Domain-Specific Entity Linking via Fake Named Entity Detection

Reporter: Jiangtao Zhang

Knowledge Engineering Group

Department of Computer Science & Technology

Tsinghua University

Outline

- **Introduction**
- Preliminaries
- Our Proposed Approach
- Experiments and Evaluation
- Conclusion

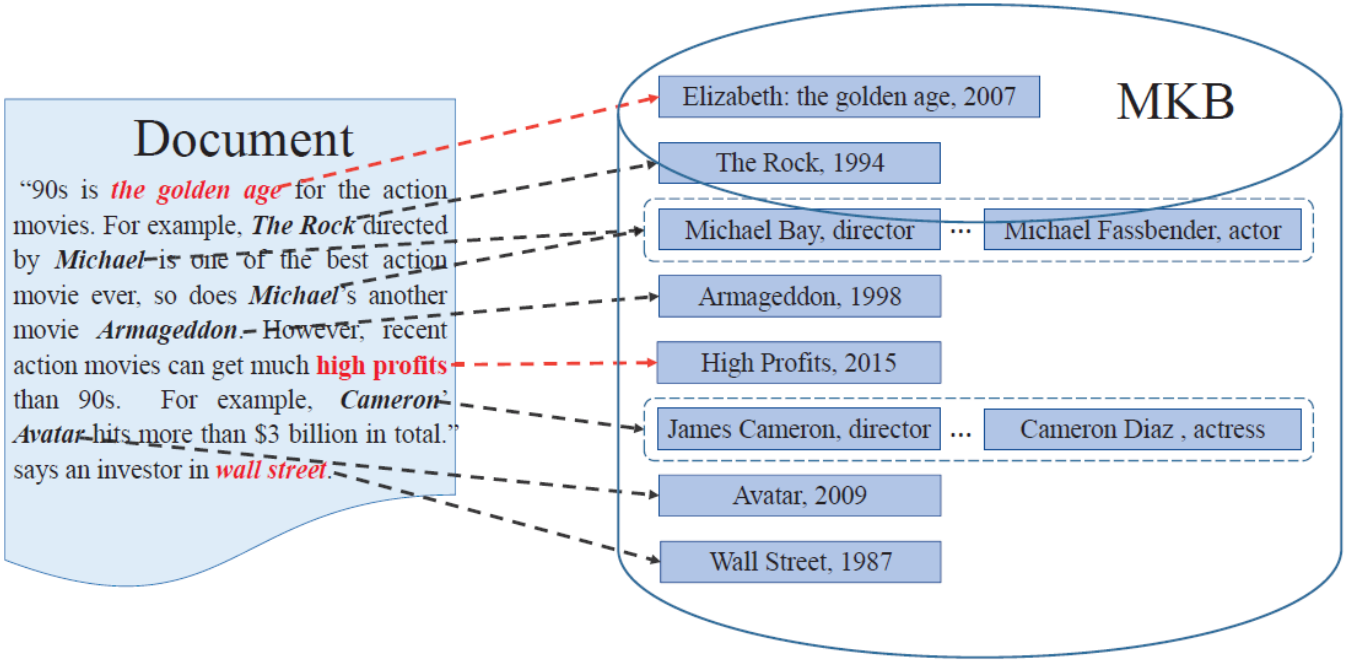
Introduction

- Motivation
  - Increasing demand for constructing and populating domain-specific knowledge bases (DSKBs)
    - IMDB, DBLP, UMLS …
  - Domain-specific entity linking (EL) techniques become more important for DSKB population
  - Traditional EL methods are ineffective for domain-specific tasks

Domain-specific EL deserves much deeper exploration by research communities

Introduction

- Challenges
  - Many Fake Named Entities exist in specific domain
  - Error propagation exists between NED and EL processes



Introduction

- Contributions
  - Among the first to explore the problem of joint NED and EL with the domain-specific knowledge base
  - Proposed an effective technique that iteratively models NED and EL processes
  - Conducted extensive experiments to evaluate the effectiveness of approach

