지식증류를 활용한 지속적 한국어 개체명 인식

장준서○, 박성식, 김학수
건국대학교 인공지능학과
{jjs970612, a163912, nlpdrkim}@konkuk.ac.kr

Continuous Korean Named Entity Recognition Using Knowledge Distillation

Junseo Jang○, Seongsik Park, Harksoo Kim
Department of Artificial Intelligence, Konkuk University

요 약

개체명 인식은 주어진 텍스트에서 특정 유형의 개체들을 식별하고 추출하는 작업이다. 일반적인 딥러닝 기반 개체명 인식은 사전에 개체명들을 모두 정의한 뒤 모델을 학습한다. 하지만 실제 학습 환경에서는 지속적으로 새로운 개체명이 등장할 수 있을뿐더러 기존 개체명을 학습한 데이터가 접근이 불가할 수 있다. 또한, 새로 모델을 학습하기 위해 새로운 데이터에 기존 개체명을 수동 태깅하기엔 많은 시간과 비용이 든다. 해결 방안으로 여러 방법론이 제시되었지만 새로운 개체명을 학습하는 과정에서 기존 개체명 지식에 대한 망각 현상이 나타났다. 본 논문에서는 지식증류를 활용한 지속학습이 한국어 개체명 인식에서 기존 지식에 대한 망각을 줄이고 새로운 지식을 학습하는데 효과적임을 보인다. 국립국어원에서 제공한 개체명 인식 데이터로 실험과 평가를 진행하여 성능의 우수성을 보인다.

주제어: Continual Learning, Knowledge Distillation, 개체명 인식

1. 서론

개체명 인식(Named Entity Recognition, NER)[1]은 주어진 텍스트에서 특정 유형의 개체들을 식별하고 추출하는 자연어처리 태스크이다. 일반적인 개체명 인식은 사전 정의된 개체명들을 식별하는 것을 목표로 한다. 이를 위해 사전 정의된 개체명이 태깅된 데이터셋을 학습에 활용한다. 하지만 실제 환경에서는 새로운 개체명이 지속적으로 등장 할 수 있다. 이 경우 새로운 학습 데이터에 이전 개체명을 모두 수동으로 태깅하는 방법 또는 이전에 사용했던 학습 데이터에 새롭게 추가할 개체명을 수동으로 태깅하여 새롭게 구축한 데이터로 모델을 학습하는 방법을 가장 쉽게 떠올릴 수 있다. 하지만 이는 많은 시간과 비용이 요구될 뿐만 아니라 이전에 사용했던 데이터가 여러 불가피한 이유로 접근이 어려우면 사용할 수 없는 방법이다. 또 다른 방법으로는 새로운 개체명이 등장할 때마다 하나의 모델에 순차적으로 학습시키는 방법이 있다. 이 경우 새로운 지식을 습득하는데 쉽지만, 이전 지식을 잊어버리는 치명적 망각 현상(catastrophic forgetting)[2]이 발생한다. 이를 해결하는 방안으로 지속학습(continual learning, CL)[3]에 대한 연구가 활발히 진행되고 있다. 지속학습은 치명적 망각 현상을 방지하고 새로운 지식을 효과적으로 학습할 수 있는 학습 기법이다.

개체명 인식에 지속학습을 적용하기 위한 직관적인 방법에는 self-training[4]이 있다. self-training은 이전의 개체명 인식 데이터로 학습된 모델을 통해 새로운 데이터를 자동으로 태깅하고 이를 새로운 모델의 학습에 활용하는 방법이다. 그러나 self-training은 이전 모델의 오류가 다음 모델에 그대로 전파된다는 문제점이 존재한다.

본 논문에서는 지식증류(knowledge distillation)[5]를 이용한 지속학습 모델인 ExtendNER[6]을 한국어 개체명 인식의 지속학습에 적용한다. 지식증류는 전이학습(transfer learning)의 일종으로 큰 모델의 지식을 작은 모델에 전이시키는 방법을 말한다. 모델이 학습한 지식을 한 모델에서 다른 모델로 이전한다는 점에서 지속학습에 활용할 수 있다. 지속학습 환경에서는 기존 개체명을 학습한 모델을 교사 모델, 새로운 개체명이 추가될 모델을 학생 모델로 정의하여 학습 과정에서 지식증류를 적용한다.

