SimKoR: 한국어 리뷰 데이터를 활용한 문장 유사도 데이터셋 제안 및 대조학습에서의 활용 방안

김재민*1, 나요한*2, 김강민1, 이상락2, 채동규O,2

한양대학교 인공지능학과1
한양대학교 컴퓨터·소프트웨어학과2
{jaemink, nayohan, kevin7133, sangrak, dongkyu}@hanyang.ac.kr
*공동 1저자, O교신저자

SimKoR: A Sentence Similarity Dataset based on Korean Review Data and Its Application to Contrastive Learning for NLP

Jaemin Kim*1, Yohan Na*2, Kangmin Kim1, Sang Rak Lee2, Dong-Kyu Chae2
Dept. of Artificial Intelligence, Hanyang University, South Korea1
Dept. of Computer Science, Hanyang University, South Korea2

요 약

최근 자연어 처리 분야에서 문맥적 의미를 반영하기 위한 대조학습 (contrastive learning) 에 대한 연구가 활발히 이뤄지고 있다. 이 때 대조학습을 위한 양질의 학습 (training) 데이터와 검증 (validation) 데이터를 이용하는 것이 중요하다. 그러나 한국어의 경우 대다수의 데이터셋이 영어로 된 데이터를 한국어로 기계 번역하여 검토 후 제공되는 데이터셋 밖에 존재하지 않는다. 이는 기계번역의 성능에 의존하는 단점을 갖고 있다. 본 논문에서는 한국어 리뷰 데이터로 임베딩의 의미 반영 정도를 측정할 수 있는 간단한 검증 데이터셋 구축 방법을 제안하고, 이를 활용한 데이터셋인 SimKoR (**Sim**ilarity **Ko**rean **R**eview dataset) 을 제안한다. 제안하는 검증 데이터셋을 이용해서 대조학습을 수행하고 효과성을 보인다.

주제어: 자연어처리, 대조학습, 평가지표, 감정분석, 평가 데이터셋

1. 서론

최근 자연어의 문맥적 의미를 반영한 임베딩을 이용해서 감정분석, 질의응답 등 다양한 자연어처리의 응용 분야에서 성능 향상을 이루고 있다. 자연어처리 분야에서 임베딩이란 단어, 문장 등의 의미를 반영하여 고차원의 공간에 있는 벡터로 변환하는 것이다. 이 때 임베딩에 문맥적 의미를 반영하기 위해 여러가지 연구가 활발히 진행되고 있다. 그 중 대표적인 방법으로 대조학습 (contrastive learning)이 있다. 대조학습은 의미적으로 유사한 데이터들이 임베딩 공간에서도 가까워지도록 학습하는 방법이다.

SimCSE [1]는 유사한 데이터가 임베딩 공간에서 가까워지도록 학습할 뿐만 아니라 의미가 전혀 유사하지 않는 데이터는 멀어지도록 학습하는 방법을 통해 성능 향상을 이루었다. 여기서 학습에 중요한 영향을 미치는 요소는 의미적으로 가까워져야 할 쌍과 멀어져야 할 쌍을 구성하는 것이다. SimCSE의 저자들은 가설 문장과 전제 문장의 의미 관계가 얽힘 (entailment), 중립 (neutral), 모순 (contradiction) 범주로 레이블을 갖고 있는 NLI [2] 데이터를 이용해서 관계가 얽힘의 경우 가까워지도록, 모순일 경우 멀어지도록 학습하였다. 특히 학습 과정에서 검증 (validation) 에 필요한 성과 지표로써 문장의 유사도를 레이블로 갖고 있는 STS [2] 를 검증 데이터셋으로 활용하였다.

한국어 자연어처리를 위한 대조학습의 경우 KoSimCSE [3] 가 있는데, 이 연구에서는 NLI데이터 대신 KorNLI [4] 를, STS 데이터 대신 KorSTS [4] 를 활용하였다. 그러나 KorNLI와 KorSTS 데이터들은 영어를 한국어로 단순 번역한 것으로 기계번역의 성능에 의존한다는 문제점을 갖고 있다.