Outline

- Introduction
- **Preliminaries**
- Our Proposed Approach
- Experiments and Evaluation
- Conclusion

Preliminaries

- Domain-Specific Knowledge Base (DSKB)
  - $DSKB = \{C, E, P, R\}$
- Mentions and Linked Entities
  - mention: a textual phrase m
  - linked entity: the correct mapping entity $e_m \in E$
- Fake Named Entity (FNE)
  - Common textual phrase: could likely be linked to but should not be linked
- Context Mention and Entity
  - context mention: all other candidate mentions in a window
  - context entity: most possible linked entity for context mention

Preliminaries

Task Definition

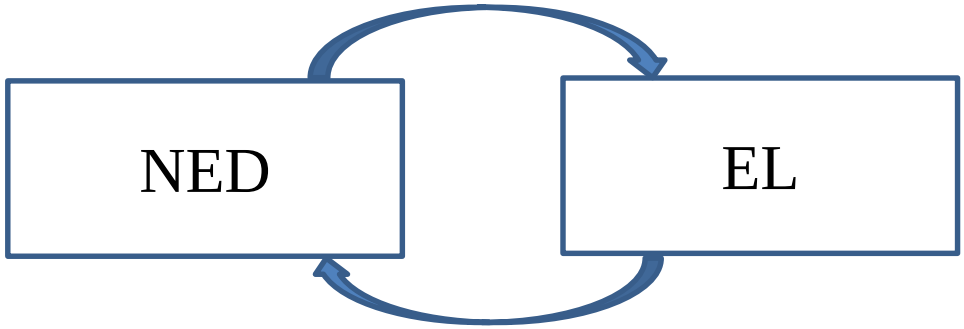
- Input: $\langle d, KB_{DS} \rangle$
  - an unstructured document d in a specific domain
  - a DSKB pertaining to the same domain
- Output: $\{\langle m, \sigma(m) \rangle | \forall m \in M_T\}$
  - Extract TNEs and filter out FNEs in d
  - Map each extracted TNE to its linked entity

Outline

- Introduction
- Preliminaries
- **Our Proposed Approach**
- Experiments and Evaluation
- Conclusion

