Coreference Resolution 과 Information Extraction모델을 통한 Knowledge Base Population데이터 구축

조수현

2020 12

Outline

1. 논문 개요(목표)
   - KBP(Knowledge Base Population)
   - Open IE and Coreference resolution
2. Dataset
3. Model
   - Open IE
   - Coreference resolution
4. Result

01
논문 개요(목표)

(1) KBP(Knowledge Base Population)
(2) Proposal
(3) Open IE and Coreference resolution

Knowledge Base Population

- **Knowledge Base Population (KBP)**

구조화되지 않은 텍스트에서 지식 기반 (KB)을 채우기 위한 기술을 개발하는 것, NLP 문제에 맞게 데이터를 가공하는 것이 아닌 추출된 지식을 통해 문제를 해결

기존 KBP 방식은 특정 도메인에서만 지식을 추출 가능한 문제점을 제시하고 본 연구를 통해 Open Domain 데이터를 추출할 수 있는 방법을 제시

- KnowledgeNet: A Benchmark Dataset for Knowledge Base Population

TAC 2015 우승 시스템인 NER -> NEL -> Relation Extraction , 모델에 Coreference Resolution 을 추가하여 4 개의 모델을 사용하여

KBP 데이터를 구축 및 여러 데이터를 비교 데이터 셋 - ACE, TAC, TACRED, FewRel, DocRED, GoogleRE, T-REx

입력 문장 : Obama called Trump.

KBP 지식 트리플 추출 ("subject", "relation", "object")

{“Obama", "per:co-worker", “Trump" }

Knowledge Base Population Data example

Knowledge Base Population

- **Knowledge Base Population (KBP)**
  - **필요 기술 (순서)**

Step 1 NER(named entity recognition)

개체명, 멘션을 추출하기 위해서 사용

Step 2 NEL(named entity linking)

텍스트에 언급 된 개체명에 고유 한 ID를 할당.

https://www.aclweb.org/anthology/D19-1069.pdf

Ex) wikipedia url or doc id

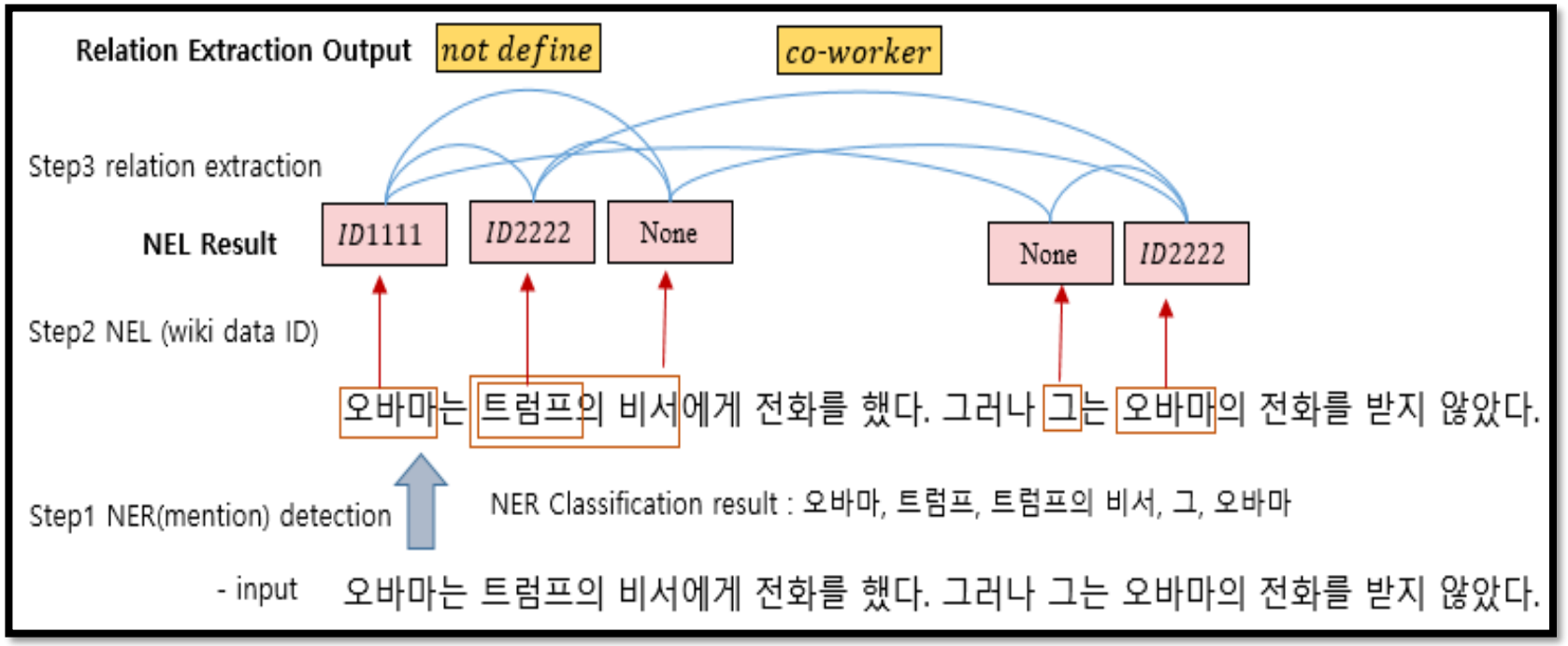
or 관계가 있는 개체 끼리 연결

Step 3 relation extraction

연결된 객체 간의 관계를 추출

Ex) Tac Relation Extraction Dataset 은 41개의 관계를 정의

나이, 제목, 조직, 생일, 부모, 형제, 배우자, 직원, 대체 이름, 수도, 자회사 등



Knowledge Base Population task

Knowledge Base Population

기존 Knowledge Base Population (KBP) 방식의 문제점

1. 새로운 단어에 적용이 불가능(NEL)

