언어 모델을 활용한 난독화된 한국어 숙소 이용 후기 해독*

김유빈[0], 성무진**
경희대학교 컴퓨터공학부
kyb0314@khu.ac.kr, mujeensung@khu.ac.kr

Deciphering Obfuscated Accommodation Reviews in Korean using Language Models

Yubin Kim[0], Mujeen Sung
School of Computing, Kyunghee University
kyb0314@khu.ac.kr, mujeensung@khu.ac.kr

요 약

한국인 관광객이 해외 숙박시설 이용 후 한국어로 솔직하지만 부정적인 후기를 남기는 경우, 업주들이 번역기를 통해 뜻을 이해하고 삭제하는 것을 방지하기 위해 난독화를 하여 후기를 올려놓는 경우가 많다. 본 논문에서는 KLUE의 Airbnb Review 데이터를 기반으로 난독화된 한국어 숙박후기를 해독하는 모델을 만들어 난독화의 종류와 모델의 사이즈에 따른 성능 비교 분석을 진행하였고, 최대 75.2의 BLEU-4 점수를 얻을 수 있었다. 특히, 난독화의 규칙이 간단할수록, 그리고 모델의 사이즈가 클수록 BLEU 점수가 높게 나타나는 것으로 분석하였으며, 크로스 데이터셋 학습 성능을 비교하고, 통합된 난독화 데이터에 대해 난독화 해독 능력을 확인함으로써 다양한 난독화 문장에 대한 해독 가능성을 확인할 수 있었다.

1. 서 론

해외여행을 하는 한국인 관광객 수가 늘어나며, 해외 숙박 예약 서비스의 사용이 함께 증가하고 있다. 그들은 숙박시설 이용 후 후기를 남김으로써 다른 이용자들에게 정보를 제공하고 있다. 특히, 솔직하지만 부정적인 후기의 경우 숙박후기에 난독화를 하여 올려놓는 경우가 많다 [그림 1]. 이는 한국어를 모르는 업주들이 번역기를 통해 뜻을 이해하고 삭제하는 것을 방지하기 위함이다 [그림 2].

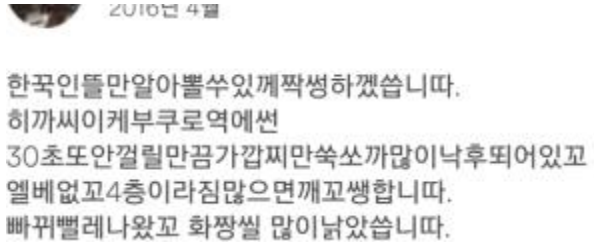
한편, 딥러닝 기반의 언어 모델들은 기계 번역 [1], 문서 요약 [2] 등 다양한 텍스트 변환 영역에서 뛰어난 성능을 보이고 있다. 이중, 시퀀스-투-시퀀스 (Sequence-to-sequence) 모델은 크게 인코더와 디코더로 구성되어 입력 문장과 출력 문장의 쌍에서 패턴을 학습한다. 이후, 새로운 입력 문장이 주어졌을 때 학습한 패턴을 기반으로 새로운 출력 문장을 생성한다 [3].

이 논문에서는, 한국어 숙박 후기 난독화를 해독하는 모델을 만든다. 이를 위해, 난독화된 한국어 숙박 후기와 해독된 숙박 후기의 쌍 데이터를 구축하고, 언어 모델들을 수집된 데이터셋에 학습한다. 이후 난독화 종류별 해독 성능과, 모델의 크기별 해독 성능을 분석한다.

2. 데이터셋 구축

2.1 데이터 출처

난독화된 한국어 숙박 후기와 해독된 숙박 후기의 쌍 데이터를 구축하기 위해, 먼저 한국어 숙박 후기 데이터를 수집하였다. 한국어 숙박 후기는 다양한 한국어 언어 모델을 평가하기 위한 벤치마크인 KLUE [6]의 Airbnb Review 데이터에서 얻었다. 그 다음, 한국어 난독화 라이브러리 kotka [7]을 사용하여 한국어 숙박 후기에 난독화를 진행하였다. 해당 라이브러리에서 지원하는 한국어 난독화의 종류는 총 3가지로, '네모네모 멈뭄미의 저주 시리즈', '야민정음', 그리고 '확률적 자음/모음 분리'이며 각 난독화를 순서대로 '네모네모', '야민정음', '자모분리'으로 표현한다 [표 1]. 3가지로 난독화한 데이터셋을 통해 한국어 난독화 종류별로 언어 모델의 해독의 난이도를 검증하려고 한다.



