현장 데이터셋과 딥러닝 기술을 이용한 대화 utterance 유사성 판별

김주희[1], 이은서[2], 남지희[1], 고나경[3], 배상환[4], 심준호[1]
[1]숙명여자대학교 컴퓨터과학전공
[2]숙명여자대학교 일반대학원 컴퓨터과학과
[3]숙명여자대학교 통계학과
[4]네이버
{ggsmainstay2210, les97, jh00, rhskrud727}@sookmyung.ac.kr
sanghwan.bae@navercorp.com, jshim@sookmyung.ac.kr

Similarity Determination of Conversational Utterances Using Field Dataset and Deep Learning Technology

Juhee Kim[1], Eunseo Lee[2], Jeehee Nam[1], Nakyeong Koh[3], Sanghwan Bae[4], Junho Shim[1]
[1]Dept. of Computer Science, Sookmyung Women's University
[2]Dept. of Computer Science, Sookmyung Women's University, Graduate School
[3]Dept. of Statistics, Sookmyung Women's University
[4]NAVER Corp.

요 약

객체 유사도를 판별하는 기술은 정보 처리의 여러 분야에서 응용되고 있다. 본 연구에서는 현장 자연어 텍스트 데이터셋과 딥러닝 모델을 이용하여 챗봇 등에서 응용되는 데이터 유사성을 판별하고, 해당 모델의 성능을 측정해보았다.

1. 연구 배경

객체 사이의 유사도를 판별하는 기술은 정보 처리의 여러 분야에서 응용되고 있다. 유사도 판별은 크게 자연어 텍스트 데이터 유사 판별과 영상 이미지 데이터 유사 판별로 나누어 살펴볼 수 있는데, 각각 저작물의 표절 여부를 확인하는 예와 사진을 통해 인물 신원을 파악하는 예 등을 각각 대표적 사례로 살펴볼 수 있다. 최근 객체 간 유사도 측정 기술로서 딥러닝 기술이 많이 이용되고 있는데, 모델 학습 과정에서 많은 연산과 자원 소모가 발생하게 되므로 이러한 딥러닝 알고리즘은 효율적으로 동작하여야 한다. 본 연구에서는 국내 기업 NAVER의 실제 자연어 텍스트 데이터를 사용하여 대화 발화(utterance) 자연어 텍스트 데이터 사이의 유사성을 측정하는 딥러닝 모델을 구현해보고 해당 모델의 성능을 측정해보았다.

객체 간 유사성을 판별하려는 연구는 다양하게 이루어져 왔다. 대표적으로 코사인 유사도를 이용한 유사도 판별 기법은 딥러닝 기술을 사용하지 않고 두 객체 간의 유사도를 측정하는 것으로, 두 객체 벡터의 코사인 값을 이용하여 결과를 낸다 [1]. 이와는 다르게 최근에는 딥러닝 알고리즘을 사용함으로써 보다 효과적으로 객체 간 유사성을 판별하려는 시도도 있다 [2].

MRPC(Microsoft Research Paraphrase Corpus) [3]는 문장 쌍으로 구성된 대표적인 데이터셋으로, 하나의 쌍을 이루는 두 문장의 의미가 서로 유사할 경우 레이블 값이 1이며 그렇지 않을 경우 레이블 값이 0이다. 이러한 데이터셋을 이용해 딥러닝 모델을 학습시킬 경우 동일한 데이터셋의 테스트 셋에 대해 모델의 성능, 즉 정확도(accuracy), 정밀도(precision), 재현율(recall), f1 스코어(f1-score) 등을 확인할 수 있게 된다.

2. 데이터셋과 전처리

본 연구에서 사용한 데이터셋은 국내 기업 NAVER 클로바에서 공개한 케어콜 데이터셋의 시스템 발화(utterance) 쌍[4]이다. NAVER 클로바 케어콜은 돌봄이 필요한 독거 어르신들에게 전화를 걸어 안부를 묻는 서비스로, 인공지능 기술을 사용하고 있다. 인공지능은 대화를 위해 어르신에게 질문 대화 형식의 발화를 건네게 되는데, 유사한 의미를 가진 발화가 중복적으로 발생하면 이는 불필요한 연산 수행과 자원 낭비로 이어지고 서비스의 대상이 되는 어르신들 또한 대화에 불편을 느끼게 되므로 이를 사전에 검출하는 과정이 필요하다. 본 연구에서는 이러한 목적을 달성하기 위해 해당 데이터를 다음과 같이 가공하는 작업을 거쳤다.

발화의 유사성을 판별하는 문제에서, 의미적으로는 동일하지 않은데도 형태상으로만 유사한 발화 쌍을 긍정으로 판별하는 경우를 방지할 수 있어야 한다.

