가상 엔터티 설명문 및 엔터티 정렬에 기반한 엔터티 링킹 전이학습

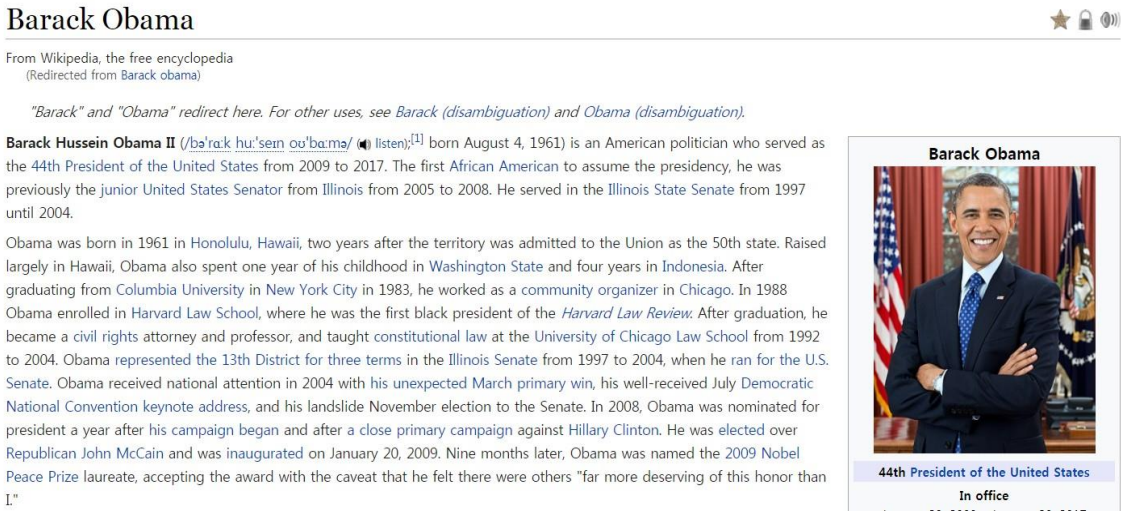
2020.10.16

최형준[O1], 나승훈[1], 김현호[2], 김선훈[2], 강인호[2]
[1]전북대학교, [2]NAVER

Contents

- Background
- Related Works
- RELIC
- Experiment Result
- Conclusion

개체명 인식



Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Barack obama)

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation) and Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ (listen);[1] born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017. The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008. He served in the Illinois State Senate from 1997 until 2004.

Obama was born in 1961 in Honolulu, Hawaii, two years after the territory was admitted to the Union as the 50th state. Raised largely in Hawaii, Obama also spent one year of his childhood in Washington State and four years in Indonesia. After graduating from Columbia University in New York City in 1983, he worked as a community organizer in Chicago. In 1988 Obama enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduation, he became a civil rights attorney and professor, and taught constitutional law at the University of Chicago Law School from 1992 to 2004. Obama represented the 13th District for three terms in the Illinois Senate from 1997 to 2004, when he ran for the U.S. Senate. Obama received national attention in 2004 with his unexpected March primary win, his well-received July Democratic National Convention keynote address, and his landslide November election to the Senate. In 2008, Obama was nominated for president a year after his campaign began and after a close primary campaign against Hillary Clinton. He was elected over Republican John McCain and was inaugurated on January 20, 2009. Nine months later, Obama was named the 2009 Nobel Peace Prize laureate, accepting the award with the caveat that he felt there were others "far more deserving of this honor than I."

Barack Obama

44th President of the United States

In office

- **Context :** ***After campaigning on the promise of health care reform, President Obama gave a speech in March 2010 in Pennsylvania***
- **Mention :** ***President Obama***
- **Entity :** ***Barack Obama*** **(http://wikipedia.org/wiki/Barack_Obama)**

엔티티 링킹은 엔티티의 중의성을 해결하기 위한 작업으로, 문서에 나타난 개체 표현과 부합하는 지식 베이스에 있는 개체를 연결해주는 기술

멘션 임베딩을 이용한 nil 멘션 탐지와 개체 연결의 통합 모델 [홍승연 et al' 20]

