

개체 멘션 군집화에 기반한 Cross-document 상호참조해결

이종현⁰¹, 나승훈¹, 김현호², 강인호², 김선훈²

¹전북대학교, ²네이버

1. Introduction

2. Proposed Model

2.1. Baseline

2.2. Coreference Resolution Model

2.3. Text Classifier

2.4. KNN using Dense retrieval

3. Experiment

4. Conclusion

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

“I voted for Nader because he was most aligned with my values,” she said.

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 개체로서 존재할 수 있으며, 개체 설명 없이는 하나의 개체로서 다루지 못함

We know 'Sebastian Thrun' is a person
but do we know which person exactly?

When **Sebastian Thrun** PERSON started at **Google** ORG in **2007** DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** NORP car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** ORG **earlier this week** DATE.

A little **less than a decade later** DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

About: Sebastian Thrun

An Entity of Type : scientist, from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org

Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech.

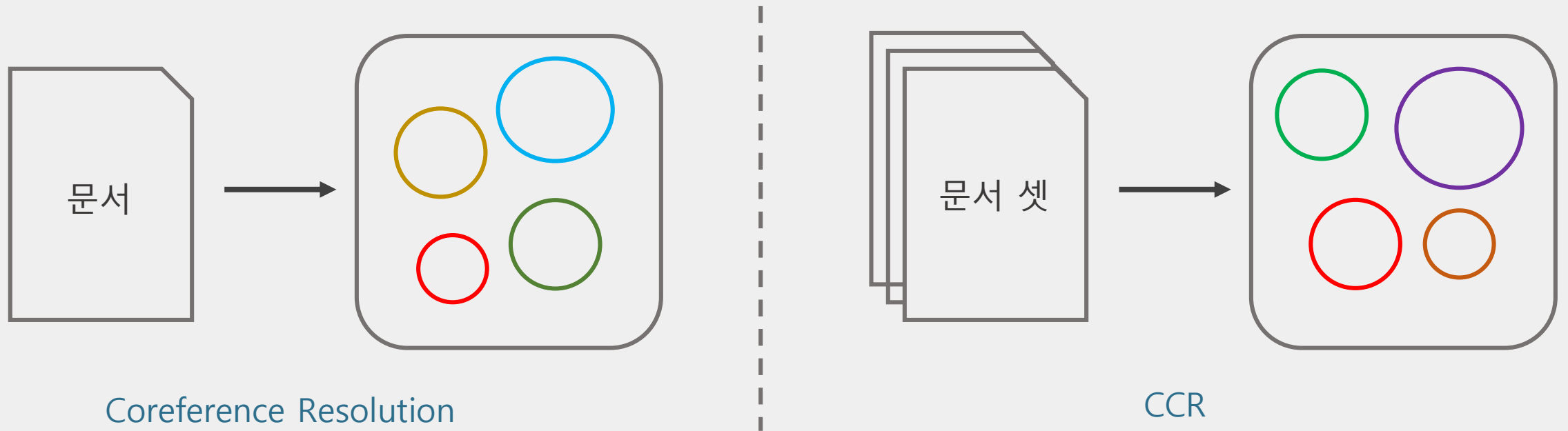
Property	Value
abstract	<ul style="list-style-type: none">Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech. Thrun led development of the robotic vehicle Stanley which won the 2005 DARPA Grand Challenge, and which has since been placed on exhibit in the Smithsonian Institution's National Museum of American History. His team also developed a vehicle called Junior, which placed second at the DARPA Grand Challenge (2007). Thrun led the development of the Google self-driving car. Thrun is also known for his work on probabilistic algorithms for robotics with applications including robotic mapping. In recognition of his contributions, and at age 39, Thrun was elected into the National Academy of Engineering and also into the Academy of Sciences Leopoldina in 2007. In 2011, Thrun received the Max-Planck-Research Award, and the inaugural AAAI Ed Feigenbaum Prize. Fast Company selected Thrun as the fifth most creative person in the business world. The Guardian recognized Thrun as one of 20 "fighters for internet freedom". ^(en)

http://dbpedia.org/page/Sebastian_Thrun

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

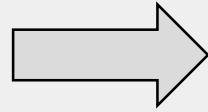
Document 1

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.

군집화



Cluster 1

맨체스터 유나이티드
맨체스터유나이티드

Cluster 5

맨유

Cluster 2

박지성 박지성

Cluster 6

전북

Cluster 3

맨체스터 시티

Cluster 4

전북현대

2. Proposed Model

■ 상호참조해결 모델

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

• 상호참조해결 모델 수식

- g_i : i 번째 멘션의 표상(Representation)

$$\mathbf{h} = \text{RoBERTa}(w)$$

$$\mathbf{g}_i = [\mathbf{h}_{\text{START}(i)}; \mathbf{h}_{\text{END}(i)}] \quad \longrightarrow \mathbf{g}_i^1$$

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

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

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

2. Proposed Model

- 상호참조해결 모델 수식

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

$$P_n(y_i) = \frac{e^{s(\mathbf{g}_i^n, \mathbf{g}_{y_i}^n)}}{\sum_{y \in \mathcal{Y}(i)} e^{s(\mathbf{g}_i^n, \mathbf{g}_y^n)}} \quad \mathbf{g}_i^n \text{ 은 } n \text{ 번째 반복에 해당하는 멘션 } m_i \text{ 의 표상}$$

- 멘션 m_i 의 $n + 1$ 번째 표상 \mathbf{g}_i^{n+1}

$$\mathbf{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot \mathbf{g}_{y_i}^n$$
$$\mathbf{f}_i^n = \sigma(W_f[\mathbf{g}_i^n, \mathbf{a}_i^n])$$

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

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

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

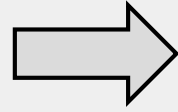
$$P = sigmoid(FFNN(\text{meanPool}(r)))$$

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

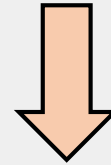
2. Proposed Model

■ 텍스트 분류기

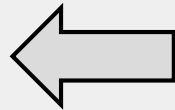
맨체스터 유나이티드가 '맨체스터 더비'를 승리하며 맨체스터 시티의 공식전 연승 행진을 저지했다.



Coref Model



Prediction

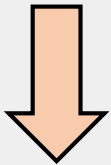


맨체스터 유나이티드

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

$$H_B = \text{RoBERTa}^B(B)$$

$$e_B = W_B \frac{\sum_i^{L_B} H_B^i}{L_B}$$

쿼리 인코더

$$H_Q = \text{RoBERTa}^Q(Q)$$

$$e_Q = W_Q \frac{\sum_i^{L_Q} H_Q^i}{L_Q}$$

$H_B \in \mathbb{R}^{L_B \times d}$, $H_Q \in \mathbb{R}^{L_Q \times d}$: RoBERTa 모델의 출력

d : RoBERTa 모델의 출력 차원(Dimension)

$e_B, e_Q \in \mathbb{R}^u$: 각각 블록 인코더와 쿼리 인코더의 출력

i : i 번째 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%
Baseline	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