개체 멘션 군집화에 기반한 Cross-document 상호참조해결

이종현[01], 나승훈[1], 김현호[2], 강인호[2], 김선훈[2]

[1]전북대학교, [2]네이버

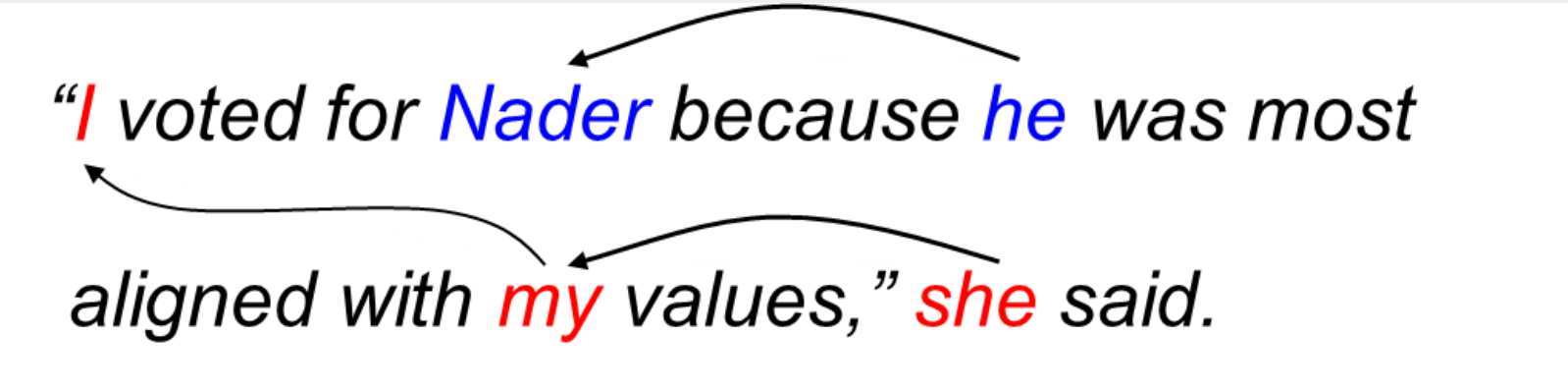
Contents

1. Introduction

■ 상호참조해결

- 같은 의미를 가지는 단어들을 Clustering (군집화)하는 것
- 문서 내에서 같은 개체에 대한 여러 표현들 간의 참조 관계를 밝히는 과정

