적대적 예제에 강건한 한국어 패러프레이즈 문장 인식 모델

김민호○, 허정, 김현, 임준호

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

Korean Paraphrase Sentence Recognition Model Robust Against Adversarial Examples

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

요 약

본 연구는 적대적 예제에 강건한 한국어 패러프레이즈 문장 인식 기술을 다룬다. 구글에서 적대적 예제를 포함하는 PAWS-X 다국어 말뭉치를 공개하였다. 이로써, 한국어에서도 적대적 예제를 다룰 수 있는 실마리가 제공되었다. PAWS-X는 개체 교환형을 대표로 하는 적대적 예제를 포함하고 있다. 이 말뭉치만으로도 개체 교환형 이외의 적대적 예제 타입을 위한 인식 모델을 구축할 수 있을지, 다양한 타입의 실(real) 패러프레이즈 문장 인식에서도 적용할 수 있는지, 학습에 추가적인 타입의 패러프레이즈 데이터가 필요한지 등에 대해 다양한 실험을 통해 알아보고자 한다.

주제어: 패러프레이즈, 적대적 예제, 딥러닝 언어모델

1. 서론

패러프레이즈 문장 인식은 다양한 응용에서 적용될 수 있는 유용한 기술이다. 실생활에서도 많이 이용되고 있는 자주 묻는 질의 응답, 즉, FAQ(Frequently Asked Questions)가 대표적인 예이다. 패러프레이즈 문장 인식 모델 구축을 위해 영어권에서는 다양한 말뭉치가 개발되고 있다. 최근 구글에서는 패러프레이즈 문장 인식 문제의 장벽으로 여겨졌던 적대적 패러프레이즈 예제 상황을 위한 말뭉치 PAWS[1]를 개발하여 공개하였다. 기존 비적대적 패러프레이즈 인식 모델이 PAWS에 대해 영어권 기준으로 정확률 40% 이하의 성능을 보이던 모델이 PAWS를 적용하여 학습하였을 경우 85%까지 성능이 향상된다 결과를 제시해 많은 관심을 받았다[1]. 몇 개월 후 한국어를 포함한 다국어를 위한 적대적 패러프레이즈 말뭉치인 PAWS-X[2]도 공개하였다. 이로 인해 한국어에서도 적대적 패러프레이즈를 다룰 수 있게 되었다.

본 논문에서는 적대적 패러프레이즈 예제가 패러프레이즈 인식 모델의 성능에 어떻게 영향을 미치는가에 대해 다양한 실험을 통해 살펴볼 것이다. 좀 더 구체적으로, PAWS-X 말뭉치를 이용하여 딥러닝 언어모델을 학습할 경우, PAWS-X 말뭉치에서 주로 다루는 교환형 타입 이외의 적대적 패러프레이즈에 대해서도 높은 성능을 보여주는가, FAQ를 포함한 다양한 타입의 실(real) 패러프레이즈 문장을 잘 인식하는가 등에 대한 실험 결과를 제시한다. 실험을 통해서 클린 패러프레이즈 말뭉치와 PAWS-X 말뭉치의 한계와 이를 해결할 수 있는 실마리를 확인한다.

2. 적대적 패러프레이즈 예제

본 논문에서 다루는 문제는 아래의 예와 같은 표현이 다른 두 문장쌍이 주어졌을 때 동일 의미(패러프레이즈) 여부를 인식하는 것이다.

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

보통의 경우 두 문장의 어휘 공유율이 높을수록 동일 의미를 가질 가능성이 높다. 적대적 패러프레이즈 예제는 높은 어휘 공유율(Overlap)을 가지지만, 동일 의미가 아닌 다른 의미를 가지도록 만든다(아래의 예 참조).

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

위 예제는 PAWS에서 주로 다루는 양방향 개체 교환형의 적대적 패러프레이즈 예제이다. 그런데, 실제 언어 환경에는 교환형 이외의 다양한 적대적 패러프레이즈 예제가 존재한다. 예를 들어, 단일 개체 대체형 타입(아래 예제 참조), 부정형 타입, 수치 값 변형 타입 등이 존재한다.

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

3. 패러프레이즈 문장 인식 모델

PAWS 논문에서는 기존의 전통적인 언어처리 기반 알고리즘, 예를 들어 BOW(Bag Of Words) 기반의 인식 알고리즘으로는 적대적 패러프레이즈를 효과적으로 처리하지 못한다고 하였으며[1], BERT와 같은 딥 뉴럴 네트워크 언어모델 기반 인식 모델이 효과적이라고 하였다. 따라서,

본 연구에서는 딥 뉴럴 네트워크 언어 모델에 기반한 패러프레이즈 문장 인식 모델을 적용한다. 좀 더 구체적으로는 KorBERT[3] 기반의 패러프레이즈 문장 인식 모델을 이용한다.

4. 학습 및 평가용 말뭉치

본 논문에서는 4가지의 말뭉치를 이용하였다(아래 표 참조): 1) PAWS-X(KR), 2) 뉴스 기반 패러프레이즈 말뭉치(뉴스-PP), 3) FAQ-패러프레이즈 말뭉치(FAQ-PP), 4) FAQ-적대적 패러프레이즈 말뭉치(FAQ-ADV). 2), 3), 4) 말뭉치는 본 연구팀에서 자체적으로 구축하였다.