본 논문은 KorSTS 대신 활용될 수 있는 한국어 리뷰 문장 유사도 데이터셋인 SimKoR를 제안한다. SimKoR의 구성에는 기계번역이 사용되지 않았으며, 기존의 한국어 리뷰 데이터셋을 활용하여 한국어 의미의 레이블로 한국어 유사도 데이터셋을 구성할 수 있었다. 제안하는 데이터셋의 우수성을 평가하기 위해 KoSimCSE를 학습시키는 과정에서 기존의 KorSTS 대신 SimKoR를 validation 데이터로 활용해서 최종 성능을 평가하였다. 실험 결과 SimKoR은 KorSTS에 비해 사람의 관여가 적게 들어간 데이터임에도 불구하고 사용되었을 때 비슷하거나 더 좋은 결과를 도출하였다. 본 논문에서 제안하는 SimKoR 데이

터셋을 아래의 링크에 공개하였다.
(Github: https://github.com/nayohan/SimKor).

2. 관련 연구

2.1. 한국어 자연어 추론 (KorNLI) 데이터셋

자연어 추론 (**NLI, Natural Language Inference**) 이란 한 쌍으로 주어진 전제 (premise) 문장과 가설 (hypothesis) 문장이 의미적으로 얽힘 (entailment), 중립 (neutral), 모순 (contradiction) 관계인지를 예측하는 문제이다. 영어 기반의 NLI 데이터셋으로는 SNLI [5], MNLI [6], XNL [7] 등이 있다.

KorNLI [4] 는 해당 문제를 한국어 문장으로 학습 및 평가할 수 있는 데이터이다. KorNLI는 영어 문장으로 이루어진 기존 데이터인 SNLI, MNLI, XNLI를 기계번역을 통해 한국어로 번역한 것이다. SNLI, MNLI는 학습 (training) 데이터로, XNLI는 데이터의 일관성을 위해 전문 번역가의 검토 후 평가 (test) 데이터로 활용하였다.

KLUE-NLI의 경우, 가설 생성 작업자들이 가설을 생성한 후 4명의 검증 작업자가 생성된 문장 쌍의 의미 관계를 판별하여 관계 레이블로 활용한 데이터셋이다.

2.2. 한국어 텍스트 의미적 유사성 (KorSTS, Korean Semantic Textual Similarity) 데이터셋

텍스트 의미적 유사성 (STS, Semantic Textual Similarity) 은 두 문장 사이의 의미적 유사성을 0 (전혀 유사하지 않음) 에서 5 (완벽히 일치함) 사이의 연속형 수치로 예측하는 문제이다. 영어로 된 데이터셋으로는 STS, STS-B 등이 있다.

KorSTS [4] 는 해당 문제를 한국어로 학습 및 평가할 수 있는 데이터셋이다. 마찬가지로 영어로 된 기존의 STS-B를 기계번역을 통해 한국어로 번역하여 약 66%는 학습 데이터로, 약 17%는 검증 데이터, 17%의 문장은 전문 번역가의 검토 후 평가 데이터로 활용하였다.

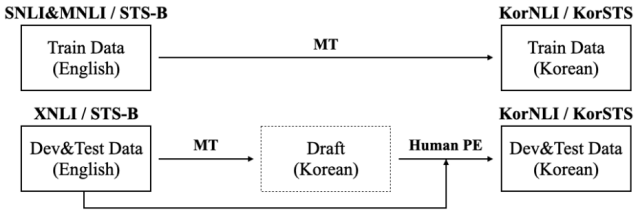

그림 1. KorNLI & KorSTS 데이터 구축 흐름 [2]

KLUE-STS는 에어비앤비 리뷰, 정책 뉴스 브리핑, 스마트 홈 기기 등의 발화 데이터를 활용하였으며, 문장 간 유사성을 추정하기 힘든 데이터는 기계번역을 통해 유사한 문장 쌍을 생성하였다.

2.3. KoSimCSE

자연어처리 분야에서 대조학습은 문맥적 의미가 동일한 데이터는 (**양성 샘플**) 임베딩 공간에서 가까워지도록 학습하고, 문맥적 의미가 비슷하지 않은 데이터 (**음성 샘플**) 끼리는 서로 멀어지도록 학습하는 방법이다.