본 논문에서는 지식증류를 활용한 지속학습이 한국어 개체명 인식에도 유효한지 확인하기 위해 한국어 언어모델인 KoElectra[7]와 국립국어원에서 제공하는 한국어 개체명 인식 데이터 중 6개의 개체명을 사용하여 모델 학습과 평가를 진행하였다. 결과적으로 지식증류를 활용하면 한국어 개체명 인식의 지속학습에서도 과거 개체명에 대한 망각 현상을 방지할 수 있음을 보인다.

2. 관련 연구

지속학습은 새로운 지식을 학습할 때 과거에 학습한 지식에 대한 망각 현상을 방지하는 것을 최우선 목표로 한다. 현재까지 이러한 망각 현상을 극복하기 위해 많은 연구가 이루어져 왔다.

지속학습은 컴퓨터비전 분야에서 주도적으로 연구가 진행됐지만 최근에는 자연어처리 분야에서도 활발한 연구가 진행되고 있다[8]. 개체명 인식의 경우 시간 경과

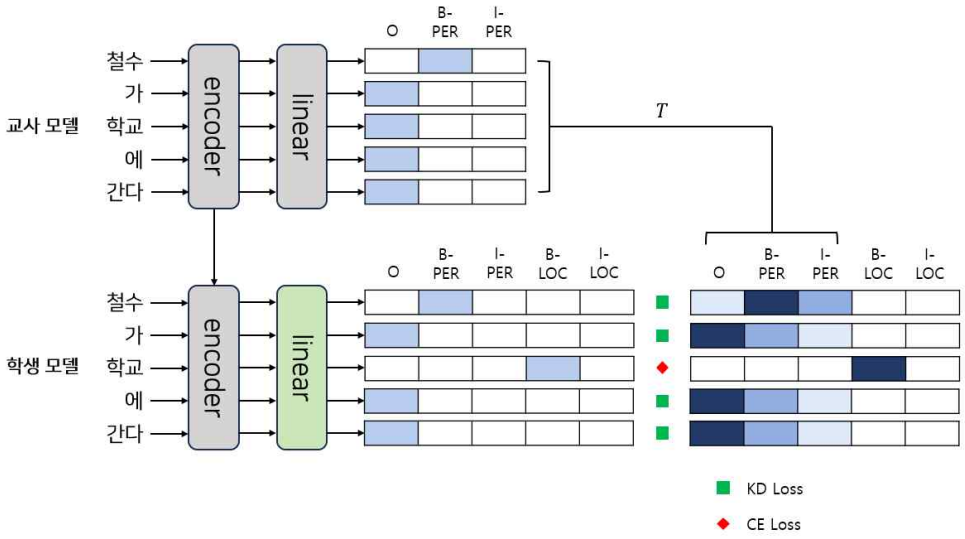

그림 1. 지식증류를 활용한 지속적 개체명 인식 학습 과정

에 따라서 식별해야 할 개체의 종류가 증가할 수 있기 때문에 이에 대응하기 위한 지속학습 방법 연구가 필요하다. [9]는 가장 처음으로 sequence labeling 모델에 transfer learning을 통해 지속학습을 시도했다. 지식을 전이 받는 target 모델은 지식을 전이해주는 source 모델의 파라미터를 전이 받고 그 위에 새로운 레이어를 추가하여 생성된다. 이후 새로운 지식이 담긴 데이터로 학습을 진행하여 이전 지식을 보존함과 동시에 새로운 지식에 대해 학습한다. 추가로 neural adapter를 통해 source 모델의 출력 확률 분포와 target 모델의 출력 확률 분포 간 차이를 학습하여 정보 간극을 줄임으로써 지식 전이를 더욱 강화한다. [4]에서는 self-training을 사용한 개체명 인식을 시도한다. Self-training은 이전 개체명이 학습된 모델로 새로운 개체명만 태깅된 데이터를 자동 태깅 해준 뒤 해당 데이터로 새로운 모델을 학습하여 지식을 전이한다. 하지만 이전 모델의 오류가 함께 전이된다는 문제점을 가지고 있다. 최근에는 ExtendNER & ADDNER[6] 그리고 L&R[10]이 개체명 인식의 지속학습을 위한 방법으로 제안됐다. AddNER은 인코더와 여러 레이어로 구성되어있고 새로운 개체명을 학습할 때마다 새로운 레이어가 추가되어 이전 모델의 확률 분포와의 KL Divergence, 새로운 데이터의 원핫 인코딩과의 Cross Entropy Loss를 통해 학습된다. ExtendNER은 AddNER과 달리 인코더와 하나의 레이어로 구성이 되어있고 개체명이 추가로 학습될 때마다 레이어의 차원이 증가한다. 학습 과정은 AddNER과 같다. L&R은 ExtendNER 모델을 사용하고 추가로 reviewing 단계를 넣어 이전 개체명들에 대한 추가 데이터 생성을 통해 망각을 방지한다.

