Natural language processing Bible

8

서수원

Business Intelligence Lab.

산업경영공학과, 명지대학교

01

개체명 인식

개체명 인식 소개

- 개체명 인식(NER)
  - 질의답변, 정보검색, 관계추출등을 위한 nlp시스템의 핵심 구성 요소이다.
  - NER은 사람(Person, PS),장소(Location,LC),기관(Organization, OG),날짜(Date,DT) 이외에도 분야에 따라 약물, 임상 절차, 생물학적 단백질 등 과 같은 명명된 개체(단어)를 식별하는 작업을 말한다.

[문장 1]
춘향아, 8월 15일에 강남에서 홍길동과 약속이 있으니까, 늦지 말고 오도록 해!

[사람] : 춘향, 홍길동
[날짜] : 8월 15일
[장소] : 강남

춘향[PS]아, **8월 15일[DT]**에 **강남[LC]**에서 **홍길동[PS]**과 약속이 있으니까, 늦지 말고 오도록 해!

개체명 인식의 필요성

- 개체명 인식(NER)
  - 기계번역의 품질을 높이며, 사용자에게 맞춤형 번역을 제공할 수 있도록 도와준다
  - ‘TWIGFARM’이란 글자를 그대로 해석하면 ‘트위그팜‘이 아닌 ‘나뭇가지 농장‘ 이라고 해석이 된다. 이러한 해석은 번역오류로 이루어지기 때문에, TWIGFARM을 회사명으로 제대로 인식할 수 있다면, 번역 품질 뿐만 아니라 사용자 경험 까지도 향상 될 수 있다.
  - 2003년에 나온 논문에 따르면 “개체명이 일반적인 명사로 잘못 해석되면 문장의 이해 자체가 어려워지고, 이를 수정하기 위해 많은 비용이 소요된다.”고 한다.

개체명 인식 시스템

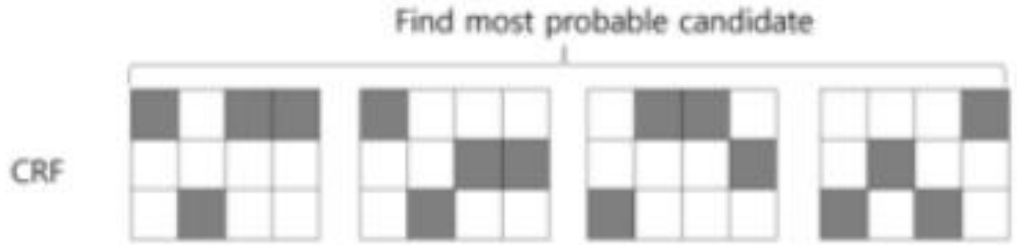
- 개체명 인식 시스템(지도학습 기반)
  - 은닉 마르코프 모델(HMM)
  - 서포트 벡터 머신(SVM)
  - 조건부 무작위(CRF)

- Zhou[12] : MUC-6, MUC-7 데이터를 활용한 HMM 기반 NER을 사용하여 각각 96.6%, 94.1%의 f-score를 달성
- Malouf[13] : HMM을 최대 엔트로피(Maximum Entropy, ME)와 비교, CoNLL2002의 스페인어 및 네덜란드어 데이터셋에서 각각 77.66%, 68.08의 f-score를 달성
- Ando[14] : 간단한 이진 관계 표현을 위해 대문자, 트리거 단어, 이전 태그 예측, 단어 모음, 사전 등과 같은 자질을 사용하여 CoNLL 2002 스페인 및 네덜란드 데이터셋에서 81.39%, 77.05%의 f-score를 달성

CRF

- CRF(conditional random field)
  - Sequential labeling 을 위한 potential function을 이용하는 softmax regression을 의미한다. RNN등장 전 Sequential labeling 에 있어 좋은 성능을 보인다.
    - ✓ 데이터의 형식이 벡터가 아니고 sequence인 data에 대한 classification 이라는 의미로 sequential labeling이라고 부른다.
      - ❖ 대표적인 sequential labeling으론 띄어쓰기 문제나 품사 판별이 있다.

- $x =$ 이것은예문입니다.
- $y = [0, 0, 1, 0, 1, 0, 0, 1]$.



- CRF는 가능성이 있는 sequence y 후보를 몇 개 선택한 뒤, 가장 적합한 하나의 label을 고르는 방식

지식 기반 시스템

- 지식 기반 시스템
  - 지식 기반 NER시스템은 어휘 자원 및 도메인 별 지식에 의존 하므로, 주석이 달린 학습 데이터가 필요하지 않다.
    - ✓ 사전 정보가 철저할 때에만 효과적이다.
    - ✓ 도메인 및 언어 별 규칙과 사전의 불완전성으로 인해 recall값은 낮다.
    - ✓ 지식기반 NER 시스템은 지식 자원을 구성하고 유지하기 위한 도메인 전문가가 필요하다.

- Collins과 Singer는 엔티티의 문맥, 개체명이 포함된 단어 등을 포함한 7가지 기능만을 사용하여 종자를 분류 및 추출[9]
- Etzioni는 8개의 제네릭 패턴 추출기를 적용하여 웹 텍스트를 여는 NER 시스템의 리콜을 개선하기 위해 비지도 학습 시스템을 제안[10]
- Nadeau는 앞선 두 연구에 기초하여 추출된 지명사전과 일반적으로 사용 가능한 지명사전을 결합하여 MUC-7에서 지명, 사람, 조직 엔티티에 대해 각각 88%, 61%, 59%의 성능을 달성[11]

NER 평가 척도

- NER 평가 척도
  - Confusion Matrix
    - ✓ 재현율, 정밀도 f1-score를 이용한다.

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

BIO tagging scheme

- BIO tagging scheme
  - 개체명을 텍스트로부터 인식 시키기 위한 기법 중 하나로, 정보추출 작업에서 자주 이용되는 기법이다.
  - B : Begin의 약자로 개체명 중 시작을 나타내는 단어에 태그한다.
  - I : Inside의 약자로 B 혹은 I 개체명 뒤에 오는 단어를 태그한다.
  - O : Outside의 약자로 개체명이 아닌 나머지 단어에 대해 태그한다.

. New York의 경우 (New, B-LOC), (York, I-LOC)로 태그 할 수 있음

학습 코퍼스

- 학습 코퍼스
  - CoNLL2002, CoNLL2003, CHEMDNER,Twitter가 대표적이다.