- 엔터티 링킹 시 Nil 엔터티를 두어 엔터티 링킹과 멘션 탐지를 동시에 수행
  - Nil 멘션은 문장 내의 엔터티가 아닌 부분을 의미
  - Nil 멘션인 경우 Nil 엔터티를, non-Nil 멘션인 경우 해당하는 엔터티를 예측
- RoBERTa를 통해 엔터티 설명을 인코딩, 출력의 첫 번째 토큰인 <s>에 해당하는 부분을 엔터티 임베딩으로 사용
  - 엔터티 임베딩은 학습되지 않도록 고정시켜 사용
- 문장 내의 멘션에 해당하는 부분을 Concatenate 한 것과 각각의 엔터티 후보를 Biaffine 연산을 통해 점수를 계산
- 엔터티 후보는 멘션-엔터티 사전을 통해 추출
  - 데이터셋에서 등장한 모든 멘션-엔터티 쌍을 통해 사전 구성
  - 사전은 각각의 멘션과 등장한 모든 엔티티로 이루어짐

멘션 임베딩을 이용한 nil 멘션 탐지와 개체 연결의 통합 모델 [홍승연 et al' 20]

- 한국어 위키백과를 사용하여 엔티티 링킹 학습
  - 위키백과 문서에서 하이퍼링크가 걸린 부분을 멘션, 해당 링크가 가리키는 문서를 엔티티로 사용
- 모든 엔터티와 멘션 쌍을 통해 각 멘션이 나타낼 수 있는 엔터티-멘션 집합 $\mathrm{E}(m)$을 구축
  - 멘션 m에 대해 전체 데이터셋에서 멘션이 m인 경우 등장한 모든 엔티티 $\{e_1, e_2, \dots\}$를 가리킴
  - Nil 엔터티를 추가하기 위해 토크나이징 된 각 데이터에 대해 n-gram 매칭을 통해 멘션을 탐지, 실제로 엔터티가 있는 경우를 non-NIL, 없는 경우를 Nil로 취급
  - Nil 멘션인 경우 Nil 엔터티인 $e_\emptyset$를 매칭
- 문장 내의 각 멘션 스팬은 시작지점 x_{start}과 끝점 x_{end}으로 구성됨
- 각 데이터는 다음과 같이 구성

토크나이징 된 문장: $\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{i}]$
멘션 스팬: $\mathrm{S} = [(\mathrm{x}_{start}, \mathrm{x}_{end}, \mathrm{e}, [\mathrm{e}_\emptyset, \mathrm{e}_1, \mathrm{e}_2, \dots]), \dots]$

멘션 임베딩을 이용한 nil 멘션 탐지와 개체 연결의 통합 모델 [홍승연 et al' 20]

- 엔터티 임베딩: 각 엔터티에 대응하는 표상을 저장해둔 것
  - RoBERTa를 통해 각 엔티티 설명을 인코딩, 문서 맨 첫번째 토큰인 <s>의 출력을 사용
  - 하나의 엔터티에 대해 해당 엔터티 ID에 맞는 표상을 매핑
  - 엔터티 링킹 학습 시 엔터티 임베딩을 고정, 학습되지 않도록 함
- 엔터티 링킹 모델은 RoBERTa 모델과 엔터티 임베딩으로 구성
  - RoBERTa를 통해 입력 문장 를 인코딩 한 후, 문장 내의 각 멘션스판의 시작 토큰과 끝 토큰을 Concatenate 한 후, 이를 엔티티 표상과 Biaffine 연산을 통해 점수를 계산

$$\mathrm{X}' = \mathrm{RoBERTa}(\mathrm{X}) = [x_1', \ldots, x_i']$$
$$\mathrm{Score}(\mathrm{e}) = \mathrm{Biaffine}(\mathrm{concat}(\mathrm{x}'_{\mathrm{start}}, \mathrm{x}'_{\mathrm{end}}), \mathrm{Emb}(\mathrm{e}))$$

멘션 임베딩을 이용한 nil 멘션 탐지와 개체 연결의 통합 모델 [홍승연 et al' 20]

