범용의 한국어 패러프레이즈 문장 인식 모델을 위한 연구

김민호○, 허정, 임준호

한국전자통신연구원
kimmh@etri.re.kr, jeonghur@etri.re.kr, h.kim@etri.re.kr, joonho.lim@etri.re.kr

Towards General Purpose Korean Paraphrase Sentence Recognition Model

Minho Kim○, Jeong Hur, Joonho Lim
Electronics and Telecommunications Research Institute

요 약

본 논문은 범용의 한국어 패러프레이즈 문장 인식 모델 개발을 위한 연구를 다룬다. 범용의 목적을 위해서 가장 걸림돌이 되는 부분 중의 하나는 적대적 예제에 대한 강건성이다. 왜냐하면 패러프레이즈 문장 인식에 대한 적대적 예제는 일반 유형의 말뭉치로 학습시킨 인식 모델을 무력화 시킬 수 있기 때문이다. 또한 적대적 예제의 유형이 다양하기 때문에 다양한 유형에 대해서도 대응할 수 있어야 하는 어려운 점이 있다. 본 논문에서는 다양한 적대적 예제 유형과 일반 유형 모두에 대해서 패러프레이즈 문장 여부를 인식할 수 있는 딥 뉴럴 네트워크 모델을 제시하고자 한다.

주제어: 패러프레이즈, 적대적 예제, 딥러닝 언어모델

1. 서론

패러프레이즈 문장 여부 인식 문제는 언어 이해 능력을 측정하는 주요 평가 지표에 속하는 중요한 과제이다. 또한, FAQ(Frequently Asked Question)와 같은 응용의 필수 요소 기술이기도 하다.

패러프레이즈 문장 여부 인식 문제는 두 개의 문장이 주어졌을 때 두 문장이 동등한 의미를 가지는지를 판별하는 것으로 정의할 수 있다. 예시[1]는 다음과 같다.

(문장1) 이세돌 9단이 알파고의 실수들에 대해 말하였다.
(문장2) 이세돌 9단은 알파고의 오류 가능성에 대해서 입을 열었다.
패러프레이즈 여부: O

딥러닝 뉴럴 네트워크 기반 패러프레이즈 문장 인식 모델이 보급된 이후 첫번째 걸림돌은 적대적 예제였다(적대적 패러프레이즈 예제에 대한 정의는 2장 참조). 왜냐하면 일반 말뭉치를 이용하여 학습한 인식 모델의 경우, 적대적 예제들에 대한 인식 성능이 매우 낮은 것으로 드러났기 때문이다. [1]에서 일반 유형의 데이터로 학습된 인식 모델이 동일 유형의 평가셋에 대해서는 85%의 성능을 보인 반면, 적대적 유형의 평가셋에 대해서는 48%의 성능으로 급격히 낮아졌다.

이 문제를 해결하기 위해 일반 유형과 PAWS-X [3]에 포함된 적대적 유형의 한국어 패러프레이즈 학습 말뭉치를 함께 적용하여 학습시켰을 경우, 양쪽 유형 모두에 대해 대응할 수 있는 딥러닝 인식 모델 생성의 가능성을 확인하였다[1].

하지만, 다른 유형, 예를 들어 의문형의 패러프레이즈 예제에 대해서는 12%의 성능 하락을 보였다[1].

본 연구는 위와 같은 성능 특성에서 착안하여 시작되었다. 즉, 본 연구의 동기는 '다양한 유형의 적대적 패러프레이즈에 대해서도 강건하면서 일반 유형의 패러프레이즈도 인식할 수 있는 범용의 패러프레이즈 문장 인식 모델의 개발이 가능한가' 이다.

본 논문에서는 3가지 유형의 적대적 패러프레이즈를 다룬다. 또한 2가지의 일반 유형도 다룬다. 이러한 유형들을 아우를 수 있는 인식 모델을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 다양한 적대적 패러프레이즈 예제를 다룬다. 3장에서는 인식 모델에 대해 기술하였으며, 4장에서는 실험에서 사용될 말뭉치를 정리해 두었다. 5장에서는 다양한 실험에 대한 결과와 고찰을 기술하였다. 마지막으로 6장에서 결론을 맺었다.

