Incremental Memory에 기반한 Cross-document 엔터티 상호참조해결

최형준[1], 나승훈[1], 김현호[2], 김선훈[2], 강인호[2]
[1]전북대학교, [2]네이버

Entity Linking

조 바이든

위키백과, 우리 모두의 백과사전.

바이든은 여기로 연결됩니다. 다른 뜻에 대해서는 바이든 (동음이의) 문서를 참조하십시오.

조지프 로비넷 바이든 주니어[1](영어: Joseph Robinette Biden Jr., 1942년 11월 20일 ~)는 2021년 1월 20일에 취임한 미국의 제46대 대통령이다. 1973년부터 2009년까지 델라웨어주 연방 상원의원으로 재직했으며, 2009년부터 2017년까지 제47대 부통령을 지냈으며 2021년부터 제46대 미국 대통령으로 재임 중이다.

2017년 1월 13일, 퇴임을 앞두고 넬슨 록펠러(1977년), 휴버트 호레이쇼 험프리(1980년, 추서)에 이어 부통령으로서는 역대 세 번째로 대통령 자유 훈장을 받았다. 바이든이 받은 것은 그 중에서도 특별훈장(Presidential Medal of Freedom with Distinction)으로, 부통령으로서는 유일하며 역대 대통령 중에서도 로널드 레이건만이 받은 훈장이다.[2][3][4]

목차 [숨기기]

1 생애
- 1.1 초기 생애
- 1.2 초기 경력
- 1.3 상원의원 활동
- 1.4 2008년 대선
- 1.5 2016년 대선
- 1.6 2020년 대선
- 1.7 대통령 재임 시절

2 정치적 견해
- 2.1 성적지향에 관한 차별 금지

조지프 로비넷 바이든 주니어
Joseph Robinette Biden Jr.

조 바이든 (2021년)

미국의 제46대 대통령

임기 2021년 1월 20일 ~

부통령 카멀라 해리스

전임: 도널드 트럼프(제45대)

- **Context :** ***조 바이든은 2021년 1월 20일에 취임한 미국의 제46대 대통령이다.***
  Mention : ***조 바이든***
- **Entity :** ***조 바이든*** **(https://ko.wikipedia.org/wiki/%EC%A1%B0_%EB%B0%94%EC%9D%B4%EB%93%A0)**

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

Entity Linking

- **Context :** ***지난 15일 오후 10시 암호화폐 8종의 상장폐지를 알린 암호화폐 거래소 코인빗이 이번엔 상장폐지 3시간을 앞두고 상장폐지 일정을 돌연 연기했다.***
  Mention : ***코인빗***
- **Entity :** ***???***

사업체 상호명이나 유명하지 않은 인명의 경우, 엔티티 링킹을 위한 정보를 얻기 어려운 경우가 존재
기존의 엔티티링킹 방법을 통해 연결하기 어려움

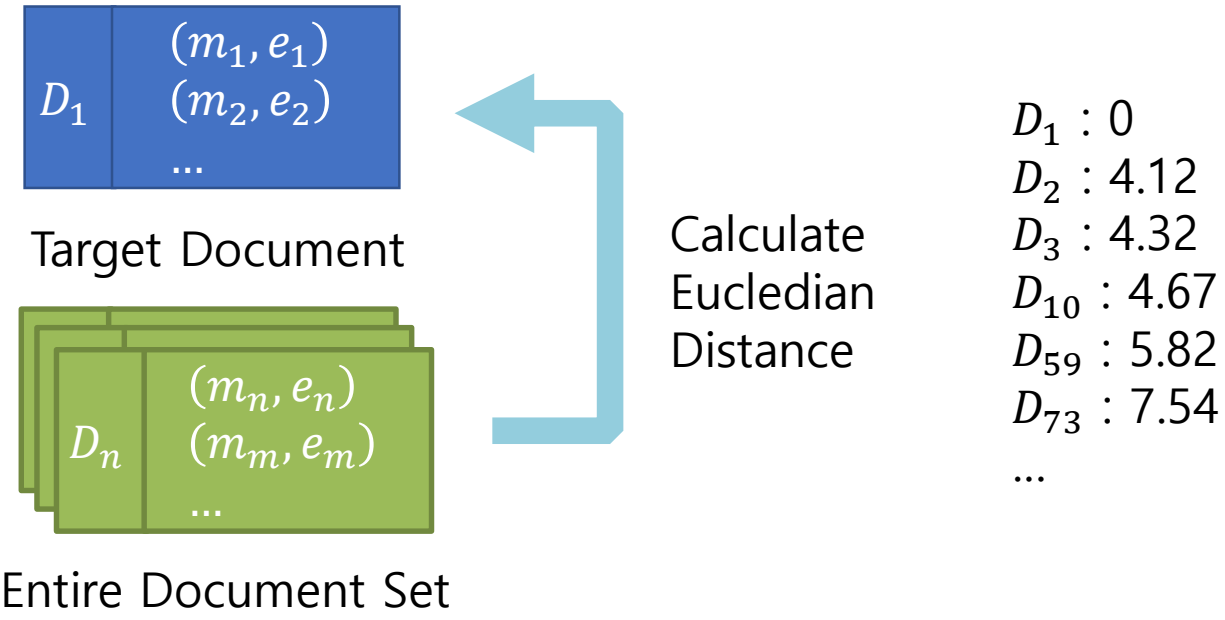
Incremental Neural Coreference Resolution in Constant Memory [Patrick Xia et al. 20]