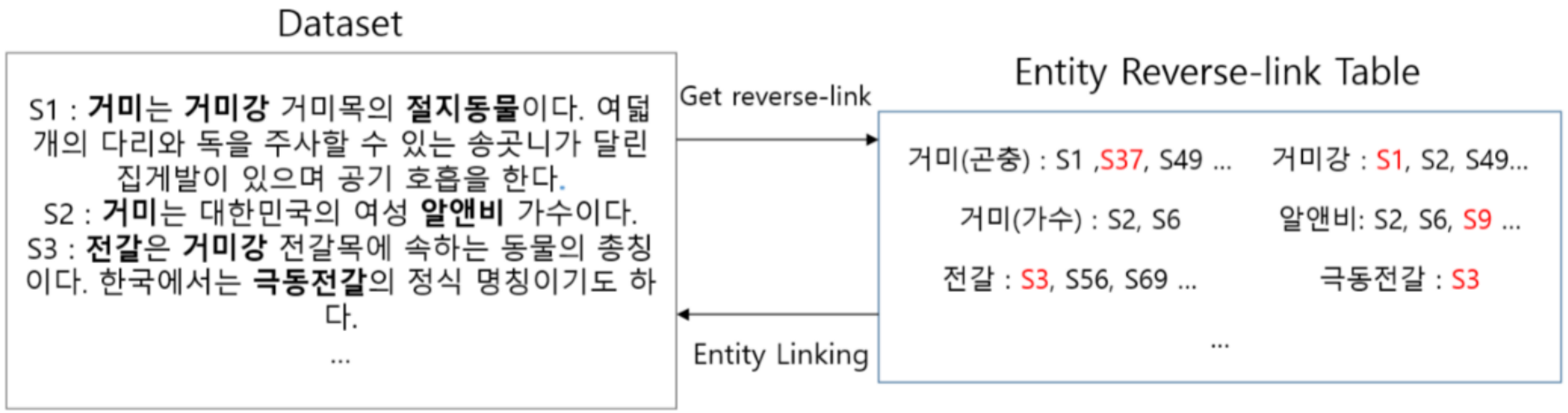
- 한계
  - 엔터티 링킹을 수행하기 위해 엔터티에 대한 정보가 필수적
  - 엔터티에 대한 정보가 매우 적거나, 얻기 어려운 환경에서는 엔터티 링킹을 수행하기 어려움
- 대응책
  - 엔터티에 대한 정보를 해결하기 위해 가상 엔터티 설명문 도입 : 엔터티 요약을 위한 여러 방법으로 확장 필요
  - 기존 엔터티 임베딩과 가상 엔터티 설명을 연결
    - 엔터티 얼라인먼트를 통해 기존 엔터티 임베딩으로 등장 한 적 없는 멘션을 연결

가상 엔터티 설명문

- 엔터티에 대한 정보를 얻을 수 없는 상황을 대응하기 위해 가상 엔터티 설명문을 구축
  - 기존의 엔터티에 대한 정확한 설명 대신 해당 엔터티가 등장한 문장을 그 엔터티의 설명으로 사용
  - 상표명, 인명 등 해당 엔터티에 대한 정보를 획득 할 방법이 없는 경우에 대한 해결책으로써 사용
  - 해당 엔터티가 등장한 문장 중 하나를 임의로 선택해서 가상 엔터티 설명문으로 사용
- 엔터티 임베딩은 이전 연구와 동일하게 RoBERTa를 통해 엔터티 설명문을 인코딩, <s>토큰에 해당하는 문장의 첫 번째 토큰 출력을 사용

$$\mathrm{M}(e) = \{m | (m, e) \in \mathrm{E}(m)\}$$
$$\mathrm{T_p} = \mathrm{RandomChoise}\big(\mathrm{M}(e)\big)$$
$$e_\mathrm{p} = \mathrm{RoBERTa}\big(\mathrm{T_p}\big)[0]$$

가상 엔터티 설명문


Dataset
S1 : 거미는 거미강 거미목의 절지동물이다. 여덟 개의 다리와 독을 주사할 수 있는 송곳니가 달린 집게발이 있으며 공기 호흡을 한다.
S2 : 거미는 대한민국의 여성 알앤비 가수이다.
S3 : 전갈은 거미강 전갈목에 속하는 동물의 총칭이다. 한국에서는 극동전갈의 정식 명칭이기도 하다.
...
Get reverse-link
Entity Linking
Entity Reverse-link Table
거미(곤충) : S1 ,S37, S49 ...
거미강 : S1, S2, S49...
거미(가수) : S2, S6
알앤비: S2, S6, S9 ...
전갈 : S3, S56, S69 ...
극동전갈 : S3
...

