한국어 문장형 수학 문제를 활용한 심층 산술 풀이모델

김강민, 전찬준*
조선대학교
could714@chosun.kr, cjchun@chosun.ac.kr

Deep Arithmetic Solver using Korean Math Word Problem

Kangmin Kim, Chanjun Chun*
Chosun University

요 약

본 논문에서는 한국어 문장형 수학 문제 풀이를 위한 데이터 세트를 생성하고 기계 번역 모델을 학습시켜 문제 풀이모델의 성능과 데이터의 유효성을 확인한다. 문장형 수학 문제란 수학적 논리 관계가 함축된 자연어 문장을 의미한다. 이러한 문제를 풀기 위해 모델은 언어의 이해와 추론 능력이 요구된다. 해외에는 해당 연구를 위한 데이터 세트가 마련되어 있지만, 국내에는 해당 한국어 데이터 세트에 대한 공개가 부족하다. 이러한 이유로 본 논문에서는 한국어 데이터 생성기를 구현한다. 데이터 생성기는 네 가지 유형, 42가지 세부 문제 유형으로 구성된다. 학습에는 15만 개의 생성 데이터를 사용하며, 모델의 평가에는 학습 문제 유형과 상관없이 실제 수학 문제집에서 등장하는 문제 311개를 사용한다. 이를 통해 모델 성능 측정과 모델이 생성 데이터의 형식에만 과적합 되진 않았는지 확인한다. 실험 결과 본 데이터 생성기가 기계 번역 기반의 수학 문제 풀이모델에 적용 가능함을 확인하였다.

Ⅰ. 서 론

문장형 수학 문제(Math Word Problem, MWP)는 수학적 논리와 수, 그리고 자연어로 구성된 문제이다. 문장형 수학 문제 풀이모델은 이러한 문제를 입력받으면 주어진 상황을 이해하고 문장 내의 필요한 정보를 추출하여 새로운 정보인 수식을 도출해야 한다. 이를 위해서 모델은 인간처럼 다양한 도메인 지식을 습득할 수 있어야 하며, 학습한 기반 지식을 사용하여 맥락 파악과 합리적인 수식을 추론하는 및 생성하는 능력이 요구된다. 언어 능력과 추론 능력을 확인할 수 있기에 최근에는 튜링 테스트(turing test)보다 수학, 과학의 문제를 풀게 하는 것이 인공지능의 실제 지식을 평가하는 데에 더 적합하다는 연구도 보고 되었다 [1]. 이런 평가와 학습에는 주로 문장형 수학 데이터 세트가 쓰인다 [2]. 하지만 기존의 연구들은 영어와 중국어 기반의 연구가 주를 이뤘고, 한국어 데이터의 경우 연구를 위한 공개 사례가 보고되지 않았다. 본 연구에서는 한국어 기반의 문장형 수학 문제 풀이모델의 학습을 위해 데이터를 만들고 그 성능과 유효성을 확인한다.

Ⅱ. 한국어 기반의 문장형 수학 문제 데이터 세트 생성 및 학습 방법

본 논문에서는 한국어 기반의 문장형 수학 문제 풀이모델을 학습시키기 위해 한국어 데이터 생성기를 직접 구현한다. 이전까지의 연구들은 주로 국외에서 이루어지고 ASDiv-A, Math23, HMWP, Dolphin18K, MAWPs 등의 영어 또는 중국어 기반의 데이터를 이용한 문제 풀이모델에 관한 연구이다 [2-6]. 해당 연구를 위한 한국어 데이터 세트는 달리 제안된 사례가 없다. 한국어 기반으로 진행하기 위해선 앞선 형태의 데이터 세트들을 일차적으로 한국어로 번역한 후, 영미권 단위 표현, 고유명사, 문화적 차이나 익숙하지 않은 상황들을 대체 표현으로 바꾸는 등의 과정이 요구된다 [7]. 우리는 데이터 세트를 번역하는 대신, 데이터 생성기를 통해 우리가 원하는 양의 데이터를 얻고 그 데이터를 기반으로 모델을 학습시킨 후 모델의 성능과 생성 데이터의 유효성 평가를 수행한다. 생성기가 만들어 내는 문제 지문과 난이도는 초등학생 수준의 학생이 이해하고 풀 수 있는 난이도로 설정하였으며, 데이터 생성기는 표 1의 문제 유형을 포함한다. 괄호 안의 숫자는 해당 유형의 세부 예시 문제 수를 나타낸다. 데이터를 만들 때 생성기는 서술형 문장을 변형한다. 문장형 수학 문제는 유형이 같더라도 지문의 표현은 다양하다. 특히 한국어는 교착어이기 때문에 접사가 다채롭게 변할 수 있는데, 이로 인해 의미는 같지만, 표현이 달라진다 [8]. 강건한 수학 문제 풀이모델을 만들기 위해서는 모델에 가능한 많은 표현을 학습시키는 것이 중요하다. 이러한 이유로 생성기는 예시 문제에서 명사, 고유명사, 문장의 순서, 접사, 산술 연산자, 피연산자를 변형하여 문제를 생성할 수 있도록 한다.