3. 지식증류를 활용한 지속적 개체명 인식

본 논문에서는 지식증류를 활용한 지속학습 방법을 한국어 개체명 인식에 적용한다. 3.1절에서는 개체명 인식을 위한 모델 구조에 대해 설명하며 3.2절에서는 지식증류를 활용하여 지속학습을 수행하는 과정에 대해 설명한다.

3.1 개체명 인식 모델

개체명 인식 모델은 KoElectra 인코더와 linear layer로 구성된다. 본 논문에서는 BIO 태깅 방식을 사용하기 때문에 학습한 개체명의 수가 n이라 할 때 $(2n+1)$개의 클래스 분류를 한다. 따라서 linear layer는 $h \times (2n+1)$의 크기를 가진다.

한국어 문장이 입력되면, 토큰 단위로 쪼개진 후 임베딩의 형태로 모델에 전달된다. 모델 인코더와 linear layer를 거치면 마지막 단계에서 토큰별 출력 확률 분포를 softmax를 통해 각 토큰별로 $(2n+1)$개의 클래스 중 가장 확률값이 높은 클래스를 선정하여 최종 출력을 얻는다.

3.2 지식증류를 활용한 지속학습 방법

지식증류는 한 모델에서 다른 모델로 지식을 전이하는 기법이다. 지식을 전이하는 모델을 교사 모델, 전이 받는 모델을 학생 모델이라고 할 때 학생 모델은 교사 모델의 출력 확률 분포와 학생 모델의 출력 확률 분포 간의 차이를 줄이는 방향으로 학습된다. 두 확률 분포 간의 차이를 계산할 때는 KL Divergence를 사용하며 식은 다음과 같다.

$$D_{KL}(p_S \parallel p_T) = \sum p_S \log \frac{p_S}{p_T} \quad (1)$$

학생 모델의 출력 분포를 p_S, 교사 모델의 출력 분포를 p_T라 할 때 학습이 진행될수록 D_{KL}이 최소화되도록 학생 모델의 파라미터가 조정된다. 교사 모델은 이전 개체명 데이터에 대해서 최적화가 완료된 상태이기 때문에

확률 분포 p_T는 정답 개체명에서만 높은 확률을 갖는 첨예한 형태를 띤다. 그러나 지식증류 과정에서는 정답이 아닌 클래스의 확률 또한 전이해야 할 지식이기 때문에 확률 분포 조절 계수 T를 추가하여 분포의 형태를 조금 더 완만하게 만들어 준다. 이를 반영한 p_T 계산식은 다음과 같다.

$$p_T = \frac{\exp(z/T)}{\sum \exp(z/T)} \quad (2)$$

T는 수동으로 설정할 수 있는 하이퍼 파라미터로 높게 설정할수록 완만한 형태의 확률 분포가 된다.

그림 1은 지식증류를 활용한 지속학습 과정을 보여준다. 이전 데이터셋은 인명(PER)만 태깅되어 있고 현재 데이터셋은 장소명(LOC)만 태깅되어 있다고 가정하고 지속학습을 진행하는 예시를 보인다. 교사 모델은 인명을 예측하도록 학습이 완료된 상태이다. 학생 모델은 교사 모델의 인코더 파라미터를 그대로 계승하며 linear layer만 이전과 현재 개체명을 학습할 수 있는 확장된 linear layer로 교체된다.

제안 방법은 두 가지 loss를 사용하여 최종 loss 함수를 계산한다. 첫째, 입력 데이터 중 라벨이 '0' 에 해당하는 토큰들의 교사 모델을 통해 얻은 출력 확률 분포와 학생 모델을 통해 얻은 출력 확률 분포 사이의 KL-Divergence이다(수식 3). 다만 학생 모델을 통해 얻은 확률 분포와 차원이 다르기 때문에 교사 모델로부터 얻은 확률 분포를 0으로 채운다. 둘째, 입력 데이터 중 라벨이 '0' 가 아닌 클래스에 해당하는 토큰들, 즉 새로운 개체명에 해당하는 토큰들의 정답 라벨 원핫 인코딩과 학생 모델을 통해 얻은 출력 확률 분포 사이의 Cross Entropy Loss이다(수식 4). 그리고 이 두 loss의 합을 통해 최종 loss를 구한다(수식 5).