SimCSE [1] 는 대표적인 자연어처리를 위한 대조학습 방법 중 하나이다. SimCSE는 양성 샘플, 음성 샘플, 극 음성 샘플 (hard negative pair) 을 구성한 후 양성 샘플은 가까워지고, 극 음성 샘플은 멀어질 수 있도록 학습하는 기법을 제안하였다. 이러한 샘플들을 구성하기 위해 NLI 데이터셋을 활용하였고, 성능 지표를 계산하기 위한 검증 데이터셋으로 STS를 활용하였다. **KoSimCSE** [3] 는 SimCSE와 동일한 프레임워크를 가지며 한국어로 사전학습된 모델을 활용하는 방법이다.

2.4 감성 분석용 말뭉치 [8]

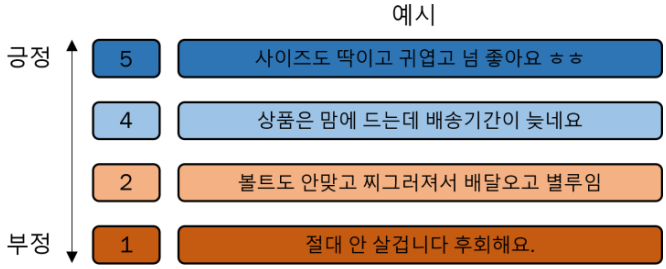

그림 2. 감성 분석용 말뭉치의 구성

해당 데이터는 네이버 쇼핑에서 제품별 후기를 평점과 함께 수집한 데이터셋으로, 긍정/부정으로 분류하기 애매한 3점에 해당하는 리뷰들을 제외한 나머지 데이터로 구성되어 있다 ([그림 2] 참고). 평점 4~5점을 긍정, 평점 1~2점을 부정으로 구분하였으며, 구체적으로 5점 81,777개, 4점 18,786개, 2점 63,989개, 1점 36,048개의 리뷰 텍스트들로 구성되어 있다. 총 20만 개의 리뷰 텍스트를 포함하고 있다.

3. 제안하는 방법

본 장에서는 이전 장에서 소개한 감성 분석용 말뭉치 데이터셋 [8] 에 구성되어 있는 평점 정보 (1,2,4,5점) 를 활용하여 문장 간의 유사도를 포함하는 데이터셋인 SimKoR를 구축하는 방법을 다룬다. 4개 종류의 평점들 중 4, 5 점은 긍정 의견을 의미하고, 1, 2 점은 부정적 의견을 의미한다고 볼 수 있다. 만약 서로 다른 문장 2개를 무작위로 추출할 경우, 각 문장에 부여된 평점에 따라 [그림 3]과 같이 16개의 유형으로 나눌 수 있다. 각 유형들은 문장들 간의 평점 차이에 따라 5개의 유사도 스코어 레이블로 정의될 수 있다. 이렇게 레이블링 된 데이터셋은 추후 한국어를 위한 대조학습 시 검증용 데이터로 활용할 수 있다.

3.1 데이터셋 생성 방법

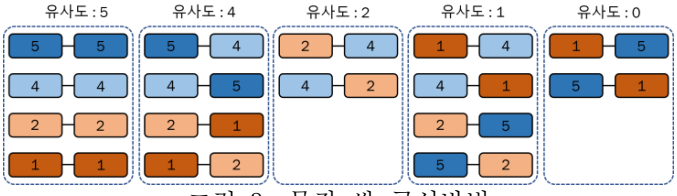

그림 3. 문장 쌍 구성방법

먼저 데이터가 갖고 있는 유사도 쌍을 활용하여 [<텍스트1>, <텍스트2>, <레이블>] 형태로 데이터셋을 구성하였다. 여기서 레이블은 평점을 기반으로 자동 생성된다. 즉, 기존의 KorSTS 데이터셋은 영어 문장의 유사도에 초점을 맞춰서 점수를 사람이 작성했다면, SimKoR데이터셋

은 평점 간의 차이를 활용하여 사람의 개입 없이도 기계적으로 레이블링 된 데이터셋을 제작할 수 있다.

