Simple Question Answering with Subgraph Ranking and Joint-Scoring

Wenbo Zhao, Tagyoung Chung, Anuj Goyal, and Angeliki Metallinou
Carnegie Mellon University
Amazon Alexa AI

NAACL 2019

Simple KBQA

- Subgraph selection(entity linking)
  - Find topic entity in KB
  - Extract neighbor entities
- Relation/Pattern matching

Scoring

- Mention score

$$\mathbb{P}(s \mid m_q) = \frac{e^{h(f(m_q), f(s))}}{\sum_{s' \in \mathcal{S}_q} e^{h(f(m_q), f(s'))}}$$

- Relation score

$$\mathbb{P}(r \mid p_q) = \frac{e^{h(g(p_q), g(r))}}{\sum_{r' \in \mathcal{R}_q} e^{h(g(p_q), g(r'))}}$$

- Ranking loss

$$\mathcal{L}_{\text{rank}} = \sum_{q \in \mathcal{Q}} \left(\sum_{s \in \mathcal{S}_q} \left[h_f(m_q, s^-) - h_f(m_q, s^+) + \lambda \right]_+ + \sum_{r \in \mathcal{R}_q} \left[h_g(p_q, r^-) - h_g(p_q, r^+) + \lambda \right]_+ \right), \quad (4)$$

Filter Candidate

- Literal closeness
  - length of the longest common subsequence
- Semantic closeness

$$
\begin{aligned}
\mathbb{P}(s,m) &= \mathbb{P}(s|m)\mathbb{P}(m) \\
&= \mathbb{P}(w_1,\dots w_n|\widetilde{w}_1,\dots \widetilde{w}_m)\mathbb{P}(\widetilde{w}_1,\dots \widetilde{w}_m) && (5) \\
&= \prod_{i=1}^{n}\mathbb{P}(w_i|\widetilde{w}_1,\dots \widetilde{w}_m)\mathbb{P}(\widetilde{w}_1,\dots \widetilde{w}_m) && (6) \\
&= \prod_{i=1}^{n}\left(\prod_{k=1}^{m}\mathbb{P}(w_i|\widetilde{w}_k)\right)\mathbb{P}(\widetilde{w}_1,\dots \widetilde{w}_m) && (7) \\
&= \prod_{i=1}^{n}\left(\prod_{k=1}^{m}\mathbb{P}(w_i|\widetilde{w}_k)\right)\prod_{j=1}^{m-1}\mathbb{P}(\widetilde{w}_{j+1}|\widetilde{w}_j)\mathbb{P}(\widetilde{w}_1), && (8)
\end{aligned}
$$

- Weighted score

$$\mathrm{score}(s,m) = \tau|\sigma|(s,m) + (1-\tau)\log\mathbb{P}(s,m)$$

Ranking Loss

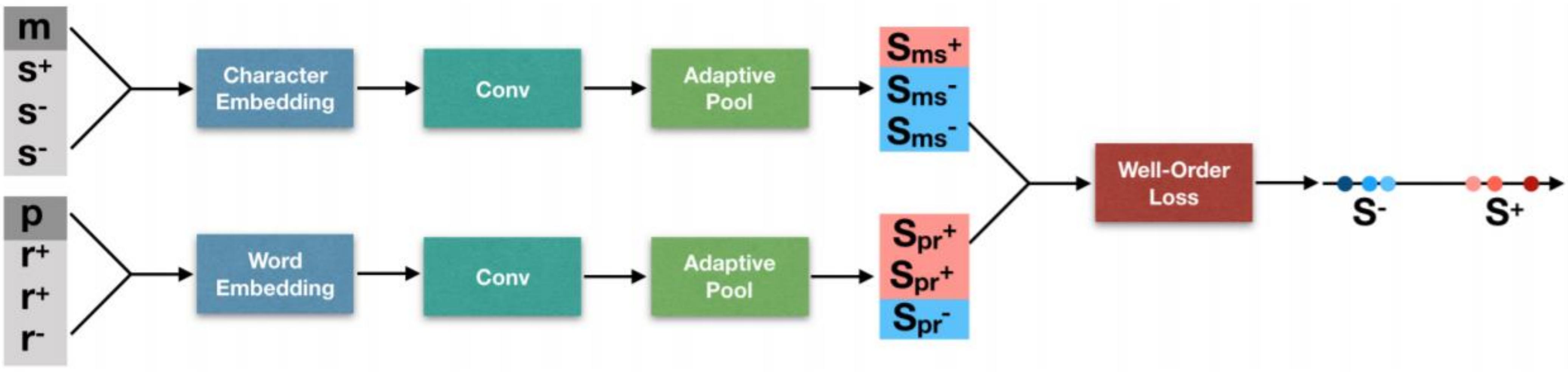


Figure 1: Model Diagram (Section 3.3) The model takes input pairs (**m**ention, **s**ubject) and (**p**attern, **r**elation) to produce the similarity scores. The loss dynamically adjusts the weights and enforces the order of positive and negative scores.

