라벨 정보를 이용한 Few-shot Learning 환경에 강건한 중첩 개체명 인식 모델

황현선[O,†], 이창기[†], 고우영[††], 강명철[††]

강원대학교[†], 국가보안기술연구소[††]

hhs4322@gmail.com, leeck@kangwon.ac.kr, {gwy876, mckang}@nsr.re.kr

A Nested Named Entity Recognition Model Robust in Few-shot Learning Environments using Label Information

Hyunsun Hwang[O,†], Changki Lee[†], Wooyoung Go[††], Myungchul Kang[††]

Kangwon National University[†], National Security Research Institute[††]

요 약

중첩 개체명 인식(Nested Named Entity Recognition)은 하나의 개체명 표현 안에 다른 개체명 표현이 들어있는 중첩 구조의 개체명을 인식하는 작업으로, 중첩 개체명 인식을 위한 학습데이터 구축 작업은 일반 개체명 인식 학습데이터 구축보다 어렵다는 문제가 있다. 본 논문에서는 이러한 문제를 해결하기 위해 Few-shot Learning 환경에 강건한 중첩 개체명 인식 모델을 제안한다. 이를 위해, 기존의 Biaffine 중첩 개체명 인식 모델의 출력 레이어를 라벨 의미 정보를 활용하도록 변경하여 학습데이터가 적은 환경에서 중첩 개체명 인식의 성능을 향상시키도록 하였다. 실험 결과 GENIA 중첩 개체명 인식 데이터의 5-shot, 10-shot, 20-shot 환경에서 기존의 Biaffine 모델보다 평균 10%p이상의 높은 F1-measure 성능을 보였다.

주제어: 중첩 개체명 인식, 정보추출, Label Embedding, Label Description, Few-shot learning

1. 서론

개체명 인식(Named Entity Recognition)은 정보추출의 기본이 되는 단계로 비정형의 텍스트에서 인물, 지명, 조직명 등 고유명사를 주축으로 미리 정의된 특정한 언어 표현을 인식하고 분류하는 작업으로, 개체명의 범위를 찾고 해당 범위가 어떤 개체명에 속하는가를 분류한다[1]. 이러한 개체명 표현은 중첩 구조로 이루어진 경우가 많으나 기존 방식의 개체명 인식 기술은 기술적인 이유로 단일 개체명 인식(Flat Named Entity Recognition) 기술이 주로 연구되어 정보추출이라는 관점에서 제한된 정보를 추출하게 된다는 단점이 있다[2].

중첩 개체명 인식(Nested Named Entity Recognition)은 기존의 단일 개체명 인식을 극복하기 위해 하나의 개체명 표현 안에 다른 개체명 표현이 들어있는 중첩 구조를 분석하는 방식의 접근법이다[2,3]. 그러나 개체명 인식 기술은 주로 단일 개체명 인식이 연구되어 중첩 개체명 인식을 위한 학습데이터는 거의 구축되어 있지 않다는 문제가 있다.

본 논문에서는 이러한 문제를 해결하기 위해 Few-shot Learning 환경에 강건한 중첩 개체명 인식 모델인 LDE(Label Description Embedding) 모델을 제안한다. LDE 모델은 기존 Biaffine 중첩 개체명 인식 모델의 출력 레이어를 라벨의 설명문을 인코딩하여 라벨의 의미 정보를 사용할 수 있는 Label Attention 레이어로 수정한 모델이다.

2. 관련 연구

개체명 인식은 주로 BIO 태그를 부착하는 순차적 레이블링 문제로 간주되어 CRF 기반의 모델들이 연구되었다[1]. [4]에서는 개체명 인식이 특정 Span을 찾는다는 부분에 집중하여 기계독해 모델을 활용하여 개체명 인식을 시도하였다.

중첩 개체명 인식은 중첩된 개체명들을 모두 인식해야 하기 때문에 기존의 CRF 기반의 모델들로는 처리하기 어렵다는 문제점이 있다. [5]에서는 문장에서 두 단어 간의 의존 관계를 분류하기 위해 Biaffine 분류기를 활용한 의존 구문 분석 모델[6]에 착안하여 Biaffine Span 분류기를 적용하여 중첩 개체명 인식을 시도하였다. [7]에서는 Biaffine 중첩 개체명 인식 모델에 Cross Span Representation 정보를 추가한 Triaffine 중첩 개체명 인식 모델을 제안하였다.