$$Loss_{KD} = KL(p_{E_i}^{M_i}, p_{E_i}^{M_{i+1}}) \quad (3)$$

$$Loss_{CE} = CE(y_{e^{i+1}}, p_{e^{i+1}}^{M_{i+1}}) \quad (4)$$

$$Loss = Loss_{KD} + Loss_{CE} \quad (5)$$

4. 실험

4.1 데이터셋

모델 학습과 평가에는 국립국어원에서 제공하는 여러 한국어 개체명 인식 데이터[1]를 종합하여 사용했다. 이 데이터에는 대략 180k개의 문장이 포함되어 있고 다양한 개체명들이 문장에 태깅되어있다. 실험에는 학습 데이터 수와 개체명 학습 순서의 경우의 수를 고려하여 전체 문장 데이터에서 태깅된 빈도수가 높은 6개의 개체명만 실험에 사용했다: 기구(ORG), 사람(PER), 문물(CVL), 날짜(DAT), 장소(LOC), 수량(QNT).

제안 모델의 학습 과정은 총 6 step으로 진행되기 때문에 매 step에 사용할 학습 데이터 D_i를 중복 없이 4,190문장씩 임의 추출하여 D_1~D_6를 구성한다. 평가 데이터셋 D_{test}는 학습 데이터와 중복이 안 되게 3,023개 문장을 임의 추출하여 구성한다.

4.2 실험 환경

표 1. 다양한 개체명 인식 지속학습 순서

	step1	step2	step3	step4	step5	step6
순서1	ORG	PER	CVL	DAT	LOC	QNT
순서2	DAT	QNT	PER	LOC	ORG	CVL
순서3	CVL	LOC	ORG	QNT	DAT	PER
순서4	QNT	ORG	DAT	PER	CVL	LOC
순서5	LOC	CVL	QNT	ORG	PER	DAT
순서6	PER	DAT	LOC	CVL	QNT	ORG

지속학습 환경에서는 데이터를 학습하는 순서도 최종 성능에 영향을 끼칠 수 있다. 따라서 지속적 개체명 인식에서도 학습하는 개체명의 순서에 따라 성능이 달라지는 현상을 보완하고 성능을 일반화하기 위해 6개의 개체명의 학습 순서를 총 6가지를 만들고 이를 평균 내어 최종 성능을 도출한다. 실험에 사용한 개체명 학습 순서의 경우의 수는 표 1과 같다.

지속적 개체명 인식 환경을 가정하기 위해 i번째 step의 학습 데이터셋 D_i에는 e^i만 태깅되어 있도록 설정한다. 각 step의 평가 단계에서는 현재 step까지 학습한 개체명이 모두 태깅되어 있도록 평가 데이터셋 D_{test}를 설정한다. 모델 인코더는 KoElectra를 사용했다. 토큰의 최대 길이는 300으로 설정하였고 학습률 5e-05, 배치 크기 16, warmup ratio 0.05, T는 2로 실험 환경 설정을 하였다. 1단계의 경우 10 epoch, 2~6단계는 20 epoch만큼 학습했다.

4.3 비교 모델

본 논문에서 성능 비교에 사용한 baseline 모델들은 다음과 같다.

Naive Learning 개체명 인식 모델에 별다른 지속학습 방법을 적용하지 않은 상태로 D_1~D_6까지 학습을 순차적으로 진행한 경우이다. 각 학습 단계에서 이전 데이터로 학습된 모델을 그대로 사용한다. 치명적 망각이 끼치는 영향을 확인할 수 있다.

Self-training M_i와 e^{i+1}만 태깅된 D_{i+1}가 주어졌을 때 M_i을 이용해 D_{i+1}에 $\{e^1$~$e^i\}$를 태깅하여 $\{e^1$~$e^{i+1}\}$가 태깅된 D_{i+1}을 만든다. 이후 D_{i+1}을 이용해 M_{i+1}을 학습한다.

Full Annotated 모든 개체명이 태깅된 상태의 모든 데이

1)https://corpus.korean.go.kr/request/reausetMain.do?lang=ko

표 2. 제안 모델과 비교 모델들의 학습 단계별 성능

	Macro f1					
모델	1	2	3	4	5	6
Full Annotated	0.8828	0.9033	0.9061	0.9130	0.9143	0.9165
Naive Learning	0.8828	0.4439	0.2984	0.2220	0.1795	0.1483
Self-Training	0.8828	0.8683	0.8649	0.8576	0.8552	0.8468
제안 모델	**0.8828**	**0.8859**	**0.8900**	**0.8891**	**0.8898**	**0.8904**

터셋 $\bigcup_{i=1}^{6} D_i$를 사용한다. 지속학습의 upper bound 성능이라고 정의할 수 있다.