Approach

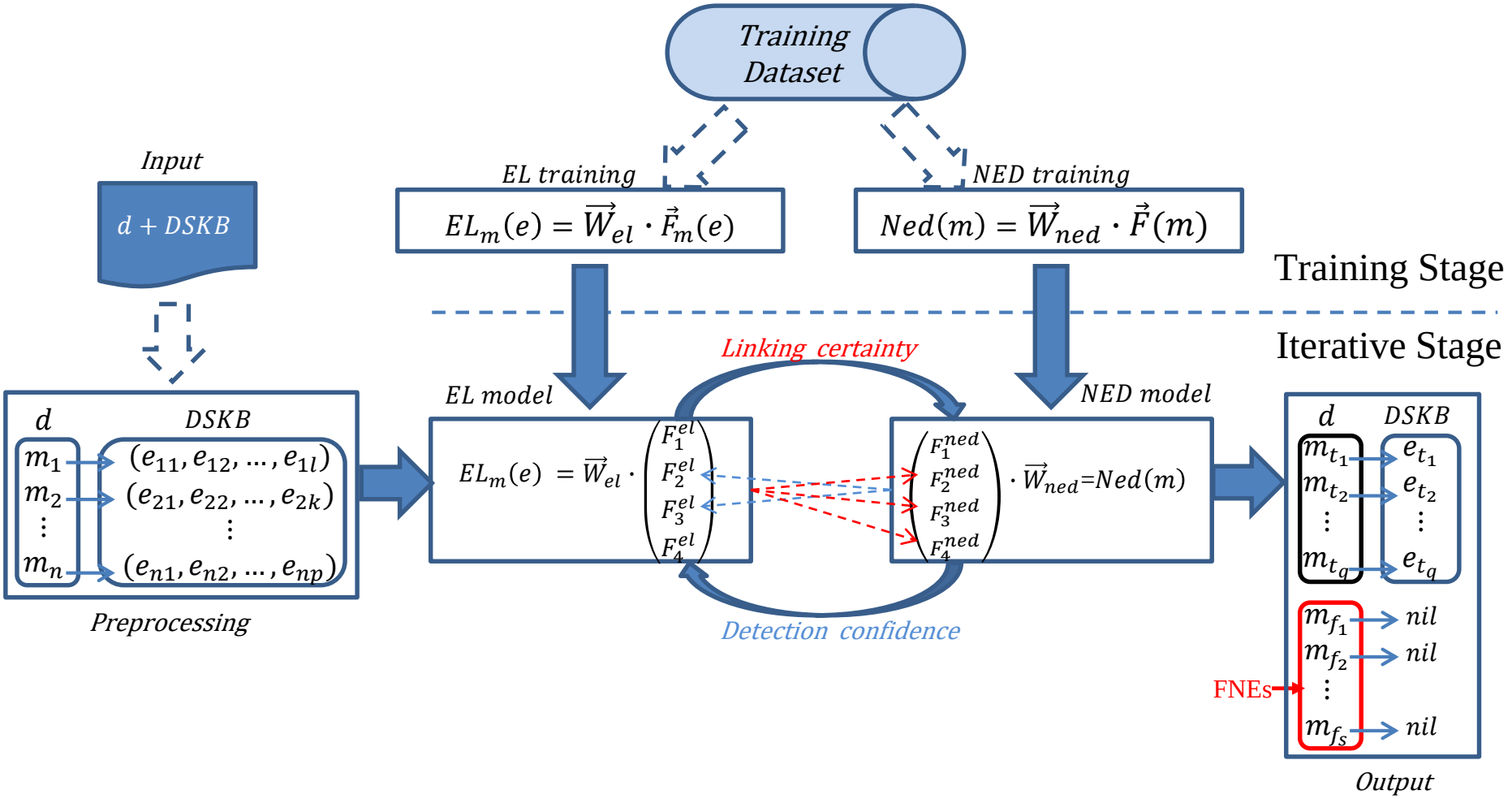
- Main Idea
  - Jointly models NED and EL iteratively
  - In each iteration, gradually improve the confidence of TNEs while reduce the confidence of FNEs by leveraging the interdependency of NED and EL



Iteration

Approach

Framework



Approach

- Features in EL Model
  - Popularity
  - Context Relatedness
  - Content Similarity
- Features in NED Model
  - Link Probability
  - Linking Certainty
  - Coherence

Approach

- Features in EL Model
  - Context Relatedness

$$ConRel_m(e) = \frac{\sum_{e_c \in C_E} SmtRel(e_c, e)}{|C_E|}$$

$$ConRel_m(e) = \frac{\sum_{m_c \in C_M} Ned(m_c) * SmtRel(e_{top}(m_c), e)}{|C_M|}$$

$$e_{top}(m_c) = \arg\max_{e_c \in E(m_c)} (El_m(e_c))$$

$$SmtRel \begin{cases} \textbf{\textit{WLM}} \longrightarrow ConRel1_m(e) \\ \textbf{\textit{Jaccard}} \longrightarrow ConRel2_m(e) \end{cases}$$

Approach

- Features in NED Model
  - Linking Certainty

$$LC(m) = El_m(e_{top}(m)) = \max\{El_m(e)|e \in E(m)\}$$

Approach

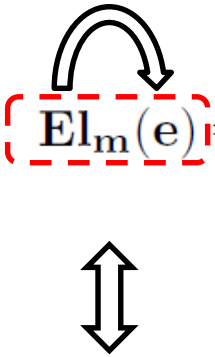
- Features in NED Model
  - Coherence

$$Coh(m) = \frac{\sum_{m_c \in C_M} Ned(m_c) * SmtRel(e_{top}(m_c), e_{top}(m))}{|C_M|}$$

$$SmtRel \begin{cases} \boldsymbol{WLM} \longrightarrow Coh1(m) \\ \boldsymbol{Jaccard} \longrightarrow Coh2(m) \end{cases}$$