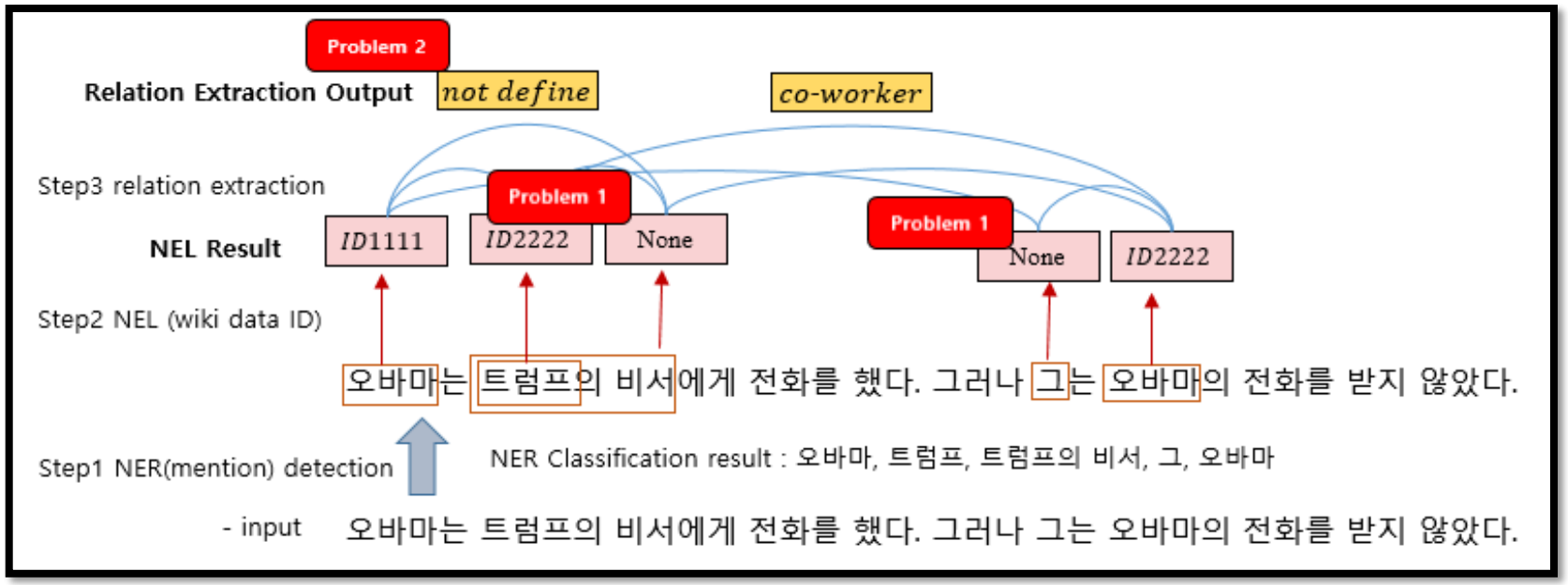
Wiki data 를 기준으로 entity linking을 하기때문에 wiki에 없는 데이터의 경우에는 적용할 수 없다.

Problem 1 : 구축된 지식에서 문장의 "트럼프의 비서"라는 개체 정보가 없는 경우

2. 오픈 도메인에는 적용 불가능 (relation extraction)

관계를 정의하여 그 관계 중 분류를 하는 문제 이기 때문에 정의되지 않은 관계는 KBP 데이터를 구축할 수 없다.

Problem 2 : "오바마"와 "트럼프의 비서"의 관계인 "전화를 했다"가 미리 정의된 관계 분류에 없는 경우



Problems with Knowledge Base Population tasks

Proposal

기존 방식의 문제점과 해결방법 제시

- **기존 방식의 문제점**

1. 새로운 단어에 적용이 불가능(NEL)

Wiki data 를 기준으로 entity linking을 하기때문에 wiki에 없는 데이터의 경우에는 적용할 수 없다.

2. 오픈 도메인에는 적용 불가능 (relation extraction)

관계를 정의하여 그 관계 중 분류를 하는 문제 이기 때문에 정의되지 않은 관계는 KBP 데이터를 구축할 수 없다.

3. 한국어 Open Information Extraction Dataset

한국어 Open Information Extraction 학습 데이터 없음

- **제안**

1. Coreference resolution task를 통해 해결

Wiki data 를 기준으로 entity linking을 하는 것이 아닌 입력 문서를 기준으로 개체간의 관계를 찾음

2. Open Information Extraction task를 통해 해결

관계를 입력 데이터에서 찾음으로 미리 관계를 정의할 필요가 없다.

