문장 유사도를 이용한 다양한 표현의 패러프레이즈 생성

박다솔[O1], 장두성[2], 차정원[3]

창원대학교[1,3], KT[2]

{dasol_p[1],jcha[3]}@changwon.ac.kr, dschang[2]@kt.com

Various Paraphrase Generation Using Sentence Similarity

Da-Sol Park[O1], Du-Seong Chang[2], Jeong-Won Cha[3]

Changwon National University[1,3], KT[2]

요 약

패러프레이즈란 어떤 문장을 같은 의미를 가지는 다른 단어들을 사용하여 표현한 것들을 의미한다. 이는 정보 검색, 다중 문서 요약, 질의응답 등 여러 자연어 처리 분야에서 중요한 역할을 한다. 특히, 양질의 패러프레이즈 코퍼스를 얻는 것은 많은 시간 및 비용이 소요된다. 이러한 문제점을 해소하기 위해 본 논문에서는 문장 유사도를 이용한 패러프레이즈 쌍을 구축하고, 또 구축한 패러프레이즈 쌍을 이용하여 기계 학습을 통해 새로운 패러프레이즈을 생성한다. 제안 방식으로 생성된 패러프레이즈 쌍은 기존의 구축되어 있는 코퍼스 내 나타나는 표현들로만 구성된 페러프레이즈 쌍이라는 단점이 존재한다. 이러한 단점을 해소하기 위해 기계 학습을 이용한 실험을 진행하여 새로운 표현에 대한 후보군을 추출하는 방법을 적용하여 새로운 표현이라고 볼 수 있는 후보군들을 추출하여 기존의 코퍼스 내 새로운 표현들이 생성된 것을 확인할 수 있었다.

주제어: 기계학습, 자연어처리, 패러프레이즈, 윤문생성

1. 서론

챗봇은 채팅(Chatting)과 로봇(Robot)이 결합되어 만들어진 용어로써, 인간의 자연 언어를 이해하고 대화가 가능한 채팅 로봇이다. 챗봇은 기본적으로 사용자가 입력한 질문을 알아듣고 챗봇이 답변하는 '요청과 응답(Request-Response)'의 구조로 동작한다. 챗봇이 사용자가 입력한 텍스트를 분석하여 의도를 파악하는 것이 챗봇의 핵심 중에 하나이다.

사용자 의도를 위한 요청에는 사람마다 다양한 표현을 할 수 있다. 사용자의 의도가 포함된 발화를 표현하는 방법은 여러 가지가 있고 이러한 표현이 '알려주다'와 '찾아주다'와 같이 단어 혹은 짧은 구 단위이면 두 표현은 동의어(Synonym)라고 한다. 반면 표현이 'A에 대해 알려주세요'와 'A에 대해 찾아주세요'와 같이 문장 단위인 것을 패러프레이즈(paraphrase)라 한다. 즉, '패러프레이즈'란 어떤 문장을 같은 의미를 가지는 다른 단어들을 사용하여 표현한 문장들을 말한다.

챗봇의 품질을 향상시키기 위해 동일한 의미를 갖는 다양한 형태의 문장인 패러프레이즈를 이용한 학습이 필요하다. 이에 대한 일환으로 특정 문장에 대한 다양한 유사 문장을 생성하고, 생성된 유사 문장을 기계에 학습시키는 방안이 고려되어야 한다.

양질의 패러프레이즈 코퍼스를 수작업으로 얻는 것은 많은 시간 및 비용이 소요된다. 이러한 문제점을 해소하기 위해 본 논문에서는 문장 유사도를 이용해 패러프레이즈 쌍을 구축하고, 또 구축한 패러프레이즈 쌍을 이용하여 기계 학습을 통해 새로운 패러프레이즈를 생성하는 방법을 제안한다.

최근 언어 사전학습(Pre-training) 기법이 개발되면서 자연어처리 분야에는 큰 진전이 있었으며 BERT[1], RoBERTa[2], XLNet[3], T5[4], Generative Pre-training(GPT)[5]과 같은 최첨단 모델이 크게 향상되었고, 최근 이를 이용한 문장 생성 연구에 관한 다양한 연구가 진행되었다.