Ranking Loss

- Well-order loss

$$
\begin{aligned}
&\inf S^{+} \geq \sup S^{-} \\
\Leftrightarrow\ &\forall i^{+} \in I^{+} : \forall i^{-} \in I^{-} : S^{+}_{i^{+}} - S^{-}_{i^{-}} \geq 0 \\
\Leftrightarrow\ &\sum_{i^{+} \in I^{+}} \sum_{i^{-} \in I^{-}} (S^{+}_{i^{+}} - S^{-}_{i^{-}}) \geq 0 \\
\Leftrightarrow\ &n_2 \sum_{i^{+} \in I^{+}} S^{+}_{i^{+}} - n_1 \sum_{i^{-} \in I^{-}} S^{-}_{i^{-}} \geq 0,
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{L}_{\text{well-order}}(S_{ms}, S_{pr}) = \\
&\left[|I^{+}| \sum_{i^{-}} S^{i^{-}}_{ms} - |I^{-}| \sum_{i^{+}} S^{i^{+}}_{ms} + |I^{+}||I^{-}|\lambda \right]_{+} + \\
&\left[|J^{+}| \sum_{j^{-}} S^{j^{-}}_{pr} - |J^{-}| \sum_{j^{+}} S^{j^{+}}_{pr} + |J^{+}||J^{-}|\lambda \right]_{+},
\end{aligned}
$$

$$
\begin{aligned}
\min_{q \in \mathcal{Q}, (s,r) \in \mathcal{S}^{n}_{q\downarrow} \times \mathcal{R}^{n}_{q\downarrow}} &\Bigg[|I^{+}| \sum_{i^{-}} h_f(m_q, s^{i^{-}}) - \\
&\quad |I^{-}| \sum_{i^{+}} h_f(m_q, s^{i^{+}}) + |I^{+}||I^{-}|\lambda \Bigg]_{+} \\
+ &\Bigg[|J^{+}| \sum_{j^{-}} h_g(p_q, r^{j^{-}}) - \\
&\quad |J^{-}| \sum_{j^{+}} h_g(p_q, r^{j^{+}}) + |J^{+}||J^{-}|\lambda \Bigg]_{+} .
\end{aligned}
\tag{13}
$$

Ranking Loss

- Jointly scoring both mention detection and relation ranking module

- Dynamically adjust scoring based on candidate size

- Ability to prune relations with incorrect candidate

Experiment

	Approach	Obj. (= **Overall Acc.**)	Sub.	Rel.
1	AMPCNN (Yin et al., 2016)	76.4		
2	BiLSTM (Petrochuk and Zettlemoyer, 2018)	78.1		
3	AMPCNN + wo-loss	77.69		
4	JS + wo-loss	81.10	87.44	69.22
5	JS + wo-loss + sub50	**85.44**	91.47	76.98
6	JS + wo-loss + sub1	79.34	87.97	84.12

Table 5: Fact Selection Accuracy (%). The object accuracy is the end-to-end question answer accuracy, while subject and relation accuracies refer to separately computed subject accuracy and relation accuracy.

Experiment

Rank Method	Top-N	Recall
Literal: $|\sigma|$ + heuristics (Yin et al., 2016)	1	0.736
	5	0.850
	10	0.874
	20	0.888
	50	0.904
	100	0.916
Semantic: $\log \mathbb{P}$	1	0.482
	10	0.753
	20	0.854
	50	0.921
	100	0.848
Joint: $0.9|\sigma| + 0.1 \log \mathbb{P}$	1	0.855
	5	0.904
	10	0.920
	20	0.927
	50	0.945
	100	0.928

Table 4: Subgraph Selection Results

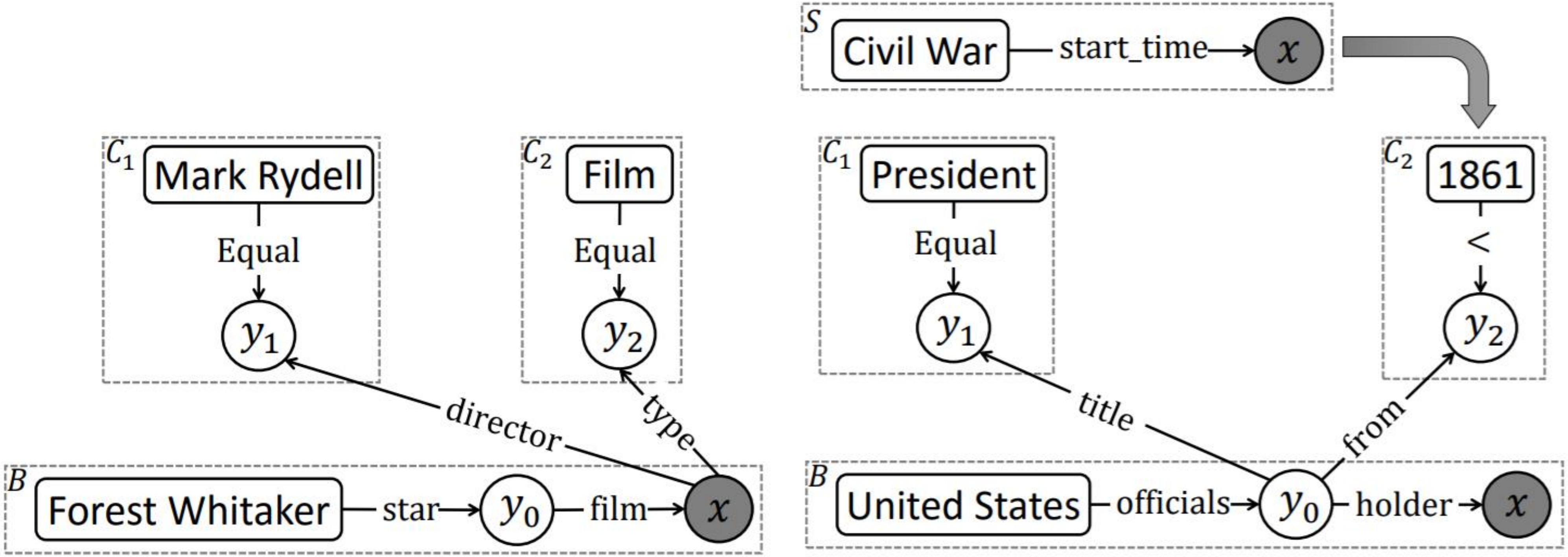
Incorrect Sub. only	8.67
Incorrect Rel. only	16.26
Incorrect Sub. & Rel.	34.50
Other	40.57