3. 한국어 Open Information Extraction Dataset

영어 Open Information Extraction 데이터를 가공하여 한국어 데이터 30만개 데이터 구축

Open IE and Coreference resolution

- **Open IE(Open Information Extraction)**

domain-independent한 특성과 정의된 관계에 국한되지 않는 특성 덕분에

large-scale의 코퍼스로부터 방대한 정보를 추출할 수 있는 효과적인 도구

예시) 한국어

- 입력 문장

오바마는 트럼프에게 전화를 했다.

- 추출 데이터 : 트리플(" 주체 " , " 관계 " , " 대상 ")

{"오바마 " , "전화를 했다.", "트럼프" }

Open IE 예시

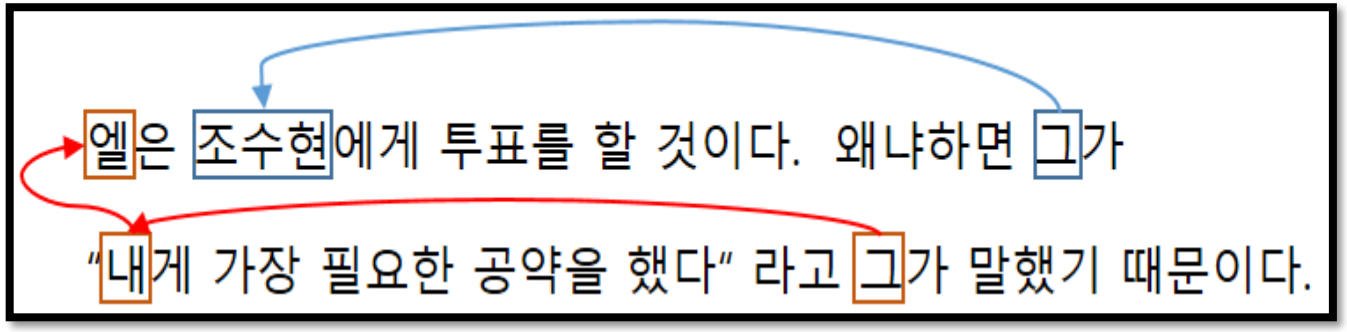
- **Coreference resolution**

텍스트에서 동일한 개체를 참조하는 모든 표현식을 찾는 작업

즉, 대명사가 지칭하는 개체를 찾는 것

입력 문장 : '엘은 조수현에게 투표를 할 것이다. 왜...'

출력 결과 : [{"엘", "내", "그"}, {"조수현", "그"}] 및 각 위치정보



[그림 5] Coreference resolution 예시

https://arxiv.org/pdf/1901.10879v6.pdf

02
Dataset

(1) 학습 데이터
(2) 한국어 데이터 가공

Dataset – 학습 데이터

Open IE 와 Coreference Resolution 학습 데이터

Open IE4 Corpus	
URL	http://knowitall.github.io/openie/
데이터	약 110만 문장
특징	데이터가 많음, 노이즈(특수문자 등)가 있음

Coreference resolution Dataset	
URL	https://www.ldc.upenn.edu/
데이터	학습(2799), 평가(343), 테스트(348) 총 3,490 개
특징	OntoNotes Relaese5.0 - LDC2013T19, 데이터 수 가 적음

Open IE Corpus	
RUL	데이터 없음, 직접 가공
데이터	약 30만 문장을 가공
특징	가공 데이터로 노이즈 포함
비고	구글 번역기 API 를 이용하여 번역 후 object, subject, relation을 직접 or 룰 기반 을 이용하여 데이터 구축

Coreference resolution Dataset	
URL	http://aiopen.etri.re.kr/service_dataset.php
데이터	645 문서, 1086문장
특징	ETRI 에서 제공 , 데이터 수 가 적음

한국어 데이터 가공

한국어 Open IE 데이터 가공


```
{
"sentence": "₩u2020 - # 59 team modena was disqualified for failing post-race inspection .",
...
"tuples": [
    {
    ...
    "arg0": "- # 59 team modena",
    "relation": "was disqualified",
    "args": [
        "for failing post-race inspection"
    ],
    "arg0_pos": [
        1,  5
    ],
    "rel_pos": [
        6, 7
    ],
    "args_pos": [
        [
        8, 11
        ]
    ]
    }
]
}
```


원본 데이터 일부


```
{
"sentence": "†-# 59 팀 모데나는 레이스 후 검사 실패로 실격 처리되었습니다.",
"tuples": [
    {
        "arg0": "59 팀 모데나는 레이스",
        "arg0_pos": [
            1, 4
        ],
        "args": [
            "후 검사 실패로"
        ],
        "args_pos": [
            [
                5, 7
            ]
        ],
        "rel_pos": [
            8,  9
        ],
        "relation": "실격 처리되었습니다."
    }
]
}
```


