기계 독해 기반
한국어 개체명 인식

민진우[1], 나승훈[2], 신종훈[3], 김영길[4], 김강일[5]
전북대학교[12], ETRI[34], GIST[5]

목차

- 서론
- 관련연구
- 기계독해 기반 한국어 개체명 인식
- 실험결과

개체명 인식

(Named Entity Recognition)

Location

Location

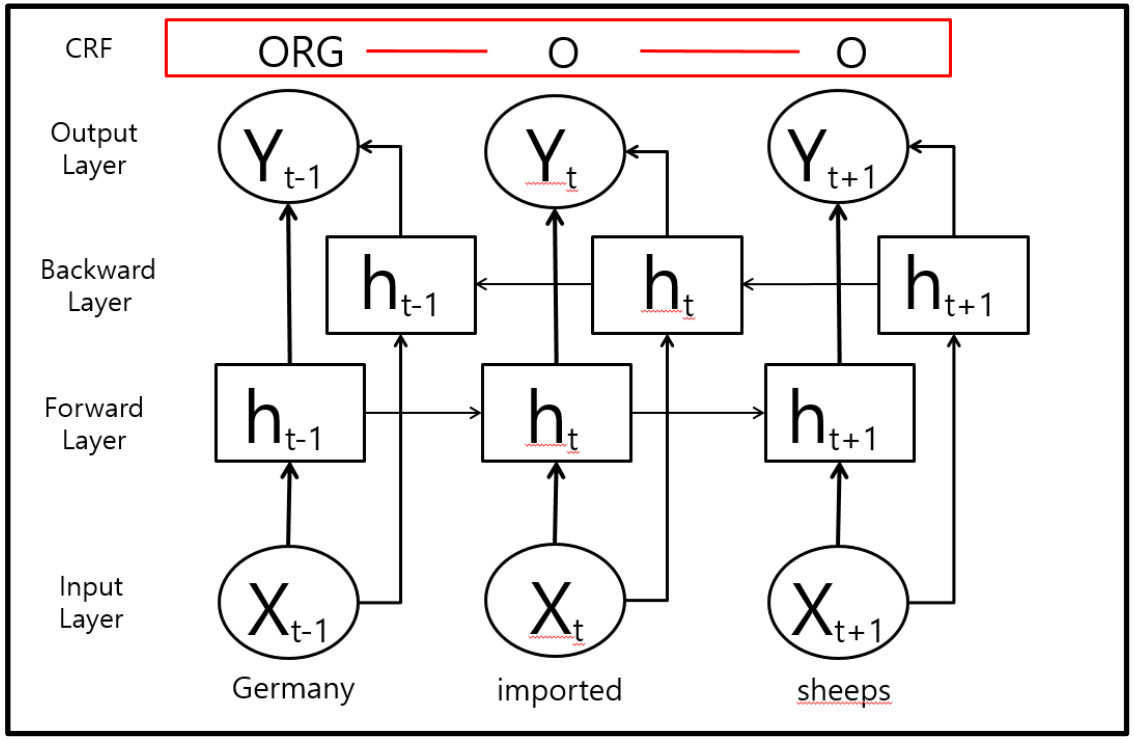
Vancouver is a coastal seaport city on the mainland of **British Columbia**.

The city's mayor is **Gregor Robertson**.

Person

- 주어진 문장 혹은 문서 내에서 인명, 지명, 기관명, 날짜, 시간 등과 같이 고유한 의미를 갖는 표현을 찾고 개체명의 종류를 분류하는 자연어 처리 분야

Bi-LSTM-CRF 기반 개체명 인식



- 딥러닝 기반 한국어 개체명 인식은 순차 태깅 문제로 보고 Bi-LSTM CRF를 이용한 연구가 주를 이룸
- 양방향 LSTM으로 얻어진 점수에 출력 태그간의 의존성을 모델링하는 CRF로부터 태그 전이 점수를 더하여 최종 태그 점수를 얻는 방식

기계 독해

(Machine Reading Comprehension)