기존 딥러닝 모델의 분류기는 입력 정보를 벡터 공간상에 표현하여 분류를 시도한다. Label Embedding은 분류될 라벨에 대한 정보도 학습하여 벡터 공간상에 표현하고 이를 분류에 활용하는 기술이다. 이러한 Label Embedding 기술은 라벨에 대한 정보도 같이 학습되어 전혀 학습하지 못한 라벨에 대한 데이터도 어느정도 처리가 가능한 모델을 만들 수 있다는 장점이 있다[8]. [9,10]에서는 순차적 레이블링 문제를 위한 CRF 기반의 모델에 Label Embedding 기술을 적용하여 CRF 레이어 대신 Label

Attention 레이어를 적용하였다. [11]에서는 라벨 정보를 이용하기 위해 이를 인코딩하여 Attention Network로 분류하는 모델을 제안하였다. 해당 모델은 BIO 태그 부착 개체명 인식 모델로 Few-shot Learning 환경에서 높은 성능을 보였다.

본 논문에서 제안한 모델은 [11]의 연구와 유사하게 Label Description을 이용하지만, 하나의 완성된 긴 문장의 Label Description을 라벨 정보로 사용하며, [11]의 모델은 중첩 개체명 인식이 아닌 일반 개체명 인식 모델이며, 라벨 정보를 입력 정보와 결합하여 사용하였으나, 본 논문에서는 중첩 개체명 인식을 위해 기본 모델로 Biaffine 모델을 사용하고 라벨 정보를 출력 레이어의 Label Attention에서만 사용한다. 본 논문에서 제안한 모델은 [5]의 연구와 유사하게 Span 기반의 Biaffine 중첩 개체명 인식 모델을 이용하지만, 출력 레이어에서 라벨 정보를 활용하여 Few-shot Learning 환경에서 높은 성능을 보인다.

3. Span 기반 중첩 개체명 인식 모델

Span 기반 중첩 개체명 인식 모델은 BIO 개체명 태그를 순차적 레이블링하는 모델과 달리 문장에 존재하는 Span 후보들을 처리하는 모델이다. 본 논문에서는 [5]의 Biaffine 중첩 개체명 인식 모델을 기본 모델로 하여 구현을 하였다.

3.1 Biaffine 기반 중첩 개체명 인식 모델

Biaffine 기반 중첩 개체명 인식 모델은 문장의 모든 Span 후보들을 분류하는 모델로 수식은 다음과 같다.

$$h = BERT(tokens_{seq}) \quad (1)$$
$$h^{start} = FFNN_{start}(h) \quad (2)$$
$$h^{end} = FFNN_{end}(h) \quad (3)$$
$$s_{i,j} = biaffine(h_i^{start}, h_j^{end}) = (h_i^{start})^T U(h_j^{end}) + W(h_i^{start} \oplus h_j^{end}) + b \quad (4)$$
$$y'_{i,j} = \arg max\,(s_{i,j}) \quad (5)$$

$tokens_{seq}$는 입력 문장의 단어들이며 이를 BERT[12]로 인코딩한다. 이후 인코딩 된 정보는 $FFNN_{start}$와 $FFNN_{end}$를 각각 거쳐 h^{start}와 h^{end}를 구한다. h^{start}와 h^{end}는 문장에 존재하는 모든 span 후보들의 시작 단어 정보와 끝 단어 정보를 나타내게 된다. 이 모든 후보 Span들에 대해 Biaffine 개체명 분류를 시도하며 상세 수식은 수식(4)와 같다. U, W, b는 모두 학습 가능한 가중치이며 최종적으로 Span i,j (i,j는 문장의 모든 단어들에 대한 인덱스)에 대해 개체명 태그를 분류하게 된다.

