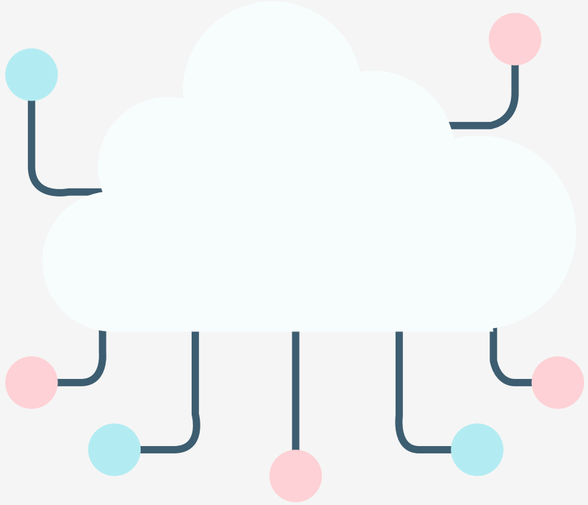
제34회 한글 및 한국어 정보처리 학술대회

Natural Language Explanations 에 기반한 한국어 자연어 추론

구두 발표

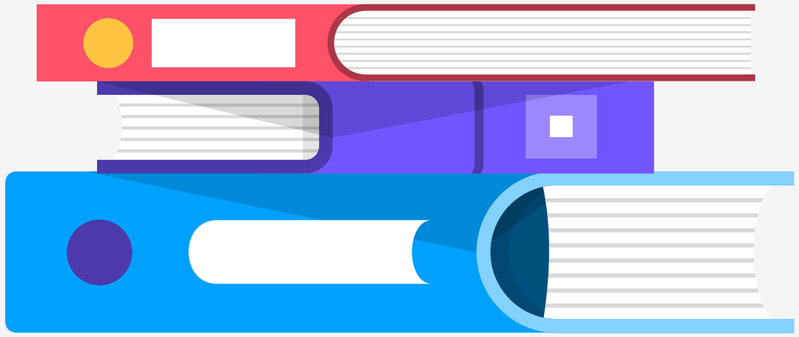
전북대학교 윤준호
hoho5702@jbnu.ac.kr

목차



01

개요



“ Natural Language Explanation 이란?

일반적으로 대규모 언어 모델들은 다량의 데이터를 오랜시간 사전학습하면서 레이블을 예측하기 위한 성능을 높여왔다.

최근 언어 모델의 레이블 예측과 더불어 언어 모델이 왜 해당 결정을 내렸는지 이해하기 위한 신뢰도 높은 Natural Language Explanation (NLE) 을 생성하는 것이 시간이 지남에 따라 주요 요소로 자리잡고 있다.

<u>설명가능한 자연어 태스크를 돕는 정보</u>

높은 정확도를 유지하면서 레이블 예측에 대한 explanation 을 생성하는 자연어 추론

*Natural-language Inference over Label-specific Explanations : NILE

+ NILE 모델에 한국어 데이터셋을 적용하여 NLE 에 기반한 한국어 자연어 추론 성능 확인
+ NLE 를 활용하지 않는 기존의 자연어 추론 태스크와 그 성능을 비교

