자연어 생성 모델을 이용한 준지도 학습 기반 한국어 사실 확인 자료 구축

Semi-Supervised Data Augmentation Method for Korean Fact Verification Using Generative Language Models

정재환○, 전동현, 김선훈, 강인호

스탠퍼드 대학교, 네이버

jaehwanj@stanford.edu, donghyeon.jeon@navercorp.com, seonhoon.kim@navercorp.com,

once.ihkang@navercorp.com

요 약

한국어 사실 확인 과제는 학습 자료의 부재로 인해 연구에 어려움을 겪고 있다. 본 논문은 수작업으로 구성된 학습 자료를 토대로 자연어 생성 모델을 이용하여 한국어 사실 확인 자료를 구축하는 방법을 제안한다. 본 연구는 임의의 근거를 기반으로 하는 주장을 생성하는 방법 (E2C)과 임의의 주장을 기반으로 근거를 생성하는 방법 (C2E)을 모두 실험해보았다. 이때 기존 학습 자료에 위 두 학습 자료를 각각 추가하여 학습한 사실 확인 분류기가 기존의 학습 자료나 영문 사실 확인 자료 FEVER를 국문으로 기계 번역한 학습 자료를 토대로 구성된 분류기보다 평가 자료에 대해 높은 성능을 기록하였다. 또한, C2E 방법의 경우 수작업으로 구성된 자료 없이 기존의 자연어 추론 과제 자료와 HyperCLOVA Few Shot 예제만으로도 높은 성능을 기록하여, 비지도 학습 방식으로 사실 확인 자료를 구축할 수 있는 가능성 역시 확인하였다.

주제어: 한국어 사실 확인, 자연어 생성 모델, HyperCLOVA

1. 서론

사실 확인 (Fact Verification)은 임의의 주장이 사실인지 아닌지를 주장과 관련된 근거를 토대로 판단하는 자연어 처리 분야이다[1]. 사실 확인 모델은 주장과 주장의 논점과 관련된 근거가 주어지면 그 둘 사이에서의 Entailment 성립 여부로 주장의 사실 여부를 추측한다. 사실 확인은 인터넷상에 존재하는 가짜 뉴스를 구분하거나 자연어 생성 모델의 출력 값을 검증 과정에서 사용된다. 여러 방면에서 활용도가 높은 사실 확인은 최근 활발히 연구되고 있는 자연어 처리 분야 중 하나이다[1].

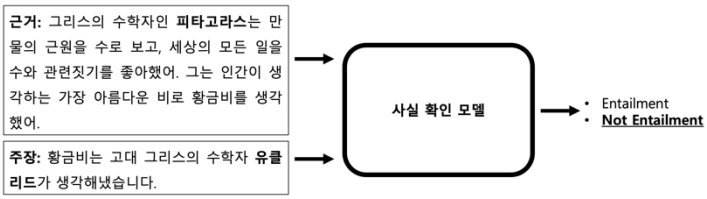

그림 1: 사실 확인 과제 예제

사실 확인 모델을 구축하는 과정에서 만나는 가장 큰 난관은 바로 학습 자료 확보이다. 하나의 사실 확인 예제로는 주장, 근거, 그리고 주장과 근거 사이의 관계가 주어져야 하는데, 근거가 주장을 뒷받침하면 Entailment 예제, 뒷받침하지 않으면 Not Entailment 예제가 된다. 서로 관련이 있는 주장과 근거를 생성하고, 또 그 둘 사이의 Entailment/Not Entailment 관계를 추론하는 과정에서 사람이 개입해야 하는 경우가 많다. 사실 확인 자료를 수작업으로 구성하는 방법은 시간과 비용 측면에서 비효율적이기에, 사실 확인 학습 자료 생성 과정을 자동화하려는 노력이 꾸준하게 이어져 오고 있다 [2].