3.2 라벨 정보를 이용한 Label Description Embedding 모델

LDE(Label Description Embedding) 모델은 Biaffine 중첩 개체명 인식 모델에서 Biaffine 분류기를 Label Attention으로 대체한 모델이며 상세 수식은 다음과 같다.

$$h = BERT(tokens_{seq}) \quad (6)$$
$$h2_k = BERT(tokens_k^{Label\ Description}) \quad (7)$$
$$h2_k^{label} = FFNN_{label}(h2_k[CLS]) \quad (8)$$
$$h3_{i,j}^{span} = FFNN_{span}([h_i; h_j]) \quad (9)$$
$$s_{i,j}^k = attention(h2_k^{label}, h3_{i,j}^{span}) = (h2_k^{label})^T h3_{i,j}^{span} \quad (10)$$
$$y'_{i,j} = \arg max\,(s_{i,j}) \quad (11)$$

입력 문장을 인코딩 하는 과정은 3.1의 Biaffine 모델과 같으며 추가적으로 Label Description들을 인코딩한다. Label Description들은 각 라벨 별로 작성하며 이때 k는 모든 라벨들에 대한 인덱스이다. Label Description 인코더는 입력 문장 인코더와 가중치를 공유하는 동일한 BERT를 사용하여 라벨에 대한 정보 $h2_k$를 만든다. 그리고 해당 $h2_k$에서 $[CLS]$ 인덱스 위치(BERT의 입력 문장 맨 처음 단어)의 벡터를 $FFNN_{label}$를 거쳐 $h2_k^{label}$로 만들고 해당 정보를 Label Embedding으로서 활용한다. BERT로 인코딩한 입력 문장 정보 h는 모든 Span 후보들(인덱스 i,j)의 시작 인덱스와 끝 인덱스의 정보를 합치고(Concatenate) $FFNN_{span}$를 거쳐 인덱스 i,j 에 대한 Span 표현 $h3_{i,j}^{span}$를 구한다. 이후 Span 표현 $h3_{i,j}^{span}$에서 각각의 모든 k들에 대해 $h2_k^{label}$와 Attention Score를 구하여 해당 Span에 대한 분류를 시도한다. 이때 Attention Score 함수는 [13]의 dot 수식을 사용하며 모델의 전체 그림은 다음과 같다.

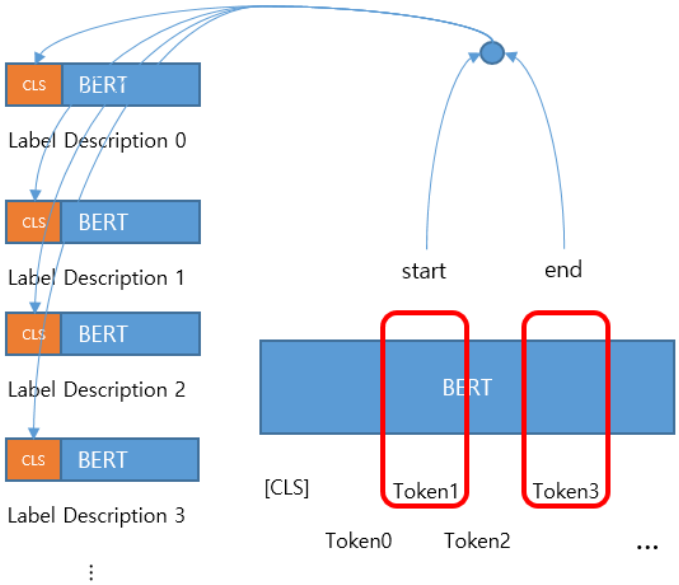

그림 1. Label Description Embedding 모델의 그림

NER tag	Description
O	O : Outside of named entities.
G#DNA	G#DNA : DNA is a fundamental molecule in the biomedical domain, serving as the genetic blueprint of all living organisms, carrying hereditary information, enabling genomics research, facilitating personalized medicine, aiding in diagnostics and forensics, and offering insights into evolutionary biology and gene editing for disease treatments.
G#protein	G#protein : Proteins are fundamental biomolecules in the biomedical domain, serving as essential building blocks of cells and tissues, catalysts for biochemical reactions, and key regulators of biological processes, playing crucial roles in health and disease.
G#cell_type	G#cell_type : In the biomedical domain, cell type refers to a specific class or category of cells sharing similar morphological, functional, and genetic characteristics within a particular organism or tissue.
G#cell_line	G#cell_line : A cell line in the biomedical domain refers to a population of cells derived from a single source and cultured in a laboratory setting, providing a valuable tool for studying various biological processes and testing experimental treatments.
G#RNA	G#RNA : RNA (Ribonucleic acid) in the biomedical domain plays a critical role as a versatile molecule responsible for translating genetic information from DNA to proteins, regulating gene expression, and serving as a potential therapeutic target in various diseases.

