문법 오류 교정을 위한 적대적 학습 방법

권순철[1 O], 이근배[1, 2]

포항공과대학교 컴퓨터공학과[1], 포항공과대학교 AI대학원[2]
theincluder@postech.ac.kr, gblee@postech.ac.kr

Adversarial Training for Grammatical Error Correction

Soonchoul Kwon[1 O], Gary Geunbae Lee[1, 2]
Pohang University of Science and Technology
Department of Computer Science[1], Graduate School of Artificial Intelligence[2]

요 약

최근 성공적인 문법 오류 교정 연구들에는 복잡한 인공신경망 모델이 사용되고 있다. 그러나 이러한 모델을 훈련할 수 있는 공개 데이터는 필요에 비해 부족하여 과적합 문제를 일으킨다. 이 논문에서는 적대적 훈련 방법을 적용해 문법 오류 교정 분야의 과적합 문제를 해결하는 방법을 탐색한다. 모델의 비용을 증가시키는 경사를 이용한 fast gradient sign method(FGSM)와, 인공신경망을 이용해 모델의 비용을 증가시키기 위한 변동을 학습하는 learned perturbation method(LPM)가 실험되었다. 실험 결과, LPM은 모델 훈련에 효과가 없었으나, FGSM은 적대적 훈련을 사용하지 않은 모델보다 높은 $F_{0.5}$ 성능을 보이는 것이 확인되었다.

주제어: 문법 오류 교정, 적대적 훈련

1. 서론

문법 오류 교정(Grammatical error correction)은 문장에 포함된 문법적 오류를 찾아내고 올바른 문장으로 고치는 과제이다. 문법 오류 교정은 그 자체로 문서 작성, 대화 시스템, 언어 학습 등 여러 분야에서 활용될 수 있다.

현재 문법 오류 교정 분야의 주류 연구들은 이 과제를 기계번역 과제로 보고, 틀린 문장을 올바른 문장으로 번역하는 것으로 문제를 해결한다. 이를 위해 기계번역에서 높은 성능을 보이는 인공신경망 기반 모델들이 사용된다[1, 2]. 그러나 이런 연구에 사용되는 합성곱 신경망(Convolutional neural network; CNN), transformer 등의 복잡한 모델을 훈련하기 위해서는 대규모의 훈련 데이터가 필요하다는 문제가 있다.

문법 오류 교정에 있어 훈련 데이터의 부족은 가장 중요한 문제 중 하나이다. 영문 문법 오류 교정에 가장 많이 활용되는 공개 데이터셋으로는 NUCLE[3], Lang-8[4], CLC FCE[5] 등이 있다. 그러나 성능이 더 높은 모델을 훈련하기 위해서는 항상 더 많은 데이터가 필요하며, 여러 연구에서는 비공개 데이터를 사용해 높은 성능을 얻었으나[6, 7] 이러한 연구는 데이터가 공개되기 전까지 재현이 불가능하다.

이를 위해 공개 데이터에서 데이터를 생성하거나 비슷한 효과를 얻기 위한 연구가 진행됐다. Fluency boost learning 연구는 오류를 생성하는 모델을 훈련하거나 모델이 생성하는 잘못된 출력을 이용해 새로운 데이터를 생성했고[7], diverse backtranslation 연구에서는 문장 생성 중 빔서치 과정에 적절한 노이즈를 줌으로서 다양한 오류 문장을 생성했다[6].

일반적으로 적절히 훈련된 모델은 의미적으로 유사한 입력에 대해 거의 같은 결과를 낼 것으로 기대된다. 예를 들어, '*He went for the store.*' 라는 틀린 문장을 교정한 결과가 '*He went to the store.*' 라면, '*He went as the store.*' 도 같은 문장으로 교정되기를 기대한다. 그러나 훈련 데이터에 과적합된 모델은 위와 같이 훈련 데이터와 비슷하지만 훈련 데이터셋에는 없는 예제에 대해 틀린 결과를 출력할 수 있다.