사실 확인 학습 자료의 Not Entailment 예제를 자동으로 생성하는 방법은 오랫동안 연구되어 왔다. Entailment 예제의 경우, 위키피디아 등에서 사실 기반의 문서 문단을 근거로 삼은 뒤, 근거의 내용 중 일부를 취합하여 주장을 형상하는 방식으로 만들 수 있다. 그러나 근거와 Not Entailment 관계를 띄는 주장을 만드는 과정에서는 근거 외부의 정보가 추가적으로 필요하다. Not Entailment를 유도하는 방법으로는 근거 밖의 정보를 추가한 중문 생성, NER을 토대로 Entailment 주장 내의 Entity E를 다른 Entity E'로 치환하는 Entity Swapping, 그리고 단순히 Entailment 주장을 부정하는 방법 등을 꼽을 수 있다. 이러한 휴리스틱 기반 Not Entailment 예제 생성은 일상생활에서 등장하는 다양한 종류의 Not Entailment 주장을 망라하지 못한다는 한계를 지닌다.

더 나아가, 지금까지 연구되어온 사실 확인 학습 자료는 영문 예제로 구성되어 있기에, 한국어를 기반으로 하는 사실 확인 과제는 영어에 비해 훨씬 더 제한적이다 [1,3,4]. Multilingual 사실 확인을 위해 개발된 학습 자료 몇 건이 존재하지만, 한국어에 초점을 맞춘 사실 확인 학습 자료는 국내에 아직 많이 부족하다.

본 논문은 수작업으로 구축된 한국어 사실 확인 자료 D를 자동으로 확장하는 2개의 data augmentation 방법론을 제안한다. 첫 번째 방법(E2C)은 주어진 근거(Evidence)를 기반으로 하는 주장(Claim)을 자연어 생성 모델을 이용하여 만드는 것이다. 본 논문은 근거에 대한 Entailment 주장을 생성하는 KoBART와 Not Entailment를 생성하는 KoBART를 각각 D 내의 Entailment 예제와 Not Entailment 예제를 기반으로 학습시켰다 [5]. 그 후, D 밖의 외부 위키피디아 문서 D'를 토대로 하는 주장을 새로 생성하여 학습 자료를 증강하였다. 두 번째 방법(C2E)은 주장(Claim)을 뒷받침하는 근거(Evidence)를

HyperCLOVA를 이용하여 만든다. [6]. 네이버가 한국어 문서를 기반으로 학습시킨 HyperCLOVA는 긴 문장을 생성할 때 Pre-trained된 지식에 대한 의존도가 높아짐이 확인되었다. 따라서 HyperCLOVA가 주장과 관련된 근거를 여러 문장에 걸쳐 서술할 때, HyperCLOVA의 생성문이 주장의 주어와는 관련 있지만, 주장의 논점과는 관련 없는 설명문이 생성되는 것이 종종 관찰되었다. 본 논문은 위의 특성을 이용해 주장에 대해 자연스러운 Not Entailment 근거를 생성해내고자 하였다. HyperCLOVA가 생성한 근거와 주장 쌍에는 D에 학습된 사실 확인 분류기를 이용하여 Entailment / Not Entailment pseudo-label을 부여했고, Not Entailment에 해당하는 예제들 중 일부는 주장의 주어에 Entity Swapping을 적용해 확실한 Not Entailment 주장으로 변환하였다.

본 논문은 위의 방식으로 확장한 사실 확인 자료의 품질을 확인하기 위해 KoELECTRA 기반 사실 확인 모델을 본 예제를 토대로 학습시켜보았다[7]. 그 결과, D에 C2E와 E2C 방식을 둘 다 적용하여 증강한 학습 자료를 사용할 경우 평가 자료에 대한 모델의 accuracy가 3% 향상됨을 확인하였다.