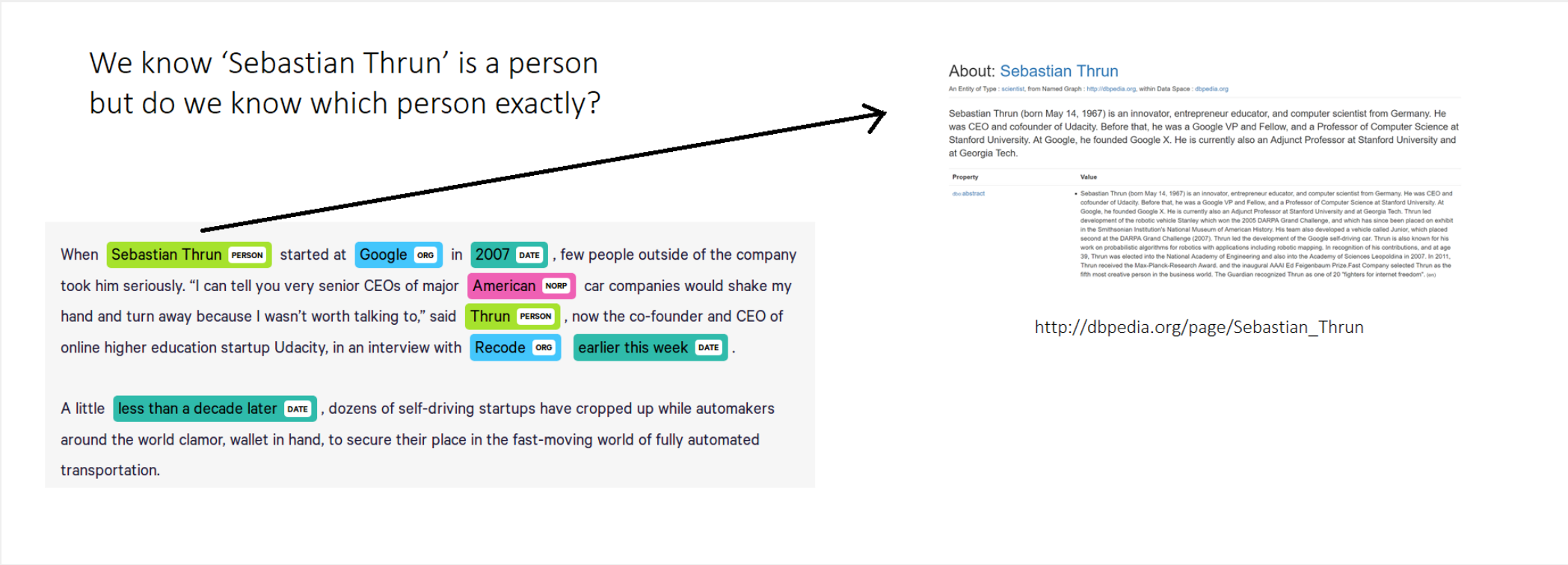
- 상호참조해결 태스크는 local document 상에서 수행



1. Introduction

■ 개체 연결 (Entity Linking)

- 주어진 문서에 출현한 멘션(Mention)을 위키피디아와 같은 지식 베이스(Knowledge Base)상의 하나의 개체와 연결하는 작업
- 개체 연결 작업의 대상이 되는 개체들은 지식 베이스 상에서 개체 설명(Entity description)을 동반해야만 하나의 개체로써 존재할 수 있으며, 개체 설명 없이는 하나의 개체로서 다루지 못함



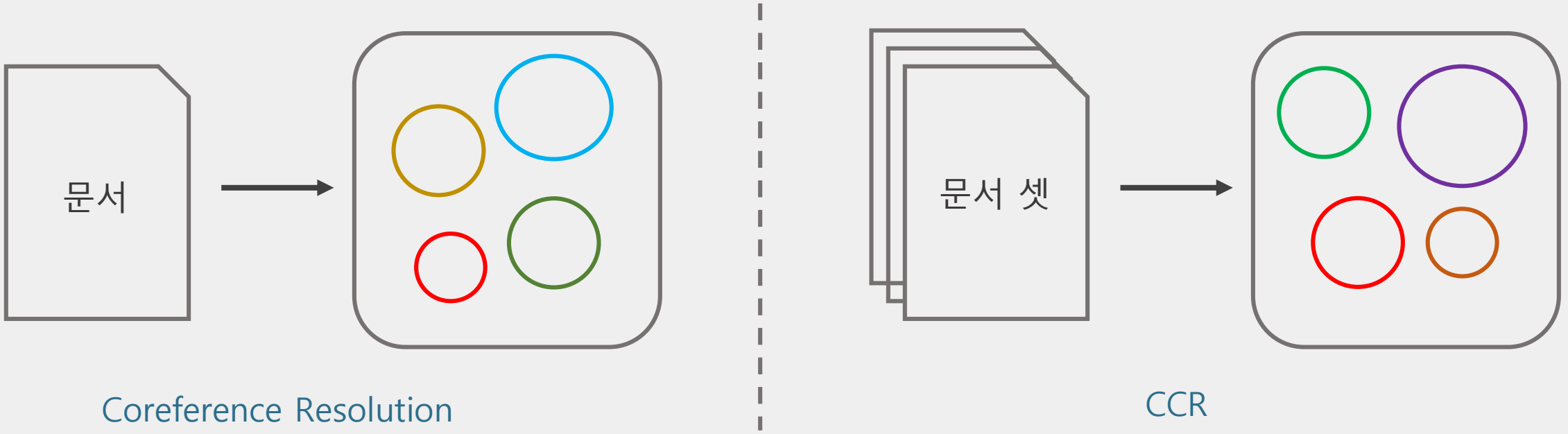
https://medium.com/analytics-vidhya/entity-linking-a-primary-nlp-task-for-information-extraction-22f9d4b90aa8

1. Introduction

■ Cross-document Coreference Resolution (CCR)

- 상호참조해결 모델을 이용해 개체 설명 없이 여러 문서에서 등장하는 같은 개체를 가리키는 멘션들을 하나로 묶어주는 군집화(Clustering) 작업

- 즉, 상호참조해결 작업을 한 문서가 아니라 전체 문서 셋 상에서 수행하는 것



2. Proposed Model

■ 베이스라인(Baseline)

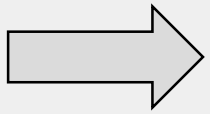
- 전체 문서 셋 내에서 등장한 멘션들의 텍스트가 같은 경우 이를 모두 하나의 군집으로 묶어내어 베이스라인 군집으로써 정의함

Document 1

맨체스터 유나이티드가 '맨체스터 더비'를 승리하며 맨체스터 시티의 공식전 연승 행진을 저지했다.