Table 6: Error Decomposition (%). Percentages for total of 3157 errors.

Constraint Detection in KBQA

[COLING16]Constraint-based Question Answering on Knowledge Base

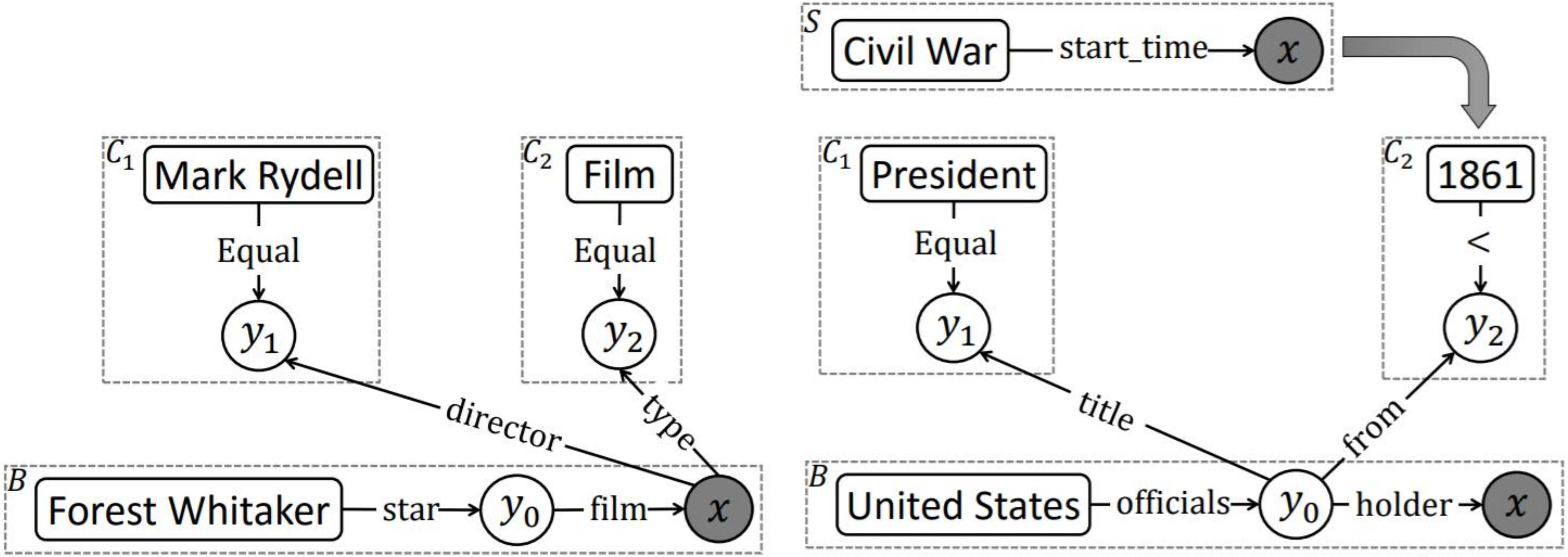
- Entity Constraint
  - Get from entity linking system. Check whether it's linked to basic graph
  - Similarity measured by path/pattern similarity



(a) MulCG with entity and type constraint for question "Which films star by Forest Whitaker and are directed by Mark Rydell". (b) MulCG with implicit temporal constraint for question "Who was U.S. president after the Civil War started".

[COLING16]Constraint-based Question Answering on Knowledge Base

- Type Constraint
  - Extract dependent word of qword, check if it's in type lexicon.