표 1. 데이터 생성기 내에 포함된 문제 유형과 유형별 예시 문제 수

문제 유형 (44)	설명
산술 연산 (12)	주어진 특성 상황에서 연산식을 구하고 원하는 값을 구하는 유형.
순서 정하기 (7)	줄을 서는 상황이 주어지고 요구하는 값을 구하는 유형.
수 찾기 (11)	a) 문제에 미지수가 주어지고 미지수가 포함된 연산식 조건을 만족하는 미지수를 찾는 유형. b) 연산을 잘못했을 때의 상황이 주어지고, 바르게 연산했을 때의 결과를 찾는 유형.
도형 (12)	도형과 특정 값이 주어졌을 때, 넓이, 둘레, 변의 길이를 구하는 유형.

실험에 쓰인 모델은 vanilla Seq2Seq 모델, attention을 추가한 Seq2Seq 모델과 트랜스포머를 사용한다 [9-11]. 두 Seq2Seq 모델들의 인코더와 디코더에는 게이트 순환 유닛(Gated Recurrent Unit, GRU)이 사용된다. 최적화에는 Adam을 사용한다. 각 모델에 적용한 하이퍼 파라미터는 표 2와 같다. 모든 모델은 400회의 Epoch를 수행하고 early stopping을 적용하여 20회 동안 검증 손실 값의 감소가 없으면 도중에 학습을 멈추도록 한다. 모델의 학습에는 15만 개의 생성된 수학 문제와 수식을 사용하며, 검증에는 3만 개의 데이터가 사용된다. 모델의 평가는 311개의 데이터로 이루어진다. 평가 데이터는 실제 문제집에서 등장하는 문제들로만 구성한다. 평가 데이터를 뽑을 때 문제 유형을 가리지 않았기에 모델 학습에

표 2. 모델별 하이퍼 파라미터 구성

하이퍼 파라미터	Seq2Seq	Seq2Seq (with attention)	트랜스포머
임베딩 크기	[256]	[256]	[256]
은닉 상태 크기	[256]	[256]	[256]
층수	[3]	[3]	[3]
학습률	[5e-4]	[5e-4]	[5e-4]
드롭아웃	[0.4]	[0.4]	[0.4]
배치 크기	[1024]	[1024]	[256]
FFN 크기	-	-	[512]
멀티 헤드 수	-	-	[8]
파라미터 개수	5.8M	7.1M	7.4M
Epochs	400	400	66

사용된 유형뿐만 아니라 학습시키지 않은 유형도 포함되어 있다. 이는 성능 평가뿐 아니라 모델이 생성 데이터 세트에 과적합 된 상태인지를 판별 및 모르는 유형이라도 얼마나 정답에 근접할 수 있는지 확인하기 위함이다.

Ⅲ. 실험 결과

표 3. 모델별 검증 및 평가 데이터 세트에서의 결과

모델	검증 정확도(%)	평가 정확도(%)
Seq2Seq	82.11	16.39
Seq2Seq (with attention)	**89.56**	20.57
트랜스포머	89.05	**34.72**

각 모델의 실험 결과는 표 3과 같다. 평가 지표로는 정확도를 사용하였으며, 예측 수식으로 얻어낸 값이 실제 값과 일치하는지를 보여준다.

평가 데이터 세트에서 가장 높은 성능을 보인 것은 트랜스포머 모델이다. 311개 중 108개를 맞추며 34.71%의 정확도를 보여주었다. attention을 적용한 Seq2Seq 모델은 트랜스포머 모델보다 검증 데이터 세트에서는 정확도가 0.51% 더 높은 결과를 보였지만, 평가 데이터 세트에서 64개를 맞추며 트랜스포머 모델 성능보다 정확도가 14.15% 낮은 20.57%를 기록하였다. Seq2Seq 모델은 평가 데이터 세트에서 51개를 맞추어 16.39%의 정확도를 보였으며 검증과 평가 두 데이터 세트에서 가장 낮은 수치를 기록하였다. 해당 실험 결과를 통해 트랜스포머 모델이 다른 두 모델보다 표현의 변화에 비교적 강건함을 확인하였다. 본 데이터 생성기 또한 모델을 생성 데이터에 과적합 시키지 않아 실제 문제에서도 유효함을 확인할 수 있었다. 하지만 더 많은 문제 유형과 언어 표현에 대한 보충이 필요할 것으로 보인다.