2. 제안 방법

본 논문에서는 패러프레이즈에 해당하는 기준을 다음과 같이 단어 치환, 구 치환, 문장 구조 변경으로 총 3가지로 구분하였다. 아래는 각 기준에 따른 패러프레이즈 예제이다.

1) 단어 치환 : 아침에 일어나자마자 좋아하는 노래를 들었다. → 아침에 일어나자마자 좋아하는 음악을 들었다.
2) 구 치환 : 날씨가 쌀쌀해서 온수로 샤워를 했다. →

날씨가 쌀쌀해서 뜨거운 물로 샤워를 했다.

3) 문장 구조 변경 : 민수는 커피를 마시고, 지수는 함께 앉아 있었다. → 지수와 함께 앉아 있는 민수는 커피를 마신다.

단어 치환은 동의어 치환 및 어미의 표현들을 다른 표현으로 생성한 경우를 의미하고, 구 치환은 단어를 구와 같이 단어 뜻과 정의를 풀어 사용하는 경우 또는 구와 구 끼리의 치환을 의미한다. 마지막으로 문장의 구조를 변경하여 나타내는 경우를 말한다.

기존의 패러프레이즈 생성은 문장 유사도를 이용하여 패러프레이즈 쌍을 구축하였다[6,7]. 패러프레이즈 3가지의 기준을 복합적으로 나타내고 있으며, 문장의 길이나 복잡도로 인해 실제 입력과 출력의 두 문장이 동일하지 않더라도 유사해질 수 있다.

본 논문에서는 단어 치환을 목표로 하였다. 표 1 내 입력 문장 내 어미와 출력 문장 내 어미가 본 연구에서 설정한 어미 치환이라고 설정하였다. 그리고 그림 1은 원시 코퍼스를 이용하여 패러프레이즈 생성의 전체 구조도이다.

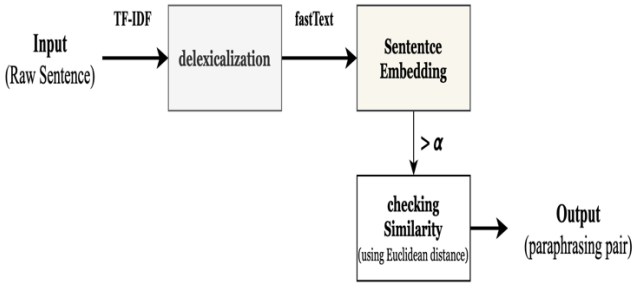

그림 1 패러프레이즈 생성을 위한 전체 구조도

어미 치환 및 다양한 어미 표현에 대한 패러프레이즈 생성을 위해 우리는 비어휘화를 진행한다. 비어휘화란 특징을 나타낼 수 있는 단어를 제외하고 문장의 표현을 강조하기 위함이다. 표 1는 입력 문장 'A 이용료 요금이 나왔는데 그게 뭔가요'[1]에 대해 가장 유사한 3 문장을 추출한 예시이다. 분류 1과 2는 추출한 문장의 전,후처리에 따라 분류하였다. 분류 1은 입력 문장과 후보 문장들에 대해서 전처리 없이 유사도를 계산하였고, 분류 2는 TF-IDF를 이용하여 특정 서비스명 또는 상품명에 대해 특수 토큰으로 치환하여 유사도를 계산한 후 다시 어휘화로 복구하는 작업을 후처리로 진행하였다. 문장 유사도는 fastText[8]를 이용하였으며, 해당 유사도는 이후 상세하게 설명한다.

