ManiFL을 이용한 한국어 개체명 인식

김완수○, 신준철, 박서연, 옥철영
울산대학교 한국어처리연구실
kimwansu@outlook.com, ducksjc@gmail.com, seoyeon9695@gmail.com, okcy@ulsan.ac.kr

Korean Named Entity Recognition using ManiFL

Kim Wansu○, Shin Joon-choul, Park Seoyeon, Ock CheolYoung
Korean Language Processing Lab. University of Ulsan, Korea

요 약

개체명 인식은 주어진 문장 안의 고유한 의미가 있는 단어들을 인명, 지명, 단체명 등의 미리 정의된 개체의 범주로 분류하는 문제이다. 최근 연구에서는 딥 러닝, 대용량 언어 모델을 사용한 연구들이 활발하게 연구되어 높은 성능을 보이고 있다. 하지만 이러한 방법은 대용량 학습 말뭉치와 이를 처리할 수 있는 높은 연산 능력을 필요로 하며 모델의 실행 속도가 느려서 실용적으로 사용하기 어려운 문제가 있다. 본 논문에서는 얕은 기계 학습 기법을 적용한 ManiFL을 사용한 개체명 인식 시스템을 제안한다. 형태소의 음절, 품사 정보, 직전 형태소의 라벨만을 자질로 사용하여 실험하였다. 실험 결과 F1 score 기준 90.6%의 성능과 초당 974 문장을 처리하는 속도를 보였다.

주제어: 개체명 인식, 모두의 말뭉치, 얕은 학습, 혼합 자질 가변 표지기(ManiFL)

1. 서론

개체명 인식은 주어진 문장 안의 고유한 의미가 있는 단어들을 인명, 지명, 단체명 등의 미리 정의된 개체의 범주로 분류하는 문제이다. 초기의 기계 학습 기반 개체명 인식 모델에는 주변 단어와의 맥락을 고려할 수 있는 CRF(Conditional Random Fields)[1]가 주로 쓰여 왔다. 최근에는 사람이 직접 자질을 설계할 필요가 없는 딥 러닝 기반 모델인 bi-LSTM-CRF와 대용량 언어 모델인 BERT[2]를 미세 조정 하는 방법이 높은 성능을 보이고 있다.

하지만 딥 러닝 모델은 대용량 학습 말뭉치와 이를 처리할 수 있는 높은 연산 능력을 필요로 하고, 모델의 실행 속도가 느려서 실용적으로 사용하기 어려운 문제가 있다. 따라서 본 논문에서는 정확률과 처리 속도 양 쪽 모두를 고려하여 딥 러닝 모델보다 빠른 속도로 학습 및 실행할 수 있도록 얕은 학습 방식에서 적용한 혼합 자질 가변 표지기[8]을 사용한 개체명 인식 모델을 제안한다.

2. 관련 연구

개체명 인식 방법은 사전과 규칙에 기반한 전통적인 방법과 개체명 태그가 부착된 말뭉치를 학습하여 분석하는 방식으로 나눌 수 있다.

규칙, 사전 기반 개체명 인식 방법은 정의되어 있는 범위 안에서는 빠르고 정확한 분석이 가능하다는 장점이 있지만 개체명은 고유명사이거나 학습 데이터에 등장하지 않은 어휘(OOV)인 경우가 많으며 새롭게 만들어지거나 문맥에 따라 다른 의미로 사용되는 경우가 있어 지속적인 규칙과 사전 갱신을 필요로 하여 개체명 인식에는 적용하기 어렵다.

학습 기반 방식은 개체명 태그가 부착된 학습 말뭉치에서 성분을 선택, 조합하여 개체명임을 잘 나타낼 수 있는 자질을 찾아서 적용하는 얕은 학습 방식[3]과 문장의 각 단어를 적절하게 표현할 수 있는 임베딩으로 변환해서 학습하면 신경망 내부에서 적절한 자질을 찾아내는 깊은 학습 방식[4, 5, 6, 7]이 있다.

최근에는 컴퓨터의 병렬 연산 성능 향상과 대량의 학습 데이터 수집이 용이해짐으로써 bi-LSTM-CRF와 같은 깊은 학습 기법[4], 트랜스포머 기반 언어 모델인 BERT 또는 ELECTRA를 사용한 방법[5, 6, 7]이 주로 연구되어 한국어에서는 사용한 말뭉치와 개체명 태그 집합에 따라 90~92%정도로 우수한 성능을 보이고 있지만 매우 많은 연산량으로 인해 모델 학습에 비용과 시간이 많이 소비된다는 단점이 있다. 본 논문에서는 학습 비용과 속도의 실용성을 중시하여 얕은 학습 방식의 자연어 기계 학습 라이브러리인 ManiFL을 사용한 개체명 인식 모델을 제안한다.