2. 관련 연구

사실 확인을 위한 학습 자료 구성 방법은 오랫동안 연구되어 왔다. FEVER와 FEVEROUS는 위키피디아 문서를 근거로 삼는 SUPPORTED, REFUTED, NOT ENOUGH INFO 예제를 수작업으로 구성한 자료로, 사실 확인 분야에서 벤치마크로 종종 참고되곤 한다[1,4]. 본 논문에서는 사실 확인 예제를 Entailment과 Not Entailment로 구성하였는데, 그 이유로는 자연어 생성 모델을 통해 생성한 REFUTED 예제와 NOT ENOUGH INFO 예제 사이의 경계가 모호하다는 점과, 사실 확인 분류기를 실제로 응용 시 모델 예측값이 REFUTED 이나 NOT ENOUGH INFO일 때의 해당 주장의 처리 방안이 비슷하다는 점을 꼽을 수 있다. DANFEVER는 FEVER 논문에서 예제를 구성한 방법을 참고하여 덴마크어 사실 확인 학습 자료를 구축하였다[8]. FEVER의 주장은 누가, 언제, 어디서 등 간단한 정보를 묻는 경우가 많아, Adversarial NLI의 저자들은 사실 확인 분류기가 틀리는 어려운 예제를 사람이 auto-regressive하게 구성하는 학습 자료 구축 방법을 제안한다 [9]. 하지만 위의 방법은 모두 수작업에 의존하기에, 시간과 비용 부담이 높다는 단점을 지닌다.

사실 확인 예제의 주장과 근거를 처음부터 만들기보다, 이미 존재하는 질의응답 학습 자료의 예제에서 주장과 근거를 추출하는 방법 역시 연구되고 있다. 질의응답 예제로 질의 Q, 응답 A, 그리고 근거 E가 주어지면, QA2D는 Q와 A를 결합한 평서문 주장 (Declarative Claim) D와 기존의 근거 E를 토대로 자연어 추론 학습 자료를 구성하는 방법을 제안하였다[10]. FAVIQ는 서로 비슷하지만, 다른 질의 $Q1$, $Q2$를 disambiguate해놓은 질의응답 학습 자료 ($Q1$, $A1$), ($Q2$, $Q2$)가 있다면, ($Q2$, $A1$)과 ($Q1$, $A2$)에 QA2D기법을 적용하면 난이도 높은 Not Entailment 주장을 만들 수 있을 것이라 제안한다[11]. 하지만 위의 두 방법은 주장과 관련된 근거가 평행하게 주어진다는 전제를 지니고 있기에, 사람이 수작업으로 찾은 근거 없이 Q와 A만으로 사실 확인 예제를 만들지 못한다는 한계를 지닌다.

QACG(Question Answer Claim Generation)는 사람의 손을 거치지 않고 순전히 zero shot으로 사실 확인 학습 자료를 구성하는 방법이다[2]. QACG는 우선 위키피디아 문서 문단을 근거 E로 설정한 후, 해당 근거를 토대로 질의 Q와 그 질의에 대한 응답 A를 추출하여 QA2D방식으로 Entailment 주장을 생성한다. Not Entailment 주장은 응답 A를 A'로 대체하여 주장을 생성하거나, E 밖의 문단 E'에 담겨있는 내용을 주장에 추가하는 방식으로 생성된다. QACG식 사실 확인 자료 구성법은 해당 논문이 소개한 FEVER 베이스라인에서 준수한 성능을 보였지만, 휴리스틱에 의존하여 Not Entailment 주장을 구성한다는 한계점을 지닌다.

3. 학습 자료 구성

본 논문이 연구한 사실 확인 학습 자료에서 한 개의 예제는 (주장, 근거, label) 식으로 구성되어 있다. 본 논문은 우선 수작업으로 구성된 기본 학습 자료를 소개한 후, 이 기본 학습 자료를 증강시키는 방법인 E2C와 C2E를 다룬다. 또한, 학습 자료의 품질을 비교 평가하기 위해 기계 번역으로 FEVER를 국문으로 번역한 자료 역시 짚고 넘어간다.

3.1. 기본 학습 자료

기본 학습 자료는 네이버 내부에서 지식백과 문서들을 기반으로 구축한 질의응답 자료를 토대로 구성되었다. 질의응답 자료 내의 예제로 근거 E, 질의 Q, 그리고 응답 A가 주어지면, QA2D 방식으로 Q와 A를 기반으로 하는 주장 C를 만든 다음, E와 C 사이의 Entailment/Not Entailment 여부를 사람이 직접 판단하였다. 위의 방식으로 사실 확인 학습 자료 예제를 14,182개 구성하였다.