평점 간의 차이를 기준으로 부여되는 유사도 레이블의 경우 다음과 같은 규칙을 적용하였다. 평점의 차이가 4일 경우 감정의 차이가 가장 큰 경우 (하나는 강한 긍정에 해당되는 5점, 다른 하나는 강한 부정에 해당되는 1점) 에 해당되므로 유사도 0을 부여한다. 평점 차이가 3점인 경우 유사도 1을, 평점 차이가 2점인 경우 유사도 2를 부여한다. 평점의 차이가 2점~4점인 경우 두 문장의 감정이 서로 달랐다면 (하나는 긍정, 다른 하나는 부정), 평점의 차이가 1점인 경우부터는 두 문장이 같은 감정에 해당된다. 평점의 차이가 1점일 경우 유사도 3이 아닌 4를 부여하는데, 그 이유는 두 문장 간의 감정이 같을 경우와 다를 경우의 차이를 명확하게 하기 위함이다. 마지막으로 평점의 차이가 0점인 경우 가장 높은 유사도 레이블인 5를 부여한다. 위 규칙을 요약하면 아래와 같다:

- 평점 차이:4점 -> 유사도: 0
- 평점 차이:3점 -> 유사도: 1
- 평점 차이:2점 -> 유사도: 2
- 평점 차이:1점 -> 유사도: 4
- 평점 차이:0점 -> 유사도: 5

문장들을 샘플링하는 방법은 우선 매 회마다 각 유형별 하나씩, 총 16개의 쌍을 생성한다. 그 결과 [그림 3]에서 볼 수 있듯이 유사도 레이블 간의 불균형이 발생한다. 즉, 유사도가 2 또는 0이 되는 경우가 다른 유사도 레이블에 비해 적게 발생하므로, 레이블 별 샘플의 개수를 맞추기 위해 유사도가 2 또는 0이 되는 텍스트 쌍에 대해서는 두배 샘플링을 진행하였다.

위와 같은 방법으로 SimKoR을 구성하는 학습, 검증, 테스트 데이터셋 (6:2:2의 비율) 을 각각 생성하였다. 학습 데이터의 경우 각 유사도 레이블 당 4,000개씩 생성하여, 총 20,000개의 학습 데이터를 중복 없이 구성하였다. 검증 데이터와 평가 데이터 또한 각각 5,000개씩 생성하였다. 타 한국어 데이터셋 대비 SimKoR이 포함하고 있는 데이터의 수는 아래의 [표 1] 에 비교되어 있다.

	KorSTS	KLUE-STS	SimKoR
Train	5,749	11,668	20,000
Validation	1,500	519	5,000
Test	1,379	1037	5,000

표 1. 각 데이터셋 별 데이터 개수

3.2 데이터셋 활용

SimKoR데이터셋은 두 문장 간의 평점 차이를 거리 정보로 활용하여 생성한 문장 유사도 데이터 셋으로, 기존의 단순 문장 유사도의 데이터셋과는 달리 문장의 감정을 확인할 수 있는 유사도 데이터 셋을 구축하였다. 기존 문장 유사도 데이터셋과 혼합하여 문장 유사도와 감정 이해 두가지를 모두 평가하는 데이터셋으로도 활용할 수 있으며, 대조학습 시 검증 데이터로 활용되는 KorSTS를 대체하여 언어 모델이 문장을 잘 이해하고 있는지를 평가 수단으로 활용할 수 있다. 또한 기존 데이터셋에 비해서 매우 쉽게 구축이 가능하며 기계번역 등 외부 요소에 의존하지 않는 장점이 있다.

4. 실험

본 논문은 KLUE [9] 의 RoBERTa-Base, BERT-Base를 활용하여 실험을 진행하였다. 해당 모델들을 베이스라인으로 활용하여, 제안하는 SimKoR데이터셋을 평가하고, 추가적으로 대조학습을 위한 검증데이터셋인 KorSTS데이터셋을 대체하여 활용해보았다.