Algorithm 1 FindClusters(Document)

```
Create an empty Entity List, E
for segment ∈ Document do
    M ← SPANS(segment)
    for m ∈ M do
        scores ← PAIRSCORE(m, E)
        top_score ← max(scores)
        top_e ← argmax(scores)
        if top_score > 0 then
            UPDATE(top_e, m)
        else
            ADD_NEW_ENTITY(E, m)
    EVICT(E)
return E
```

- 주어진 문서 내의 멘션과 메모리 사이의 Pair score를 계산, 이를 통해 메모리에 엔티티를 추가하거나 메모리를 업데이트
  - Pair score가 0 이상인 경우 같은 엔티티로 취급, 현제의 멘션을 메모리에 업데이트
  - 0 미만인 경우 메모리상에 동일한 엔티티가 없는 경우로, 메모리 상에 새로운 엔티티를 추가

Cross-document Entity Coreference Resolution



- Dense Retrieval 을 통해 각 문서를 벡터화, 각 문서간 거리를 계산
  - 두 문서간의 거리는 **유클리드** 거리로 측정
  - Dense Retrieval 모델은 RoBERTa와 FNN으로 구성되고, 질의와 질의에 맞는 문서 사이의 유사도가 최대가 되도록 미리 학습된 모델을 사용

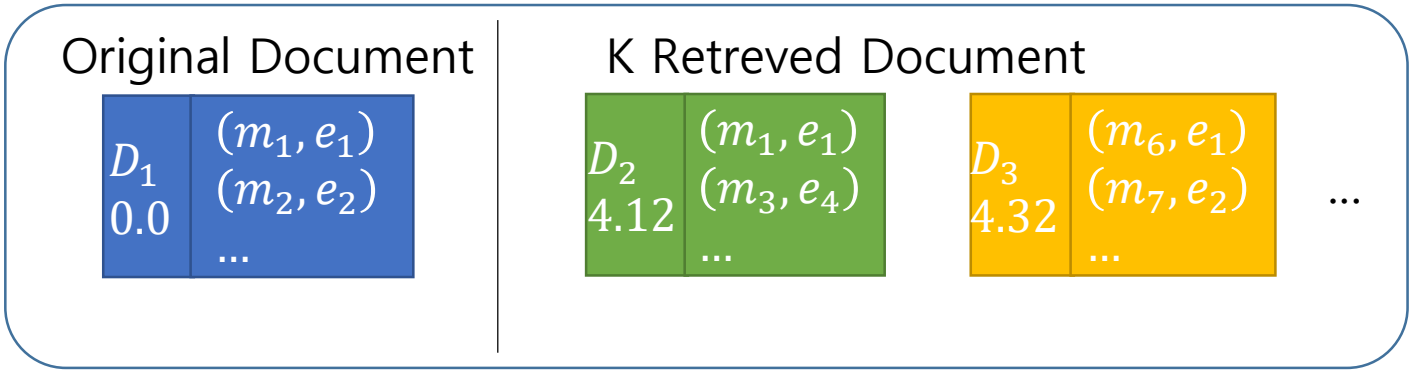
$$dis(D_i, D_j) = \sqrt{\left(DR(D_i) - DR(D_j)\right) \cdot \left(DR(D_i) - DR(D_j)\right)}$$