Document 2

박지성이 맨체스터 유나이티드 엠베서더 자리를 포기하고 전북현대 어드바이저로서 유소년 시스템 발전에 기여한다. 박지성은 "전북 소속이기 때문에 맨유 엠베서더 일은 당연히 할 수 없다."고 설명했다.



Cluster 1

맨체스터 유나이티드

맨체스터유나이티드

Cluster 2

박지성 박지성

Cluster 3

맨체스터 시티

Cluster 4

전북현대

Cluster 5

맨유

Cluster 6

전북

2. Proposed Model

■ 상호참조해결 모델

- 멘션의 텍스트가 서로 다른 두 군집을 하나의 군집으로 병합하기 위해 상호참조해결 모델을 이용

• 상호참조해결 모델 수식

- $\boldsymbol{g}_i : i$ 번 째 멘션의 표상(Representation)

$$\boldsymbol{h} = RoBERTa(w)$$

$$\boldsymbol{g}_i = [\boldsymbol{h}_{START(i)}; \boldsymbol{h}_{END(i)}] \rightarrow \boldsymbol{g}_i^1$$

- 두 멘션 m_i 와 m_j 의 연관성을 계산하는 점수 함수(Scoring function) $s(i,j)$

$$s(i,j) = FFNN([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \circ \boldsymbol{g}_j])$$

∘ 는 요소별 곱 (element-wise multiplication) 을 의미

2. Proposed Model

- 상호참조해결 모델 수식
  - 모델의 추론 과정은 총 N 번의 반복을 통해 진행
  - n 번 째 반복에서의 멘션 m_i 에 대한 선행사(antecedent)의 확률 분포 $P_n(y_i)$

$$P_n(y_i) = \frac{e^{s\left(\boldsymbol{g}_i^n, \boldsymbol{g}_{y_i}^n\right)}}{\sum_{y \in \mathcal{Y}(i)} e^{s\left(\boldsymbol{g}_i^n, \boldsymbol{g}_y^n\right)}}$$

$\boldsymbol{g}_i^n$ 은 n 번 째 반복에 해당하는 멘션 m_i 의 표상

  - 멘션 m_i 의 $n+1$ 번 째 표상 $\boldsymbol{g}_i^{n+1}$

$$\boldsymbol{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot \boldsymbol{g}_{y_i}^n$$

$$\boldsymbol{f}_i^n = \sigma\left(W_f [\boldsymbol{g}_i^n, \boldsymbol{a}_i^n]\right)$$

$$\boldsymbol{g}_i^{n+1} = \boldsymbol{f}_i^n \circ \boldsymbol{g}_i^n + (1 - \boldsymbol{f}_i^n) \circ \boldsymbol{a}_i^n$$

2. Proposed Model

- 상호참조해결 모델의 결과로 두 멘션 m_i, m_j 가 같은 개체를 가리킨다고 예측 되었을 때, 두 멘션이 포함된 서로 다른 두 군집을 병합하는 방식으로 점차 서로 다른 두 군집을 병합해 나감

2. Proposed Model

■ 텍스트 분류기

- 상호참조해결 모델의 결과로 두 군집의 병합이 잘못된 경우 정밀도(precision)의 하락으로 이어질 수 있으므로 RoBERTa 기반의 텍스트 분류기를 추가적인 검증 모델로써 사용

- 두 멘션 텍스트 $text_1, text_2$ 의 결합 점수

$$\boldsymbol{r} = RoBERTa(\{[CLS], text_1, [SEP], text_2, [SEP]\})$$

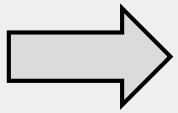
$$P = sigmoid(FFNN(\text{meanPool}(r)))$$

- 텍스트 분류기의 출력 확률이 $p = (0.4)$ 이상인 경우 두 멘션이 포함된 서로 다른 두 군집을 병합하도록 함

2. Proposed Model

■ 텍스트 분류기

맨체스터 유나이티드가 '맨체스터 더비'를 승리하며 맨체스터 시티의 공식전 연승 행진을 저지했다.