[그림 1] 한글 난독화 숙소 후기 [4]

[그림 2] 난독화된 문장에 대한 잘못된 번역 [5]

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)
** 교신저자 (Corresponding Author)

[표 1] 한국어 난독화 종류 3가지 예시

원본 문장	샤워장이 공용이고 방음이 조금 미약하긴 하지만 나름 여러가지 방면에서는 만족 합니다.
↓	
난독화 종류	난독화된 문장
네모네모	샤뭐잠미 곰묨미고 밤믐미 조금 미먁하긴 하지만 나름 며러가지 밤면메서는 만족 합니다.
야민정음	샤워장이 끙용이끄 방음이 조금 미약하긴 하거만 나름 억러가거 방띤에서는 만족 합니다.
자모분리	샤워장ㅇㅣ ㄱㅗㅇ용ㅇㅣㄱㅗ 방ㅇㅡㅁㅇㅣㅈㅗㄱㅡㅁ 미약하긴 하ㅈㅣ만 나름 ㅇㅕㄹㅕㄱㅏ지 방ㅁㅕㄴㅇ에ㅅㅕ는 만족 ㅎㅏㅂㄴㅣ다.

2.2 데이터 전처리

KLUE Airbnb Review 데이터 중 중복된 데이터를 제외하고 학습 4336개, 검증 308개의 문장을 얻었다. 이를 원본 데이터로 하여 원본 문장과 3종류의 난독화 문장을 JSON 파일로 구축하였다. 검증 데이터를 검증:평가 = 1:1로 나눈 후, 각 난독화 종류별로 학습(4336), 검증(154), 평가(154) 데이터셋을 얻었다. 구축된 전체 데이터셋에 대한 자세한 데이터셋의 통계는 [표 2]에 표기되어 있다.

[표 2] 데이터셋 통계

난독화 종류	학습	검증	평가
네모네모	4,336	154	154
야민정음	4,336	154	154
자모분리	4,336	154	154

3. 실험 및 분석

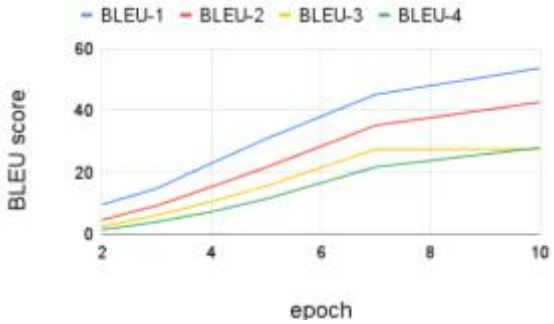
3.1 실험 세팅

본 논문에서 사용한 모델은 KETI-AIR/ke-t5-ko로, 시퀀스-투-시퀀스 모델을 한국어와 영어 코퍼스를 이용하여 사전 학습한 모델이다. 파라미터 사이즈에 따라 small (60 million), base (220 million), large (770 million)의 3가지 모델을 사용하여 실험을 진행하였다. 난독화된 문장을 모델의 입력 문장으로 하여 모델이 해독한 출력 문장을 원본 문장과 비교하였으며, BLEU의 n-gram (1~4)을 통해 순서쌍들이 얼마나 겹치는지 측정하였다.

하이퍼파라미터를 설정하기 위해 '네모네모'의 데이터셋에서 small 모델로 learning rate를 각각 {2e-5, 5e-5, 1e-4}로 설정한 후 성능을 비교한 결과 1e-4에서 가장 좋은 성능을 보였다. 이때, epoch은 기본값인 3으로 진행하였다. epoch의 경우 2부터 10까지의 범위에서 성능을 비교한 결과 10에서 가장 좋은 성능을 보였다 [그림 3]. 이러한 결과를 통해 learning rate는 1e-4로, epoch은 10으로 실험을 진행하였다.

3.2 실험 결과

모델의 사이즈와 난독화 종류에 따른 평가 데이터셋의 BLEU 점수는 [표 4]에 작성되어 있다. 표의 세로축은 모델의 사이즈와 학습 데이터의 난독화 종류를 나타내며, 세로축은 n-gram (1~4)에 대한 BLEU 점수를 나타낸다.



[그림 3] epoch별 성능 비교. small 모델 사용.