3. 제안 모델

3.1. ManiFL 설명

ManiFL[8]은 얕은 학습 방식의 자연어처리 도구로, 소프트맥스 분류기를 미니배치 경사 하강법으로 학습하는 모델로 아래와 같은 특징을 가지고 있다.

- 고정된 라벨 후보 집합을 가진 기존 기계 학습 도구의 다중 클래스 분류기와 다르게, 상황마다 각기 다른 정답 라벨 후보 집합을 가질 수 있다.
- 자질의 각 사례마다 개별 가중치를 가질 수 있다. 예를 들면, 앞 단어를 보는 자질이 있을 때 앞 단어가 A일 때의 가중치와 B일 때의 가중치를 별도로 학습할 수 있다.
- 지정한 수치 이하의 빈도인 자질의 값을 확률적으로 0을 출력하도록 만드는 드롭아웃을 적용하여 얕은 학습에서도 과적합을 방지할 수 있다.

표 1 개체명 인식 학습에 사용하는 자질

번호	자질 종류 설명	예문 '국가대표'의 자질
1	앞 라벨, 현재 형태소 모든 글자, 품사	OG-B 국가대표NNG
2	앞 라벨, 현재 형태소 처음과 마지막 글자, 품사	OG-B 국표NNG
3	앞 라벨, 현재 형태소 마지막 두 글자, 품사	OG-B 대표NNG
4	앞 라벨, 현재+다음 형태소 처음과 마지막 글자, 품사	OG-B 국표NNG 팀NNG
5	앞 라벨, 이전+현재 형태소 처음과 마지막 글자, 품사	OG-B 컬링NNG 국표NNG
6	앞 라벨, 현재+다음 형태소 마지막 두 글자, 품사	OG-B 대표NNG 팀NNG
7	앞 라벨, 이전+현재 형태소 마지막 두 글자, 품사	OG-B 컬링NNG 대표NNG
8	앞 라벨, 현재+다음 2개 형태소 처음과 마지막 글자, 품사	OG-B 국표NNG 팀NNG 이JKS
9	앞 라벨, 이전 2개+현재 형태소 처음과 마지막 글자, 품사	OG-B 대국NNP 컬링NNG 국표NNG
10	앞 라벨, 현재 형태소 첫 글자, 다음 2개 형태소 품사	OG-B 표NNG NNG JKS
11	앞 라벨, 이전, 현재 형태소 첫 글자, 다음 형태소 품사	OG-B NNG 표NNG NNG
12	앞 라벨, 이전, 현재 형태소 첫 글자, 다음 2개 형태소 품사	OG-B NNG 표NNG NNG JKS

3.2. 학습에 사용한 자질

본 논문에서는 ManiFL을 사용한 형태소 단위의 개체명 인식 모델을 제안한다. 위 <표 1>은 <예문>에서 '국가대표'의 개체명 학습 자질을 나타낸 것이다. 각 자질들은 현재 형태소와 이전 최대 2개, 다음 최대 2개 형태소에서 주로 첫 글자와 마지막 글자, 품사로 구성되어 있다. 그리고 CRF, HMM 등에서 사용하는 Viterbi 알고리즘으로 라벨열의 최적 해를 찾는 대신에 모든 자질에 직전 라벨을 포함시키는 방식을 사용하였다. 직전 라벨 자질은 학습 시에는 학습 말뭉치에 존재하는 정답 라벨을 직접 사용하였고, 평가 시에는 모델을 통해 인식된 직전 라벨을 사용하였다.

