times n}, \text{where},$$
$$2\times [X' = \mathrm{FFN}(\mathrm{MultiHead}(X', X', X'))],$$
$$X' = X + W^{(pe)},$$

## 3. Pointer-equipped Logical Form Decoder

$$\mathbb{V}^{(dec)} = \{start,\ end,\ e,\ p,\ tp,\ u\_num,\ \mathrm{A1}, \cdots, \mathrm{A20}\}$$

$$S = [s_1, \cdots, s_m] \triangleq Z \in \mathbb{R}^{d_e \times m}, \text{where},$$
$$2\times [Z = \mathrm{FFN}(\mathrm{MultiHead}(H, H, \mathrm{MultiHead}^{mask}(Z, Z, Z)))].$$

$$p_j^{(tk)} = \mathrm{softmax}(\mathrm{FFN}(s_j; \theta^{(tk)}))$$

## For different entry semantic categories

- For predicate $p$ and type $tp$, two parameter-untied $\mathrm{FFN}(\cdot)$ are used as

$$p_j^{(p)} = \mathrm{softmax}(\mathrm{FFN}([s_j; h^{(ctx)}]; \theta^{(p)})),$$

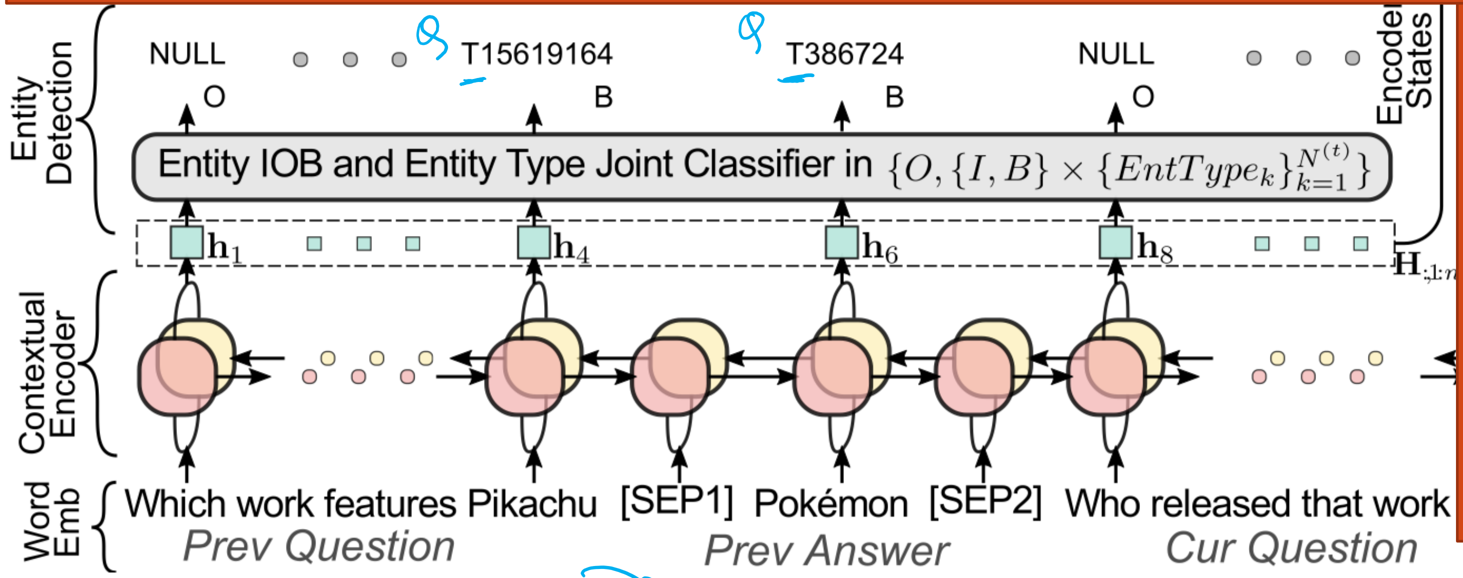
$$p_j^{(t)} = \mathrm{softmax}(\mathrm{FFN}([s_j; h^{(ctx)}]; \theta^{(t)})),$$

- For entity $e$ and number $u\_num$, two parameter-untied pointer-networks with learnable bilinear layer are employed to point toward the targeted entity and number, which are defined as follows.

$$p_j^{(e)} = \mathrm{softmax}(s_j^T W^{(e)} H_{:,1:n-1}),$$

$$p_j^{(n)} = \mathrm{softmax}(s_j^T W^{(n)} H_{:,1:n-1}),$$

# Type-aware Entity Detection




classes

Lord of the Rings

O

Joint Prediction Space of entity IOB-tagging and entity type:

60K

NER

$$\mathbb{E} = \{O, \{I, B\} \times \{ET_k\}_{k=1}^{N^{(t)}}\}$$

B I I

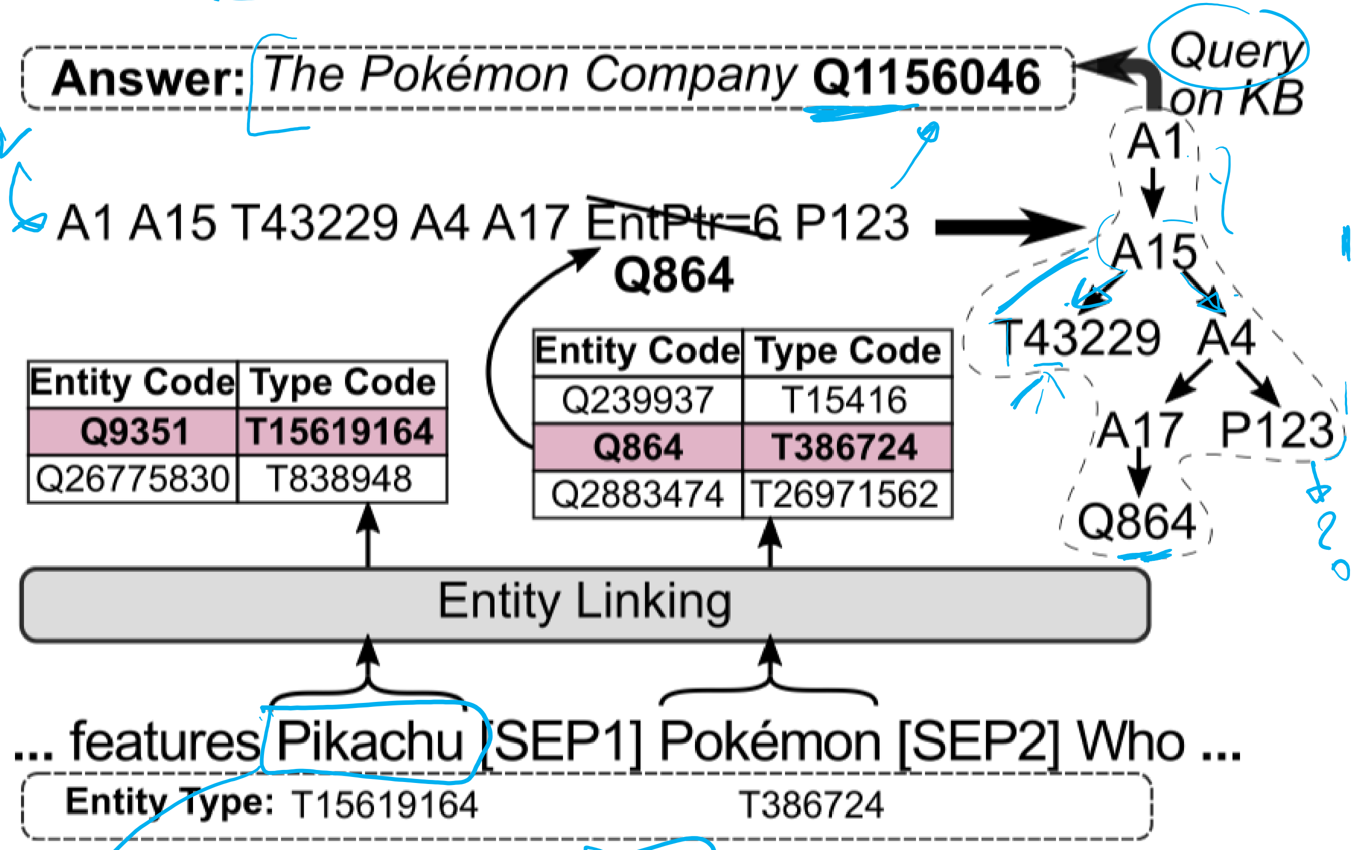
Prediction Probability:

$$p_i^{(ed)} = \text{softmax}(\text{FFN}(h_i; \theta^{(ed)})), \forall i \in [1, n-1]$$

# Entity linking and substitution

Steps (Bottom-up illustrated in Right Figure):

1. Entity Mentions Location
2. Inverted Index applied for entity candidates
3. Filtering candidates w.r.t. predicted entity type
4. Replacing the pointer value with the highest-scored candidate
5. Executing the complete logical form to query an answer.




search in all E

E in Wikid

# Learning and inference

- Loss Functions

$$L = \alpha L^{(sp)} + L^{(ed)},$$

linear combination

$$\left\{\begin{aligned} L^{(sp)} &= -\frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\frac{1}{m}\sum_{j=1}^{m}\log \boldsymbol{p}_{\boldsymbol{j}}^{(\boldsymbol{tk})}{}_{[tk'=y_j^{(tk)}]} \\ &+ \sum_{c\in\{p,t,e,n\}} I_{(y_j^{(tk)}=c)} \log \boldsymbol{p}_{\boldsymbol{j}}^{(\boldsymbol{c})}{}_{[c'=y_j^{(c)}]} \\ L^{(ed)} &= -\frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\frac{1}{n-1}\sum_{i=1}^{n-1}\log \boldsymbol{p}_{\boldsymbol{i}}^{(\boldsymbol{ed})}{}_{[ed'=y_i^{(ed)}]} \end{aligned}\right.$$

- Inference Phase
  - Grammar-guided Decoding
  - Beam Search
  - Early-stage Execution

# Putting it together

- Alleviating the error propagation problem
  - The approach reduces errors in entity detection and linking subtask by predicting the type of each entity mention
- Making the best of supervisions
  - The two subtasks, i.e., pointer-equipped semantic parsing and entity detection, are closely related
- Taking into account the context of entity mention
  - The approach is naturally beneficial to coreference resolution for conversational QA due to rich contextual features captured by pointer for entity mention

# Conclusions

- Conversational question answering is one of the most important future directions in QA
- Fueled by the rise of conversational assistants
- Benchmarks and methods – over KGs and text - still in infancy
- Context resolution is key challenge
- Much more than ellipsis and coreference resolution!!

Google assistant
Alexa
Cortana
Siri

ConvQuestions, CSQA

CoQA, QuAC

reformulating

question completion

# Take home messages

- Question answering is an extremely active area
- Key problems open
  - 1 Complex questions
  - 2 Conversational questions
  - 3 Heterogeneous sources KG + text
  - 4 User feedback
- Prototype your systems!! demos
- Read more at: [https://arxiv.org/pdf/2004.11980.pdf](https://arxiv.org/pdf/2004.11980.pdf)

KG
→ try to work on full KG
→ NED is part of the QA prob
→ Do not ignore qualifiers!!
5B

tutorial paper
SIGIR 2020