위키피디아 링크 데이터를 이용한 한국어 뉴럴 멘션 탐지 및 개체명 연결

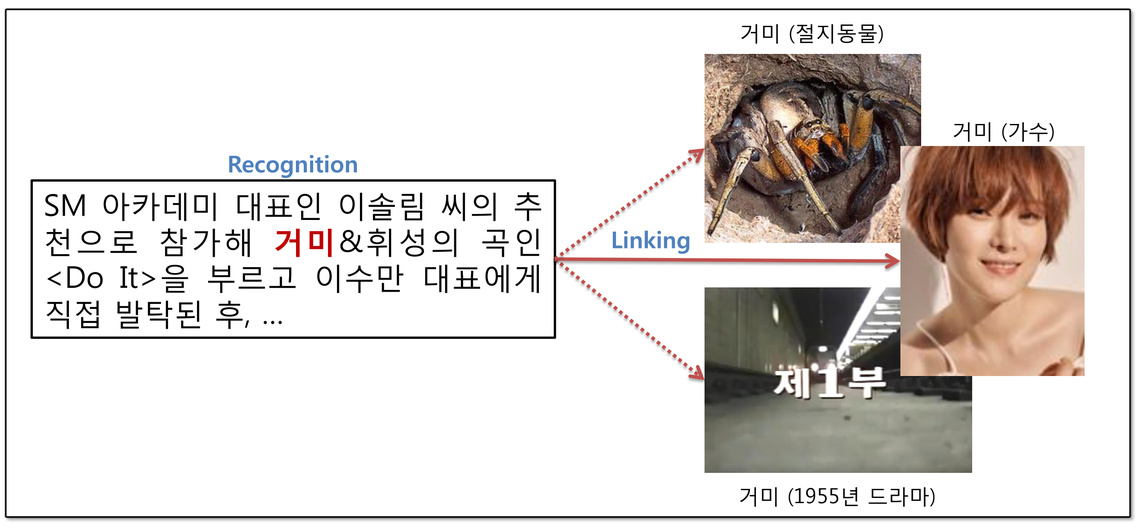
Cognitive Computing Lab.

이영훈, 나승훈
전북대학교
dldudgns73@jbnu.ac.kr, nash@jbnu.ac.kr

Ⅰ. 서론

개체명(Named Entity) 문장에서 고유한 의미를 가지는 단어 또는 구절을 의미. 최근 이러한 개체명을 주어진 문장에서 추출하기 위한 개체명 인식(Named Entity Recognition)과 위키피디아와 같은 지식기반(Knowledge base) 상의 하나의 의미와 연결하여 특정 개체와 무엇인지 식별하는 개체명 연결(Named Entity Linking) 연구가 활발히 진행 중임.

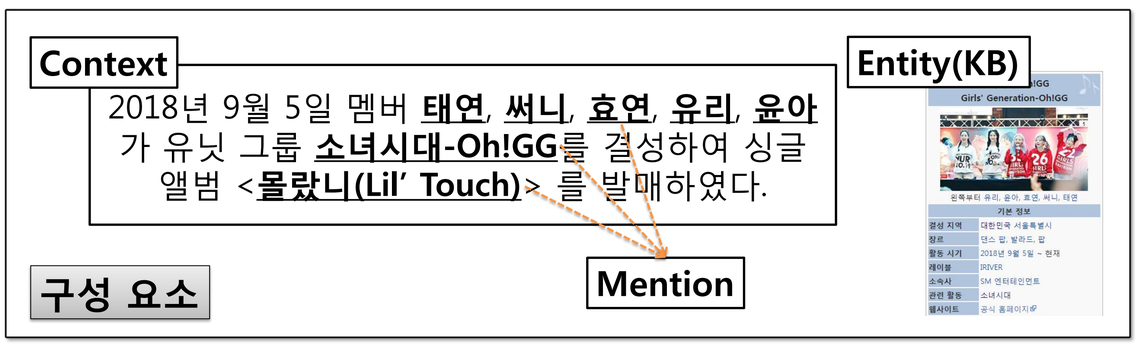
개체명 인식과 개체명 연결은 지식기반의 확장이나, 정보 검색 시스템, 질의응답 시스템과 같이 여러 자연언어 처리 태스크에서 자질 등으로 활용되는 등의 중요한 요소임.



Ⅱ. 제안 방법

Dataset

실험에 사용된 데이터는 한국어 위키피디아의 문서에서 개체가 링크(Hyper Link)되어 있는 데이터를 실험 데이터로 사용하였음.



