제어가능한 한국어 패러프레이즈 생성을 위한 제약들

[1] 최승권, [1] 권오욱, [1] 김영길
[1] 한국전자통신연구원 언어지능연구실
choisk@etri.re.kr, ohwoog@etri.re.kr, kimyk@etri.re.kr

Constraints for Controllable Korean Paraphrase Generation

Sung-Kwon Choi[1], Oh-Woog Kwon[1], Young-Gil Kim[1]
[1]Language Intelligence Research Section, ETRI

요 약

언어학적 다양성을 가지는 고품질의 한국어 패러프레이즈 생성을 위해 패러프레이즈의 생성을 제어할 수 있는 제약이 필요하다. 원문을 패러프레이즈로 변경하기 위한 생성용 제약으로 6 개의 제약을 제시한다: 키워드 치환, 키워드 확장, 품사 변경, 패턴 변경, 구조 변경, 키워드 리스트, 생성 길이. 원문으로부터 패러프레이즈를 생성할 때 제약이 적용되는 정도를 시뮬레이션해 보았다. 10 어절 이하의 원문은 평균 2.05 번의 제약이 적용되면 패러프레이즈가 생성되었으며 키워드 치환, 마스킹에 의한 키워드 확장과 패턴 변경에 관한 제약이 가장 많이 적용되는 것을 확인하였다.

1. 서론

패러프레이즈(paraphrase)는 원문과 동일한 의미를 가지지만 다른 말로 바꾸어 표현한 것이라고 말할 수 있다. 패러프레이즈가 원문의 다양한 다른 표현이기 때문에 데이터 확장 측면에서 패러프레이즈 생성 기술은 자동 번역, 자동 통역, 정보 검색, 정보 추출, 자동 요약, 질의 응답, 자연어 생성, 표절 인식 등 자연어 처리 분야에서 활용할 수 있는 중요한 데이터 증강(data augmentation) 기술 중 하나라고 할 수 있다[1]. 고품질의 패러프레이즈 생성을 위해서는 고품질의 지도학습(supervised learning)용 패러프레이즈 데이터가 필요하다. 그러나 현재 구축되어 있는 패러프레이즈를 분석해 보면, 데이터의 크기에 비해 언어학적으로 다양한 구조와 의미적 유사성을 동시에 가지는 고품질의 패러프레이즈 데이터가 아직도 부족하다[2]. 그래서 최근에 언어학적 다양성을 확보하기 위해 제어 가능한 패러프레이즈 생성 모델이 개발되고 있다.

이런 점에서, 본 논문은 언어학적 다양성을 가지는 고품질의 한국어 패러프레이즈 생성을 제어할 수 있는 제약을 기술하는 것을 목적으로 한다.

2. 관련 연구

신경망 기반의 패러프레이즈 생성 연구는 패러프레이즈의 품질과 다양성(diversity)을 확보하기 위해 단어 수준의 생성 모델로부터 구 수준의 패러프레이즈 생성 모델로 발전하고 있다[3]. 영어권에서 제약(constraints)에 기반하여 패러프레이즈의 생성을 제어하는 모델을 개발하기 시작하였는데, [4]에서는 키워드와 품사를 제약 조건으로, [5]에서는 생성 단어 길이, 통사 구조, 의미 유사성, 패턴을 제약 조건으로, [6]에서는 품사 태깅, 통사 구조, 마스킹 탬플릿, 샘플문, 생성 키워드를 제약 조건으로 하여 패러프레이즈 생성을 제어하는 모델을 제안하였다.

한국어에서 제약에 기반하여 패러프레이즈의 생성을 제어하는 방법으로는 문법구조 및 의미 거리 계산법을 이용하여 원문과 의미가 가장 유사한 문장을 생성하는 모델[7]과 실질형태소의 키워드를 추출하여 키워드순서를 유지/변경하여 패러프레이즈를 생성하는 모델[8]을 제시하였다.

3. 한국어 패러프레이즈의 유형 분류

영어 패러프레이즈 유형 분류로 어휘 교체나 구조 변경이 제안되었다[9]. 영어와 달리 한국어는 어순의 자유성과 교착어라는 특징이 있어 이러한 특징을 반영하는 한국어 패러프레이즈 유형 분류가 필요하다. 한국어 패러프레이즈 유형으로 4 가지 유형을 분류하였다. 첫번째 유형은 치환으로서 주변 어휘는 유지하면서 원문의 동의어나 어미들을 교체하여 패러프레이즈를 만드는 방법이다. 두번째 유형은 확장으로서 주변 어휘는 유지하면서 원문의 어휘를 구/어절/관용어 등의 더 큰 언어 단위로 교체하는 방법이다. 세번째 유형은 변형으로 품사순서 변경, 태변경, 어순 변경등