따라서 보다 더 높은 수준에서의 유사성을 판별하기 위해 먼저 케어콜 데이터셋의 시스템 측 발화 각각에 대해 BERT 모델을 사용했다. 이 과정으로 각각의 발화에 대해 10개의 유사한 다른 발화가 매치되었다. <그림1(a)>는 이 과정의 결과 일부를 예시로 보여준다.

	A	B
1	original sentence	candidate sentences
2	간밤에 잠은 잘 주무셨어요?	간밤에 잠은 잘 주무셨나요?
3		저야 늘 좋죠 어제 잠은 잘 주무셨어요?
4		네, 어제 잠은 잘 주무셨어요?
5		어제 밤에 잠은 잘 주무셨어요?
6		그러셨군요. 간밤에 잠은 잘 주무셨어요?
7		어제 잠은 잘 주무셨어요?
8		어제 잠은 잘 주무셨나요?
9		어제 밤에 잠은 잘 주무셨나요?
10		감사합니다. 더 주무세요.
11		네, 푹 주무셨나봐요.

<그림1(a)> 유사 발화 예시

다음으로는 이와 같이 필터링된 발화 쌍 각각에 대해서 수작업 레이블링을 진행하였다. 각각의 발화 쌍에 있는 두 객체가 의미적으로 동일하면 1을, 그렇지 않다면 0을 부여했다. 이 작업은 총 9천여개의 발화 쌍에 대하여 수행하였다. <그림1(b)>는 이 과정의 결과 일부를 예시로 보여준다.

	A	B	C
1	original sentence	candidate sentences	label
2	간밤에 잠은 잘 주무셨어요?	간밤에 잠은 잘 주무셨나요?	1
3		저야 늘 좋죠 어제 잠은 잘 주무셨어요?	1
4		네, 어제 잠은 잘 주무셨어요?	1
5		어제 밤에 잠은 잘 주무셨어요?	1
6		그러셨군요. 간밤에 잠은 잘 주무셨어요?	1
7		어제 잠은 잘 주무셨어요?	1
8		어제 잠은 잘 주무셨나요?	1
9		어제 밤에 잠은 잘 주무셨나요?	1
10		감사합니다. 더 주무세요.	0
11		네, 푹 주무셨나봐요.	0

<그림1(b)> 유사 발화 라벨링 처리 예시

3. 딥러닝 모델과 실험

본 연구에서는 두 객체 사이의 유사성을 판별하기 위해 딥러닝 알고리즘을 사용했다. 사용한 모델은 LSTM+CNN[5]으로, 컨텍스트 모델링을 위한 Bi-LSTM(Bidirectional Long Short Term Memory networks)모델에 패턴 인식을 위한 19-레이어 CNN(Convolutional Neural Network)모델을 추가한 형태를 가진다. 모델 제작 후에는 가공한 데이터셋을 학습시켜 성능 실험을 진행했는데, ADAM과 교차 엔트로피를 각각 최적화 함수와 손실 함수로 설정했으며, 0.0002의 학습률과 20의 에포크를 사용하였다[6].

모델 성능 평가 지표로는 정확도와 f1 스코어를 사용했으며, 반복 실험 수행 후 평균 성능치 결과는 <표1>과 같다.

	정확도	f1 스코어
성능치(%)	82.62	74.62
표준편차	0.0044	0.0057

<표1> 모델 성능 측정 결과

4. 결론

본 연구에서는 현장 자연어 텍스트 데이터셋과 딥러닝 모델을 이용하여 자연어 텍스트 객체 간 유사도를 판별하는 작업을 수행했다. 딥러닝 모델을 이용한 객체 간 유사도 측정은 모델 학습 과정에서 많은 연산과 자원 소모가 발생하게 된다. 따라서 추후 연구로는 적은 데이터셋으로도 해당 딥러닝 모델을 효율적으로 동작시킬 수 있는 방법에 대해 실험할 것이다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF2022H1D8A303739411)

참고문헌

[1] 이동주, 심준호, “대수적 특성을 고려한 벡터 유사도 측정 함수의 고찰”, 한국전자거래학회지, 제17권 제4호, 한국전자거래학회, 2012.

[2] Y. Kim, H.-J. Lee, and J. Shim, “Developing Data-conscious Deep Learning Models for Product Classification”, Applied Sciences, Vol. 11, Issue 12, MDPI, 2021.

[3] B. Dolan and C. Brockett, “Automatically Constructing a Corpus of Sentential Paraphrases”, Third International Workshop on Paraphrasing(IWP 2005), 2005.

[4] 곽동현, 배상환, 함동훈, “세상 빠르고 안전한 챗봇 만들기”, DEVIEW 2021, https://deview.kr/2021/sessions/474, Naver Corp.

[5] H. He and J. Lin, “Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement”, In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[6] 김주희, 남지희, 심준호, “딥러닝 텍스트 분류기에서의 Square Loss와 Cross-Entropy의 실증적 평가”, 한국전자거래학회 2022년 춘계학술대회, 2022.