$$DR(D_i) = FNN(RoBERTa(D_i))$$

Cross-document Entity Coreference Resolution

K=5

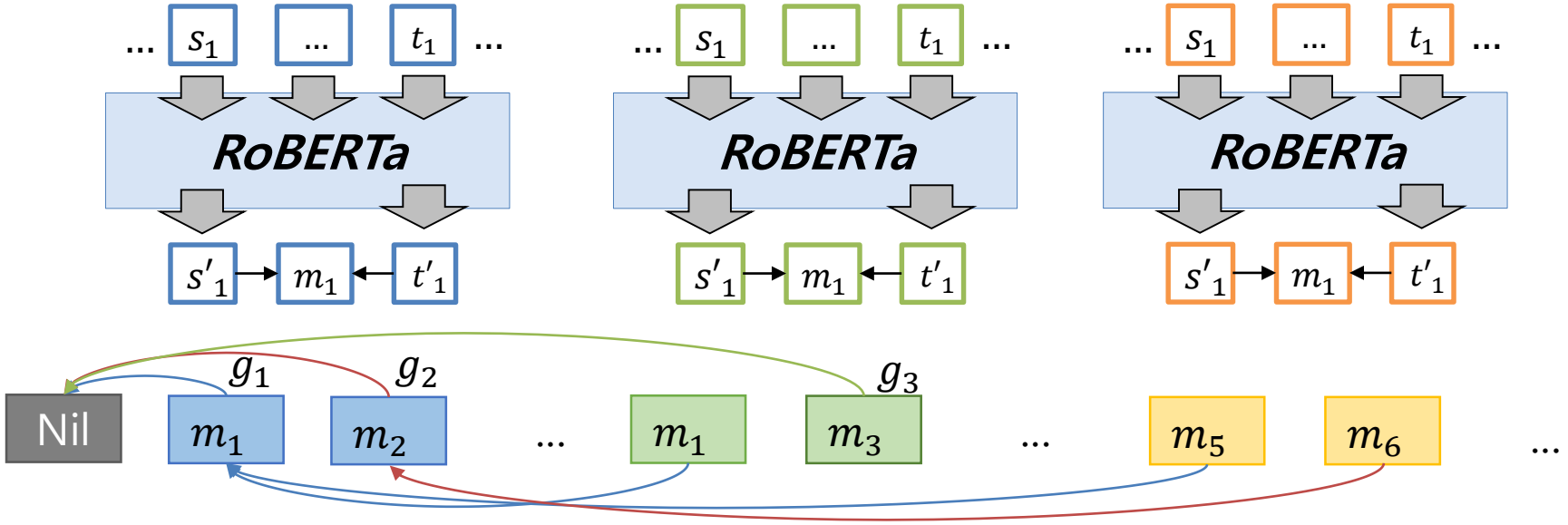
D_1 : 0
D_2 : 4.12
D_3 : 4.32
D_{10} : 4.67
D_{59} : 5.82
D_{73} : 7.54
…



- 가장 가까운 k개의 문서를 Retrieve Coreference Resolution 시행
  - 각 멘션 m은 surface form, 멘션 스팬의 시작점 s, 끝점 t로 구성됨
  - ***Surface form***은 해당 **멘션의 문자열**을 의미하고, 같은 Surface form을 가진 경우, 같은 cluster로 취급
  - Coreference Resolution은 k개 내의 문서 내에 등장한 멘션들을 같은 엔티티인 경우 묶어 서로 다른 Surface form을 가진 Cluster를 **병합**하는 것이 목적
- Surface Form Normalization
  - 멘션 내에서 **특수문자**, **공백문자**를 제거, **한자**를 일반적인 한글로 변환하여 Surface form을 교정

◆문대통령 文대통령 문대통령 -> 문대통령

Cross-document Entity Coreference Resolution



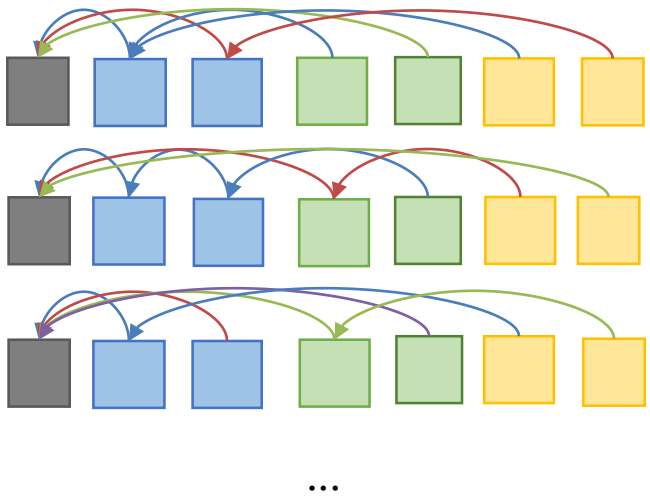
- 각 문서는 개별적으로 RoBERTa로 인코딩, 문서 내의 멘션의 표상은 멘션 스팬의 시작점과 끝점에 해당하는 토큰을 **Concatenate** 하여 사용
- 각 멘션 표상은 이전에 등장한 멘션과 **Biaffine attention**을 통해 점수를 계산, 어떤 멘션과 연결 해야 할 지를 결정
  - 이전 멘션에서 자기와 같은 엔티티인 멘션이 없을 경우 **Nil**을 선택
  - 존재 할 경우 **가장 첫 번째에 등장한 멘션**을 선택