과 같이 위치에 의해 구조를 변경하는 방법이다. 네 번째 유형은 치환. 확장, 변형을 혼합하여 사용하는 복합 방법이다. 한국어 패러프레이즈의 유형과 그 예 [10]를 들면 다음과 같다:

<표 1> 한국어 패러프레이즈 유형과 예

유형		원문<=>패러프레이즈
치환	동의어 치환	범국민대회 **참가자**들에게 <=> 범국민대회 **참여자**들에게
	연결/종결어미 치환	아무리 시간이 없**어도** <=> 아무리 시간이 없**더라도**
확장	어휘 풀어쓰기	인간은 언어를 통해 **의사소통한다**. <=> 인간은 언어를 통해 **서로의 생각을 전달하고 의사표현을 할 수 있다**.
	관용어 활용	직장 구하기가 **어렵다** <=> 직장 구하기가 **하늘에 별 따기 같다.**
변형	품사순서	축구를 좋아하는 **사람 다섯**이 모였다. <=> 축구를 좋아하는 **다섯 사람**이 모였다.
	태 변경	모기가 아기를 **물었다**. <=> 아기가 모기**한테 물렸다**.
	격틀 변경	징계 **절차에 착수하다** <=> 징계 **절차를 시작하다.**
	어순변경	**다행스럽게도** 그는 시험에 붙었다. <=> 그는 **다행스럽게도** 시험에 붙었다.
	구조변경	**철수는 훌륭하게 시연하였다.** <=> **철수의 시연은 훌륭하였다**.
	문장 통합	이 섬은 아름답다. **또한** 살기 좋을 것 같다. <=> 이 섬은 **아름다울 뿐만 아니라** 살기 좋을 것 같다.
	문장 분리	원유 값이 올랐**기 때문에** 기름 값도 오를 것이다. <=> 원유 값이 올랐다. **따라서** 기름 값도 오를 것이다.
복합	치환+확장+변형	이 **사회**는 **대학 교육**을 **개인의 성공**의 사다리로 간주해 왔다. <=> **사회**에서 **개인의 성공**은 **대학 교육**에 의하여 달성되는 것으로 믿어진다.

4. 제어가능한 한국어 패러프레이즈 생성용 제약들

제약을 통해 패러프레이즈를 제어한다는 것은 입력 문장 x 와 제약들 c 에 의해 패러프레이즈 $y = (y_1, y_2, \ldots y_T)$를 생성하는 다음과 같은 확률 모델 식으로 정의할 수 있다:

$$p(y|x,c) = \prod_{t=1}^{T} p(y_t|y_{<t}, x, c; \theta)$$

위의 식에서 θ는 원문 대 패러프레이즈 쌍들로 구성된 패러프레이즈 코퍼스로부터 조건부 우도(conditional likelihood) 최대화에 의해 학습된 모델 파라미터를 의미하며, 제약 c는 원문-패러프레이즈로 구성된 패러프레이즈 코퍼스로부터 자동으로 추출한다. 표 1 의 한국어 패러프레이즈 유형을 토대로 제약을 기술하면 다음과 같다. 제약에 대한 표현은 정규 표현식(regular expression)으로 쓰였다:

1) constraint: substitute_keyword

생성할 패러프레이즈에 나타나야 하는 키워드를 치환하는 제약으로 키워드가 없는 패러프레이즈를 제어할 수 있다.

예) input: 범국민대회 참가자들에게
constraint: substitute_keyword = 참여자
output: 범국민대회 참여자들에게

2) constraint: extend_keyword

원문과 동일하지 않은 마지막 어휘들을 모두 마스킹함으로써 패러프레이즈의 확장을 제어할 수 있다.

예) input: 직장 구하기가 어렵다.
constraint: extend = * 하늘에 별 따기 같다$
output: 직장 구하기가 하늘에 별 따기 같다.

3) constraint: change_pos

형태소 분석된 결과를 대상으로 품사 시퀀스를 기술하는 제약으로 품사 시퀀스가 들어가지 않는 패러프레이즈를 제어할 수 있다.

예) input: 축구를 좋아하는 사람 다섯이 모였다.
constraint: pos = * NUM N*
output: 축구를 좋아하는 다섯 사람이 모였다.

4) constraint: change_pattern

의존 분석(dependency analysis)된 결과를 토대로 생성할 패러프레이즈에 나타나야 하는 어휘 패턴을 기술하는 제약으로 태 변경, 격틀 변경 등을 제어할 수 있다.

예) input: 징계 절차에 착수하다.
constraint: change_pattern = * NP!를 시작하!다 *
output: 징계 절차를 시작하다.

5) constraint: change_structure

생성할 패러프레이즈에 나타나야 하는 구조를 기술하는 제약으로 어순 변경, 구조변경, 문장 통합, 문장 분리 등을 제어할 수 있다.

예) input: 철수는 훌륭하게 시연하였다.
constraint: change_pattern = NP_subj[N_mod N] Pred
output: 철수의 시연은 훌륭하였다.