Approach

- Iterative Process



$$\mathbf{El_m(e)} = w_1^{el} * Pri_m(e) + w_2^{el} * \frac{\sum_{m_c \in C_M} \mathbf{Ned(m_c)} * WLM(\mathbf{e_{top}(m_c)}, e)}{|C_M|} +$$
$$w_3^{el} * \frac{\sum_{m_c \in C_M} \mathbf{Ned(m_c)} * Jac(\mathbf{e_{top}(m_c)}, e)}{|C_M|} + w_4^{el} * ConSim_m(e)$$

$$\mathbf{Ned(m)} = w_1^{ned} * LP(m) + w_2^{ned} * El_m(\mathbf{e_{top}(m)}) +$$
$$w_3^{ned} * \frac{\sum_{m_c \in C_M} \mathbf{Ned(m_c)} * WLM(\mathbf{e_{top}(m_c)}, \mathbf{e_{top}(m)})}{|C_M|} +$$
$$w_4^{ned} * \frac{\sum_{m_c \in C_M} \mathbf{Ned(m_c)} * Jac(\mathbf{e_{top}(m_c)}, \mathbf{e_{top}(m)})}{|C_M|}$$

$$e_{top}(m_c) = \arg\max_{e_c \in E(m_c)} (El_m(e_c))$$

Approach

- Iterative algorithm

Input: $M; \forall m \in M, E(m); \overrightarrow{W}_{el}; \overrightarrow{W}_{ned}$
Output: M_T; $\forall m \in M_T, \langle m, Ned(m) \rangle, \langle m, e_{top}(m), El_m(e_{top}(m)) \rangle$
repeat
  for each $m \in M$ **do**
    for each $e \in E(m)$ **do**
      $El_m(e) = \overrightarrow{W}_{el} \cdot \overrightarrow{F}_m(e)$;
    end
    $e_{top}(m) = \arg\max_{e \in E(m)}(El_m(e))$;
  end
  for each $m \in M$ **do**
    $Ned(m) = \overrightarrow{W}_{ned} \cdot \overrightarrow{F}(m)$
  end
until convergence;

Outline

- Introduction
- Preliminaries
- Our Proposed Approach
- **Experiments and Evaluation**
- Conclusion