표 1. GENIA 데이터의 개체명 태그의 Description

Groups	GENIA									
	1-shot	5-shot	10-shot	20-shot	Train (1%)	Train (10%)	Train (25%)	Train (50%)	Train (100%)	Test
# 1	0.00%	22.83%	18.54%	22.92%	17.86%	17.46%	17.43%	17.91%	17.97%	21.73%
# 2	12.50%	21.62%	21.50%	18.50%	-	-	-	-	-	-
# 3	53.33%	27.72%	12.26%	26.51%	-	-	-	-	-	-
# 4	21.05%	24.44%	28.57%	18.53%	-	-	-	-	-	-
# 5	42.86%	24.74%	26.84%	24.12%	-	-	-	-	-	-

표 2. GENIA 데이터 별 중첩된 개체명들의 비율

	1-shot	5-shot	10-shot	20-shot	1%	10%	25%	50%	100%
Biaffine	5.75 (±2.58)	30.74 (±2.95)	31.93 (±1.93)	50.65 (±2.39)	57.64	73.56	75.63	77.13	78.20
LDE	11.60 (±2.51)	45.07 (±3.57)	47.90 (±2.27)	61.46 (±1.62)	66.05	74.37	76.21	77.40	79.01

표 3. 모델 별 GENIA 중첩 개체명 인식 데이터에 대한 F1-measure 성능

4. 실험

제안한 LDE 모델의 성능 평가를 위해 중첩 개체명 인식 데이터 중 하나인 GENIA[14]을 사용하였다. GENIA는 생명 과학 분야에서 사용되는 중첩 개체명 인식을 위한 데이터이며 개체명 타입은 유전자, 단백질 등이다. 각각의 개체명 태그는 [3]의 연구와 동일하게 카테고리들을 통합하여 총 5개의 개체명 태그를 사용하였다. 본 논문에서 제안하는 LDE 모델을 위해 GENIA 개체명에 대한 Label Description들을 작성하였으며 상세한 내용은 표 1과 같다. Label Description에는 다양한 단어들이 들어가도록 작성하였으며 최대한 긴 문장이 되도록 하였다. 문장 인코딩에 사용되는 인코더는 생명 과학 분야의 데이터를 사전학습한 BioBERT-v1.1[15]을 사용하였으며 비교실험을 위해 [5]의 Biaffine 모델을 자체적으로 구현하였다(3.1절). 실험은 Biaffine 모델과 LDE 모델을 동일하게 설계하였으며 마지막 개체명 분류를 위한 ***FFNN***들만 최적의 성능을 나타내는 하이퍼파라미터를 모델별로 찾았다. 실험데이터는 [5,7]의 연구와 마찬가지로 학습데이터(90%), 평가데이터(10%)로 나누었으며 Few-shot Learning 환경을 위해 1-shot(5문장), 5-shot(25문장), 10-shot(50문장), 20-shot(100문장), 1%(167문장), 10%(1,670문장), 25%(4,173문장), 50%(8,346문장), 100%(16,691문장)로 학습데이터를 나누었다. 이때 k-shot 데이터는 각 개체명 태그(‘O’ 태그 제외) 별로 k문장씩 학습데이터로 사용한 것이다. Few-shot 학습데이터에 대한 편향된 성능을 피하기 위해 5-그룹으로 학습데이터를 나누었으며 1%, 10%, 25%, 50%의 학습데이터는 충분히 다양한 라벨들을 가지고 있다고 판단하여 한 가지의 그룹만 실험을 하였고, 각 데이터들의 중첩된 개체명 비율은 표 2와 같다.