변경 데이터 일부

03
Model

(1) Architecture Overview
(2) Open IE Model
(3) Coreference Resolution Model

Architecture Overview

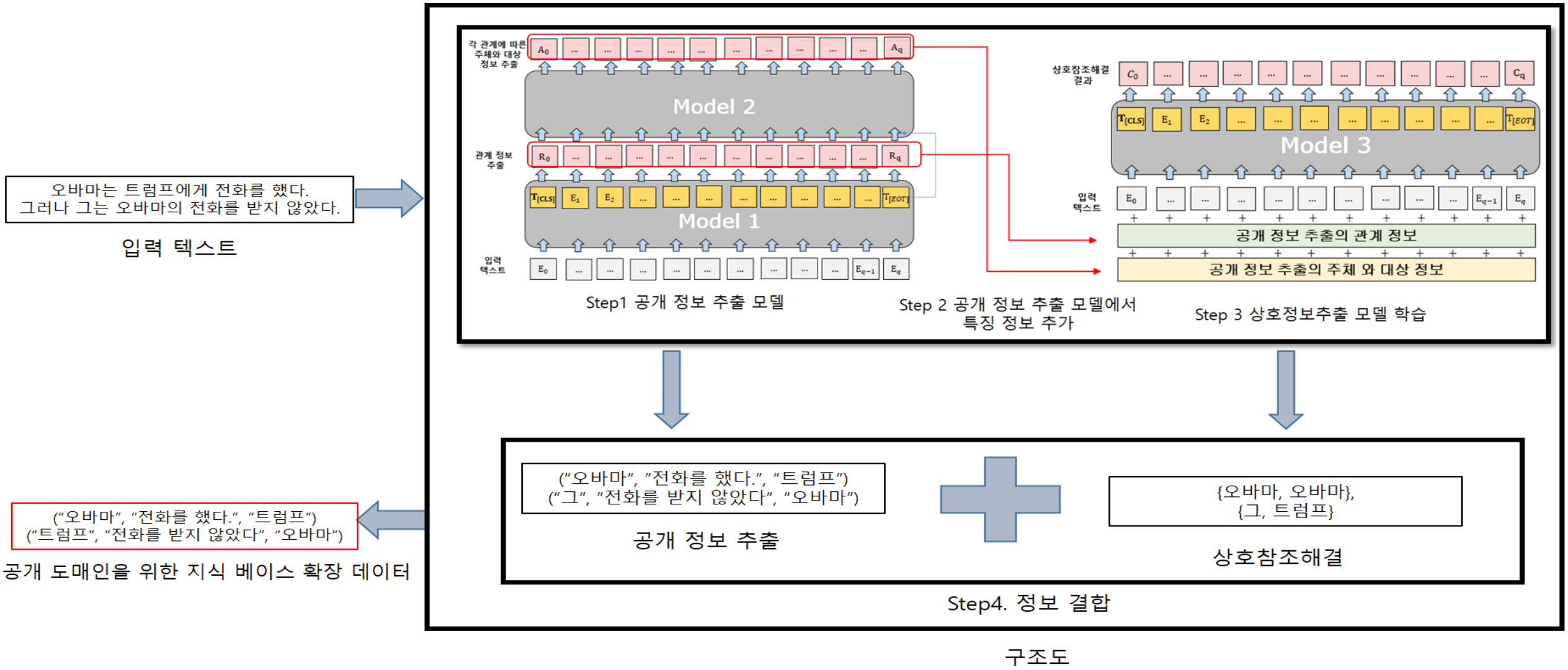


그림 12. Our KBP(knowledge base population) Structure

Open IE Model

Open IE 모델

1. 관계 정보를 예측 후 주체와 대상을 예측

- Step1 : BERT모델에서 문장 내의 관계 정보 위치를 예측

- Step2 : BERT의 hidden layer정보와 관계 정보에서 주체와 대상을 예측

2. Step2부분의 성능 향상을 위해 layer를 변형

CNN Layer를 추가하여, Step1의 결과인 문장 표현 벡터와 관계 정보의