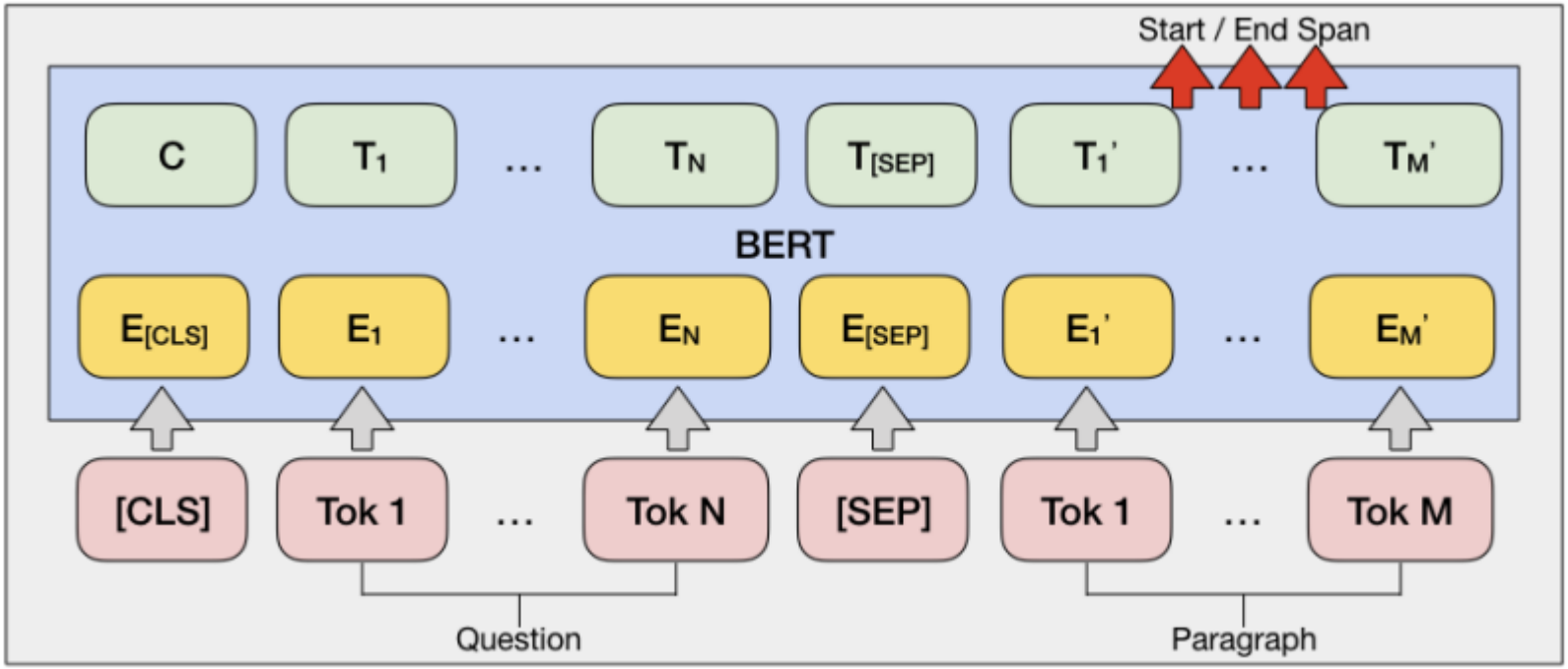
Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

- 시스템이 지문에 대한 내용을 이해하고 지문에 포함된 질문이 주어졌을 때 해당 질문에 대한 답변이 될 수 있는 정보를 지문에서 찾는 과정

Bert with MRC



- BERT의 입력으로 Question과 paragraph를 하나의 sequence로
  - BERT 내부의 transformer에서 query와 paragraph사이의 attention 및 self attention 이루어짐
  - 출력층에서 pointer network만 추가된 형태
  - 기존 모델보다 간략화된 형태이지만 높은 성능

기계독해 기반 한국어 개체명 인식

- 2개의 모듈로 구성 => 두 모듈은 joint 학습이 아닌 별도 학습

<Span 생성 모듈>

(Span의 시작, Span의 끝)으로 구성되는 후보 개체 span을 생성

<Span 타입 결정 모듈>

MRC 프레임워크를 사용하여 적절한 후보 개체의 타입을 결정

Span 생성 모듈

- 후보 개체를 찾는 과정을 해당 단어 i 를 시작으로 하여 Span 범위의 끝을 찾는 텍스트 범위 추출 문제로 변환
  - Span의 끝 위치에 대한 점수 $score_{end}(j|S_i)$를 Biaffine 함수를 통해 계산하고 손실 함수를 정의

$$score_{end}(j|S_i) = X_i^T U_{end} X_j + w_{end}^T X_j$$

$$L_{span}^{end} = -\sum_{i=1}^{n} \frac{exp(score_{end}(S_i.end|S_i))}{\sum_{j=1}^{n} exp(score_{end}(j|S_i))}$$

  - i 번째 단어를 시작으로 하는 S_i가 존재하지 않으면 S_i.end는 문장의 0번째 인덱스 즉,루트이다.
  - 디코딩 과정에서 최대한 많은 개체 후보를 포함하기 위해 반대로 i 번째 단어를 끝으로 하는 Span E_i의 시작 위치에 대한 점수 및 손실함수를 동일한 방법으로 정의

$$score_{start}(j|E_i) = X_i^T U_{start} X_j + w_{start}^T X_j$$

$$L_{span}^{start} = -\sum_{i=1}^{n} \frac{exp(score_{start}(E_i.start|E_i))}{\sum_{j=1}^{n} exp(score_{start}(j|E_i))}$$