표 1. 기본 학습 자료 예제

근거	수성은 태양과 가장 가까운 행성이다. 항상 밝은 태양 가까이 있어 관측하기가 쉽지 않다. 해가 진 직후 서쪽 하늘이나, 해가 뜨기 직전 동쪽 하늘에서만 볼 수 있다.작은 궤도와 빠른 공전 속도를 가져 공전 주기는 88 일밖에 되지 않는다.
질의 1	수성의 공전 주기는 얼마나 되나요?
응답 1	88 일입니다.
주장 1	수성의 공전 주기는 88 일 입니다.
label 1	Entailment
질의 2	수성은 해가 뜨기 전 몇 시쯤 보이나요?
응답 2	해가 뜨기 전 새벽에 동쪽 하늘에서 잠시 보이다가 해가 뜨면 보이지 않아요. 2 분 30 초 정도 보이지요.
주장 2	수성은 해가 뜨기 전 새벽에 동쪽 하늘에서 잠시 보이다가 해가 뜨면 보이지 않아요. 2 분 30 초 정도 보이지요

label 2	Not Entailment

표 1.의 주장 1은 근거를 토대로 검증을 할 수 있기에 Entailment 관계가 성립한다. 반대로, 주장 2의 "2분 30초 정도 보이지요."는 근거를 토대로 검증이 되지 않기에 Not Entailment가 적절하다.

3.2. 근거 기반 주장 생성 (E2C)

근거(Evidence) 기반 주장(Claim) 생성 방식(E2C)은 자연어 생성 모델이 근거를 입력으로 받으면 근거와 관련된 주장을 생성하는 방식으로 학습 예제를 만든다. 본 논문에서 하나의 자연어 생성 모델을 기반으로 Entailment, Not Entailment 주장을 둘 다 생성을 시도하였을 때, 모델이 Not Entailment를 의도하고 생성한 예제들이 실제로는 Entailment 주장인 경우를 종종 확인하였다. 따라서, 본 논문은 각각의 주장을 생성하는 자연어 생성 모델을 따로 구성하였다. E2C 방식에서는 우선 기본 학습 자료 D를 Entailment 예제 $D_{Entailment}$와 Not Entailment 예제 $D_{Not\ Entailment}$로 나눈 후, 각각 자료를 토대로 Entailment 주장 자연어 생성 모델 $M_{Entailment}$과 Not Entailment 주장 자연어 생성 모델 $M_{Not\ Entailment}$을 학습한다. $M_{Entailment}$ 과 $M_{Not\ Entailment}$가 학습이 완료된 후, D 외부의 위키피디아 문서 문단을 근거로 삼는 주장을 $M_{Entailment}$ 과 $M_{Not\ Entailment}$를 이용해 추가 생성한다.

본 논문에서는 E2C 기법의 자연어 생성 모델로 HyperCLOVA을 사용할 때, 입력 근거와 모델이 출력한 주장 사이의 어휘 유사도가 매우 높아 사실 확인 과제에 유용한 샘플이 되지 않으리라 판단하였다. 따라서, 본 논문은 입력 근거와 생성 주장의 어휘 유사도를 낮추기 위해, Entailment 예제를 생성하는 KoBART와 Not Entailment 예제를 생성하는 KoBART 모델을 각각의 label에 해당하는 기본 학습 자료 내 예제에 finetuning 하여 사용하였다. 본 논문은 위의 KoBART 모델들 2개를 토대로 E2C 기법을 적용하여 사실 확인 예제 22,973개를 추가로 구성하였다.