*https://aclanthology.org/2020.acl-main.771.pdf

02

방법

NILE 및 한국어 적용

데이터셋

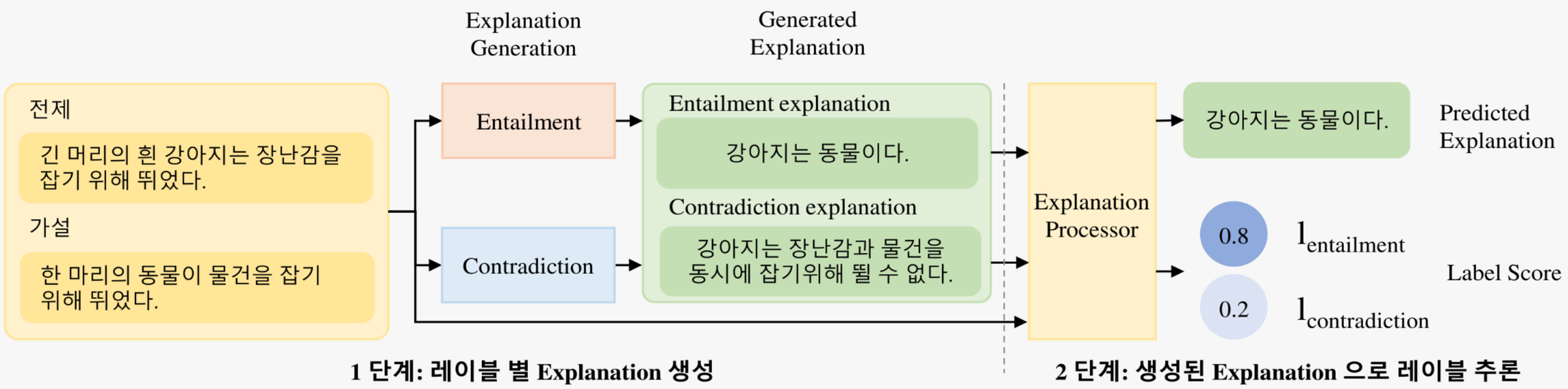
ColumbiaNLP 의 FigLang2022SharedTask (huggingface)
을 NAVER Papago 를 이용하여 한국어로 번역 후 활용

– 구성

[1]가설, [2]전제, [3]레이블 ∈ { Entailment, Contradiction }, [4]레이블에 대한 설명

파이프라인 구조

1. 예측 가능한 레이블에 해당하는 explanation 생성
2. 생성된 explanation 을 이용하여 레이블 score 생성 후 자연어 추론



* NILE 프레임워크 Overview

가. 전제와 가설 쌍을 입력으로 받아 예측 가능한 모든 레이블에 해당하는 explanation 생성
나. 생성된 explanation 이 전제와 가설 쌍과 함께 (또는 혼자) Explanation Processor 에 입력
다. 레이블 score 를 생성하고 최종적으로 레이블을 예측

파이프라인 구조

1. explanation 생성

레이블 별로 데이터셋을 분리

레이블 별 데이터를 이용하여 각 언어 모델을 미세 조정 후 explanation 생성

⟶ 데이터셋의 예측 가능한 레이블에 해당하는 모든 explanation 얻을 수 있음.

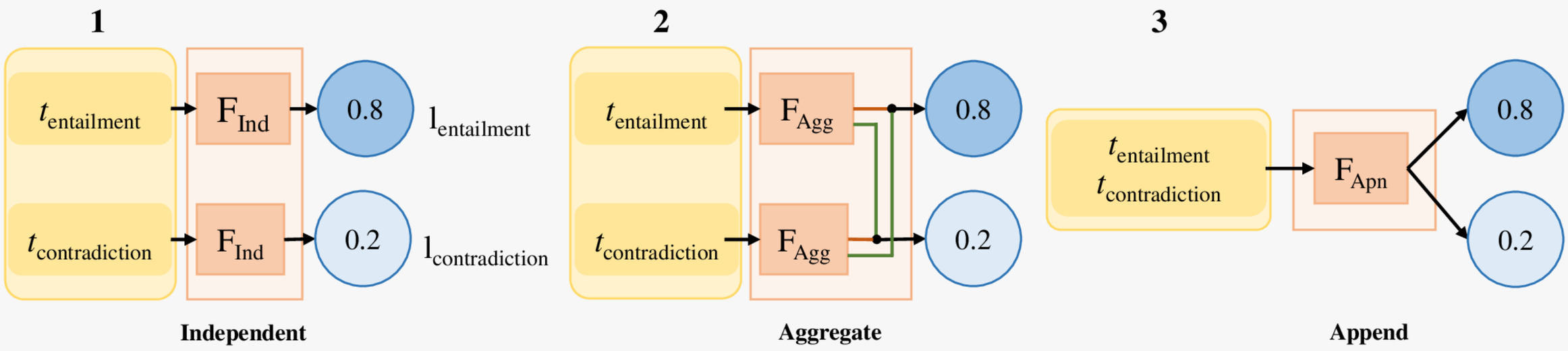
2. 자연어 추론

입력에 따라 NILE-NS 와 NILE-PH

NILE-NS: 생성된 explanation 과 전제와 가설 쌍을 모두 입력으로 이용

NILE-PH: 생성된 explanation 만을 입력으로 이용

또한 explanation 가공에 있어 세 가지 Explanation Processor 구조 이용



* 세 가지 Explanation Processor 구조

[1]Independent: 각 레이블에 해당하는 explanation 에 대해 독립적으로 레이블 score 생성

[2]Aggregate: 각 레이블에 해당하는 explanation 에 대해 중간 score 인 V_1 과 V_2 를 생성 후 최종 레이블 score 생성