Span 타입 결정 모듈

- 개체에 대한 타입을 계산하기 위해 기계 독해 프레임워크를 도입
  - MRC framework 사용 : [문맥, 쿼리(q), 정답(a)]의 형태
    - 문맥 X는 원문이고 쿼리(q)는 후보 개체 span이며 이 두 가지가 주어졌을 때 개체 타입을 추출하는 것을 목표.
    - 쿼리 형태

$$l_1, \cdots, l_L [CLS], X_1, X_2, \cdots, X_{n-1}, X_n, [SEP],$$

$$X_{S_j.start}, X_{S_j.start+1}, \cdots, X_{S_j.end-1}, X_{S_j.end}, [SEP]$$

  - 기계독해 기반의 Context 인코딩
    - 특수 토큰을 사이에 두고 문장에 대한 전체 토큰 시퀀스에 j번째 단어를 시작으로 하는 Span의 토큰들과 합하여 BERT를 통해 인코딩
    - 인코딩된 표상에 L개의 개체명 타입 임베딩과 결합한 다음 추가적인 트랜스포머 레이어에서 인코딩하여 얻어진 개체명 타입 임베딩은 컨텍스트와 Span의 정보가 반영

Span 타입 결정 모듈

- 개체에 대한 타입을 계산하기 위해 기계 독해 프레임워크를 도입
  - 개체 타입의 결정
    - 아래 수식과 같이 포인터 네트워크를 통해 적절한 타입을 결정

$$score_{label}(l|S_i) = \frac{exp(h_{label}^{\top} \times h_l)}{\sum_{l' \in L} exp(h_{label}^{\top} \times h_{l'})}$$

    - Span S_i의 l번째 레이블에 대한 점수를 구하는 수식으로 여기서 h_l은 l 번째 레이블에 대한 개체 타입 임베딩. 만약, 후보 개체가 실제 개체가 아니면 "O"를 레이블의 위치를 포인팅 하고 그렇지 않으면 적절한 타입 임베딩의 위치를 포인팅하여 타입을 결정
  - 학습 및 디코딩
    - 학습 과정에서는 최대 Span 길이를 설정하여 문장 내의 모든 가능한 Span을 쿼리로 하여 레이블을 결정하는 방식으로 학습
    - 디코딩 과정에서는 각 단어를 시작으로 하여 Top K개의 후보 Span을 추출 후 동일하게 각 단어를 끝으로 하여 K개의 후보 Span을 추출하여 총 2n * K개의 후보 Span를 추출
    - 총 2n * K개의 Span 중 포인팅된 위치가 0(root)가 아닌 후보 개체들만을 이용하여 Span 타입 결정 모듈에서 개체명 타입을 결정

실험 결과

- 데이터 셋
  - CRF를 이용한 자동 형태소 분석 결과(F1: 97.60%) 이용

	Train	Dev	Test
ETRI 개체명 인식 데이터 셋	4250	250	500

- 실험 결과
  - 보듯이 본 논문에서 제안한 기계 독해 기반 개체명 인식 모델이 BERT 기반 LSTM-CRF 대비 F1 : 1.08%p 높은 성능을 보이나 같은 언어 모델인 RoBERTa 기반 모델 대비 2.13%p 성능하락

	F1
문자 LSTM-CRF [KCC '16]	86.53%
문자 LSTM-CRF + 사전자질 [HCLT '16]	89.34%
(BERT) LSTM-CRF [KCC '19]	91.58%
(RoBERTa) LSTM-CRF [KSC '19]	**94.79%**
(RoBERTa) 기계독해 기반 개체명 인식	92.66%

결론

- 결론
  - Span 생성 모듈에서 후보 개체 표현에 대한 Span을 생성 한 후 span 타입 결정 모듈에서 MRC 프레임워크를 사용하여 적절한 후보 개체의 타입을 찾는 기계 독해 기반 개체명 인식 모델을 제안 후 적용하여 실험 결과를 얻음
  - 현재 기계 독해 기반 개체명 인식 모델은 추가적인 자질 없이 BERT 인코딩 표상만을 활용. 추가적인 자질을 활용한 동일 자질 성능 비교 필요

- 향후 연구
  - 기계 독해 프레임워크를 개체명 연결, 의미역 결정 등 다양한 자연어 처리 태스크에 적용할 예정

Q&A

감사합니다.