4.1 실험 설계

한국어를 위한 대조학습은 KorNLI데이터셋을 학습 데이터셋으로 활용한다. 우리는 검증 데이터셋인 KorSTS와 비교실험을 진행하기 위해 KorNLI를 학습 데이터셋으로 고정한 후, KorSTS 대신 제안하는 SimKoR데이터셋을 검증 데이터셋으로 활용하여 성능을 평가하였다. 실험 조건은 조기종료 (early stop) 를 5 epoch 으로 설정하였으며 validation을 위한 지표로서 임베딩 벡터 간의 코사인 유사도를 이용하였다. 코사인 유사도의 최고점이 갱신될 때 마다 모델이 저장되게 하였으며, 5 epoch 이상 진행되었음에도 최고점이 갱신되지 않는다면 early stop 조건을 만족한 것으로 학습을 종료한다.

SimKoR데이터셋을 대조학습의 검증데이터셋으로 활용하여 BERT와 RoBERTa 모델의 학습을 완료한 후, 감정 분석용 말뭉치 데이터셋에 대해서 지도학습 기반의 fine-tuning을 진행하였다.

4.2 실험 결과

우리는 SimkoR가 대조학습에서의 검증 데이터셋으로서의 역할과 새로운 한국어 감정분석 데이터셋으로서의 연구적 발전을 위해 기준점을 제시하기 위한 실험을 진행했다. 대조학습의 validation 데이터셋으로 SimKoR, KorSTS, Mixed를 각각 활용해서 모델을 학습시킨 후, 두 가지의 감정분석 데이터셋에 fine-tuning 하여 실험하였다. 여기서 Mixed는 SimKoR과 KorSTS를 모두 활용했다는 것을 의미한다.

4.2.1 감성 분석용 말뭉치에 대한 fine-tuning 결과

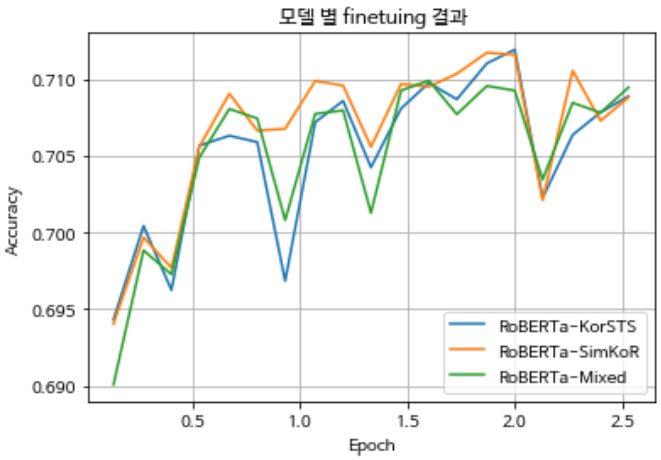

그림4. 모델 별 감성 분석용 말뭉치 fine-tuning 결과

Model	Training method	Valid dataset for CL	Accuracy
BERT-Base	CL+ Finetuning	KorSTS	0.711
	CL+ Finetuning	SimKoR	0.712
	CL+ Finetuning	Mixed	0.711
RoBERTa-Base	CL+ Finetuning	KorSTS	0.711
	CL+ Finetuning	SimKoR	0.711
	CL+ Finetuning	Mixed	0.709

표 2. 검증데이터셋 별 대조학습한 모델의 감정 분석용 말뭉치 fine-tuning 결과. CL은 contrastive learning의 약자이다.

감정분석 말뭉치 데이터로 fine-tuning을 수행하였다. 이에 대한 실험 결과는 [표 2]에 정리되어 있다. [표 2]에서 볼 수 있듯이 SimKoR을 사용했을 경우와 KorSTS를 사용했을 경우에 비슷한 성능을 내고 있음을 볼 수 있다. 즉, 기존의 KorSTS가 하던 대조학습의 검증데이터셋 역할을 SimKoR이 동일하게 해내고 있는 것을 볼 수 있다. SimKoR이 감정과 감정에 대한 크기를 기반으로 구성된 데이터셋인 만큼, 감정분석에 적절한 임베딩을 학습하기에 적합한 검증 데이터셋임을 의미한다.

4.2.2 SimKoR에 대한 모델 별 fine-tuning 결과

앞선 실험과 같이 validation 데이터셋을 활용하여 RoBERTa-Base의 대조학습 및 SimKoR에 대한 fine-tuning을 수행하고 정확도를 비교해보았다.

