상대적 가중치 자질을 반영한 CRF 기반의 개체명 인식

정진욱
nlfactory@naver.com

Named Entity Recognition based on CRF reflecting relative weight

Jin-Wook Jeong

요 약

본 논문은 개체명 인식을 위해 CRF 모델을 이용해 분류를 수행했다. 개체명 후보를 개체명으로 식별에서 중의성 문제가 필요하다. 본 논문에서는 이러한 중의성 문제 해결을 위해 학습 셋으로부터 패턴과 형태적 특성을 고려해 개체명 후보를 최대로 선택하고 선택된 개체명 후보의 중의성과 정확도를 높이기 위해 주변의 문맥 자질과 분별 확률 모델인 CRF를 이용해 중의성 문제를 해결한다.

1. 서 론

개체명 인식(NER: Named Entity Recognition)은 입력된 자연어 문장에 나타난 특정 어휘가 사람, 장소, 기관, 날짜, 시간과 같은 미리 정의된 개체명 중 어떤 개체명인지를 식별해 태깅(tagging)하는 작업이다. 개체명 인식은 정보 추출에 있어서 하위 분야에 해당하며 최근 많은 관심을 받는 질의응답 시스템의 필수 기술이다.

잘 알려진 바와 같이 개체명에 해당하는 어휘는 중의성이 존재한다. 개체명 인식에서 개체명 사전을 이용할 수 있지만, 중의성 문제로 제대로 된 성능을 발휘할 수 없다. 본 논문은 개체명 인식 률을 높이기 위해 개체명 후보 선택의 커버리지를 넓게해 인식 가능성을 높이고, 개체명 후보 주변에 나타나는 문맥 자질(context feature)들과 확률 모델인 CRF를 이용해 개체명 인식을 수행한다.

2. 관련연구

개체명 성능을 높이기 위해 많은 방식이 도입되고 있다. 개체명 인식 초창기에는 개체명 사전이나 언어 문법(linguistic grammar) 기반으로 접근이 이뤄지다 현재는 지도 학습(supervised) 방식이나 빈지도 학습(semi-supervised) 학습 기반도 활용됐다. 최근에는 통계 모델(statistical model) 기반인 CRF와 LSTM 기반으로 개체명 인식 성능을 높이려는 연구가 있다.

이와 관련해 본 논문의 관련 연구로 최윤수 등[1]은 개체명 인식 자질 부족 문제를 활용하기 위해 워드 임베딩 자질로부터 추출한 벡터들에 대한 군집 정보를 CRFs의 워드 임베딩 자질로써 사용했다. 이 연구에서 사용한 자질의 종류로 형태소 자질, POS 태그, 형태소 길이, 어절의 위치, 개체명 사전에 존재 여부 값, 명사 유무 값을 활용했다.

박용민 등[2]이 있다. 이 연구는 도서, 영화, 노래, 음악, TV 프로그램에 대한 개체명 식별을 위한 데이터로 뉴스 기사를 활용했다. 이 연구는 주변 문맥 단어 및 거리를 이용하여 SVM을 활용해 제목 후보들을 추출하고 고유명사나 미등록어에 대한 사전을 구축하는 연구다.

Xuezhe 등[3]의 연구는 시퀀스 레이블 시스템을 구축하기 위한 지식을 이용하지 않고 뉴럴넷인 LSTM, CNN 그리고 CRF를 결합한 모델을 이용해 WSJ 코퍼스에 대한 개체명 인식을 수행했다.

본 논문에서는 개체명 인식의 중의성 문제 해결을 위해 패턴 학습을 통해 개체명 후보를 선택하고 중의성을 해결 하기위해 개체명 주변의 자질을 상대적으로 반영한 CRF 모델을 통해 애노테이션을 수행한다.

3. 규칙을 활용한 개체명 후보 선택

본 논문에서는 개체명을 분류하기 위한 전 단계로 개체명 후보를 선택한다. 개체명 후보는 개체명 식별을 위한 전 단계에 수행되는데 개체명이 될 수 있는 최대 후보가 될 수 있도록 개체명 후보를 최대한 선택함으로써 정답 커버리지를 높인다.

개체명 후보를 선택하려는 방법은 크게 두 가지다. 첫 번째로 학습 셋으로부터 패턴을 학습한다. 이를 위해 학습 셋에 존재하는 애노테이션된 개체명을 수집한다. 패턴 생성의 예로 <나무병원:OG>이라는 복합명사에 해당하는 개체명이 존재하면 복합명사의 형태소를 분리해 *병원의 형태로 분리하는 방식으로 패턴들을 수집한다. 이렇게 수집한 패턴에 부합한 문자열을 개체명 후보로 선택한다.

개체명 후보로 고려되는 대상으로 조사와 어미라는 형태소 앞에 위치하는 문자열을 개체명 후보로 선택한다. 이때 추상 명사인 경우 개체명 후보로 제외한다.

학습한 패턴과 형태적 특성을 고려해 수집된 개체명 후보는 개체명 후보의 커버리지를 최대한 높일 수 있게 한다.

4. 중의성 해결

4-1 중의성 해결을 위한 상대적 가중치 자질 수집

개체로 식별하기 위해 개체명 후보 중 동음이의로서 중의성이 있는지를 판단할 필요가 있다. 특히 중의성은 단음절 혹은 2음절인 단어는 중의성이 존재할 가능성이 높다.