- 예문: 대한민국/NNP 컬링/NNG 국가대표/NNG+팀__01/NNG+이/JKS 은메달/NNG+을/JKO 따__01/VV+았/EP+다/EF+./SF

3.3. 학습 과정

<그림 1>은 본 논문에서 제안하는 개체명 인식 모델에서 개체명을 학습하는 과정이다. 가장 먼저 UTagger[9]로 형태소 분석이 되어 있는 개체명 학습 말뭉치에서 위 <표 1>에 제시된 것과 같이 자질을 추출한다. 다음으로 말뭉치 전체에서 각 자질의 라벨 별 출현 빈도를 측정한다. 예를 들면 말뭉치 전체에서 3번 자질이 'OG-B 대표NNG'의 형태로 추출된 횟수를 정답 라벨 별로 집계한다. 다음으로, 자질의 각 사례 별 가중치를 학습하는 단계로 구성되어 있다. 가중치 학습은 소프트맥스 함수와 교차 엔트로피 손실함수를 사용하여 수행한다. 예문의 '국가대표/NNG'의 정답 개체명 태그인 'OG-I'일 확률을 계산하는 수식 1은 아래와 같다. z는 각 자질의 가중치와 출현 횟수의 곱이고, n은 학습 말뭉치에서 실제로 출현한 개체명 태그의 종류 수이다.

$$P_{OG-I} = \frac{e^{z}_{OG-I}}{\sum_{i=1}^{n} e^{z}} \qquad \text{(수식 1)}$$

모든 형태소에 대해서 위와 같은 방법으로 각 개체명 태그의 출현 확률을 계산하고, 실제 정답 개체명 태그를 1, 그 외의 오답 개체명 태그를 0으로 하여 각 개체명 태그 출현 확률(0~1)과의 오차를 계산한 것을 미니배치 단위로 모은다. 다음으로, 미니배치 단위로 모은 오차의 평균과 학습률(learning rate)을 곱한 값을 가중치에 더하여 반영한다. 따라서 정답 개체명 태그를 잘 예측할 수 있는 자질의 가중치는 커지고, 계속 틀리는 자질의 가중치는 작아진다.

또한 이 수식에 따라 출현 횟수가 매우 적은데 정답과 일치할 확률이 높은 자질의 가중치가 지나치게 커져 학습 말뭉치에 과적합되는 것을 방지하도록 빈도 기반 드롭아웃을 적용하였다. 지정한 확률로 드롭아웃을 시도하면 0과 빈도 제한 값 사이의 임의의 자연수가 자질의 빈도보다 큰 경우에는 드롭아웃을 수행하여 일시적으로 해당 자질의 빈도를 0으로 간주하여 학습을 진행한다. 따라서 빈도 제한값보다 적게 발생한 자질을 임의의 확률로 비활성화하여 학습 말뭉치에 과적합되는 것을 방지한다.

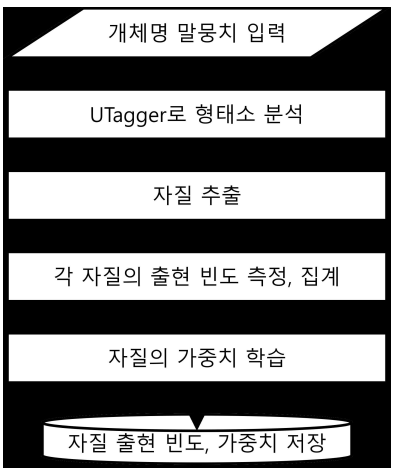

그림 1 개체명 말뭉치 학습 과정

위 학습 과정을 지정한 횟수동안 모델의 성능이 향상되지 않거나, 최대 반복 횟수에 도달할 때까지 반복 수행하여 가중치를 학습한다.

4. 실험 결과

본 논문에서 제안한 개체명 인식 모델의 학습과 성능 평가에는 모두의 말뭉치(개체명 분석 말뭉치 1.0)를 사용하였다. 모두의 말뭉치는 ETRI의 '세부분류 개체명 가이드라인 2018'을 기본으로 하여 국립국어원에서 수정한 지침에 준하여 15가지의 의미 분류 체계에 따른 태그를 부착한 말뭉치이다[10]. 본 논문에서는 그 중 PS(인명), LC(지명), OG(단체명), DT(날짜), TI(시간)의 5가지 범주만을 대상으로 한 실험과, 추가로 15가지 범주 전체를 대상으로 한 실험을 진행하였다. 여러 형태소에 걸쳐있는 개체명을 구분할 수 있도록 범주 뒤에 '-B', '-I'를 붙이고, 개체명이 아닌 형태소에는 'O'를 붙인다. 학습과 성능 평가에 사용한 말뭉치의 분량은 아래 <표 2>와 같다.