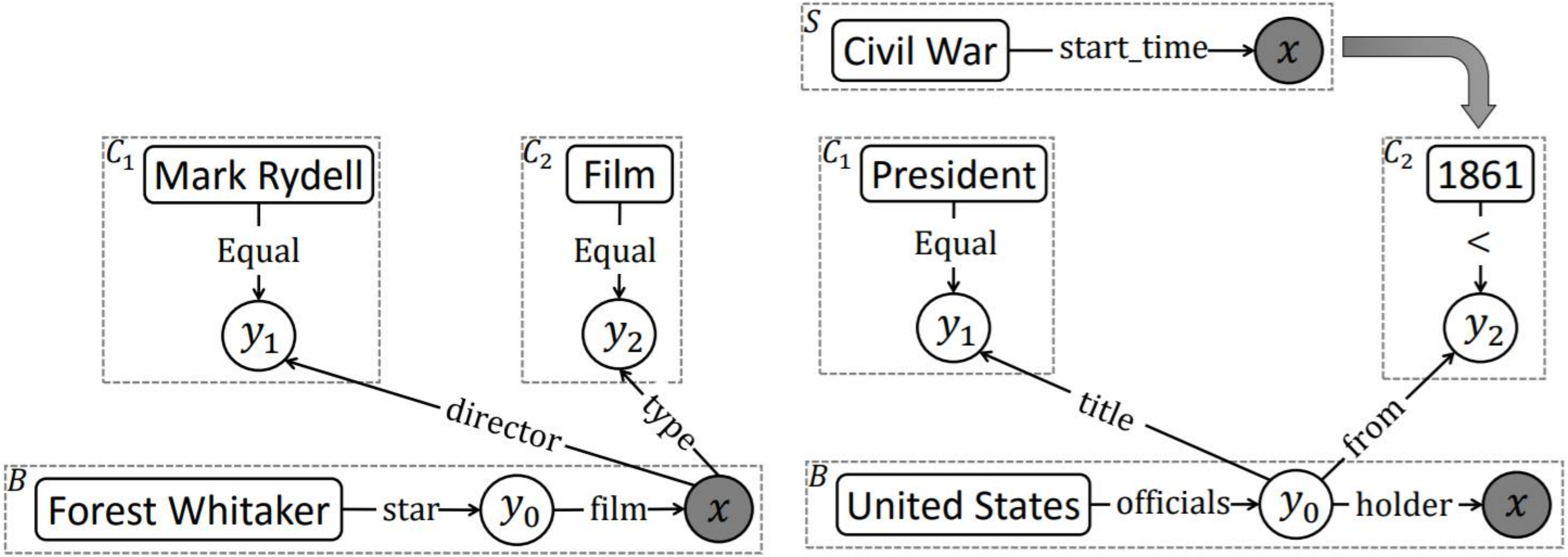


(a) MulCG with entity and type constraint for question "Which films star by Forest Whitaker and are directed by Mark Rydell".

(b) MulCG with implicit temporal constraint for question "Who was U.S. president after the Civil War started".

[COLING16]Constraint-based Question Answering on Knowledge Base

- Explicit Temporal Constraint
  - Based on NER and lexicon

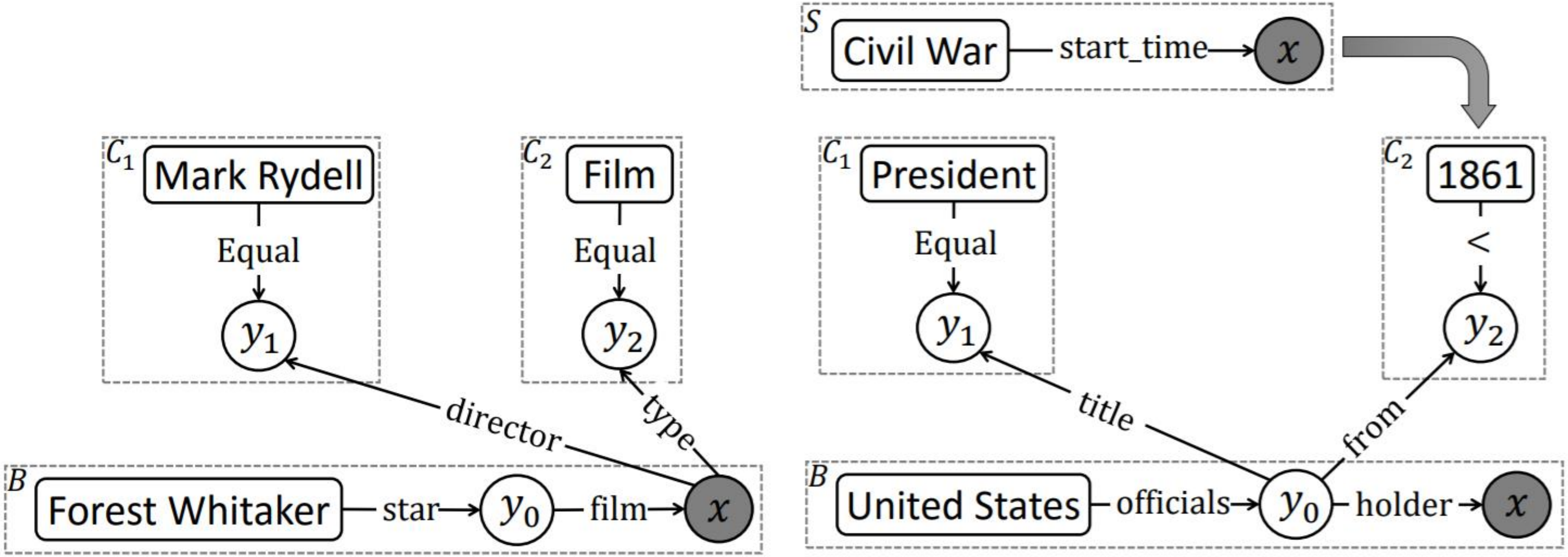


(a) MulCG with entity and type constraint for question "Which films star by Forest Whitaker and are directed by Mark Rydell".