추출된 특징 정보를 활용해 모델의 NLU를 향상

- Transformer Layer와 CNN Layer의 결과를 토큰 단위로

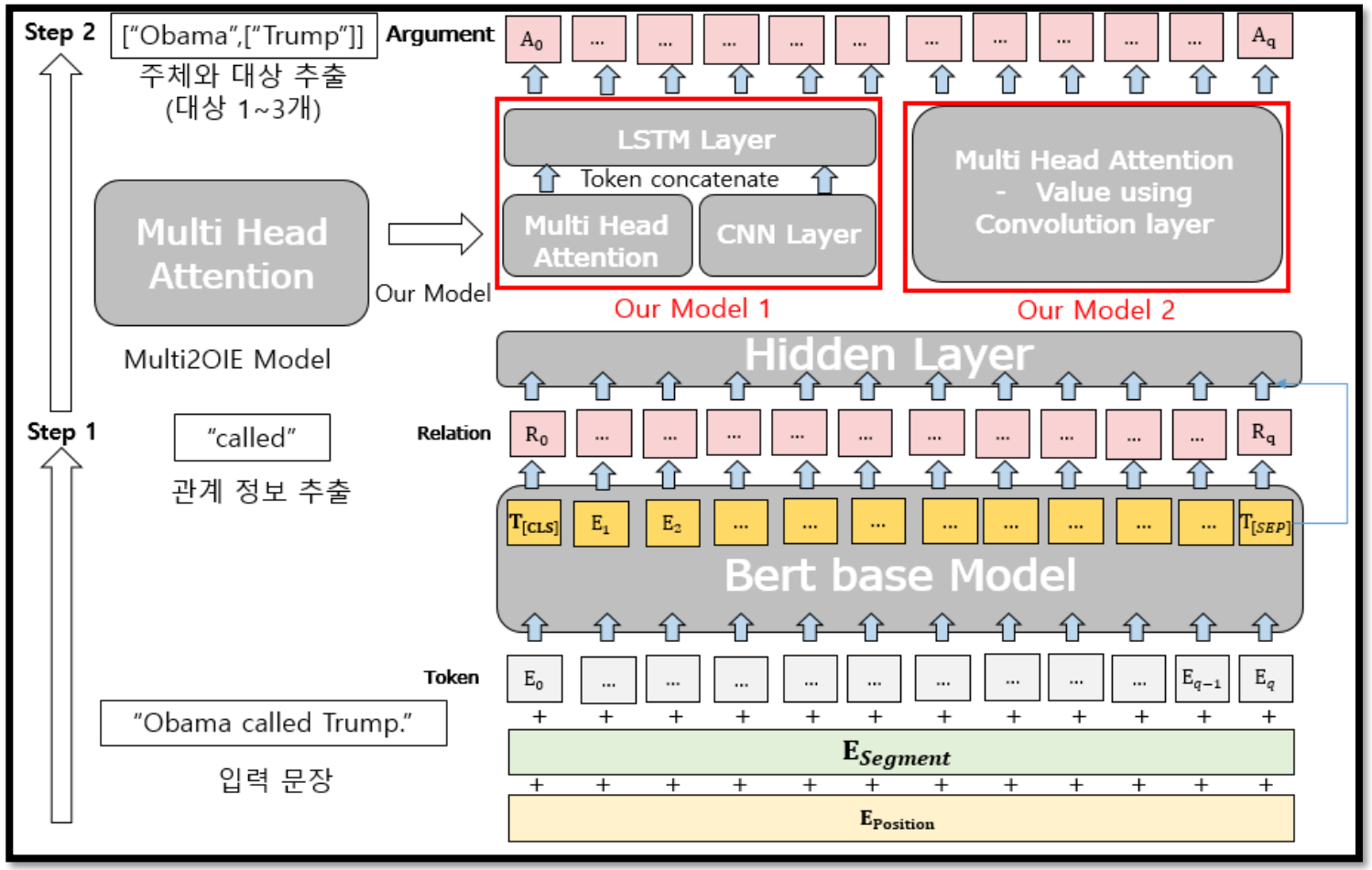
concatenate 후 LSTM입력으로 활용

- Multi Head Attention의 각 Attention 에서 CNN Layer를 통해

문장을 특징 추출한 V값을 사용

3. SOTA 모델

참조 모델



Multi2OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT

그림 10. Our Open IE(Information Extraction) 모델

Coreference Resolution Model

Coreference Resolution 모델

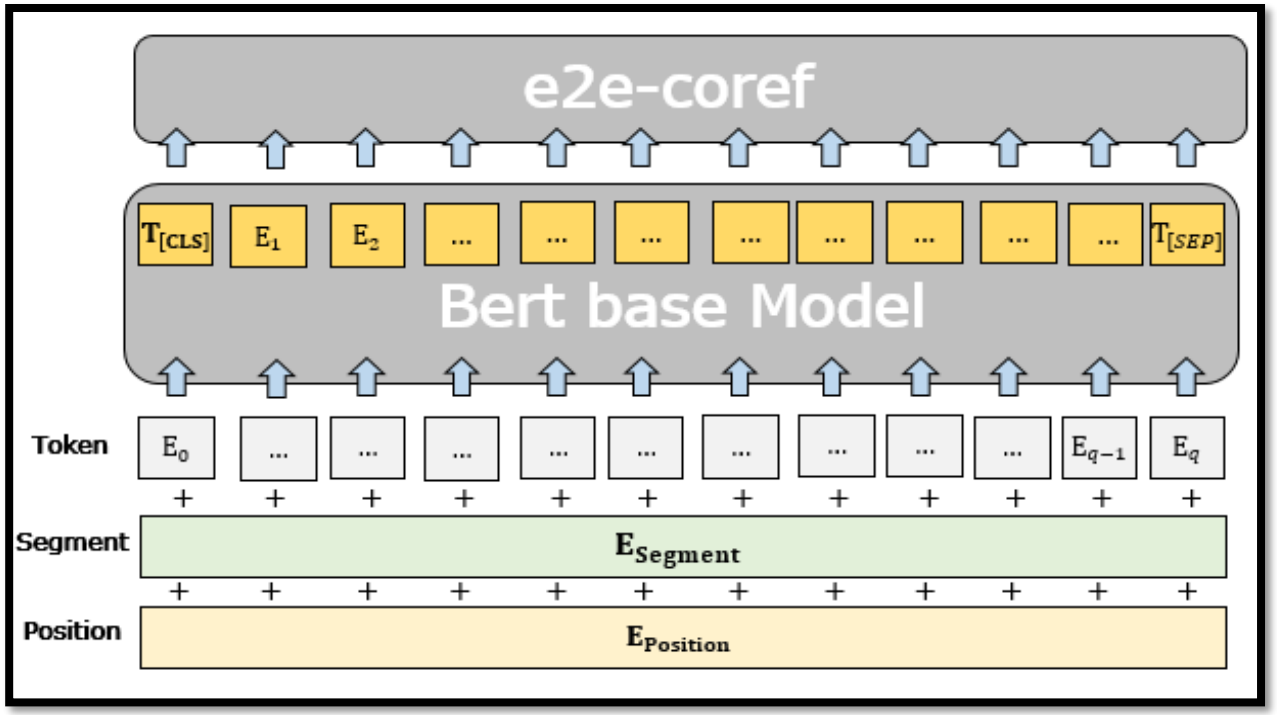