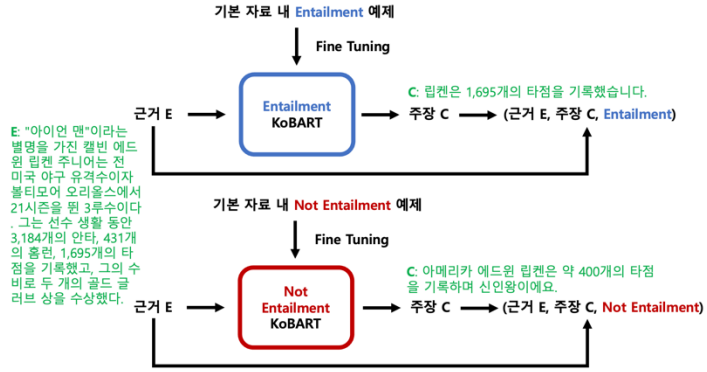

그림 2. E2C 학습 자료 생성 방법

표 2. E2C 학습 자료 예제

근거	"아이언 맨"이라는 별명을 가진 캘빈 에드윈 립켄 주니어는 전 미국 야구 유격수이자 볼티모어 오리올스에서 21 시즌을 뛴 3 루수이다. 그는 선수 생활 동안 3,184 개의 안타, 431 개의 홈런, 1,695 개의 타점을 기록했고, 그의 수비로 두 개의 골드 글러브 상을 수상했다.
Entailment KoBART 기반 주장 1	립켄은 1,695 개의 타점을 기록했습니다.
label 1	Entailment
Not Entailment KoBART 기반 주장 2	아메리카 에드윈 립켄은 약 400 개의 타점을 기록하며 신인왕이에요.
label 2	Not Entailment

3.3. Claim 기반 Evidence 생성 (C2E)

주장(Claim) 기반 근거(Evidence) 생성 방식(C2E)은 입력으로 들어온 주장에 대한 근거를 초대형 언어 모델(HyperCLOVA)가 학습한 Pre-trained 지식을 이용하여 생성한다. HyperCLOVA를 이용하여 주장을 뒷받침하는 근거를 여러 문장으로 생성하면 다음의 2가지 특징들이 관찰된다: 1) HyperCLOVA가 생성하는 근거는 주장의 주어와 관련된 Pre-trained 지식에 중점을 맞춰 서술하여, 해당 근거가 입력 주장의 논점과 관련이 없는 경우가 종종 발생한다 2) HyperCLOVA의 생성문은 사실을 기반으로 하는 Pre-trained 지식을 토대로 구성되기에, 사실이 아닌 주장이 입력으로 주어져도 생성되는 근거는 사실을 기반으로 한다. (예. 입력 주장: 버락 오바마는 제 19대 대한민국 대통령이다. HyperCLOVA생성 근거: 버락 오바마는 미국의 44대 대통령으로, 대통령 임기 전에는 일리노이주 연방 상원의원으로 선출 되었었다.) 이렇듯 주장의 주어와는 관련되지만, 주장의 논점과는 초점이 다른 근거는 자연스러우며 어려운 Not Entailment 예제를 만들 것이다. HyperCLOVA가 출력한 (주장, 근거)에 대한 pseudo label은 기본 학습 자료 D를 토대로 학습된 사실 확인 분류기를 이용하여 생성한다.

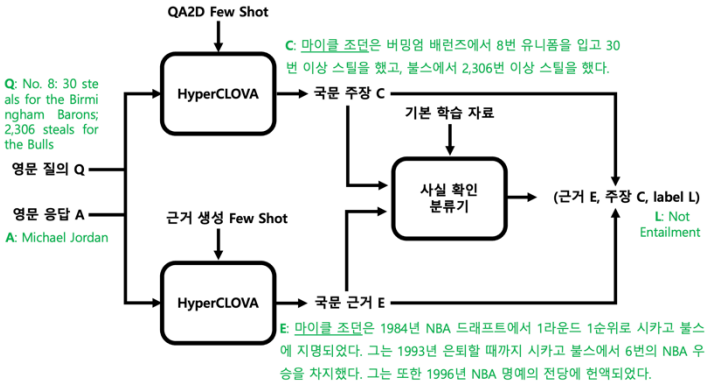

그림 3. C2E 학습 자료 생성 방법

본 논문은 인터넷 상에 공개된 JEOPARDY 학습 자료의 질의응답 샘플을 seed data로 삼아 학습 자료를 생성하였다[12]. JEOPARDY는 미국의 유명 퀴즈 프로그램 JEOPARDY에 등장했던 질의 Q와 응답 A를 모아둔 학습 자료인데, 퀴즈 프로그램 참가자가 외부 문서 없이 순전히 Q에 의

존해서 단답형 응답 A를 추론할 수 있을 만큼 Q가 자세하게 구성되어 있다. 이는 Jeopardy의 Q 내부에 HyperCLOVA가 근거 생성 시 참고할 만한 정보가 충분히 압축되어 있음을 의미한다. 본 논문이 JEOPARDY 자료를 선정한 또 다른 이유는 바로 해당 자료가 영문이라는 점이다. 국문 질의와 응답을 토대로 주장과 근거를 각각 생성하면 둘 사이의 어휘 유사도가 높게 기록되지만, HyperCLOVA에게 영문 질의 응답을 입력해주고 그에 대한 주장과 근거를 국문으로 생성할 때 둘 사이의 중복 어휘 출현 빈도가 낮아짐을 확인하였다. 이에 따라 좋은 품질과 난이도를 가진 데이터를 구축할 수 있다.