<표 1> 실험 말뭉치 구성

말뭉치 종류	TRAIN	TEST
PAWS-X(KR)	49127	1972
뉴스-PP	9665	1209
FAQ-PP		1000
FAQ-ADV		1000

5. 실험 결과

첫 번째 실험은 일반 패러프레이즈 말뭉치로 학습한 인식 모델들이 다양한 종류의 말뭉치에서 어떠한 성능을 보여주는지 확인하기 위한 것이다. 결과는 표 2에 주어져 있다. PAWS 논문에서와 유사하게 일반 패러프레이즈 말뭉치(뉴스-패러프레이즈, FAQ-패러프레이즈)에서는 우수한 성능을 보이지만 PAWS-X 와 FAQ-적대적 패러프레이즈 데이터에 대해서는 아주 낮은 성능을 보여주었다.

<표 2> 실험 결과

		학습데이터		
	말뭉치 종류	뉴스-PP	PAWS-X	뉴스-PP + PAWS-X
평가 데이터	뉴스-PP	85.19%	64.27%	84.53%
	PAWS-X	47.85%	78.90%	80.45%
	FAQ-PP	92.60%	89.30%	80.60%
	FAQ-ADV	41.20%	51.20%	85.50%

두 번째 실험은 적대적 패러프레이즈 말뭉치로만 학습한 인식 모델을 평가한 것이다(표 2 참조). 학습과 동일한 타입(개체 교환형)의 적대적 패러프레이즈 말뭉치에 대해서는 첫번째 실험 모델에 비해 월등히 높은 성능(+31.05%: 47.85% → 78.90%)을 보여 주었다. 적대적 패러프레이즈를 포함하고 있지 않는 말뭉치 중에 하나인 FAQ-패러프레이즈 말뭉치에서는 어느정도 성능이 유지되었다(-3.03%: 92.60% → 89.30%). 하지만, 뉴스-패러프레이즈 말뭉치에서는 급격한 성능 하락을 보여주었다. 또한, 단일 개체 대체형의 적대적 패러프레이즈를 다루고 있는 FAQ-적대적 패러프레이즈 말뭉치에 대해서는 약간의 성능 향상(+10.00%: 41.20% → 51.20%)이 있었다. 하지만, 해당 타입을 효과적으로 인식했다고 할 정도는 아닌 수준(51.20%)의 결과를 보여주었다.

세 번째 실험에서는 뉴스-패러프레이즈와 PAWS-X를 통합한 말뭉치로 학습한 모델을 평가하였다. 평가데이터 자신이 속한 말뭉치로 학습한 모델과 비교해보면, 뉴스 패러프레이즈 말뭉치에서는 약간의 성능 하락(-0.66%: 85.19% → 84.53%))이 있었고, PAWS-X 말뭉치에서는 약간의 성능 향상(+1.55%: 78.90% → 80.45%)이 있었다. 참고로, PAWS-X 논문[2]에서는 79.90%의 성능이 보고 되었다. 다른 타입의 말뭉치에서 학습한 모델과 비교해 보면 높은 성능 향상(뉴스-패러프레이즈:+20.26%: 64.27% → 84.53%, PAWS-X:+32.60%: 47.85% → 80.45%)을 보여주었다. FAQ-적대적 패러프레이즈 말뭉치의 경우 뉴스-패러프레이즈 모델 대비 +44.30%(41.20% → 85.50%), PAWS-X 모델 대비 +34.30%(51.20% → 85.50%)의 높은 성능 향상을 보여주었다. 매우 고무적인 결과이다. 하지만, FAQ-패러프레이즈 말뭉치의 경우 뉴스-패러프레이즈 모델 대비 -12.00%(92.60% → 80.60%), PAWS-X 모델 대비 -8.70%(89.30% → 80.06%)의 성능 하락이 있었다.

5. 결론

본 논문에서는 적대적 패러프레이즈 예제를 포함한 다양한 언어환경에서도 강건한 딥 뉴럴 네트워크 패러프레이즈 문장 인식 모델의 개발을 위한 다양한 실험적 결과를 제시하였다. PAWS 논문에서 따로 언급되지 않았던 PAWS(정확히는 PAWS-X(KR)) 말뭉치만 가지고 인식 모델을 학습했을 때의 성능 결과도 제시하였다. PAWS-X 기반 인식 모델은 자신과 FAQ-패러프레이즈 말뭉치를 제외한 나머지 말뭉치들에서는 성능이 좋지 않았다. 특히, 단일 개체 대체형 적대적 패러프레이즈 말뭉치에서는 낮은 인식 성능(51.20%)을 보였다. 이에 반해, 뉴스-패러프레이즈와 PAWS-X 통합 말뭉치 기반 인식 모델에서는 전체적으로 성능의 상승이 있었다. 하지만, FAQ-패러프레이즈 말뭉치에서는 성능 하락이 있었다. 향후 연구에서는 좀 더 다양한 타입의 적대적 패러프레이즈 말뭉치를 학습에 적용하여 모든 타입에도 강건한 일반 패러프레이즈 인식 모델을 연구하고자 한다.

* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No. 2013-0-00131, (엑소브레인-총괄/1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발].

참고문헌

[1] Zhang, Y., Baldridge, J., & He, L. (2019, April 2). PAWS: Paraphrase Adversaries from Word Scrambling. arXiv.org.

[2] Yang, Y., Zhang, Y., Tar, C., & Baldridge, J. (2019). PAWS-X - A Cross-lingual Adversarial Dataset for Paraphrase Identification. CoRR.

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