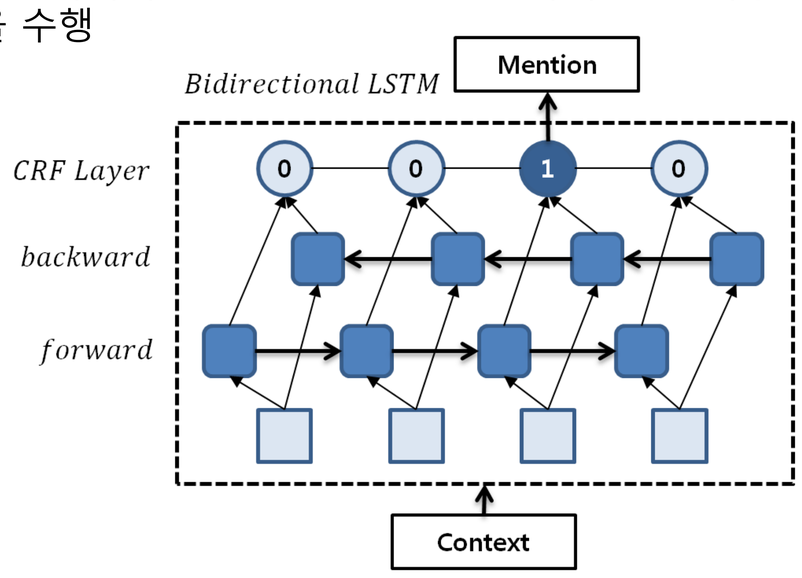
Proposed System

Bi-LSTM CRF을 이용한 개체명 연결 모델과 문맥-언급-개체 모델을 이용한 개체명 인식 모델을 이용하여 Pipeline-System을 구축

개체명 인식 모델에서 예측한 Span을 활용하여 개체 연결의 대상으로 인식하고, 인식한 Mention에 대해 개체명 연결 모델에서 각 후보 개체들의 점수를 구하게 되며, 가장 높은 점수의 개체를 선택하여 개체명 연결하여 개체 모호성 해결

ⅰ.개체명 인식 모델 (NER)

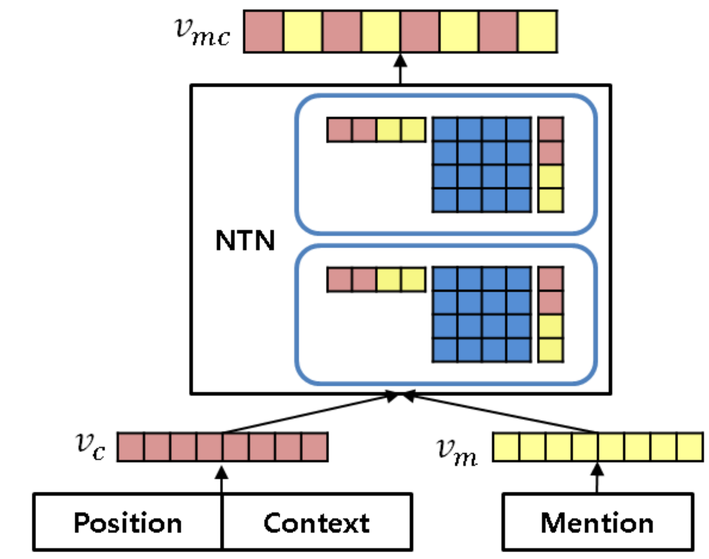
개체명 인식에서 우수한 성능을 보여주는 Bi-LSTM-CRF 모델을 이용, 문자열로부터 Bi-LSTM을 적용하여 LSTM 기반 표상을 얻고, 태그 간의 전이 가중치 고려하여 순차 입력 열에 개체명 타입을 태깅. 본 연구에서는 개체의 Mention을 Detection하여 연결 대상으로 인식하는 역할을 수행