Ⅳ. 결 론

본 논문은 생성한 데이터 세트와 기계 번역 모델을 이용하여 한국어 기반의 문장형 수학 문제 풀이에 적용할 수 있을지에 대한 실험을 진행하였다. 한국어 기반 수학 문제 풀이모델을 만들기 위해 직접 데이터 생성기를 구현하여 서술형 문제와 수식을 생성하였다. 모델은 Seq2Seq 모델과 attention을 적용한 Seq2Seq, 트랜스포머를 사용하였으며, 평가 결과 트랜스포머에서 34.72%로 실험한 모델 중 가장 높은 정확도를 기록하였다. 이는 데이터 생성기가 유효하며 학습된 모델이 실제 문제에서도 적용 가능한 것으로 보인다. 하지만 실험을 통해서 여전히 학습 데이터와 실제 문제 사이의 괴리가 있음 또한 확인하였다. 현재 구현된 데이터 생성기만으로는 일반화된 모델을 얻기엔 어려움이 있다. 이를 해결하기 위해서는 더욱 많은 데이터의 유형과 변형을 추가하고 모델의 크기를 확장 시킬 필요가 있다. 향후 연구에서는 상위 난도의 데이터 추가와 그에 따라 고도화된 언어모델을 이용해 문제 풀이 성능을 측정하고자 한다.

ACKNOWLEDGMENT

본 연구는 2021년 정부(과학기술정보통신부)의 재원으로 국가과학기술연구회 융합클러스터사업 (No. CCL21031-100) 및 연구개발특구진흥재단의 '기술사업화 협업 플랫폼' 사업으로 수행되었습니다. (No. 2022-DD-RD-0065)

참 고 문 헌

[1] Clark, P., and Etzioni, O. "My computer is an honor student—but how intelligent is it? standardized tests as a measure of AI," *AI Magazine,* vol. 37, pp. 5-12, Apr. 2016.

[2] Shen-yun, M., Chao-Chun, L., and Keh-Yih, S. "A diverse corpus for evaluating and developing english math word problem solvers," *in Proc. the 58th Annual Meeting of the Association for Computational Linguistics(ACL),* pp. 975-984, July, 2020.

[3] Yan, W., Xiaojiang, L., and Shuming, S. "Deep neural solver for math word problems," *in Proc. the 2017 Conference on Empirical Methods in Natural Language Processing(EMNLP),* pp. 845 - 854, Sep. 2017.

[4] Jinghui, Q., Lihui, L., Xiaodan, L., Rumin, Z., and Liang L. "Semantically-aligned universal tree-structured solver for math word problems," *in Proc. the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP),* pp. 3780 - 3789, Nov. 2020.

[5] Danqing, H., Shuming, S., Chin-Yew, L., Jian, Y. and Wei-Ying, M. "How well do computers solve math word problems? large-scale dataset construction and evaluation," *in Proc. the 54th Annual Meeting of the Association for Computational Linguistics(ACL),* pp. 887-896, Aug. 2016.

[6] Rik, K., Subhro, R., Aida, A., Nate, K., and Hannaneh, H. "MAWPS: a math word problem repository," *in Proc. the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL),* pp. 1152-1157, June, 2016.

[7] 우창협, 권가진, "한국어 수학 문장제 문제 자동 풀이," 제 30회 한글 및 한국어 정보처리 학술대회, pp. 310-315, Oct. 2018.

[8] 최정진, "한국어의 교착성과 교착소 중심의 문법," 어문학, pp. 123-155, Mar. 2014.

[9] Sutskever, I., Vinyals, O., and Le, Q. V. "Sequence to sequence learning with neural networks," *in Proc. the 27th International Conference on Neural Information Processing Systems(NeurIPS),* vol. 2 pp. 3104-3112, Dec. 2014.

[10] Luong, M. T., Pham, H., and Manning, C. D. "Effective approaches to attention-based neural machine translation," *in Proc. the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP),* pp. 1412-1421, Sep. 2015.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. "Attention is all you need," *in Proc. the 31st International Conference on Neural Information Processing Systems(NeurIPS),* pp. 6000-6010, Dec. 2017.