(b) MulCG with implicit temporal constraint for question "Who was U.S. president after the Civil War started".

[COLING16]Constraint-based Question Answering on Knowledge Base

- Implicit Temporal Constraint
  - Extract clause from dependency tree with predefined clause
  - Answer the clause to get the constraint node

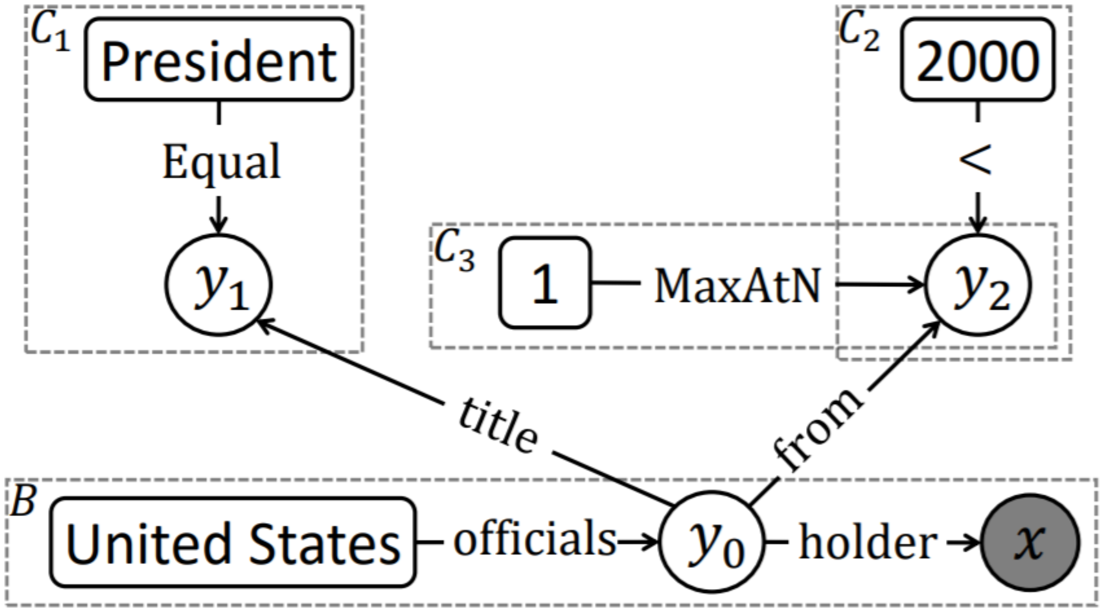


(a) MulCG with entity and type constraint for question "Which films star by Forest Whitaker and are directed by Mark Rydell".

(b) MulCG with implicit temporal constraint for question "Who was U.S. president after the Civil War started".

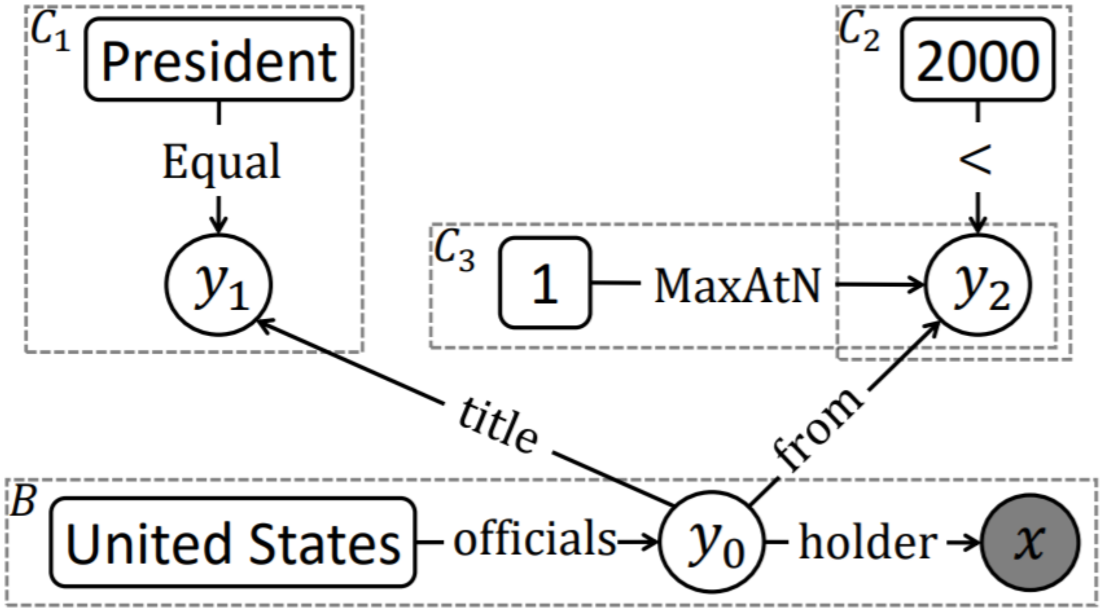
[COLING16]Constraint-based Question Answering on Knowledge Base

- Ordinal Constraint
  - Predefined lexicon for ordinal number and superlative word
  - Use word embedding similarity of speculative word and the binding path's last word