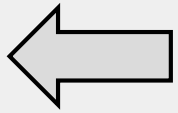


Coref Model

Prediction

맨체스터 유나이티드

맨체스터 시티



Surface form classifier

Prediction

맨체스터 유나이티드

맨체스터 시티

- Coref Model 이 두 멘션을 하나의 cluster 로 묶으려는 경우 (즉, 두 멘션의 pair score 가 높은 경우) 이를 텍스트 분류기의 입력으로 주어 두 멘션의 텍스트만을 보고 분류기가 두 멘션이 pair 관계가 될 수 있는지 판단
- 이를 통해 precision 의 하락을 어느 정도 방지 할 수 있음

2. Proposed Model

■ Dense retrieval 을 이용한 KNN (K-Nearest Neighbor)

- 비슷한 여러 문서를 한 번에 입력으로 받아 상호참조해결 작업을 수행
- Dense retrieval 을 이용하여 쿼리(query) 문서 D_q 를 이용해 K 개의 유사한 문서들을 검색하여 쿼리 문서 D_q 와 K 개의 문서 모두 상호참조해결 모델의 입력으로 사용
- 문서 검색을 위한 dense retrieval 모델은 한국어 위키피디아에서 자체 수집한 약 430,000 개의 문서를 ICT(inverse cloze task) 방식으로 사전 학습하여 사용

블록 인코더

$$\boldsymbol{H}_B = RoBERTa^B(B)$$

$$\boldsymbol{e}_B = \boldsymbol{W}_B \frac{\sum_i^{L_B} \boldsymbol{H}_B^i}{L_B}$$

쿼리 인코더

$$\boldsymbol{H}_Q = RoBERTa^Q(Q)$$

$$\boldsymbol{e}_Q = \boldsymbol{W}_Q \frac{\sum_i^{L_Q} \boldsymbol{H}_Q^i}{L_Q}$$

$\boldsymbol{H}_B \in \mathbb{R}^{L_B \times d}$, $\boldsymbol{H}_Q \in \mathbb{R}^{L_Q \times d}$: RoBERTa 모델의 출력

d: RoBERTa 모델의 출력 차원(Dimension)

$\boldsymbol{e}_B, \boldsymbol{e}_Q \in \mathbb{R}^u$: 각각 블록 인코더와 쿼리 인코더의 출력

i : i 번 째 토큰

L_B, L_Q : 각각 블록과 쿼리의 길이

$W_{B,Q} \in \mathbb{R}^{u \times d}$: 학습 가능한 파라미터

3. Experiment

■ 실험 데이터 및 세팅

- 위키피디아에서 자체 수집한 문서들을 이용하여 실험 진행

학습 셋	개발 셋	평가 셋
29,500	500	1,079

- Hyperparameter
  - Optimizer: AdamW
  - learning rate: 2e-5
  - $N = 2$
  - $W_f \in \mathbb{R}^{768}$

3. Experiment

■ 실험 결과

- CoNLL-2011 에서 공식으로 지정한 성능 지표 중 하나인 B^3 를 사용

모델	Precision	Recall	F1
Optimal performance	93.78%	97.50%	95.60%
Basline	**93.91%**	92.26%	93.08%
Coref Model	91.26%	95.24%	93.21%
Coref Model (+ text classifier)	92.62%	95.03%	93.81%
Coref Model (+ text classifier + KNN)	91.76%	**96.66%**	**94.15%**

- Optimal performance는 서로 다른 군집들이 최대한 잘 병합되었을 때, 즉 병합된 군집들에 대한 군집화 성능이 가장 높은 경우를 의미

4. Conclusion

■ 결론

- CCR 문제를 기존의 상호참조해결 모델을 이용해 해결하고자 함

- 텍스트 분류기와 dense retrieval 을 이용하여 멘션 군집을 병합하는 과정에서의 정밀도 하락을 방지하여 최종 성능을 높이고자 함

- 결과적으로 B^3 성능 지표에서 94.15% 의 F1 score 를 달성함

- 향후에는 군집 내에 가리키는 개체가 서로 다른 멘션들이 존재하는 경우 한 군집을 여러 군집으로 분리하는 작업을 통해 성능을 더욱 높일 계획임