표 1 입력 문장 'A 이용료 요금이 나왔는데 그게 뭔가요'에 대한 유사도 높은 3문장의 예시

분류	문장 유사도 top-3 문장
1	A 이용료 요금이 뭔가요
	A 이용료예요
	저번 달 B 이용료 요금 얼만가요
2	이번달 요금 중 A 이용료가 뭐예요
	A 이용료라고 나왔는데 그게 뭐예요
	A 이용료에 대해서 문의하고 싶어요

표 1의 결과를 보면, 분류 1에서는 다양한 서비스명에 관련된 문장이 높은 유사도를 가지며, 또 입력 문장에 포함된 동일한 서비스명이 포함된 문장들이 실제 유사하지는 않지만, 문장 벡터와의 계산을 통해 높은 유사도를 가지는 결과를 확인할 수 있었다. 또, 분류 2에서는 입력 문장과 동일한 서비스명이 포함되고, 다양한 어미들의 표현들이 높은 유사도를 가지기 때문에 추출된 문장들은 실제 입력 문장과 유사한 문장들임을 확인할 수 있었다.

이러한 결과를 토대로 어미에 관련된 다양한 표현을 학습하기 위해서 코퍼스 내 특징적인 단어들을 비어휘화하는 것이 필수적이며, 코퍼스 내 TF-IDF를 이용하여 특징적인 단어들을 추출한다. 추출된 단어를 동일한 특수 토큰인 〈spc〉으로 치환하여 비어휘화한 문장을 출력한다. 표 2는 비어휘화 전후 문장 예시이며 표 내 A,B,C는 서비스명 및 상품명을 나타낸다.

표 2 비어휘화 적용 전후 문장 예시

비어휘화 전 문장	비어휘화 후 문장
A가 안 돼요	〈spc〉가 안 돼요
집에 있는 B는 해지해줘	집에 있는 〈spc〉는 해지해줘
저 C 차단 좀 해주세요	저 〈spc〉 차단 좀 해주세요
D 여부 확인해 주세요	〈spc〉 여부 확인해 주세요

이후, 비어휘화된 문장을 통한 문장 벡터를 추출한다. 한국어가 조사, 어미 등을 붙여서 말을 생성하는 언어인 교착어라는 점과 설정 도메인에 대한 언어 모델을 학습하는 것에 초점을 두어 본 논문에서는 문장 벡터를 쉽게 학습 및

[1] 보안 문제로 인해 서비스 이름을 A로 변경함

생성할 수 있는 fastText를 이용한다. 문장 벡터를 생성하기 전 비어휘화 토큰을 특정 토큰으로 추가한 후 T5의 토크나이저를 이용하여 문장을 토큰화를 한다. 토큰화된 문장과 fastText의 print-word-vector 함수를 이용하여 문장 벡터를 생성한 후 유클리드 거리 계산한다.

유클리드 거리 값은 두 문장 사이의 얼마나 유사한지에 대한 척도이다. 식 (1)은 유클리드 계산 식이고, p와 q는 각 문장의 임베딩 값을 의미하고, i는 임베딩의 인덱스를 의미한다. 아래의 식은 문장 벡터가 n차원으로 설정되어 있다.

$$\text{euclidean distance} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

유클리드 계산 값이 가까울수록 낮은 값을 나타낸다. 유클리드 거리 계산 값을 임계값으로 설정하고 임계값 보다 낮은 문장 쌍을 추출하여 패러프레이즈 쌍을 후보를 추출한다. 추출된 문장들 중 비어휘화된 부분을 기존의 단어로 복원시키는 후처리를 적용하여 최종 패러프레이즈 쌍을 구축한다.

해당 코퍼스는 준완전 학습 코퍼스(silver standard corpus)[9]이며, 정답 패러프레이즈 코퍼스(gold standard corpus)가 아니라 비어휘화와 문장 유사도를 이용하여 구축한 패러프레이즈이다. 문장 벡터 유사도의 결과로 구축된 패러프레이즈는 오류가 포함되어 있을 수 있지만, 오류가 없는 것으로 간주한다.

위의 방식으로 구성된 패러프레이즈 쌍은 코퍼스 내 나타나는 표현들로만 구성된다는 단점이 존재한다. 이런 단점을 해소하기 위해 구축한 패러프레이즈 쌍을 학습 코퍼스로 사용하여 더 다양한 패러프레이즈 표현을 생성하는 실험을 진행하였다. 이 결과를 통해 새로운 표현의 후보군을 추출하여 실험 결과를 확인한다.