모델의 사이즈가 클수록, n이 작을수록 좋은 성능을 보이는 것을 확인할 수 있다. [표 1]의 난독화 종류별 예시 문장에 대해 base 모델로 해독된 문장을 [표 5]에 표기하였다. 모든 난독화에 대해 올바르게 해독한 단어와 잘못 해독한 단어가 모두 존재한다. BLEU 점수와 공통적으로 '네모네모'와 '야민정음'의 규칙으로 난독화된 문장은 대체로 올바르게 해독된 단어의 비율이 높은 반면 '자모분리'의 규칙으로 난독화된 문장은 잘못 해독된 단어의 비율이 높음을 확인할 수 있다.

[표 4] 모델의 크기와 난독화의 종류에 따른 BLEU 점수. 난독화 종류별 최고 점수는 볼드체로 표기

모델	학습/평가	BLEU-1	BLEU-2	BLEU-3	BLEU-4
small (60M)	네모네모	62.7	52.9	45.1	38.7
	야민정음	51.5	43.2	36.2	30.4
	자모분리	23.8	12.4	7.2	4.6
base (220M)	네모네모	80.4	72.2	65.4	59.7
	야민정음	83.2	75.9	69.3	63.5
	자모분리	41.9	30.0	22.3	16.4
large (770M)	네모네모	**87.4**	**81.3**	**75.8**	**71.0**
	야민정음	**89.4**	**84.2**	**79.3**	**75.2**
	자모분리	**49.3**	**38.5**	**30.3**	**24.0**

[표 5] "표 1" 난독화 종류별 문장의 해독 문장 예시

난독화 종류	해독된 문장
네모네모	샤워실이 공용이고 방음이 조금 이용하기 하지만 나름 여러가지 방면에서는 만족 합니다.
야민정음	샤워장이 조용하고 방음이 조금 미미하긴 하지만 나름 여러모로 방엔 만족 합니다.
자모분리	샤워부스와 공용이고 방이 좀 다소 다소않하지만 방이는 만족합니다.

4. 분석

4.1 난독화 종류에 따른 난독화 해독 능력

'네모네모'는 모든 자음 'ㅇ'을 자음 'ㅁ'으로 변경하는 난독화이다. '야민정음'은 특정 음절을 비슷한 모양의 다른 음절로 변경하는 난독화로 용법이 다양하다. '자모분리'는 일부 글자의 자음과 모음을 분리하는 난독화이다. '네모네모'과 '야민정음'는 기본적인 글자의 형태는 유지하면서 특정 자음이나 글자만을 변경하는 난독화로, 비교적 간단한 규칙을 가진다. 본 논문의 언어 모델은 '네모네모'과 '야민정음'의 규칙으로 난독화된 단어를 원본의 단어와 비슷하게 해독하는 것을 확인할 수 있다. 반면 '자모분리'는 일부 글자의 자음과 모음이 분리됨으로써 기본적인 글자의 형태를 갖추지 못하게 되는 난독화로, 언어 모델이 원본의 단

어와 완전히 다른 단어로 해독하는 것을 확인하였다. 이에 대해, 글자의 형태를 갖추지 못하는 특성 때문에 해독이 어려웠을 것으로 예상된다.

'네모네모'와 '야민정음'의 학습데이터를 사용한 모델의 BLEU 점수는 large 모델을 기준으로 n-gram에 관계없이 70점 이상이지만, '자모분리'의 학습데이터를 사용한 모델의 점수는 1-gram에서도 50점 미만의 점수를 나타낸다.

4.2 모델 사이즈에 따른 난독화 해독 능력

본 논문의 입력값인 난독화 데이터셋의 경우, 무작위의 난독화가 아닌 특정 규칙에 대한 난독화이기 때문에 모델의 사이즈가 클수록 좋은 성능을 보일 것으로 예상하였다. 실험 결과, 난독화의 종류에 관계없이 공통적으로 모델의 사이즈가 클수록 모든 난독화 데이터셋에서 좋은 성능을 보였다. 모델의 사이즈가 커질수록 파라미터 수가 늘어나 다운스트림 테스크에서의 성능이 점점 좋아지기에 [8], 입력값을 원하는 결과값으로 변환시키는데 있어 패턴을 잘 학습하여 좋은 성능을 보이는 것으로 판단된다.