1. 데이터의 부족으로 정확도를 더 높일 방법 필요

2. SOTA 모델

참조

- SpanBERT: Improving Pre-training by Representing and Predicting Spans

- Facebook github : https://github.com/mandarjoshi90/coref



Facebook github Coreference Resolution Model

Coreference Resolution

BERT for Coreference Resolution: Baselines and Analysis

Step 1 Open IE 모델을 학습

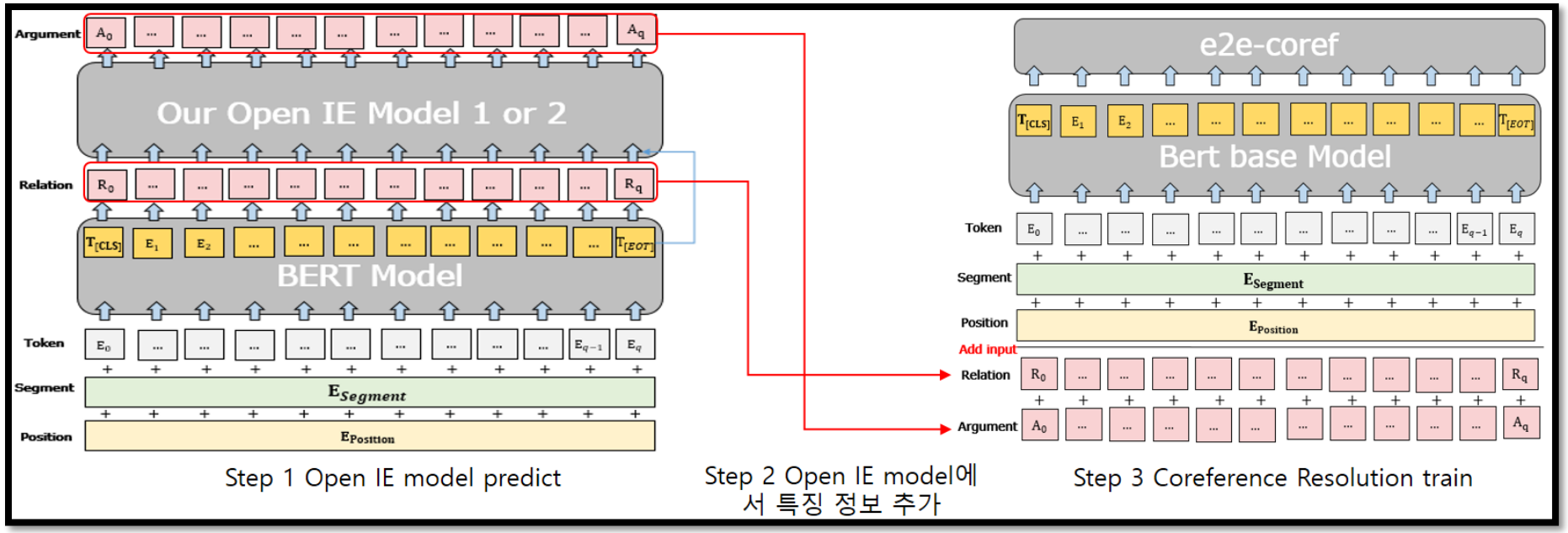
Step 2 학습한 Open IE 모델로 Coreference Resolution 데이터를 예측