3. 실험 및 결과

본 연구에서는 고객센터 상담 도메인으로 설정한다. 본 연구의 실험은 해당 도메인의 코퍼스를 이용하여 패러프레이즈 코퍼스를 구축하였다. 해당 코퍼스를 학습 코퍼스로 사용하여 더 다양한 표현을 생성하는 실험을 진행한다.

3.1 유사도를 이용한 패러프레이즈 구축

고객센터 상담 도메인은 20만 문장으로 구성되어 있다. TF-IDF 임계값을 0.15로 설정하여 비어휘화에 적용 가능한 단어를 추출한다. 우리는 비어휘화된 문장 벡터를 추출하기 위한 fastText 언어 모델 학습은 윈도우 크기는 5, 토큰의 최소 발생 횟수는 5로 설정하였으며 토큰 벡터의 사이즈는 512으로 설정하였다.

각 문장 벡터를 생성하여 유클리드 거리 계산을 통해 패러프레이즈를 구축하였다. 유클리드 거리 계산 임계값을 0.03으로 설정하였고, 생성된 패러프레이즈 수는 307,122개이다. 생성된 패러프레이즈는 비어휘화된 부분을 기존의 단어로 다시 복원시키는 후처리를 진행하였다. 표 3은 생성된 상담 의도 도메인의 패러프레이즈 코퍼스 예이다. 표 내 A, B는 서비스명 또는 상품명을 나타낸다.

표 3 생성된 패러프레이즈 코퍼스 예

입력 문장	구축된 패러프레이즈
다음주 쯤에 A 철거하는 거 예약할 수 있어요?	A 철거시키고 싶어서요.
	A 치워 주는 거 신청해줘요.
	A 빨리 철거할 수 있는 날짜 좀 잡아주세요.
B 신청한 거 취소하려구요.	저 B 신청한 거 취소하려고요.
	B 신청한 거 취소해주세요.
	B 취소해주세요.

정성 평가를 통해 기준을 7가지로 설정하였다. 1) 문장의 전체적 의미가 다른 경우, 2) 어미만 다른 경우, 3) 단어만 다른 경우, 4) 어미와 단어 모두 다른 경우, 5) 입력 문장 내 일부분이 출력된 경우, 6) 입력 및 출력 문장이 동일한 경우 (특수 기호와 띄어쓰기 제외), 7) 구조 변경이다. 우리는 100개의 샘플링을 통해서 정성평가를 진행하였다. 표 4는 정성평가의 결과이다.

표 4 정성평가 결과

정성 평가 기준	해당 문장 수
문장의 전체적 의미가 다른 경우	18
어미만 다른 경우	26
단어만 다른 경우	10
어미와 단어 모두 다른 경우	23
입력 문장 내 일부분이 출력된 경우	20
입력 및 출력이 동일한 경우 (특수 기호와 띄어쓰기 제외)	1
구조 변경	2

정성 평가의 결과에 따르면, 입력 문장과 다르게 어미를 표현한 문장은 총 3가지 기준에서 나타나며 100문장 중 61문장이다. 어미만 다른 경우가 26문장, 어미와 단어 모두 다

른 경우 23문장 그리고 입력 문장 내 일부분이 출력된 경우 20문장 중 12문장에 해당하였다. 문장의 전체적 의미가 다른 경우는 패러프레이즈 문장 내 동일한 서비스명이 나타나고, 중복되는 단어들이 많아 높은 유사도를 가지게 되는 경우가 해당된다.

3.2 구축한 패러프레이즈와 기계 학습을 이용

그림 2는 기계 학습을 이용한 패러프레이즈 생성 흐름도이다. 사용된 모델은 t5-small[10] 모델이고, 국립국어원에서 공개한 모두의 말뭉치의 유사 문장 말뭉치[11]와 제안 방법으로 구축된 코퍼스를 이용한다. 총 2단계의 미세조정(fine-tuning)단계를 거친다. 1) 모두의 말뭉치 내 유사 문장 말뭉치를 이용한 모델, 2) 1)의 결과물인 생성한 패러프레이즈 코퍼스를 이용한 모델, 각 단계별 학습 모델은 다음 단계의 학습 모델로 사용한다.