C2E 방식에서는 우선 JEOPARDY 내의 Q와A를 QA2D 방식으로 병합하여 주장 C를 생성하고, Q와 A의 연관성을 설명하는 "근거 생성 Few Shot"을 입력 받은 HyperCLOVA는 Q와 A 사이의 상관관계를 설명하는 근거 E를 생성한다. 주장 C와 근거 E 사이의 Entailment 여부에 대한 pseudo-label은 기본 학습 자료로 학습된 Larva Large 모델을 사실 확인 분류기로 생성한다.

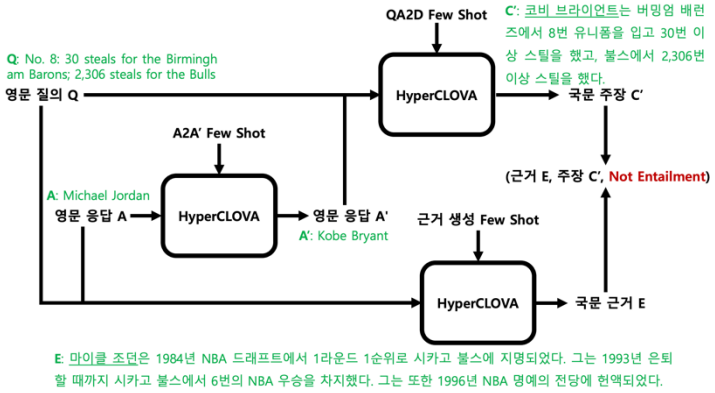

그림 4: C2E 기반 거짓 주장 C'생성 방법

C2E의 Not Entailment 예제 중 대다수는 표 3.의 근거 2와 주장 2-1처럼 주장의 논점과 근거의 논점이 다른 경우이기에, 주장과 근거가 서로 상충하는 예제는 자연스럽게 생성될 수 없다. 따라서 본 논문은 Not Entailment 예제 중 절반의 주장을 Entity Swapping을 통해 표 3.의 주장 2-2 형식의 거짓 주장C'으로 변환한다. JEOPARDY 자료 내의 질의 Q에 대한 적절한 응답 A는 하나이기 때문에, A를 비슷하지만 다른 응답 A'으로 치환할 때, Q와 A'로 구성된 주장 C'는 거짓인 주장이 된다. Q와 A를 기반으로 HyperCLOVA가 구성한 근거 E는 대부분 참인 문장이니, E와 C' 사이에는 Entailment 관계가 성립될 확률이 매우 낮다. 본 논문은 그림 4 방식으로 주장 C'를 생성하는데 사용될 seed data Q 와 A로는 C2E 방식에서 Not Entailment 예제들을 생성한 Q 와 A를 재사용하였다. HyperCLOVA에 아래의 A2A' 퓨샷과 A를 입력하여 A'를 생성하였고, Q와 A'를 QA2D 방식으로 결합하여 거짓 주장 C'를 구성하였다. 본 실험에서는 Not Entailment의 Claim을 C : C' = 1:1 비율로 구성하였다. (A2A' Few Shot 예: 입력: 메탈리카와 비슷하지만 다른 것은? 답: 린킨 파크. 입력: 부쿠레슈티와 비슷하지만 다른 것은? 답: 부다페스트. 입력: 얌이랑 비슷하지만 다른 것은? 답: 감자. 예: 트램펄린이랑 비슷하지만 다른 것은? 답: 롤러코스터.)