2. 적대적 패러프레이즈 문장

적대적 패러프레이즈 문장 예제란 두 문장에 대해 단지 1~2개를 제외한 나머지 모든 구성 성분들이 동등한 의미를 가지는 경우이다. 아래의 예제[1]에서와 같이 문장1과 문장2에서 '경찰청장'이 '질병관리청장'으로 바뀌 것만 제외하고는 다른 구성 성분은 모두 동일하다.

(문장1) 경찰청장은 아이유에게 홍보대사 임명장을 수여하였다.
(문장2) 질병관리청장은 아이유에게 홍보대사 임명장을 수여하였다.
패러프레이즈 여부: X (단일 개체 대체형 적대적 패러프레이즈)

비록 두 문장의 개별 구성성분들은 매우 유사할지 모르지만, 동등한 의미를 가지진 않는다. 즉, 인식 모델이 보기에는 두 문장의 개별 구성성분의 유사성의 매우 높기 때문에 잘못된 판단을 내릴 가능성이 높아진다. 그래서 이런 예제들이 인식 모델에 대해 적대적이라 일컬어진다.

이 예제는 특정 단일 개체가 다른 개체로 변경된 단일 개체 대체형의 적대적 유형이다.

두 번째 유형은 개체 역할 교환형의 적대적 패러프레이즈이다. 이 유형은 PAWS에서 주로 다루어졌다. 예시[1]

는 아래와 같다. 동일 문장 내의 두 개체의 위치를 서로 바꿈으로써 역할을 바꾸어준 경우이다.

(문장1) 경찰청장은 아이유에게 홍보대사 임명장을 수여하였다. (문장2) 아이유는 경찰청장에게 홍보대사 임명장을 수여하였다.
패러프레이즈 여부: X (개체 역할 교환형 적대적 패러프레이즈)

본 논문에서 다루는 3 번째 유형은 부정형 적대적 패러프레이즈이다. 예시는 아래와 같다. 이 예시는 긍정형을 부정형으로 바꾼 경우이다. 이 반대의 경우도 가능하다. 즉, 부정문을 긍정문으로 바꾼 경우도 이 유형에 포함된다.

(문장1) 경찰청장은 아이유에게 홍보대사 임명장을 수여하였다. (문장2) 경찰청장은 아이유에게 홍보대사 임명장을 수여하지 않았다.
패러프레이즈 여부: X (부정형 적대적 패러프레이즈)

3. 패러프레이즈 문장 인식 모델

본 논문에서는 BERT 기반의 딥 뉴럴 네트워크 기반한 사전 학습 언어모델, 좀 더 구체적으로는 KorBERT[4]를 이용하여 인식 모델을 구축하였다. 잘 알려져 있듯이 인식 모델에 대한 입력은 아래와 같다.

[CLS] 문장1 [SEP] 문장2 [SEP]

[CLS] 토큰에 대한 KorBERT의 마지막 레이어에서 얻은 임베딩 벡터를 선형(linear) 뉴럴 네트워크의 입력으로 제공하는 인식 모델을 적용하였다.

4. 학습 및 평가용 말뭉치

<표 1> 실험 말뭉치 구성

말뭉치 종류	TRAIN	TEST
PAWS-X(KR)	49127	1972
뉴스-PP	9665	1209
FAQ-PP	200	800
FAQ-ADV	-	1000
ADV-SUBS	2000	7999
ADV-NEG	2000	5000