그림 2 기계 학습을 이용한 패러프레이즈 생성 흐름도

표 5는 단계별 학습에 사용된 코퍼스의 통계이다. 1단계에서 제안한 방법으로 구축된 고객센터 상담 도메인 코퍼스가 아닌 코퍼스로 학습을 진행하는 이유는 고객센터 상담 도메인 내 유사 표현이 아닌 일반적인 패러프레이즈 표현을 학습하기 위함이다. 표 6은 단계별 정량 평가를 진행하였고, 정량 평가의 성능 측정 지표는 Rouge-1, Rouge-2, Rouge-L로 설정하였다.

표 5 단계별 학습에 사용된 코퍼스의 통계

단계 분류	학습 코퍼스	검증 코퍼스	평가 코퍼스
1	218,680	27,335	30,000 (고객센터 상담 도메인)
2	247,122	30,000	

표 6 단계별 학습의 정량 평가 결과

단계 분류	Rouge-1	Rouge-2	Rouge-L
1	0.1123	0.0200	0.0488
2	0.4067	0.1980	0.2569

우리는 100개의 샘플링을 통해서 정성평가를 진행하였다. 표 7은 정성 평가의 결과이다. 단어 및 어미가 변경되어 생성된 경우는 패러프레이즈 생성이 잘된 경우이고 100문장 중 56문장에 속한다. 그 중 기존 가상상담 도메인 코퍼스에 존재하지 않은 문장이 24문장이다. 그리고 문장의 전체적 의미가 다른 경우에 대한 예제로는 '결합해주세요'와 '결합해지해주세요'와 상반되는 문장임에도 불구하고 유사한 문맥을 가지는 어미들에 대해서는 높은 유사도를 가지기 때문에 이러한 패러프레이즈의 구축 오류로 인한 결과이다. 유사 문맥에 나타나는 상반되는 문장 또는 구에 대한 문제는 향후 해결해야 할 연구로 남겨둔다.

표 7 정성평가 결과

정성 평가 기준	해당 문장 수
문장의 전체적 의미가 다른 경우	40
어미만 다른 경우	7
단어만 다른 경우	9
어미와 단어 모두 다른 경우	23
입력 문장 내 일부분이 출력된 경우	17
입력 및 출력이 동일한 경우 (특수 기호와 띄어쓰기 제외)	3
구조 변경	1

4. 새로운 패러프레이즈 표현 분석

딥러닝 모델의 사용은 기존의 구축되어 있는 코퍼스 내 표현이 아닌 새로운 표현을 학습하는 것이 목적이다. 생성된 패러프레이즈 내 새로운 표현을 분석하고자 했다. 새로운 표현에 대한 후보군을 추출하는 방법으로는 어절 단위의 3-gram을 이용해 기존 구축되어 있는 코퍼스에 포함하지 않으면서 펄플렉시티(perplexity, PPL)가 설정한 임계값 이상인 것으로 설정한다. 펄플렉시티는 언어 모델을 평가하기 위한 내부 평가 지표이고, PPL의 수치가 낮을수록 언어 모델의 성능이 좋다는 것을 의미한다. PPL의 값을 15로 설정하여 6,445쌍을 구축하였으며 그 중 24,122개의 새로운 표현 후보군을 생성하였다. 표 8은 새로운 표현 후보군의 예시이고, 표 내 A는 서비스명 또는 상품명이다. 표 9는 정성 평가를 통해 새로운 표현이 확인된 문장들이다. 표 9 내 A,B,C 또한 서비스명 또는 상품명이다.