표 3은 GENIA 중첩 개체명 인식 데이터에 대한 모델별 F1-measure 성능 표이다. Few-shot 데이터들에 대한 성능은 5-그룹들의 성능의 평균과 표준편차를 구한 것이다. 학습데이터를 100% 모두 사용하였을 때 Biaffine 모델 78.20%, LDE 모델 79.01%로 LDE 모델의 성능이 더 높지만 큰 차이가 없는 것을 볼 수 있으며, 이는 GENIA의 데이터가 Biaffine 모델 기준으로 충분한 양이기 때문이라고 생각한다. 학습데이터 50%, 25%, 10%의 경우 LDE 모델이 0.27%p, 0.59%p, 0.81%p 더 높은 성능을 나타내며 특히 학습데이터를 1%만 사용하였을 때 8.41%p 더 높은 성능을 보여 학습데이터가 적을 수록 LDE 모델이 더 높은 성능을 보임을 알 수 있다. K-shot 학습데이터는 각 개체명 태그 별로 k문장씩 학습데이터로 사용한 결과이며 20-shot, 10-shot, 5-shot, 1-shot 실험에 대해 LDE 모델이 평균적으로 10.81%p, 15.97%p, 14.33%p, 5.85%p 더 높은 성능을 보임을 알 수 있다.

5. 결론

본 논문에서는 Few-shot Learning 환경에 강건한 중첩 개체명 인식 모델을 위해, Span 기반 중첩 개체명 인식 모델인 Biaffine 모델의 출력 레이어를 라벨 의미 정보를 활용하도록 변경하였다. 특히 Label Description을 라벨 정보로 사용하여 좀 더 깊은 의미 정보를 사용하였으며 실험 결과 학습데이터가 극히 적은 환경에서 기존 모델보다 높은 성능을 보임을 확인하였다.

제안된 LDE 모델은 라벨에 대한 정보를 함께 사용하기 때문에 새로운 라벨이 등장하는 새로운 도메인의 데이터도 어느정도 처리가 가능하다는 특징이 있다. 추후 연구로는 이러한 특징을 살려서 다양한 도메인의 학습데이터를 이용하여 모델의 성능을 최적화하는 연구를 진행할 예정이다.

참고문헌

[1] 이창기. "Long Short-Term Memory 기반의 Recurrent Neural Network 를 이용한 개체명 인식." 한국정보과학회 학술발표논문집 (2015): 645-647.

[2] 송영숙, 정유남, 유현조. "한국어 중첩 개체명 분석을 위한 연구." 한국어 의미학 76 (2022): 65-101.

[3] Finkel, Jenny Rose, Christopher D. Manning. "Nested named entity recognition." Proceedings of the 2009 conference on empirical methods in natural language processing. 2009.

[4] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, Jiwei Li. "A Unified MRC Framework for Named Entity Recognition." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[5] Yu, Juntao, Bernd Bohnet, Massimo Poesio. "Named Entity Recognition as Dependency Parsing." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[6] Dozat, Timothy, Christopher D. Manning. "Deep Biaffine Attention for Neural Dependency Parsing." International Conference on Learning Representations. 2016.

[7] Zheng Yuan, Chuanqi Tan, Songfang Huang, Fei Huang. "Fusing Heterogeneous Factors with Triaffine Mechanism for Nested Named Entity Recognition." Findings of the Association for Computational Linguistics: ACL 2022. 2022.

[8] Z Akata, F Perronnin, Z Harchaoui, C Schmid. "Label-embedding for image classification." IEEE transactions on pattern analysis and machine intelligence 38.7 (2015): 1425-1438.

[9] Cui, Leyang, Yue Zhang. "Hierarchically-Refined Label Attention Network for Sequence Labeling." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[10] Hongjin Kim, Harksoo Kim. "Integrated Model for Morphological Analysis and Named Entity Recognition Based on Label Attention Networks in Korean." Applied Sciences 10.11 (2020): 3740.

[11] Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. "Label Semantics for Few Shot Named Entity Recognition." Findings of the Association for Computational Linguistics: ACL 2022. 2022.
[12] Kenton, Jacob Devlin Ming-Wei Chang, Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT. 2019.
[13] Luong, Minh-Thang, Hieu Pham, Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.
[14] JD Kim, T Ohta, Y Tateisi, J Tsujii. "GENIA corpus—a semantically annotated corpus for bio-textmining." Bioinformatics 19.suppl_1 (2003): i180-i182.
[15] J Lee, W Yoon, S Kim, D Kim, S Kim, CH So, J Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

감사의 글

이 연구는 ETRI부설연구소의 위탁연구과제[2023-033]로 수행한 연구결과입니다.