[COLING16]Constraint-based Question Answering on Knowledge Base

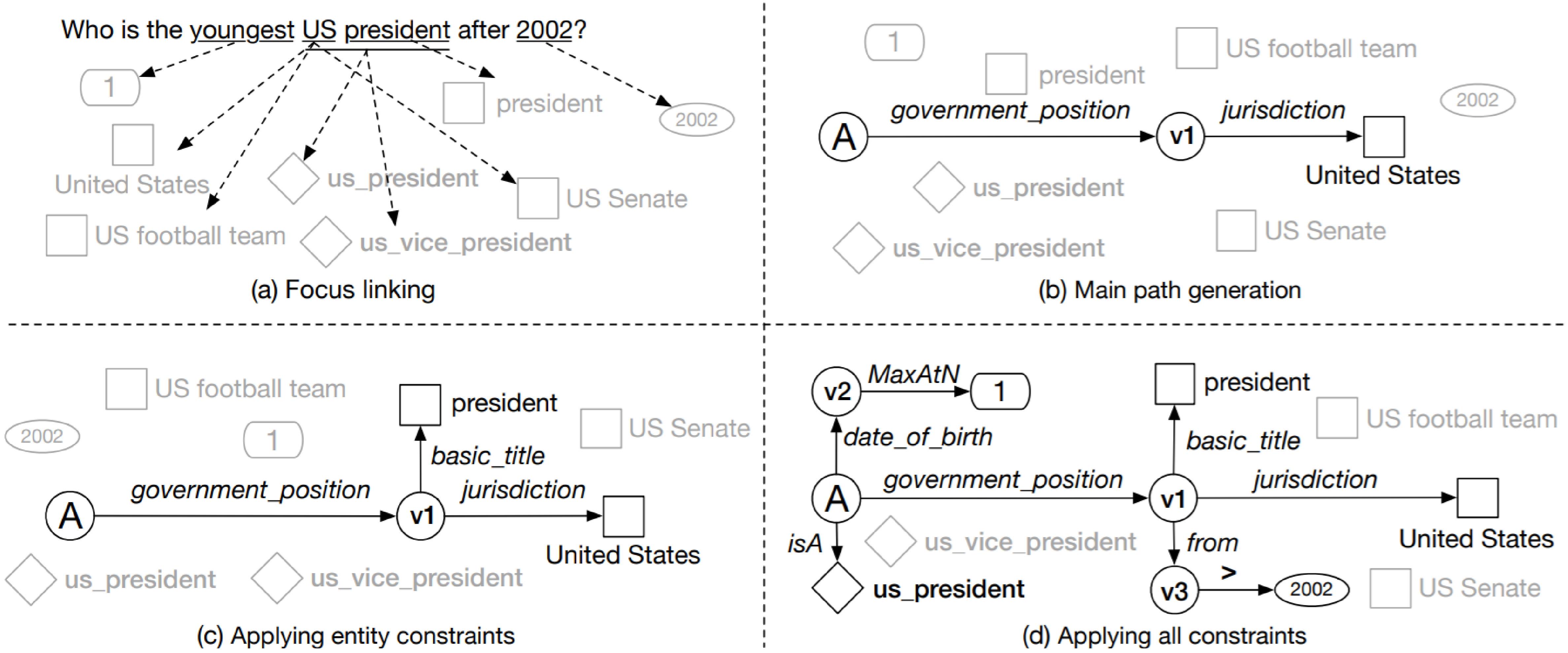
- Aggregation Constraint
  - Based on predefined constraints (How many, number of, …)



[EMNLP18] Knowledge Base Question Answering via Encoding of Complex Query Graphs

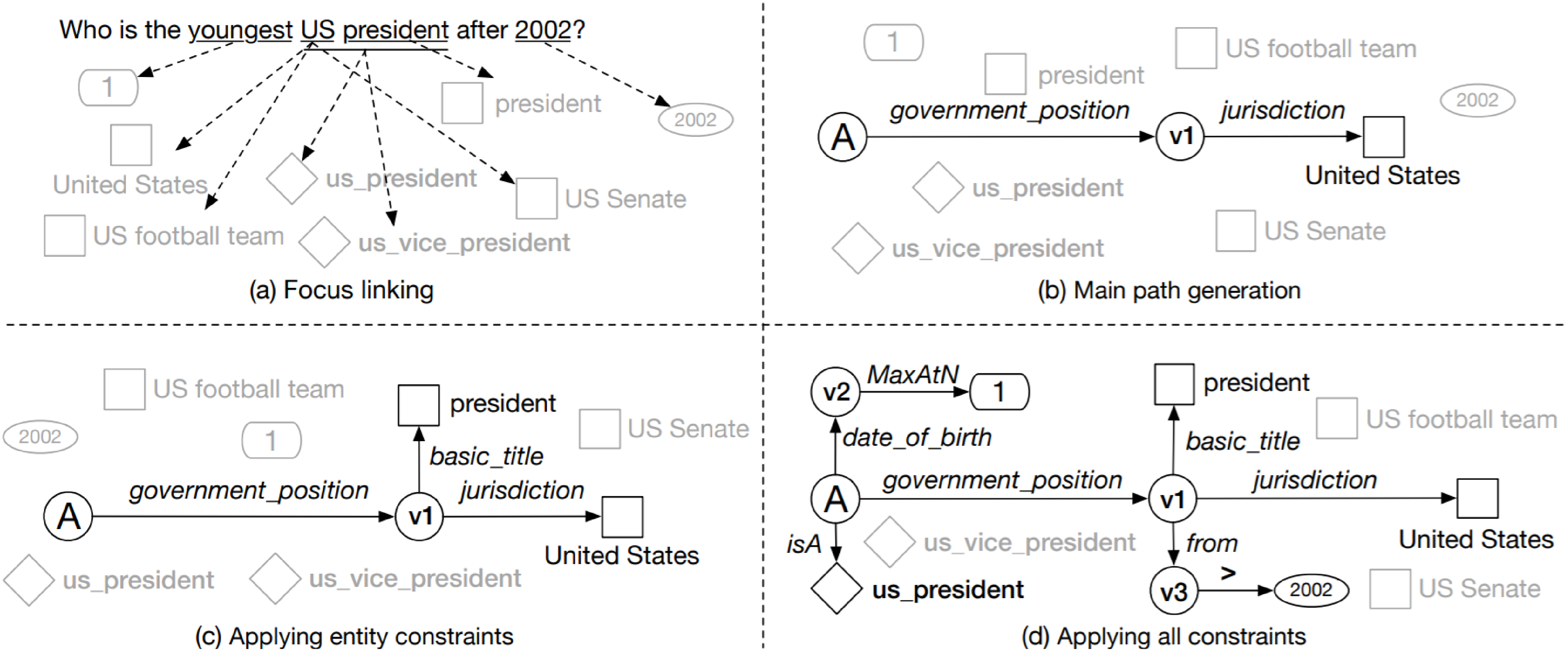
- Type Constraint
  - Build a rich hierarchy of Freebase type
  - Derive implicit type from inferential chain, check if entities and implicit type are in the same domain