표 2 학습 및 평가 말뭉치의 분량

	문장	어절	형태소
학습 말뭉치	120,065	1,600,817	3,571,639
평가 말뭉치	30,017	399,396	891,578

모두의 말뭉치(개체명 분석 말뭉치 1.0)에서 학습용 말뭉치를 학습 할 때의 ManiFL 설정 값은 아래와 같다.

- 사용한 CPU 스레드 수: 12
- 최대 반복 횟수(epochs) : 20
- 미니배치 크기: 3571(형태소 수 / 1000)
- 학습률: 0.5
- 드롭아웃 시도 확률: 90%
- 드롭아웃 빈도 제한 값: 20

위 설정으로 학습 후, 평가용 말뭉치에 5가지 범주를 대상으로 개체명 인식을 한 결과는 아래 <표 3>와 같이 micro average F1 점수 기준으로 90.6%로 나왔다. 평가 기준으로 사용한 micro average F1 score는 라벨의 빈도가 불균형한 경우에 성능을 측정하는 지표이다. 각 라벨의 정밀도, 재현률을 계산할 때 사용한 TP, FP, FN의 수를 합계를 낸 것으로 전체 라벨의 정밀도, 재현률을 계산한 것이다.

표 3 5개 범주 대상 개체명 인식 실험 결과(단위: %)

	precision	recall	F1-score
DT	96.90	96.60	96.75
LC	81.40	87.75	84.46
OG	86.35	82.88	84.58
PS	91.99	93.22	92.60
TI	95.14	94.12	94.63
micro avg.	90.53	90.68	90.60

학습의 전체 과정은 21분 49초가 소요되었고, 그 중에서 학습 말뭉치 파싱 등의 과정을 제외하고 순수하게 가중치를 학습하는 데에만 소요된 시간은 11분 56초이다.

개체명 인식 평가에는 1개의 스레드로 전체 과정에 33.49초가 소요되었다.

추가로 [7]의 실험과 비교할 수 있도록 15가지 범주 모두에 대해서 개체명 인식을 한 결과는 아래 <표 4>와 같이 micro average F1 점수 기준으로 90.45%로 나왔다.

표 4 모든 범주 대상 개체명 인식 실험 결과(단위: %)

	precision	recall	F1-score
AF	83.94	70.24	76.48
AM	88.34	80.73	84.36
CV	90.14	87.16	88.62
DT	97.01	97.12	97.07
EV	89.14	80.54	84.62
FD	85.23	62.79	72.31
LC	81.88	89.54	85.54
MT	83.25	70.25	76.2
OG	85.88	85.06	85.47
PS	92.26	93.87	93.06
PT	77.14	47.16	58.54
QT	96.39	97.67	97.02
TI	96.3	95.73	96.01
TM	85.91	73.36	79.14
TR	82.77	67.27	74.22
micro avg.	91.20	89.71	90.45

본 논문과 가장 유사한 실험 환경인 박서연(2021)[3]의 모두의 말뭉치와 CRFsuite를 사용하여 F1 점수로 90.54%를 얻은 연구와 비슷한 성능을 보였지만, [3]에서는 형태소의 글자, 품사 태그 이외에 어휘 의미망의 상위어, 의존 관계와 같은 다양한 자질이 추가로 존재한다. 본 논문에서는 이러한 추가 자질 없이도 조금 높은 90.6%의 성능이 나온 원인에는 말뭉치 오류 일부 개선, 기본 자질 구성의 차이, 단순 오차 범위 등 다양한 원인이 있을 수 있지만 가장 큰 차이로는 ManiFL과 CRFsuite의 차이가 있다. 불리한 점으로는 ManiFL에는 viterbi 알고리즘으로 최적의 라벨 조합을 찾는 기능이 없어서 직전 라벨을 자질에 넣는 것으로 대신한 것이고, 유리한 점으로는 자질의 개별 사례마다 별도의 가중치를 학습할 수 있는 것과 과적합을 방지할 수 있는 드롭아웃 기능이 있다.

또한 대용량 언어 모델인 ELECTRA에 모두의 말뭉치를 학습에 사용한 [7]의 실험과 비교하면 국립국어원 문어체 개체명 분석 말뭉치만을 사용한 결과인 F1-score 87.63%보다 앞서는 것은 물론, 위키피디아 등에서 수집한 문장을 추가로 학습하여 실험한 결과인 90.71%와 비교할만한 90.45%의 성능을 보였다. [7]의 실험에서 추가 말뭉치 학습으로 크게 성능 향상을 보인 날짜(DT), 인물(PS), 식물(PT), 수량(QT), 시간(TI) 중에서 식물(PT)를 제외한 나머지 4개 범주에서 비슷한 말뭉치를 사용하였을 때는 성능이 앞서고, [7]의 실험에서만 위키피디아 말뭉치를 추가로 학습하였을 때에도 비슷한 성능을 보여 본

논문에서 제안하는 방법이 개체명 인식을 효율적으로 학습할 수 있다는 것을 보였다.