적대적 훈련(Adversarial training)은 훈련 데이터와 비슷하지만, 모델이 잘 처리하지 못하는 데이터를 훈련 데이터에 포함시키는 방법이다[8]. 이를 통해 훈련 데이터를 증강하는 효과를 얻을 수 있고, 결과적으로 과적합 문제를 해결할 수 있다.

이 논문에서는 문법 오류 교정 과제에 적대적 훈련 방법을 적용했다. 과적합 문제를 개선하는 적대적 훈련을 통해 어떤 종류의 변동이 문법 오류 교정에 있어 도움이 되는지를 증명하였다.

2. 관련 연구

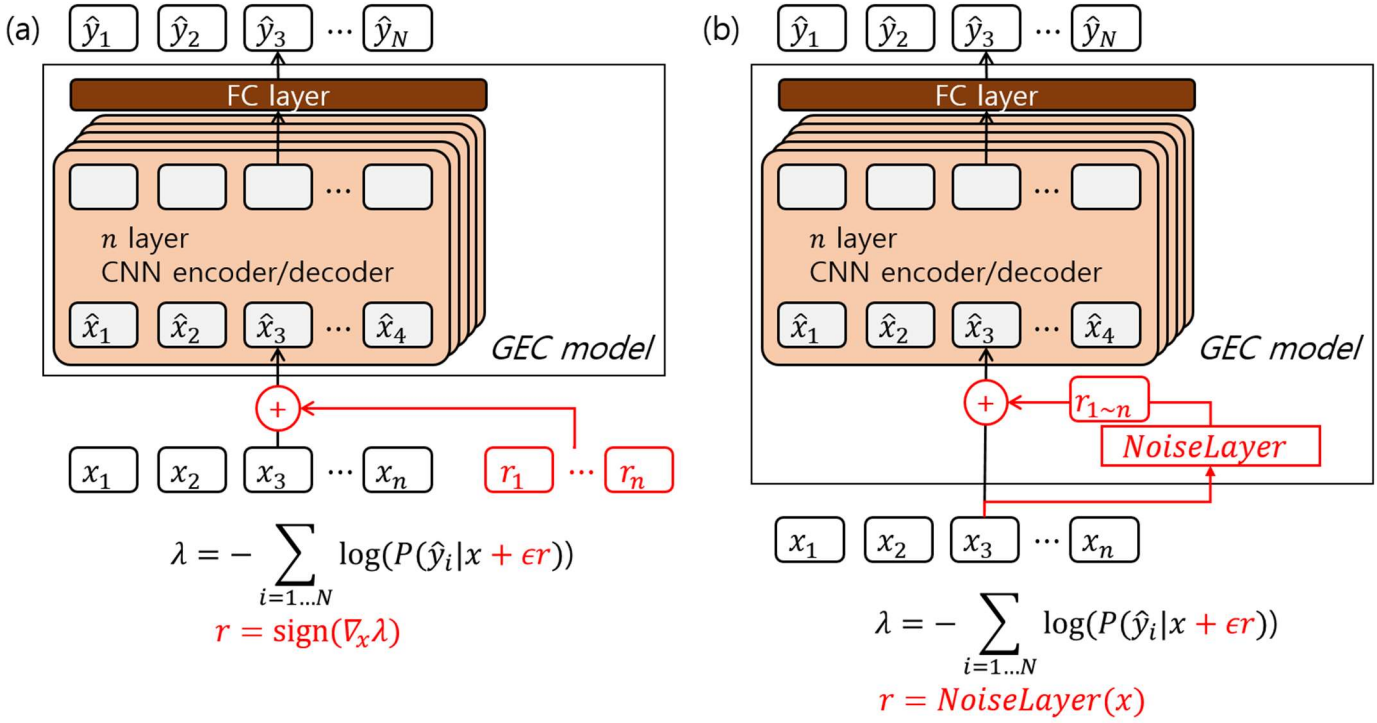

그림 1. 적대적 훈련에 사용한 두 가지 변동: Fast Gradient Sign Method(a)와 Learned Perturbation Method(b)

적대적 훈련은 실제 데이터와 유사하지만 모델의 비용을 증가시키는 데이터를 훈련 과정에 삽입하는 방법이다. 데이터를 연속적으로 바꿀 수 있는 이미지 분야에서 많이 사용되었으나[8], 자연어 처리 분야에서도 주제 분류 등에서 이용되었다[9]. 적대적 훈련으로 생성하는 데이터는 1. 실제 데이터의 분포와 비슷한 데이터이어야 하고, 2. 모델이 잘 처리하지 못하는 데이터이어야 한다.