6) constraint: list_keywords

생성할 패러프레이즈에 나타나야 하는 키워드 리스트를 기술하는 제약으로 복합을 제어할 수 있다.

예) input: 이 사회는 대학 교육을 개인의 성공의 사다리로 간주해 왔다

constraint: list_keywords = [사회, 개인, 성공, 대학 교육, 달성]

output: 사회에서 개인의 성공은 대학 교육에 의하여 달성되는 것으로 믿어진다.

7) constraint: length

생성할 패러프레이즈의 단어 길이를 기술하는 제약으로 짧거나 긴 패러프레이즈를 제어할 수 있다.

예) input: 직장 구하기가 어렵다.

constraint: length = 5

output: 직장 구하기가 하늘에 별 따기

5. 시물레이션

제약기반 패러프레이즈 생성 시스템이 현재 구현되고 있어 본 논문에서는 원문이 패러프레이즈로 변하기 위해서 어떤 제약들이 몇번 적용하면 패러프레이즈가 생성될 수 있는지를 시물레이션으로 제시한다. 이 시물레이션은 향후 대량의 한국어 패러프레이즈 코퍼스가 구축되면 제약을 학습한 후에 적용되는 과정과 유사하기 때문이다. 실험 데이터는 한국어 패러프레이즈 코퍼스인 국립국어원 유사 문장 말뭉치로부터 임의로 원문과 패러프레이즈를 추출하였다. 각 원문은 사람이 작성한 4 개의 패러프레이즈로 구성된다.

5 개의 한국어 원문과 20 개의 패러프레이즈를 대상으로 시물레이션 결과를 살펴보면 다음과 같았다:

<표 2> 한국어 패러프레이즈 시뮬레이션 결과

원문	5 문장, 9,8 어절
패러프레이즈	20 문장, 9.7 어절
Constraints	
substitute_keyword	17
extend_keyword	8
change_pos	0
change_pattern	14
change_structure	2
list_keywords	0
length	0
계	41

한국어 원문으로부터 패러프레이즈를 생성하기 위해서 10 어절 이하의 원문에서는 평균 2.05 번의 제약이 적용되면 패러프레이즈가 생성되는 것을 알 수 있다.

6. 결론

본 논문은 언어학적 다양성을 가지는 고품질의 한국어 패러프레이즈 생성 모델을 구현하기 위해 패러프레이즈 생성을 제어할 수 있는 다양한 제약이 무엇인지를 확인하고자 하였다. 제어가능한 한국어 패러프레이즈 생성을 위한 제약으로 6 개의 제약을 제시하였으며 substitute_keyword, change_pattern, extend_keyword, 순으로 적용됨을 시물레이션을 통해 알 수 있었다.

Acknowledgement

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

참고문헌

[1] 전현규, 정윤경, 텍스트 바꿔 쓰기 과제를 위한 분류 모델 기반의 손실 함수 설계와 평가. 정보과학회논문지 제 48 권 제 10 호, 1132-1141 쪽, 2021.

[2] Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shanyderman, Noam Slonim and Liat Ein-Dor, Quality Controlled Paraphrase Generation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Vol 1, pp 596-609, 2022.

[3] Jianing Zhou and Suma Bhat, Paraphrase Generation: A Survey of the State of the Art. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp.5075-5086. 2021.

[4] Xuhang Xie, Xuesong Lu, Bei Chen, Multi-task Learning for Paraphrase GenerationWith Keyword and Part-of-Speech Reconstruction. In Findings of the Association for Computational Linguistics: ACL 2022, pp. 1234 – 1243, 2022.

[5] Nazanin Dehghani, Hassan Hajipoor, Jonathan Chevelu, and Gwénolé Lecorvé, Controllable Paraphrase Generation with Multiple Types of Constraints. In Controllable Generative Modeling in Language and Vision Workshop at NeurIPS, 2021.

[6] Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen, GCPG: A General Framework for Controllable Paraphrase Generation. In Findings of the Association for Computational Linguistics: ACL 2022, pp. 4035 – 4047, 2022.

[7] 서혜인, 정상근, 정지수, 유사구조 및 유사의미문장 생성 방법, 제 32 회 한글 및 한국어 정보처리 학술대회 논문집, 162-166 쪽, 2020.

[8] 홍성태, 차정원, 키워드를 이용한 패러프레이즈 문장 생성, 한국컴퓨터종합학술대회 논문집, 377-379 쪽, 2022.

[9] Rahul Bhagat and Eduard Hovy, What is a Paraphrase?, Computational Linguistics, Volume 39, Issue 3, pp. 463-472, 2013.

[10] 팜티튀린, '바꿔 쓰기' 활동을 통한 한국어 문법 교수.학습 방안 연구, 서울대학교 대학원 국어교육과 한국어교육전공 교육학석사학위논문, 2012.