엔터티 얼라인먼트

- 가상 엔터티 설명문과 엔터티 id를 연결해주는 작업
- 등장한 적 없는 멘션에 대해 기존 엔터티 정보를 매핑
  - 기존의 미리 학습된 엔티티 링킹 모델을 활용
- 다음과 같은 과정을 통해 엔터티 id를 매핑
  1. 등장 한 적 없는 멘션을 후보로 엔터티 링킹 시행
  2. 모델의 출력에서 가장 높은 확률을 가진 엔터티를 등장 한 적 없는 멘션에 링킹되는 엔터티로 취급
  3. 멘션-엔터티 사전 $\mathrm{E}(m)$ 확장
  4. 이 과정에서 Nil 엔티티 $e_{\emptyset}$를 예측 할 경우, 해당 멘션과 가상 엔티티 설명문 e_p를 $\mathrm{E}(m)$에 확장

실험

- 엔터티 링킹 학습 데이터는 총 6만개의 문장으로 구성, 3만개를 train, 2만개를 test, 1만개를 dev 셋으로 사용
  - 2217개의 데이터로 구성된 네이버의 엔터티 링킹 데이터셋을 추가로 실험
  - 뉴스 기사 등에서 수작업으로 멘션 부분과 엔티티 id를 표시
  - 가계 이름, 인명 등 저명성이 떨어지는 엔터티가 포함됨
  - 모든 엔터티에 대해 가상 엔터티 설명문을 적용
  - 70%를 train, 20%를 test, 10%를 dev로 사용
- 엔터티 링킹 학습 데이터는 총 6만개의 문장으로 구성, 3만개를 train, 2만개를 test, 1만개를 dev 셋으로 사용
- 가상 엔터티 설명문으로 변환되는 비중을 10~50%, 100%로 조정하며 성능 측정

실험 – 가상 엔터티 설명문 변환 비중 별 엔터티 링킹 성능

변환 비중	정밀도	재현율	F1	Nil 탐지 F1
Baseline	85.96%	85.78%	85.87%	89.05%
10%	85.53%	84.69%	85.11%	88.68%
20%	84.95%	84.10%	84.52%	88.48%
30%	84.56%	83.59%	84.07%	88.31%
40%	84.39%	82.94%	83.66%	88.27%
50%	84.33%	82.31%	83.31%	88.21%
100%	84.43%	82.19%	83.29%	88.18%

- Nil 탐지 F1은 해당 멘션이 Nil 멘션 인지 아닌지를 판별하는 F1 점수를 의미
  - 모델이 예측한 것이 Nil 엔터티인지 다른 엔터티인지를 통해 측정
- 가상 엔터티 설명문은 온전한 엔터티 정보에 비해 해당 엔터티의 정보를 잘 나타내지 못하기 때문에 성능이 낮아지나, 급격히 낮아지지 않음

실험 – 전이학습 실험 결과 : 네이버 데이터셋

데이터셋	정밀도	재현율	F1	Nil 탐지 F1
dev	49.92%	66.49%	57.02%	65.52%
test	69.42%	85.97%	76.81%	81.52%

- 데이터 편차가 매우 크고 데이터 셋이 작기 때문에 dev 셋과 test 셋 사이의 성능 편차가 큼
- 아무런 정보가 주어지지 않은 환경에서도 엔터티 링킹이 어느정도 수행 가능

결론 및 향후 연구

- 결론
  - 데이터셋으로 부터 가상 엔터티 설명문을 구축, 이를 통해 엔터티의 정보를 획득 할 수 없는 환경에서의 엔터티 링킹을 시도.
  - 가상 엔터티 설명문을 사용하여 엔터티 링킹 성능 하락폭이 크지 않음을 보임
- 향후 연구
  - 각 엔터티를 하나의 페이지로만 매핑 하는 것이 아닌, 기존 엔터티 링킹 모델의 확률값을 이용하여 기존의 엔터티 임베딩과 새로 만들어진 엔터티 임베딩 중 적합한 것을 선택하도록 시도 해볼 예정

Q/A