본 논문에서는 총 6가지의 말뭉치를 이용하여 학습 및 평가를 시행하였다. [1]에서 다룬 4가지 말뭉치에 단일 개체 대체형과 부정형의 적대적 패러프레이즈 말뭉치를 추가하였다. 구성은 다음과 같다: 1) PAWS-X(KR), 2) 뉴스 기반 패러프레이즈 말뭉치(뉴스-PP), 3) FAQ-패러프레이즈 말뭉치(FAQ-PP), 4) FAQ-적대적 패러프레이즈 말뭉치(FAQ-ADV), 5) 단일 개체 대체형 적대적 패러프레이즈 말뭉치(ADV-SUBS), 6) 부정형 적대적 패러프레이즈 말뭉치(ADV-NEG). FAQ-적대적 패러프레이지 말뭉치(FAQ-ADV)는 의문문 형태의 단일 개체 대체형의 적대적 패러프레이즈이다. 즉, 평서문형 단일 개체 대체형 적대적 패러프레이즈(FAQ-SUBS)와 같은 부류의 단일 개체 대체형의 적대적 패러프레이즈이다. 그래서, 별도의 학습 데이터를 포함시키지 않았다. 요약하면, 학습용(TRAIN) 데이터에는 개체 역할 교환형, 단일 개체 대체형, 부정형의 3가지 유형의 적대적 예제를 다루었으며, 일반 평서문(뉴스-PP)과 의문문(FAQ-PP)의 2가지 일반형을 포함하고 있다.

5. 실험 결과

첫 번째 실험에서는 [1]에서 좋은 성능을 보여 주었던 뉴스 패러프레이즈와 PAWS-X를 통합한 말뭉치로 학습한 모델에서 다른 유형의 평가셋들이 어떤 성능을 보이는지를 평가하였다(<표 2> 결과 1번째 열).

뉴스 패러프레이즈, PAWS-X, FAQ-패러프레이즈, FAQ-적대적 패러프레이즈에 대해서는 80% 이상의 인식 정확도를 보여주었다. 하지만, 단일 개체 대체형과 부정형 적대적 패러프레이즈에 대해서는 매우 낮은 인식 정확률을 보여주었다. 사실상 적대적 예제이기 때문에 정답의 역으로 평가한다면 95%이상으로 인식하였다. 즉, 적대적 예제가 목표한 바대로 결과가 나온 것이다. 학습에 포함되지 않은 유형의 적대적 예제에 대해서는 대처하지 못한 결과를 얻었다고 할 수 있다.

<표 2> 실험 결과

		학습데이터			
	말뭉치 종류	1)+2)	1)+2)+5)+6)	1)+2)+3)+5)+6)	1)+2)+3)+5)+6)
	모델 크기	base	base	base	large
평가 데이터	**뉴스-PP**	84.53%	84.45%	83.79%	87.34%
	PAWS-X	80.45%	79.45%	79.15%	84.85%
	FAQ-PP	80.60%	75.70%	98.63%	95.75%
	FAQ-ADV	85.50%	92.80%	36.50%	85.10%
	ADV-SUBS	4.19%	92.96%	88.85%	94.96%
	ADV-NEG	3.12%	98.72%	98.88%	99.30%

* 1) PAWS-X, 2) 뉴스-PP, 3) FAQ-PP, 4) FAQ-ADV, 5) ADV-SUBS, 6) ADV-NEG

두 번째 실험은 첫 번째 실험에서 문제가 있었던 단일 개체 대체형과 부정형의 적대적 예제를 학습데이터에 추가하여 인식 모델을 학습하였다(<표 2> 결과 2번째 열). 기존 각각 4.19%와 3.12%였던 단일 개체 대체형과 부정형의 적대적 예제에 대한 인식 성능이 92.96%와 98.72%로 향상되었다. 즉, 적대적 예제에 대응할 수 있는 모델이 생성되었다고 할 수 있다. 추가되었던 학습데이터가 긍정적인 효과만 있었던 것은 아니다. FAQ-패러프레이즈에 대해서는 인식 성능이 하락하였다(-4.9%: 80.60% → 75.60%).

세 번째 실험에서는 두 번째 실험에서 성능 하락이 있었던 FAQ-패러프레이즈를 위한 학습데이터를 추가하여 인식 모델을 학습하였다(<표 2> 결과 3번째 열). 기존 75.70%에서 98.63%로 인식 성능이 22.93% 향상되었다. 이와는 정반대로 FAQ-적대적 패러프레이즈에 대해서는 급격한 성능 하락이 발생하였다(92.80%→36.50%).