$$deg(t_1 \subseteq t_2) = \frac{|cover(t_1) \cap cover(t_2)|}{|cover(t_1)|}$$



[EMNLP18] Knowledge Base Question Answering via Encoding of Complex Query Graphs

- Time Constraint
  - Paired time predicate: for “in” operation, link both “from” and “to” predicates



Old is Gold: Linguistic Driven Approach for Entity and Relation Linking of Short Text

Ahmad Sakor,Isaiah Onando Mulang,Kuldeep Singh, Saeedeh Shekarpour,Maria-Esther Vidal,Jens Lehmann, and Sören Auer

L3S Research Center,Hannover, Germany

Fraunhofer IAIS, Sankt Augustin, Germany

University of Dayton, Dayton, USA

TIB, Hannover, Germany

NAACL 2019

Architecture

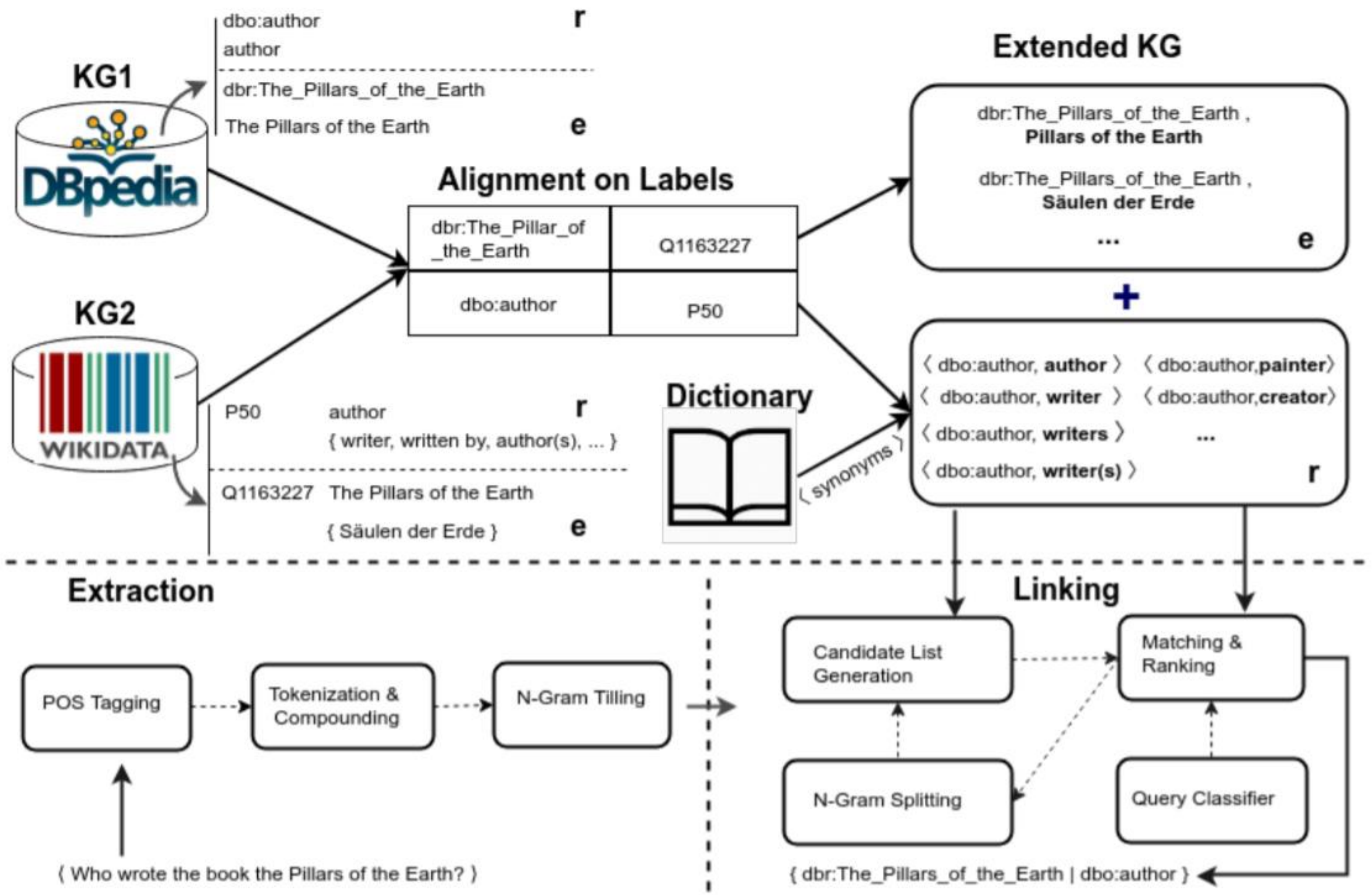


Figure 2: **Overview of Falcon Approach** . Falcon consists of two building blocks: 1) An extended knowledge graph which is built by merging information from various knowledge sources such as DBpedia, Wikidata, Oxford Dictionary, and WordNet. 2) Falcon architecture that has several modules focusing on surface form extraction and linking them to KG.

Mention Detection

- POS Tagging
  - using SpaCy
  - Identify verb and noun phrase

Mention Detection

- Tokenization and Compounding
- Break sentence into potential tokens by removing all stopwords
- Compound word
  - Lexeme that contains two or more stems
  - Example: Barack Obama, high school

- Words without any stop words between them are considered compound word.
  - These tokens would be merged first

N-gram Tiling

- Split/merge based on verb
- Example: “Who wrote the book The Pillars of the Earth”
- Outputs of previous steps:

  wrote book Pillars Earth

- Starts with the first token from either side of the verb and ends at the last non-stop words

Candidate Generation

- Potential relation candidates (“wrote”)
- Potential entity candidates(“book The Pillars of the Earth”)

- Search extended KG using Elasticsearch

Candidate Ranking

- Build a triple representing sentence <subject, predicate, object>