Step 3예측한 데이터를 Coreference Resolution 학습

$$Y_i = f(T_e, S_e, P_e, R_e, A_e)$$

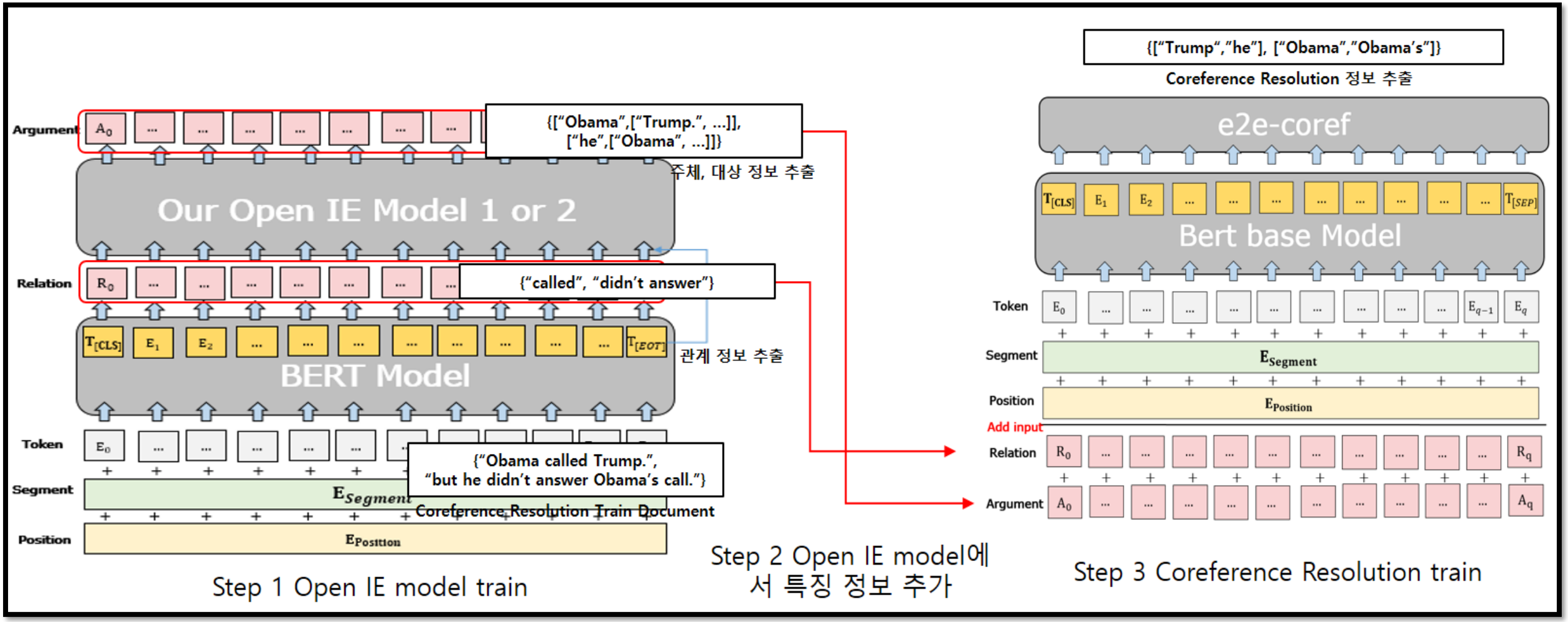
$$T_e = LayerNorm\big(GeLU(W_0, Token)\big)$$
$$S_e = LayerNorm(GeLU(W_1, Segment))$$
$$P_e = LayerNorm(GeLU(W_2, Position))$$
$$R_e = LayerNorm(GeLU(W_3, f_{oie-relation}(T_e, S_e, P_e))$$
$$A_e = LayerNorm(GeLU\big(W_4, f_{oie-argument}(T_e, S_e, P_e)\big))$$

수식 (7) 변경된 Coreference resolution Model



[그림 11] Our Coreference Resolution Model

Coreference Resolution



Our Coreference Resolution Model

04
Result

(1) Model Result
(2) Sample Result
(3) Result

Model Result

Open IE Model Result

- Open IE Step 2모델의 주체, 대상 정보 추출

Step2 모델의 입력으로 입력 문장의 토큰 단위 표현 벡터와 관계 정보의 concatenate의 특징 정보를 1D-CNN Layer를 이용하여 추출하여 모델의 정확도를 향상

제안 모델1은 Multi-Head Attention과 병렬로 CNN Layer를 수행

제안 모델2는 Multi-Head Attention 모델의 V값을 CNN Layer를 이용해 특징 추출된 값을 사용

	Argu-loss-e1	relation-loss-e1	F1-e1	Argu-loss-e2	relation-loss-e2	F1-e2
bert-base	0.209	0.695	85.8	0.107	0.436	86.7
bert-large	0.152	0.574	86.3	0.081	0.352	86.5
spanbert-base	0.171	0.432	85.7	0.074	0.295	86.0
spanbert-large	0.204	0.558	86.8	0.099	0.356	86.6
Our bert-base1	0.062	0.361	86.0	0.039	0.207	86.7
Our bert-large1	0.059	0.347	88.9	0.034	0.190	87.5
Our spanbert-base1	0.106	0.518	88.2	**0.056**	**0.270**	**88.9**
Our spanbert-large1	0.064	0.353	87.6	0.037	0.195	88.4
Our bert-base2	0.081	0.348	88.2	0.051	0.241	88.3
Our bert-large2	**0.076**	**0.335**	**89.2**	0.045	0.222	88.4
Our spanbert-base2	0.138	0.402	87.2	0.054	0.252	88.0
Our spanbert-large2	0.100	0.358	88.4	0.046	0.227	88.4