이와 같은 현상은 FAQ-패러프레이즈와 FAQ-적대적

패러프레이즈가 비슷한 인식 공간을 가지고 있는데, 한쪽(FAQ-패러프레이즈)의 학습 신호만이 제공되어서 다른 쪽에 대해서 큰 악영향을 끼친 것이 아닌가 의심된다.

네 번째 실험에서는 FAQ-패러프레이즈와 FAQ-적대적 패러프레이즈의 인식 공간을 분리하기 위해서 기존 base 모델보다 더 큰 인식 모델인 large 모델을 적용하였다. 그 결과 36.50%였던 인식 성능이 85.10%로 향상되었다. 뿐만 아니라 대부분의 다른 유형에 대해서도 인식 성능이 향상되었다. 전체 통합 인식 성능 입장에서도 매크로 기준으로 89.81%의 인식 성능을 보였다. 즉, 실험에 포함된 6 가지의 문장 패러프레이즈 유형을 모두 대응할 수 있는 인식 성능을 보였다고 판단된다.

6. 결론

본 논문에서는 다양한 유형의 문장 패러프레이즈에 대해서도 강건한 성능을 보이는 범용의 인식 모델의 개발이 가능한지에 대해 여러가지 실험을 통해 시험하였다. 딥 뉴럴 네트워크, 좀 더 구체적으로는 KorBERT 사전학습 모델을 기반으로 인식 모델을 구축하였다. 이전 모델에서 낮은 성능을 보이는 각 유형의 데이터를 학습에 점차적으로 포함시키는 방법으로 접근하였다. 뉴스 패러프레이즈와 개체 역할 교환형의 적대적 패러프레이즈 말뭉치을 기반으로한 기존 연구[1]의 인식 모델은 다른 유형, 즉, 단일 개체 대체형과 부정형의 적대적 패러프레이즈 평가셋에 대해서는 정답과는 정반대로 인식하는 문제를 보였다. 앞서 언급한 4가지 유형의 말뭉치로 학습한 모델의 경우 이전 모델의 문제를 해결할 수 있었다. 하지만, FAQ-패러프레이즈 평가셋에 대해서는 성능 하락을 가져왔다. FAQ-패러프레이즈 말뭉치를 추가한 5가지 유형의 말뭉치로 학습한 인식 모델에서는 FAQ-적대적 패러프레이즈 평가셋에 대해 급격한 성능 하락을 가져왔다. 이번에는 급격한 성능 하락의 원인으로 추측되었던 패러프레이즈 인식 공간 공유의 문제를 해결하기 위해 더 큰 인식 모델을 적용하였다. 이 인식 모델은 6가지 유형의 문장 패러프레이즈 평가셋에 대해 전체적으로 높은 인식 성능을 보였다(매크로 기준 89.81%). 당초 목표였던 범용의 패러프레이즈 인식 모델의 개발을 향해 진일보할 수 있었다. 하지만, 본 논문에서 언급되지 않은 또 다른 유형에 대해서도 대응할 수 있을지, 인식 모델 자체의 변화가 필요한 것은 아닌지에 대한 추가 연구가 필요할 것으로 보인다.

* 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 [No. 2013-2-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술개발].

참고문헌

[1] 김민호, 허정, 김현, 임준호, "적대적 예제에 강건한 한국어 패러프레이즈 문장 인식 모델", 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp. 453-454, 2020.

[2] Zhang, Y ., Baldridg e, J ., & H e, L ., "PAWS: Paraphrase Adversaries from Word Scrambling", arXiv.org, 2019

[3] Yang, Y ., Zhang , Y ., Tar, C ., & Baldridg e, J ., "PAWS-X - A Cross-lingual Adversarial Dataset for Paraphrase Identification", CoRR, 2019.

[4] KorBERT: http://aiopen.etri.re.kr/service_dataset.php