위 두 조건을 만족하는 데이터를 만들기 위해서 [8]은 훈련 데이터 x에 일정 변동(perturbation) η를 주는 것으로 새로운 데이터 $\hat{x}$를 만들었다:

$$\hat{x} = x + \eta \ \ (\|\eta\|_\infty < \epsilon)$$

이 때 가해지는 변동은 모델의 비용 함수를 증가시켜 해당 데이터를 잘 처리하지 못하게 만드는 변동이 선택된다. 하이퍼 파라미터 ϵ는 변동의 최대 크기로, 생성된 데이터 $\hat{x}$가 실제 데이터 x와 너무 많이 다르지 않게 하는 제한 역할을 한다. ϵ이 너무 작을 경우 $\hat{x}$가 x와 너무 유사해 모델의 비용 함수를 별로 바꾸지 못하며, ϵ이 너무 클 경우 실제 데이터와 크게 다른 분포의 데이터를 생성하게 된다.

3. 문법 오류 교정을 위한 적대적 훈련

현재 문법 오류 교정 과제를 위한 훈련 데이터는 모델의 복잡도에 비해 부족한 수준이며, 이로 인해 발생하는 과적합 문제를 개선하기 위해 적대적 훈련 방법이 시도되었다. 문법 오류 교정을 위해서 두 종류의 변동이 실험되었으며, 두 방법 모두 모델의 비용 함수를 크게 만드는 변동을 얻는 방법이다.

3.1 Fast Gradient Sign Method

Fast gradient sign method(FGSM)는 모델의 비용 함수에 대한 입력의 경사의 부호를 변동으로 주는 방식이다[8]. 이 방식의 변동은 이미지 처리 분야에서 눈에 보이지 않는 작은 변동을 주면서 모델의 비용을 크게 증가시키는 데 성공적이었다[8, 10]. 모델의 교차 엔트로피 비용(cross entropy loss)을 λ라 했을 때, FGSM에서 사용하는 변동은 다음과 같다(그림 1a.):

$$\hat{x} = x + \epsilon r$$
$$r = \text{sign}(\nabla_x \lambda)$$

변동 r은 비용 함수를 증가시키는 방향이며, x를 입력했을 때보다 높은 비용을 생성하는 입력 $\hat{x}$를 생성하게 된다. 이는 더 낮은 비용을 생성하는 모델 파라미터를 찾는 경사하강법(Gradient descent)과 유사성을 갖는 방법이다.

3.2 Learned Perturbation Method

표 1. 훈련 및 테스트 데이터

데이터셋	문장 수	도메인
NUCLE	57,151	학생 에세이
Lang-8	2,167,655	SNS
CoNLL-2014	1,312	학생 에세이

Learned perturbation method(LPM)은 입력에 따라 어떤 변동을 주었을 때 비용이 증가하는지 학습하는 방법이다. LPM에서 사용하는 변동은 다음과 같다(그림 1b.):

$$\hat{x} = x + \epsilon r$$
$$r = NoiseLayer(x)$$

단층 인공신경망 $NoiseLayer$에 의해 비용을 증가시키는 변동 r을 생성하게 되며, $NoiseLayer$는 비용 λ을 증가시키는 방향으로 학습된다.

4. 실험 방법

훈련 데이터로는 영문 공개 데이터인 NUCLE[3]과 Lang-8[4]을 사용했으며, 테스트 데이터로는 CoNLL-2014 Shared Task[11] 데이터를 사용했다(표 1). 성능 기준으로는 MaxMatch $F_{0.5}$ score를 사용했다[3].