- Increase the weight of the triple if they exists in KG
- Consider question headwords

N-Gram Splitting

- If we don’t get any results from generated mention, split the N-gram
- English morphology: The compound words in English have their headword always towards right side
- Split tokens from the right

Experiment

Table 1: Performance of the Falcon Framework compared to various entity linking tools.

System	Dataset	P	R	F
KEA (Waitelonis and Sack, 2016)	QALD-7	0.06	0.06	0.06
EARL (Dubey et al., 2018)	QALD-7	0.58	0.60	0.58
FOX (Speck and Ngomo, 2014)	QALD-7	0.59	0.57	0.57
Babelfy (Moro et al., 2014)	QALD-7	0.40	0.55	0.44
AIDA (Hoffart et al., 2011)	QALD-7	0.61	0.58	0.59
DBpedia Spotlight (Mendes et al., 2011)	QALD-7	0.68	0.72	0.69
TagMe (Ferragina and Scaiella, 2012)	QALD-7	0.64	0.76	0.67
Falcon	QALD-7	**0.78**	**0.79**	**0.78**
KEA (Waitelonis and Sack, 2016)	LC-QuAD	0.001	0.001	0.001
EARL (Dubey et al., 2018)	LC-QuAD	0.53	0.55	0.53
FOX (Speck and Ngomo, 2014)	LC-QuAD	0.53	0.51	0.51
Babelfy (Moro et al., 2014)	LC-QuAD	0.43	0.50	0.44
AIDA (Hoffart et al., 2011)	LC-QuAD	0.50	0.45	0.47
DBpedia Spotlight (Mendes et al., 2011)	LC-QuAD	0.60	0.65	0.61
TagMe (Ferragina and Scaiella, 2012)	LC-QuAD	0.65	0.77	0.68
Falcon	LC-QuAD	**0.81**	**0.86**	**0.83**
(Singh et al., 2018c)	LC-QuAD3253	0.69	0.66	0.67
Falcon	LC-QuAD3253	**0.73**	**0.74**	**0.73**

Experiment

Table 2: Performance of the Falcon Framework compared to various Relation Linking tools.

QA Component	Dataset	P	R	F
SIBKB (Singh et al., 2017)	QALD-7	0.29	0.31	0.30
ReMatch (Mulang' et al., 2017)	QALD-7	0.31	0.34	0.33
EARL (Dubey et al., 2018)	QALD-7	0.27	0.28	0.27
Falcon	QALD-7	**0.58**	**0.61**	**0.59**
SIBKB (Singh et al., 2017)	LC-QuAD	0.13	0.15	0.14
ReMatch (Mulang' et al., 2017)	LC-QuAD	0.15	0.17	0.16
EARL (Dubey et al., 2018)	LC-QuAD	0.17	0.21	0.18
Falcon	LC-QuAD	**0.42**	**0.44**	**0.43**
(Singh et al., 2018c)	LC-QuAD3253	0.25	0.22	0.23
Falcon	LC-QuAD3253	**0.56**	**0.57**	**0.56**