[표 2] Open IE 영어 데이터 모델 별 성능 결과 정리 표

	Argu-loss-e1	relation-loss-e1	F1-e1	Argu-loss-e2	relation-loss-e2	F1-e2	Argu-loss-e3	relation-loss-e3	F1-e3	Argu-loss-e4	relation-loss-e4	F1-e4
Our electra-small1	0.330	1.056	54.1	0.285	0.762	60.0	0.281	0.705	57.3	0.279	0.690	59.4
Our electra-base1	**0.311**	**0.868**	**64.0**	0.269	0.649	64.0	0.263	0.620	59.7	0.260	0.606	60.2
Our albert-base1	0.294	0.788	9.2	0.267	0.618	11.8	0.260	0.589	7.0	0.256	0.570	12.0
Our albert-large1	0.281	0.768	34.8	0.254	0.592	17.2	0.240	0.546	17.1	0.224	0.497	18.8
Our electra-small2	0.344	0.810	66.0	0.288	0.684	60.8	0.283	0.665	63.8	0.280	0.656	60.2
Our electra-base2	0.320	0.743	54.1	0.270	0.635	62.5	0.263	0.612	58.9	0.260	0.598	59.1
Our albert-base2	0.315	0.758	57.6	0.278	0.648	62.1	0.273	0.625	58.0	0.270	0.614	57.4
Our albert-large2	**0.296**	**0.726**	**66.7**	0.263	0.614	50.3	0.254	0.583	53.5	0.247	0.560	54.4

[표 3, 4] Open IE 한국어 데이터 모델 별 성능 결과표

Model Result

Coreference Resolution Model Result

- 영어 Coreference Resolution 데이터 F1 score약 0.5 ~ 1% Point 상승

Open IE 모델의 결과를 추가 특징 정보로 추가

부족한 상호참조해결 데이터의 학습을 추가 정보를 제공해 줌으로 써 해결

문장 내에서 주체 정보, 대상 정보 및 관계 정보를 추가 특징 정보를 입력으로 줌으로써 모델의 NLU 상승

	precision	recall	F1
Our Electra-base	96.1	7.44	13.47
Our Albert-base	84.58	8.02	14.34
Our Albert-large	**95.13**	**23.47**	**37.60**

[표 6] Coreference Resolution 한국어 데이터 모델별 성능 결과표

	precision	recall	F1
Google BERT	74.6	71	72.3
Google BERT Implementation SpanBERT	75.0	71.9	73.9
Google BERT Implementation SpanBERT -1seq	75.3	73.5	74.4
SpanBERT-large	76.4	74.2	75.3
BERT-base-c2f-coref	-	-	73.9
BERT-large-c2f-coref	-	-	76.9
SpanBERT-base-c2f-coref	-	-	77.7
SpanBert-large-c2f-coref	-	-	79.6
Our BERT-base Model	**78.12**	**72.17**	**75.02**
Our BERT-large Model	**79.41**	**75.59**	**77.47**
Our SpanBert-base Model	**78.32**	**77.36**	**77.84**
Our SpanBert-large Model	**81.42**	**78.64**	**80.1**

[표 5] Coreference Resolution 영어 데이터 모델별 성능 결과표

https://github.com/mandarjoshi90/coref

Sample Result

- **실제 데이터 구축 예시**

1. confidence, average_confidence

주체, 대상, 관계 정보의 confidence score 합계와 평균 confidence score

2. Relation_start_pos, subject_start_pos, object_start_pos

주체 대상 관계 정보와 어절 단위 인덱스

3. subject_coreference, object_coreference

주체와 대상의 문서 내에서 발생한 상호참조해결 개체 제공

- Wikipedia 데이터를 통해 KBP 데이터 3,000,000건 구축


```
{
    "file_name": "history.txt",
    "sentence": "but he didn't answer Obama's call.",
    "confidence": "2.34911185503006",
    "average_confidence": 0.7830372850100199,
    "relation": "didn't answer",
    "subject": "he",
    "object": [
        "Obama's"
    ],
    "relation_start_pos": "2",
    "subject_start_pos": "1",
    "object_start_pos": [
        "4"
    ],
    "subject_corefernece": [
        "he",
        "Trump"
    ],
    "object_corefernece": [
        [
            "Obama",
            "Obama ' s"
        ]
    ]
}
```


KBP 데이터 추출 결과

Result

결론

- Open Domain 데이터에 대한 KBP 데이터 구축이 가능하다.

기존 KBP데이터 구축 방법인

1. NEL 문제 시 사전에 구축된 지식이 없는 경우

2. Relation Extraction 에서 사전에 정의된 관계가 없는 경우

* 위의 두 문제를 해결하는 방법을 제시하고 모든 도메인의 데이터에 대해 KBP 데이터를 구축하는 방법을 제시

* 정확한 KBP데이터를 구축하기 위해 Open IE 모델과 Coreference Resolution 모델을 제시

향후 연구

- 상호참조와 관계의 모호성 해결이 필요

Creference Resolution 모델에서 추출된 개체의 경우 어떤 개체가 대명사가 지칭하고있는 핵심어 인지를 파악이 필요

OpenIE 모델을 이용한 관계 추출 시 텍스트 내에서의 관계를 구체적으로 나타낼 수 있는 방법이 필요

- 여러 모델을 사용하는 것이 아닌 하나의 모델로 KBP데이터를 구축할 수 있는 방법 필요

여러 모델을 사용하여 KBP데이터를 구축하는 리소스의 부담이 있어 Multi task Learning과 같은 연구가 필요

감사합니다