기준 모델로는 CNN을 조건부 적대적 생성 신경망(Conditional generative adversarial network; CGAN)으로 훈련한 기존 연구를 사용했다[1, 12]. 생성기의 인코더와 디코더는 3층의 CNN-attention으로 이루어져 있으며, 각 층마다 은닉층은 500차원, 출력층은 512차원으로 되어 있다. 판별기의 인코더는 생성기의 인코더와 같은 구조를 사용했다. 기타 파라미터로, dropout은 0.2, momentum은 0.99를 사용했다.

적대적 훈련을 위해서 훈련 과정의 모든 훈련 데이터에 FGSM 또는 LPM 방식으로 얻은 변동을 주었다. 변동의 크기를 제한하는 ϵ은 [0.001, 0.01, 0.1, 1]이 실험되었고, FGSM은 ϵ=0.01일 때, LPM은 ϵ=0.001일 때 가장 높은 성능을 보였다. 테스트 데이터에서 정답을 추정할 때에는 입력 데이터에 대해 변동을 사용하지 않았다.

5. 결과 및 결론

실험 결과는 표 2와 같다. 적대적 훈련을 사용하지 않은 baseline과 비교했을 때, FGSM과 LPM 모두 성능 향상이 있었다. FGSM과 LPM 두 경우 모두 baseline보다 precision은 높고 recall은 낮은 결과를 보여줬으며, 이는 적대적 훈련에 의해 과적합이 완화됨에 따라 모델 파라미터와 각 뉴런의 활성화의 크기가 작아지고, 수정되는 단어가 적어지면서 문법 교정의 결과가 보수적이 된 것이라고 추측된다.

표 2. 변동의 종류에 따른 성능 변화

Model	Precision	Recall	$F_{0.5}$
Baseline	0.6103	0.2377	0.4646
FGSM	0.6481	0.2263	0.4721
LPM	0.6543	0.233	**0.4806**

이 연구를 통해 FGSM이 모델 훈련에 일부 도움을 줄 수 있다는 것을 보일 수 있었다. 특히 훈련 데이터나 모델 복잡도를 증가시키지 않고도 훈련 방법을 통해 성능을 상승시킬 수 있다는 결과를 확인했다.

6. Acknowledgement

본 연구는 삼성 리서치의 산학협력과제의 지원을 받아 수행되었음.

참고문헌

[1] Chollampatt, Shamil, Hwee Tou Ng. "A multilayer convolutional encoder-decoder neural network for grammatical error correction." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[2] Zhao, Wei, et al. "Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data." arXiv preprint arXiv:1903.00138. 2019.

[3] Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu. "Building a large annotated corpus of learner English: The NUS corpus of learner English." Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013.

[4] Mizumoto, Tomoya, et al. "Mining revision log of language learning SNS for automated Japanese error correction of second language learners." Proceedings of 5th International Joint Conference on Natural Language Processing. 2011.

[5] Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock. "A new dataset and method for automatically grading ESOL texts." Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011.

[6] Xie, Ziang, et al. "Noising and denoising natural language: Diverse backtranslation for grammar correction." Proceedings of the 2018 Conference of the North American Chapter of the Association

for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.
[7] Ge, Tao, Furu Wei, and Ming Zhou. "Fluency boost learning and inference for neural grammatical error correction." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
[8] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572. 2014.
[9] Miyato, Takeru, Andrew M. Dai, and Ian Goodfellow. "Adversarial training methods for semi-supervised text classification." arXiv preprint arXiv:1605.07725. 2016.
[10] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." arXiv preprint arXiv:1611.01236. 2016.
[11] Ng, Hwee Tou, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, Christopher Bryant. "The CoNLL-2014 shared task on grammatical error correction." Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. 2014.
[12] 권순철, 유환조, 이근배, "적대적 생성 신경망을 이용한 문법 오류 교정", 제31회 한글 및 한국어 정보처리 학술대회 논문집. pp 488-491. 2019.