Model	Contrastive Learning			Accuracy
	Valid dataset	# of Eval-steps	Validation Score	
RoBERTa-Base				0.409
	KorSTS	3,000	0.439	0.428
	SimKoR	1,500	0.441	0.430
	Mixed	1,750	0.4406	0.434

표 3. 검증데이터셋 별 대조학습한 모델의 SimKoR fine-tuning 결과.

실험 결과, [표 3]에서 볼 수 있듯이 단순 RoBERTa-Base 모델의 SimKoR에 대한 정확도 수치는 0.4096이다. 반면, 대조학습을 적용한 후 결과를 살펴보면 검증 데이터셋 종류에 관계 없이 전부 성능이 향상됨을 보여 대조학습의 필요성을 알 수 있었다. KorSTS와 SimKoR를 검증 데이터셋으로 활용한 모델 결과를 비교하였을 때, SimKoR를 통해 효율적인 학습이 가능함을 보여주었다. 그 이유는 [표 3]의 Evaluation-step에서 확인할 수 있다. 대조학습의 검증 데이터셋을 KorSTS로 활용하였을 때 3,000번의 스텝을 학습해야 하는 반면 SimKoR을 활용했을 경우 1,500 스텝만으로도 학습을 종료하며, 성능 향상도 이뤄냈다. Mixed의 경우 가장 좋은 성능을 보였는데 이는 텍스트의 의미적 측면과 감정적 측면을 모두 잘 잡아냈다고 해석할 수 있다.

5. 결론 및 향후 연구 방향

제안한 SimKoR데이터셋은 실제 사용자의 평점 및 리뷰 텍스트 데이터를 기반으로 구축된 문장 유사도 레이블 데이터셋이다. 기존의 KorSTS 데이터셋에 비해 기계번역의 성능에 의존적이지 않으며, 훨씬 적은 비용으로 기계적인 구축이 가능하다는 장점이 있다. 또한 KorSTS 데이터셋이 동일한 문장 간의 의미가 유사한지를 평가하는 것에 더해 우리의 SimKoR데이터는 문장의 감정이 유사한지를 평가할 수 있다. 추후 해당 데이터셋을 활용하여 문장 유사도 문제를 풀 수 있을 것으로 기대된다.

최근 자연어처리에서 큰 이목을 차지하고 있는 대조학습을 한국어에 적용하기 위한 측면에서도 본 연구가 기여할 수 있다. 예를 들면 문장이 가지는 의미뿐만 아니라 해당 문장의 감정을 동시에 평가하여 감정분석을 위한 자연어 임베딩의 평가 지표 계산에 활용될 수 있다. 결론적으로 SimKoR를 통해 감정분석을 위한 대조학습 등의 다양한 한국어 자연어처리 모델 연구가 활발히 진행되기를 기대한다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 (1) 정보통신기획평가원의 지원(No.2020-0-01373, 인공지능대학원지원(한양대학교))과 (2) 한국연구재단의 지원을 받아 수행된 연구임 (No.2021R1A2C1094863)

참고문헌

[1] Tianyu Gao et al., “SimCSE: Simple Contrastive Learning of Sentence Embedding”. In EMNLP. 2021.
[2] Alex Wang et al., “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.”, In EMNLP findings. 2018.
[3] B. Kim et al., “Simple Contrastive Learning of Sentence Embeddings using SKT KoBERT”, https://github.com/BM-K/KoSimCSE-SKT
[4] Jiyeon Ham et al., "Kornli and korsts: New benchmark datasets for korean natural language understanding.". In EMNLP findings. 2021.
[5] Samuel R et al., “A large annotated corpus for learning natural language inference.”, In EMNLP findings. 2015.
[6] Adina Williams et al., “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.” In NAACL. 2018.
[7] Alexis Conneau et al., “XNLI: Evaluating Cross-lingual Sentence Representations.” In EMNLP findings. 2018.
[8] 이민철, 감성 분석용 말뭉치, https://github.com/bab2min/corpus/tree/master/sentiment
[9] Sungjoon Park et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).