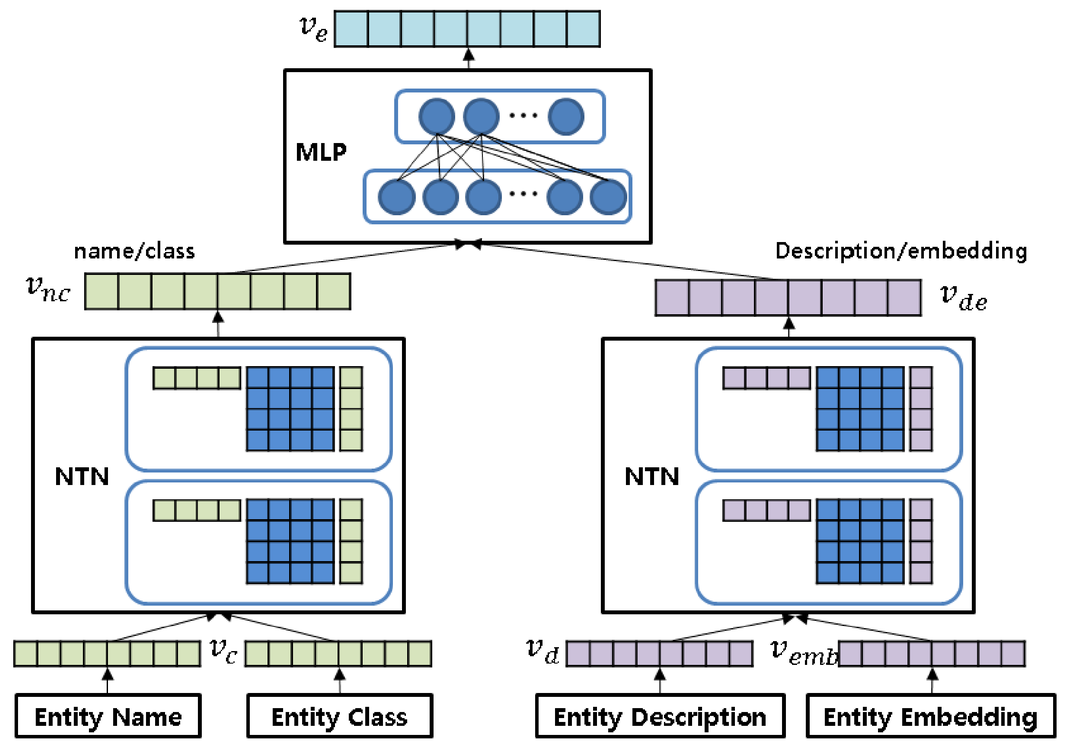
ⅱ.개체명 연결 모델 (NEL)

Mention/Context와 Entity 사이의 유사도를 이용하여 개체 모호성을 랭킹 태스크로 변환하여 문제 해결

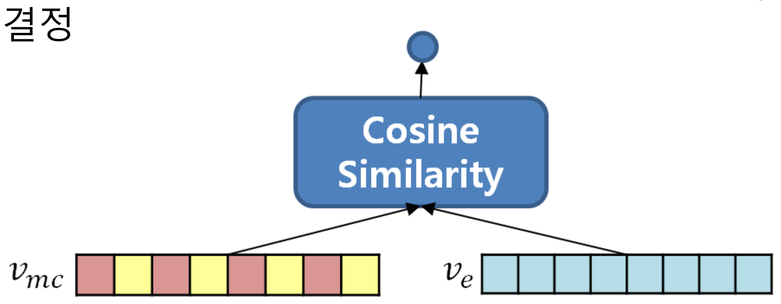
- **Mention/Context Representation** 형태소 분석을 거친 순차 입력 열에 CNN과 Max-Pooling을 이용하여 각 문장의 표현을 얻어내고, NTN(Neural Tensor Network)를 통하여 두 표현을 합친 Mention/Context Vector를 얻음.



- **Entity Representation** 개체의 이름과 분류표현, 지식기반 상의 개체의 설명, 개체의 Embedding을 각각 NTN과 MLP(Multi-layer Perceptron)를 이용하여 하나의 Vector로 표현



- **Scoring** Mention/Context Pair와 후보 Entity 간의 Cosine 유사도를 통하여 점수를 부여하고, 가장 높은 점수의 후보 Entity를 최종 연결 개체로 결정



Ⅲ. 실험 결과

- **실험 평가** Wikipedia 덤프 파일을 이용하여 동일 Mention에 대해 다른 개체를 가르키는 후보들을 추출 주어진 문장에 대해 NER 모델에서 예측한 Span이 정답 Span과 일치하고, NEL의 정답 후보 중 가장 높은 점수를 가지는 후보가 정답과 일치할 경우, 정답으로 가정하고 F1-Score를 측정하였음.

모델	Precision	Recall	F1
NER	84.43	84.53	84.48
NER-NEL Pipeline	79.41	66.79	72.56