V_1: 입력을 뒷받침하는 evidence | V_2: 입력에 반하는 evidence

[3]Append: 각 레이블에 해당하는 explanation 을 하나의 시퀀스로 구성 후 레이블 score 생성

시퀀스 =〉 "수반: $t_{entailment}$ 모순: $t_{contradiction}$"

03

결과

"가설과 전제 그리고 explanation 을 하나의 시퀀스로 묶어 활용"

베이스라인: 전제와 가설만을 이용한 기존의 자연어 추론

모델		레이블 정확도
Baseline	-	82.5
NILE-PH	Independent	48.1
	Aggregate	49.0
	Append	59.2
NILE-NS	Independent	74.9
	Aggregate	73.8
	Append	**83.5**

한국어 데이터셋 실험 결과

1. NILE-NS 의 Append 구조가 가장 높은 레이블 정확도를 가짐.

2. 한국어 데이터셋을 이용한 NILE-PH 는 **베이스라인**과의 레이블 정확도에서 큰 차이를 보임.

모델		레이블 정확도
Baseline	-	82.5
NILE-PH	Independent	48.1
	Aggregate	49.0
	Append	59.2
NILE-NS	Independent	74.9
	Aggregate	73.8
	Append	**83.5**

한국어 데이터셋 실험 결과

모델		레이블 정확도
Baseline	-	89.2
NILE-PH	Independent	78.2
	Aggregate	77.4
	Append	80.9
NILE-NS	Independent	83.4
	Aggregate	86.6
	Append	**89.8**

영어 데이터셋 실험 결과

역시 NILE-NS 의 Append 구조가 가장 높은 레이블 정확도를 가짐.

영어 데이터셋을 이용한 NILE-PH 는 **베이스라인**과의 레이블 정확도와 비교 시 한국어 데이터셋에 비해 더 작은 차이를 보임.

FigLang2022SharedTask 데이터셋은 비유적인 표현들을 포함하고 있음.
(풍자, 직유, 은유 그리고 관용어)

해당 비유적인 표현들이 단순 기계 번역을 통해 한국어 데이터셋으로 만들어지는 과정에서 데이터셋의 품질이 저하된 것이 아닐까?

	Text
영어	Premise: She gets to waste her days sucking back tequila and lorazepam and here you are, still struggling to not earn a salary. Hypothesis: She gets to waste her days sucking back tequila and lorazepam and here you are, still struggling to bring home the bacon. Explanation: To bring home the bacon means to earn a salary, but in this sentence the person is struggling not to earn a salary
한국어	전제: 그녀는 데킬라와 로라제팜을 빨아들이며 하루하루를 허비하게 되는데, 당신은 여전히 월급을 못 벌기 위해 고군분투하고 있어요. 가설: 그녀는 데킬라와 로라제팜을 빨아들이며 하루하루를 허비하게 되는데, 당신은 여전히 베이컨을 집으로 가져오려고 애쓰고 있다. 설명: 돈을 벌어야 한다는 것은 월급을 벌어야 한다는 것을 의미하지만, 이 문장에서 그 사람은 월급을 벌지 않기 위해 고군분투하고 있다.

영어 문장은 'To bring home the bacon'을 '생활비를 벌다' 라는 의미의 관용구로 사용하고 있으나 가설을 한국어로 번역하는 과정에서 '베이컨을 집으로 가져오다' 로 직역되면서 관용구의 의미를 상실하게 되었다.

04

결론

NILE 의 모델을 이용하는 것이 일반적인 자연어 추론 태스크에 비해 한국어와 영어에서 각각 1.0 그리고 0.6 의 성능이 향상되는 결과

이를 통해 Natural Language Explanation 을 생성하여 활용하는 것이 레이블 예측에 긍정적인 영향을 주었다고 판단

한계점

1. 실험을 위한 적합한 데이터셋의 부재
2. 데이터셋 중 정보성이 떨어지는 사례를 필터링 후 사용
3. 생성된 Natural Language Explanation 에 대한 평가
4. 한국어 생성 모델의 변경

발표를 들어주셔서

감사합니다