4.4 실험 결과 및 분석

실험의 결과 지표로는 Macro f1 score를 선정하였다. 각 step에서 나타나는 점수는 6개의 학습 경우의 수에서 얻은 f1 score를 평균 내어 얻은 결괏값이다. f1 score는 'O' 태그를 제외한 나머지 개체명들의 'B' 태그와 'I' 태그에 대한 점수만 측정하였다.

표 2를 보면 각 모델의 step별 결과를 확인할 수 있다. Naive learning은 step이 진행될수록 성능이 하락함과 동시에 마지막 step에서 다른 모델들과의 성능 차이도 매우 심하다. 이는 이전 개체명에 대한 지식을 잊는 치명적 망각 현상이 발생했기 때문으로 보인다. Self-training은 naive learning보다 각 step별로 더 높은 성능을 보인다. 별다른 지속학습 방법을 사용하지 않고 순차적으로 데이터를 학습하는 naive learning과 달리 이전 모델을 사용해 새로운 데이터를 자동 태깅함으로써 이전 지식 보존에 도움이 된 것으로 보인다. 하지만 여전히 step이 진행될수록 성능이 하락하는 모습을 보인다. 반면 제안 모델은 step이 진행될수록 성능이 상승하다 일정 부분에서 수렴하는 양상을 보인다. Self-training은 이전 모델의 오류가 그대로 다음 모델에 전이가 되지만, 제안 모델은 지식증류를 사용하여 이전 모델의 확률 분포를 다음 모델에 전이했기 때문에 self-training과 같은 성능 하락을 방지할 수 있다. 또한 upper-bound 모델인 full-annotated보다 최종 단계에서 대략 2% 포인트 밖에 성능 차이가 나지 않은 것으로 보아 지식증류를 활용한 지속학습이 성능적으로도 우수하다는 점을 알 수 있다.

5. 결론

본 논문에서는 새로운 개체명이 지속해서 등장하는 환경에서 이전 학습 데이터에 대한 접근 없이 효과적으로 이전 지식을 보존하고 새로운 지식을 학습하기 위해 지식증류를 활용한 지속학습 기법을 한국어 개체명 인식 태스크에 적용해 보았다. 실험 결과 지속학습 비교 모델인 self-training 모델보다 모든 학습 step에서 성능 향상을 보였고 upper-bound 모델인 full annotated와는 미세한 성능 차이를 보였다. 이에 따라 한국어 개체명 인식 태스크에서도 치명적 망각 현상을 충분히 방지하며 지속학습이 효과적으로 진행됐음을 확인할 수 있다. 앞으로는 다국어 문장에 대해 개체명을 지속해서 학습할 수 있는 다국어 개체명 지속학습을 연구할 예정이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

참고문헌

[1]김홍진, 김학수, "계층적 레이블 임베딩을 이용한 세부 분류 개체명 인식", 제33회 한글 및 한국어 정보처리 학술대회 논문집, pp.251-256, 2021.

[2]Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville and Yoshua Bengio, An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks, arXiv:1312.6211, 2013.

[3]Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà, Continual Lifelong Learning in Natural Language Processing: A Survey, In Proceedings of the 28th International Conference on Computational Linguistics, pages 6523-6541, 2020.

[4]Isaac Triguero, Salvador García and Francisco Herrera, Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study, Knowl Inf Syst 42, 245-284, 2015.

[5]Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, 2015.

[6]Monaikul, N., Castellucci, G., Filice, S., & Rokhlenko, O., Continual Learning for Named Entity

Recognition, Proceedings of the AAAI Conference on Artificial Intelligence, 35(15), 13570-13577, 2021.

[7]Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, arXiv:2003.10555, 2020.

[8]Shmelkov, K.; Schmid, C.; and Alahari, K., Incremental Learning of Object Detectors without Catastrophic Forgetting, In 2017 IEEE International Conference on Computer Vision (ICCV), 3420-3429, 2017.

[9]Chen, L. and Moschitti, A., Transfer learning for sequence labeling using source model and target data, In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 6260-6267, 2019.

[10]Xia, Y.; Wang, Q.; Lyu, Y.; Zhu, Y.; Wu, W.; Li, S.; and Dai, D., Learn and Review: Enhancing Continual Named Entity Recognition via Reviewing Synthetic Samples, In Findings of the Association for Computational Linguistics: ACL 2022, 2291-2300, 2022.