Experiments

Data

- User comments on movies

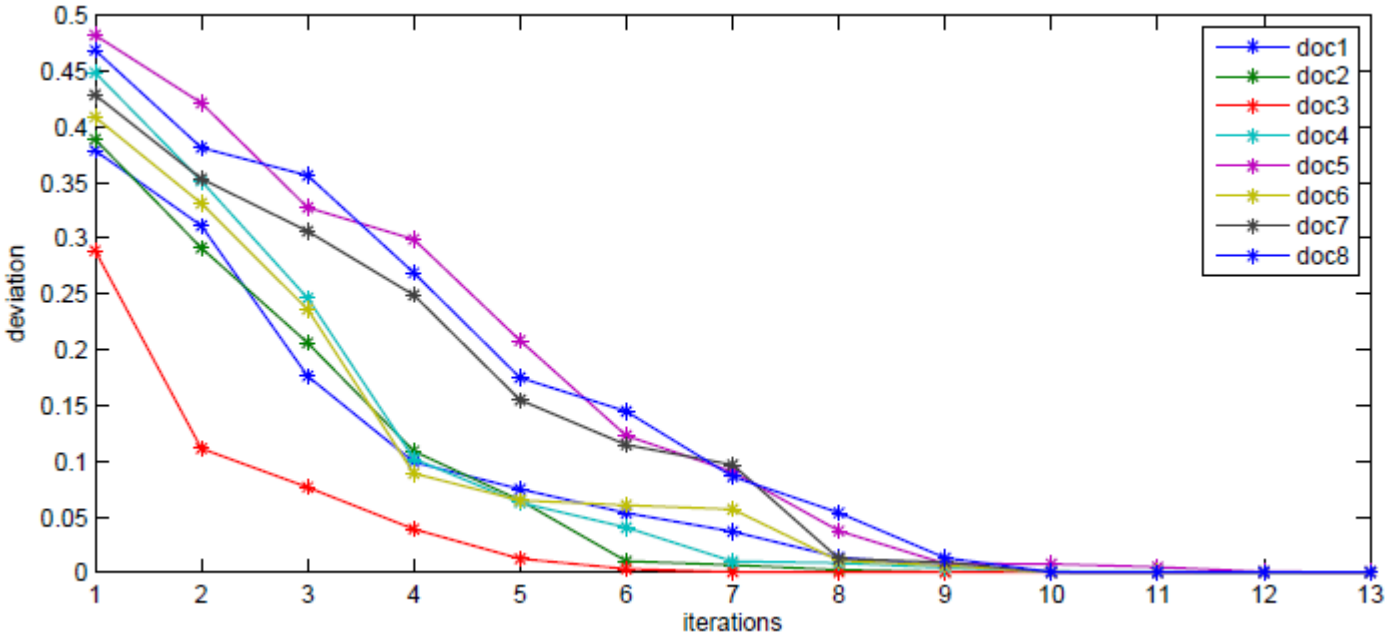
Documents	$\lvert FNEs \rvert$	$\lvert TNEs \rvert$	CEs	$\overline{\lvert M \rvert}$	$\overline{\lvert E(m) \rvert}$
843	2529	11848	42105	17.05	2.92

- Movie-Knowledge-Base
  - A high quality knowledge base about movies, TV series and celebrities which integrates several English and Chinese movie data sources from Baidu Baike and Douban
  - 23 concepts, 91 properties, more than 700,000 entities and 10 million triples

Experiment

- Parameter Setting
  - Iteration Deviation Threshold

$$\max_{m_i \in M} (Ned(m_i^{(j)}) - Ned(m_i^{(j-1)}) \leq \varepsilon$$



Experiments

- Baseline Methods
  - Prior Probability-based method (POP)
  - Vector Similarity-based method (V Sim)
- Our model with 4 different settings
  - No Training + No Iterating (NoT + NoI)
  - Training + No Iterating (T + NoI)
  - No Training + Iterating (NoT + I)
  - Training + Iterating (T +I)

Experiments

- Result and Analysis

Approach	NED			EL	Overall NED + EL		
	precision	recall	F1	accuracy	precision	recall	F1
POP	0.776	0.643	0.703	0.792	0.615	0.509	0.557
VSim	0.724	0.715	0.719	0.825	0.597	0.590	0.594
NoT+NoI	0.761	0.738	0.749	0.849	0.646	0.627	0.636
T+NoI	0.808	0.754	0.780	0.864	0.698	0.651	0.674
NoT+I	0.826	0.748	0.785	0.852	0.704	0.637	0.669
T+I	**0.847**	**0.788**	**0.816**	**0.875**	**0.741**	**0.690**	**0.714**

Table 2. Comparison of experiment results

Outline

- Introduction
- Preliminaries
- Our Proposed Approach
- Experiments and Evaluation
- **Conclusion**

Conclusion

- The current state-of-the-art entity linking research primarily focus on general knowledge bases
- We propose a novel approach that dedicates to address the two domain-specific issues: *fake named entities* and the *interdependency* by jointly modeling NED and EL iteratively
- We strongly believe that domain-specific EL deserves much deeper exploration by researchers.

Thanks!