4.3 크로스 데이터셋을 통한 학습 성능 비교

각각의 난독화 데이터셋으로 학습한 모델에 대해 난독화 종류가 다른 평가 데이터를 통해 크로스 데이터셋의 학습 성능을 비교해보았다. 모델의 사이즈에 따라 성능을 비교한 결과, base 모델이 large 모델과 비슷한 성능을 보이거나 더 좋은 성능을 보였다. base 모델을 사용하여 크로스 데이터셋의 학습 성능을 비교한 결과를 [그림 4]에 작성하였으며, 이때 BLEU-2의 점수를 표기하였다.

학습 데이터와 평가 데이터의 난독화 종류가 같은 경우에는 모델의 사이즈가 클수록 난독화 해독 능력이 좋았으나, 학습 데이터와 평가 데이터의 난독화 종류가 다른 경우에는 파라미터가 클수록 학습 데이터의 패턴을 더 잘 파악하게 되기 때문에 large 모델에서의 성능이 base 모델에서의 성능과 비슷한 것으로 예상된다.

학습 \ 평가	네모네모	야민정음	자모분리
네모네모	72.2	32.3	3.3
야민정음	22.5	75.9	5.2
자모분리	14.3	29.2	30.0
통합	36.5	38.0	24.5

[그림 4] 크로스, 통합 데이터셋 성능 비교. base 모델 사용, BLEU-2 점수 표기.

4.4 통합된 난독화 데이터의 난독화 해독 능력

모든 난독화 종류에 대한 난독화 데이터를 통합하여 통합된 데이터셋으로 학습한 모델의 난독화 해독 능력을 확인해보고자 하였다. 서로 다른 원본 문장에 대해 3가지 난독화 중 하나를 적용하여 통합된 난독화 학습 데이터셋을 구축하였으며, 각 난독화 문장의 비율은 1:1:1이다. 통합된 난독화 학습 데이터셋을 base 모델에서 learning rate는 1e-4로, epoch은 10으로 학습시켰다.

통합된 데이터셋으로 학습한 모델에 대해 난독화 종류별로 평가 데이터에 대한 성능을 확인하였으며, 크로스 데이터셋의 성능과 비교하여 [그림 4]에 작성하였다. 학습 데이터와 평가 데이터의 난독화 종류가 같은 모델의 성능과 비교하였을 때, 통합된 데이터로 학습한 모델이 더 낮은 성능을 보인다. 이는 학습 데이터가 여러 난독화 종류로 난독화되었기 때문에 모델이 난독화 해독을 위한 공통적인 특성을 파악하기 어려웠을 것으로 예상된다. 학습 데이터와 평가 데이터의 난독화 종류가 다른 모델의 성능과 비교하였을 때, 통합된 데이터로 학습한 모델이 더 높은 성능을 보인다. 이를 통해 여러 난독화 데이터에 대한 해독 가능성을 확인할 수 있었다.

5. 결론

본 논문에서는 한국어 숙박 후기 난독화를 해독하는 모델을 만들어 여러 규칙에 의해 난독화된 문장들을 해독할 수 있음을 보였다. 난독화 종류에 따른 성능 비교 분석과 모델 사이즈에 따른 성능 비교분석을 진행하여 한국어 난독화 문장을 원본 문장으로 해독할 수 있음을 보였다.

크로스 데이터셋 성능 비교를 통해 학습 데이터와 난독화 종류가 다른 평가 데이터에 대한 해독 가능성을 확인하였으나, 난독화 종류가 같은 평가 데이터에 대한 해독 결과에 비해 낮은 성능을 보였다. 또한 통합된 난독화 데이터에 대한 난독화 해독 성능을 통해 여러 난독화 데이터에 대한 해독 가능성을 확인하였으나, 한 가지 난독화 데이터에 대한 해독 능력에 비해 낮은 성능을 보였다. 이와 같은 문제를 해결하기 위해 추후 여러 난독화가 동시에 이루어진 문장에 대한 언어 모델에 대한 연구를 진행할 계획이다.

6. 참고문헌

[1] Ashish Vaswani., et al. "Attention Is All You Need", NIPS 2017

[2] Yang Liu, Mirella Lapata. "Text Summarization with Pretrained Encoders", EMNLP 2019

[3] Colin Raffel., et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", Journal of Machine Learning Research 2020

[4] https://www.joongang.co.kr/article/20947709

[5] https://translate.google.com/?hl=ko&tab=TT

[6] Sungjoon Park., et al. "KLUE: Korean Language Understanding Evaluation", NeurIPS Datasets and Benchmarks 2021

[7] https://github.com/Astro36/kotka

[8] Tom B. Brown,. et al. "Language Models are Few-Shot Learners", Johns Hopkins University, OpenAI