$$m = Concat(s', t')$$

$$score_{i,j} = \frac{\exp\left(Biaffine(m_i, m_j)\right)}{\sum_{j<i} \exp\left(Biaffine(m_i, m_j)\right)}$$

$$\text{Biaffine}(\mathbf{e}, \mathbf{d}) = \mathbf{d} \cdot \mathrm{U} + \mathbf{d} \cdot \mathrm{W_d} + \mathbf{e} \cdot \mathrm{W_e} + \mathbf{b}$$

Cross-document Entity Coreference Resolution



$g_1: \{m_1, m_5\}, g_2: \{m_2, m_6\}, g_3: \{m_3\}$

$g_4: \{m_1, m_7\}, g_5: \{m_8, m_9, m_{10}\}$

$g_6: \{m_3, m_{11}\}, g_7: \{m_{12}, m_{13}, m_{14}\}$

...

...

$g_1: \{m_1, m_5, m_7\}$
$g_2: \{m_2, m_6\}$
$g_3: \{m_3, m_{11}\}$
$g_5: \{m_8, m_9, m_{10}\}$
$g_7: \{m_{12}, m_{13}, m_{14}\}$

- Coreference Resolution 결과를 통해 Surface form 병합
  - 각 k개 문서에서 Coreference Resolution결과 멘션들이 묶일 경우, 해당 멘션을의 Surface form을 하나의 Cluster로 병합
  - 서로 다른 문서집합에서 overlapping 된 멘션의 경우, 두 Cluster를 하나로 병합
  - 최종적으로 모든 문서집합에 대해 병합이 완료되었을 때, 각 Cluster를 하나의 엔티티 Cluster로 취급

실험 구성

- 한국어 위키백과를 통해 6만개의 문서와 뉴스기사 등의 문서를 통해 구성된 네이버 데이터셋으로 데이터셋 구성
  - 네이버 데이터셋은 뉴스기사 등의 문서를 사람이 직접 엔티티를 태깅하여 구성함
  - 위키백과 데이터셋은 사전 학습을 하기위해 사용, 성능 측정은 네이버 데이터셋을 사용
  - 네이버 데이터셋과 위키백과 데이터 사이의 중첩되는 엔티티는 존재하지 않음

지난해 CJ ENM과 합작해 일본판 '프듀 ' 를 제작했던 일본 엔터사 요시모토흥업에서 중국 방송사와 ...

Cross-document Entity Coreference Resolution - Naver Dataset Performance

Model	Precision	Recall	F1
Before Correction Baseline	**93.94%**	85.29%	89.41%
After Correction Baseline	93.91%	92.26%	93.08%
Entity Clustering Oracle	93.72%	98.25%	95.93%
Zeroshot CECR	92.47%	95.02%	93.72%
Zeroshot CECR without Classsifer	89.32%	95.08%	92.11%
Futher trained CECR	92.05%	96.69%	94.31%
Futher trained CECR without Classsifer	90.61%	96.73%	93.57%

Coreference Resolution
Entity Clustering

- 결론
  - Coreference Resolution 모델을 적용함으로써 Surface form 단일 클러스터 대비 B^3 F1성능이 93.08%에서 94.31%로 **1.23%p** 향상됨
- 향후 계획
  - 현재 모델로는 Surface form이 같지만, 엔티티가 다른 경우를 분리 할 수 없기 때문에, 클러스터를 **분할**하거나, **instance 단위**에서 클러스터링을 할 수 있도록 개량이 필요
    - » B^3 precision 에 비해 recall이 높기 때문에 클러스터 분할 만으로도 성능이 충분히 향상 될 수 있을 것으로 예상
  - 전체 문서 내의 멘션에 대해 클러스터링만 진행했기 때문에, 그 클러스터가 어떤 엔티티인지 판별하는 **labeling**을 추가적으로 진행해야 함