이 때문에 규칙 기반의 방식으로 수집된 개체명 후보에 대해 중의성 해결이 필요하다. 예를 들어 '태봉'이라는 개체명 후보가 존재할 때 실제 문장에서는 다음과 같이 등장할 수 있다.

- <태봉:LC>은 901년 궁예에 의해 건국되어
- <태봉:PS>이가 제 몫을 헤다 말고

태봉은 장소(LC)를 의미하기도 하지만 사람(PS)을 의미하기도 한다. 중의성 문제를 해결하기 위한 접근 방법으로 형태적 특성을 고려한다. 이를 위해 접미사에 대한 음소(ㅇㅐㄴㅡ ?ㄴ)에 해당한다.
개체명 후보 주변에 존재하는 음절을 고려한다. 이때 개체명 후보에 멀수록 페널티를 부여하기 위해 개체명 후보에 가까울수록 가중치를 높게 해 개체명 후보 주변에 나타나는 인접 자질의 중요도를 반영했다. 이때 개체명 후보 주변에 나타난 자질이더라도 여러 개체명에 동시에 자주 나타나는 경우 중의성 자질로 분류해 패널티를 부여하거나 제외했다.

4-2 CRF 모델 구축

중의성 해결을 위해 분별 확률 모델(Discriminative Regression Models) 중 하나인 CRF(Conditional Random Field)를 이용한다. CRF에서 레이블을 선택하는 식은 다음과 같다..

$$y^* = \mathrm{argmax}_y p(y|x)$$

연속된 문자열에 대한 레이블(label)을 결정을 위해 벡터 x에 대해 p(y|x) 확률을 최대화하는 레이블 y*을 선택되도록 한다.

5. 실험

대회에서 제공한 개체명 태깅 말뭉치를 활용해 실험했다. 사용된 개체명 종류는 다음과 같다.

표 1 실험에서 사용된 개체명과 예

개체명	축약어	태깅예
Person	PS	<김:PS>기자, <이순신:PS>
Location	LC	<한국:LC>, <63빌딩:LC>
Organization	OG	<정부:OG>, <청와대:OC>
Date	DT	<10월 1일:DT>, <지난 1일:DT>
Time	TI	<3시 30분:TI>, <3시간 전:TI>

애노테이션은 크게 두 가지 원칙을 따른다. 애노테이션 방식은 정보통신단체 표준의 개체명 태그 세트 및 태깅 말뭉치 표준 권고안에 따라 최소 태깅 원칙을 태깅 지침으로 따른다. 최소 태깅 원칙이란 한 문장에서 개체명이 연이어 나올 때 최소 단위로 태깅하는 것을 의미한다.

<3월 20일:DT>에서 <4월 20일>까지

최소 태깅 원칙의 예외로 고유명사에 포함된 개체명인 경우 고유 명사를 먼저 개체명 후보로 선정한다. 예를 들어 '대한민국 임시정부'라는 고유 명사가 있을 때 고유 명사 내에 장소(LC)인 대한민국이 있더라도 고유 명사를 먼저 고려해 <대한민국 임시정보:OG>이라고 인식된다.

하지만 복합명사이지만 해당 복합명사가 개체명이 아닌 경우 분리해 식별한다. 예를 들어 "한미 양국"이라면 "<한:LC><미:LC> 양국"이라고 식별한다. 만약 테스트 세트가 표준 권고안의 태깅 지침을 따르지 않을 경우 제안한 방법에 페널티가 발생해 성능 측정이 제대로 이뤄지지 않을 수 있다.
구축된 CRF 모델을 활용해 개체명에 속하지 않았다면 애노테이션을 하지 않는 방식으로 게 된다. 본 논문에서 제안한 방법으로 구축된 NER을 이용해 대회에서 제공한 학습데이터를 학습 셋과 테스트세트를 7:3으로 나눠 실험한 결과 약 Precision 73.77%의 성능을 확보했다. 실험의 제한사항으로 학습 셋은 정답 셋이 아니여서 태깅이 되어있지 않거나 일부 태깅 오류의 문제가 있었다. 학습 셋을 개선하고 학습 데이터를 늘리면 테스트 셋에 대한 성능을 보다 높일 수 있을 것으로 기대된다.

6. 결론

본 논문은 개체명 식별을 위해 패턴을 학습을 통해 개체명 후보를 선택했고 개체명 선택의 중의성 문제를 해결하기 위해 연속 문자열의 특성을 반영하기 위해 CRF 모델을 적용했다. 본 논문에서 제안하는 패턴 학습의 방법은 개체명 사전에 존재하지 않는 개체명 후보라 할지라도 개체명 식별이 가능하며, 동음이의어일 지라도 자질의 상대적 가중치를 적용해 중의성을 해결할 수 있다. 후속 연구로 축된 개체명 인식기는 질의응답 시스템에 활용할 수 있도록 개체명 인식의 범위를 확장해 적용할 예정이다.

참 고 문 헌

[1] Yunsu Choi, Jeongwon Cha, "Korean Named Entity Recognition and Classification using Word Embedding Features)", Journal of KIISE, Vol. 43, No. 6, pp. 678-685, 2016.
[2] Yongmin Park, Jae Sung Le, "Named Entity Recognition and Dictionary Construction for Korean Title: Books, Movies, Music and TV Programs", Vol.3, No.7 pp.285-292, 2014.
[3] Xuezhe Ma and Eduard Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, pp. 1064-1074, 2016.