표 8 새로운 표현 후보군 예시

새로운 표현 후보군 예시	
다시 신청했는데 부가세	변경 사용료 한도
원래 이달에 냈는데	해지해 주세요 A랑
번호 등록해줘 서비스	결제할때 그 비밀번호

표 9 정성 평가를 통한 새로운 표현이 확인된 문장 예시

입력 문장	구축된 패러프레이즈
제 폰으로 A 결제 이용할 수 있어요?	이 폰에서 B 결제할 수 있어?
카드 분실해서 정지 요청합니다	카드 발급받자마자 분실해서 정지 신청 해주세요.
C 상품에 대한 정책 확인 좀 할려구요	C에 대한 할인 혜택 같은거 좀 보려구요 문의 좀 할려구요
지금 자동이체 되어 있는데 꼭 정해진 날짜에만 인출이 될 수 있는거에요?	그렇죠 지금 자동이체 되어 있는데 이게 꼭 이 날에 인출이 될 수 있는거예요?

5. 결론

본 논문에서는 문장 유사도를 이용한 패러프레이즈 쌍을 구축하고, 또 구축한 패러프레이즈 쌍을 이용하여 기계 학습을 통해 새로운 패러프레이즈를 생성한다.

위의 방식으로 구성된 패러프레이즈 쌍은 기존의 구축되어 있는 데이터셋으로 패러프레이즈 쌍을 구축하였기 때문에 문장 내 문법적 에러가 없는 표현들의 패러프레이즈를 구축할 수 있다는 장점이 있으나 입력 코퍼스 내 나타나는 표현들로만 구성된 페러프레이즈 쌍이라는 단점이 존재한다.

이런 단점을 해소하기 위해 구축한 패러프레이즈 쌍을 학습 코퍼스로 사용하여 더 다양한 패러프레이즈 표현을 생성하는 실험을 진행하였다. 새로운 표현에 대한 후보군을 추출하는 방법으로는 어절 단위의 n-gram을 이용해 기존 구축되어 있는 코퍼스에 포함하지 않으면서 PPL이 설정한 임계값 이상인 것으로 설정한다. 기존의 코퍼스 내 새로운 표현들이 생성된 것을 확인할 수 있었다.

현재 패러프레이즈 분류 내 '단어 치환' 내 다양한 어미 표현에 대한 연구를 진행하였으며, '구 치환' 및 '문장 구조 변경'에 관련한 연구들을 향후 연구로 설정한다. 그리고 기존의 코퍼스에 포함되어 있지 않는 새로운 표현들을 더 다양하게 생성하는 문제도 해결해볼 것이며, 유사 문맥에 나타나는 상반되는 구 또는 어미에 대해서 분별할 수 있는 연구를 진행할 계획이다. 또, 제안한 방법이 특정 도메인 내 코퍼스 뿐 아니라 패러프레이즈 구축의 일반화를 위한 연구도 진행할 예정이다.

Acknowledgement

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00113, 이종정보 활용 및 데이터융합을 통한 데이터증식 기술 개발)

참고문헌

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[2] Liu, Zhuang, et al. "A Robustly Optimized BERT Pre-training Approach with Post-training." China National Conference on Chinese Computational Linguistics. Springer, Cham, 2021.

[3] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32, 2019.

[4] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683, 2019.

[5] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8, 2019.

[6] Islam,A.,andInkpen,D."Semanticsimilarityofshorttexts",Proceedings of the International Conference on Recent Advances in Natural Language Processing(RANLP 2007), Borovets, Bulgaria, pp.291-297, 2009.

[7] Fernando,S., and Stevenson,M."A semantic similarity approach to paraphrase detection", Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloquium, 2008.

[8] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5, pp.135-146, 2017.

[9] I. Mani and T. Maybury, Advances in Automatic Text, The MIT Press, 1999. N. Kang, E. M. van Mulligen, and J. A. Kors, "Training Text Chunkers on a Silver Standard Corpus: Can Silver Replace Gold?," BMC Bioinformatics, Vol.13, No.1, pp.17-22, 2012.

[10] KETI AIRC, KE-T5: Korean English T5, mar, 2021,

https://github.com/AIRC-KETI/ke-t5.
[11] https://corpus.korean.go.kr/request/corpusRegist.do,
2021.09.24.