이번 실험 결과에서 ManiFL을 사용해서 CRF보다 더 적은 종류의 자질을 사용하면서도 CRF 기반 모델은 물론 ELECTRA 등의 대용량 언어 모델을 사용하는 모델과 비슷한 성능을 내는 모델을 빠르게 학습할 수 있어 성능과 속도를 겸비한 실용적인 개체명 인식 시스템이라고 할 수 있을 것이다.

5. 결론

본 논문은 얕은 학습 기반의 자연어 처리 도구인 ManiFL을 사용하여 한국어 개체명 인식 모델을 제안하였다. 형태소의 음절, 품사, 직전 형태소의 라벨 자질을 사용하여 깊은 학습 모델을 사용한 것과 비교할 수 있는 성능이 나옴을 실험을 통해 확인하였다. 기존의 얕은 학습 기반 모델과 비교했을 때의 장점으로는 깊은 학습 기반 모델에서 사용하는 드롭아웃을 얕은 학습 모델에서 유사하게 구현할 수 있어 훨씬 적은 종류의 자질로 동등한 성능을 낼 수 있음을 확인하였다.

속도 면에서는 깊은 학습 모델에 기반한 방법보다 빠른 것은 물론, 기존의 얕은 학습 기반의 자연어 처리 도구에서 미비한 점인 학습 과정을 병렬로 처리할 수 있어 학습 속도가 빠르다는 장점이 있다.

실험 결과 국립국어원 모두의 말뭉치(개체명 분석 말뭉치 1.0)에 UTagger로 형태소 분석을 거친 말뭉치 약 12만 문장을 11분 56초에 가중치 학습을 수행할 수 있는 것을 확인하였고, 평가용 말뭉치 약 3만 문장을 33.49초에 개체명 인식을 하고 micro average F1 점수로 90.6%의 성능을 보여 얕은 학습 모델의 빠른 속도와 깊은 학습 모델과 비교할 수 있을 만한 성능을 겸비한 실용적인 개체명 인식 시스템을 구축할 수 있음을 보였다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2013-2-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술개발)과 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2020R1I1A1A01073665)의 연구 결과임.

참고문헌

[1] John Lafferty and Andrew McCallum and Fernando Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, Proceedings of the International Conference on Machine Learning (ICML-2001), 2001.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[3] 박서연, 옥철영, 한국어 어휘 의미망을 활용한 CRF 모델 기반 개체명 인식. 정보과학회논문지, Vol. 48, No. 5,, pp. 556-567, 2021. 5.

[4] 박동주, 안창욱, “개체명 비율 사전을 결합한 Bidirectional LSTM-CRF 기반 개체명 인식”, 한국정보과학회 학술발표논문집, pp. 721-723, 2019.6.

[5] SK텔레콤, “KoBERT와 CRF로 만든 한국어 개체명인식기”, https://github.com/eagle705/pytorch-bert-crf-ner

[6] 박광현, 나승훈, 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정", 한국정보과학회 학술발표논문집, pp. 584~586, 2019. 6.

[7] 조우진, 신규표, 이원재, 손승현, 송현우, 이지훈, 이혁준, 조세현, “한국어 형태소 분석기를 이용한 KoELECTRA 기반의 개체명 인식 기법”, 한국정보과학회 학술발표논문집, pp. 1897~1899, 2021. 6.

[8] 신준철, 김완수, 옥철영, “ManiFL : 얕은 학습 기반의 더 나은 자연어처리 도구”, 제 33회 한글 및 한국어 정보처리 학술대회, 2021.(제출)

[9] Joon-Choul Shin, C. Y. Ock, "Korean Homograph Tagging model based on Sub-Word Conditional Probability", Journal of KIPS : Software and Data Enginering, Vol. 3, No. 10, pp. 407-420, Oct. 2014. (in Korean)

[10] “국립국어원 개체명 분석 말뭉치 2020(버전 1.0)” . URL: https://corpus.korean.go.kr/.