표 3. C2E 학습 자료 예제

근거 1	1963년 11월 1일, McDonald's는 전세계 매장에서 100억번째 햄버거를 판매했다. 이는 1955년 11월 1일 미국 일리노이주 오크브룩에 첫 매장을 연 지 45년 만에 달성한 기록이다.
주장 1 C: Q + A	1963년 **맥도날드** 햄버거는 100억번째 햄버거를 판매했다.
label 1	Entailment
근거 2	**마이클 조던**은 1984년 NBA 드래프트에서 1라운드 1순위로 시카고 불스에 지명되었다. 그는 1993년 은퇴할 때까지 시카고 불스에서 6번의 NBA 우승을 차지했다. 그는 또한 1996년 NBA 명예의 전당에 헌액되었다.
주장 2-1 C: Q + A	**마이클 조던**은 버밍엄 배런즈에서 8번 유니폼을 입고 30번 이상 스틸을 했고, 불스에서 2,306번 이상 스틸을 했다.
label 2-1	Not Entailment
주장 2-2 C': Q + A'	**코비 브라이언트**는 버밍엄 배런즈에서 8번을 달고 30개의 스틸을 기록했고, 불스에서 2,306개의 스틸을 기록했다.
label 2-2	Not Entailment

C2E 방식의 근거를 HyperCLOVA 기반 생성이 아닌 주장의 키워드 기반 검색으로 구성을 한다면, 지식 백과 상에서 선정된 문단이 실제 주장과 큰 관련이 없는 경우를 종종 확인하였다. 또한, 검색된 문서와 주장 사이의 Entailment / Not Entailment 여부가 불분명하여, 본 논문에서는 C2E 방식으로 주장에 대한 근거를 생성하였다.

3.4. FEVER 학습 자료 국문 번역

본 논문에서는 E2C와 C2E 방식으로 구성된 자료의 품질을 평가하기 위해 위 자료를 토대로 학습된 모델 성능을 국문 FEVER 번역본에 학습된 모델의 성능과 비교한다.

FEVER 학습 자료 예제로는 주장과 근거, 그리고 그 둘 사이의 관계를 나타내는 label (Supported, Refuted, Not Enough Info)이 주어진다. FEVER의 Not Enough Info 예제는 따로 정답인 근거가 주어지지 않기에, 본 논문은 FEVER의 Supported와 Refuted 예제를 국문으로 기계 번역하여 학습 자료를 구성한다.

표 4. FEVER 학습 자료 예제

근거 1	아드리엔 바일론 (1983년 10월 24일 ~)은 미국의 싱어송라이터, 레코딩 아티스트, 배우, 댄서, 텔레비전 퍼스널리티이다.
주장 1	애드리엔 베일론은 회계사예요.
label 1	Not Entailment
근거 2	IO, 또는 IO Chicago (이전에는 "Improv 올림픽"으로 알려졌던)는 IO West 라고 불리는 로스앤젤레스에 지점을 두고 시카고 중심부에 있는 즉흥 극장 및 훈련 센터이다.
주장 2	로스앤젤레스에는 IO 극장의 지점이 있습니다.
label 2	Entailment

4. 실험 및 평가

4.1. 실험 준비

본 논문에서는 3.1의 기본 학습 자료를 Baseline으로 설정한 후, E2C, C2E, FEVER 방식으로 구성된 학습 자료를 각각 기본 학습 자료에 추가하여 사실 확인 분류기를 학습하였을 때 평가 자료에 나타나는 모델 accuracy의 변화를 관찰한다. 또한, 위에서 생성한 학습 자료 4개를 모두 병합한 학습 자료 TOTAL을 기반으로 하는 모델의 성능 역시 측정한다. E2C 방식에서 사용한 KoBART 모델은 기본 학습 자료에 대해 fine tuning 되었고, default 하이퍼 파라미터 설정값 아래에서 학습을 시켰다. C2E 방식에서의 HyperCLOVA 모델로는 네이버에서 학습한 39B 크기의 HyperCLOVA 모델을 사용하였고, 사실 확인 분류기로는 MNLI를 기반으로 학습한 Larva large 모델에 기본 학습 자료를 전이 학습한 모델을 활용하였다. FEVER 학습 자료를 국문으로 번역할 때는 네이버 파파고의 영한 번역기를 사용하였다. 위의 방식으로 구성된 학습 자료의 크기는 다음과 같다:

표 5. 사실 확인 학습 자료 구성

	기본	E2C	C2E	FEVER	TOTAL
Entailment%	47.5%	54.0%	50.0%	50.0%	50.6%
# Samples	14,182	22,973	25,523	25,817	88,495

평가 자료로는 네이버 지식백과 질의응답 자료를 기반으로 수작업으로 구성되었는데, 그 구성은 다음과 같다:

표 6. 사실 확인 평가 자료 구성

	평가 자료
Entailment %	72.3%
# Samples	1,485

사실 확인 분류기는 주장과 근거가 주어지면, 그 둘 사이의 관계를 Entailment / Not Entailment로 이진 분류하는 과제를 수행한다. 분류기 모델로는 KoELECTRA Base 모델과 KoELECTRA Small 모델을 선정하였고, default 하이퍼 파라미터 설정값 아래에서 본 논문에서 구성한 학습 자료에 대해 fine-tuning 하였다.

4.2. 실험 평가

사실 확인 분류기 모델의 성능을 기록하기 위해서는, 모델 체크 포인트 값들 중 accuracy가 가장 높게 기록된 버전을 토대로 accuracy, recall, precision, f1 값을 기록하였다.

표 7. 기본 학습 자료 와 KorNLI-MNLI 비교

	기본	KorNLI-MNLI
Accuracy	**0.715**	0.696
Recall	0.676	**0.686**
Precision	**0.890**	0.850
F1	**0.769**	0.759

본 논문은 연구의 Baseline이 될 기본 학습 자료의 품질을 평가하기 위해, 기본 학습 자료를 토대로 학습된 KoELECTRA base 모델과 KorNLI[13]의 MNLI를 토대로 학습한 KoELECTRA base 모델의 성능을 비교한다. KorNLI-MNLI는 1개의 문장으로 구성된 자연어 추론 과제에 최적화된 학습 자료지만, 사실 확인 과제 예제는 여러 문장으로 구성된다는 데 차이가 있다. KorNLI-MNLI 원본 자료는 Contradiction : Entailment : Neutral 샘플이 1:1:1의 비율로 392,700개 있지만, 본 논문은 label을 Entailment와 Not Entailment로 구성하였기에 KorNLI-MNLI의 Contradiction, Entailment, Neutral 샘플을 1:2:1의 비율로 결합하여 학습 자료 크기를 261,800개로 축소하였다.

실험 결과, 기본 학습 자료의 크기가 KorNLI-MNLI자료 크기의 5.4%에 불과하지만, 평가 자료에 대해서는 더 높은 accuracy와 f1을 기록하여, 기본 학습 자료가 사실 확인 과제에 더 알맞은 Baseline임을 확인하였다.

표 8. KoELECTRA Small 사실 확인 분류기 성능

	기본	E2C +기본	C2E +기본	FEVER +기본	TOTAL
Accuracy	0.700	0.699	0.704	**0.711**	0.729
Recall	0.664	0.649	0.680	**0.702**	0.735
Precision	0.878	**0.890**	0.868	0.859	0.857
F1	0.756	0.751	0.762	**0.773**	0.791

표 9. KoELECTRA Base 사실 확인 분류기 성능

	기본	E2C +기본	C2E +기본	FEVER +기본	TOTAL
Accuracy	0.715	**0.739**	**0.739**	0.730	0.744
Recall	0.676	0.728	**0.739**	0.724	0.749
Precision	**0.890**	0.879	0.868	0.868	0.866
F1	0.769	0.796	**0.798**	0.790	0.803

본 논문에서 소개한 학습 자료를 토대로 학습된 KoELECTRA Small 모델과 Base 모델 둘 다 평가 데이터에서 낮은 Recall과 높은 Precision을 기록하였는데, 이는 평가 자료에서 모델이 False negative를 많이 내놓음을 의미한다.

KoELECTRA Small 모델은 FEVER 국문 번역본을 토대로 학습되었을 때 가장 높은 accuracy (0.711)를 기록하였는데, 이는 기본 학습 자료에 대한 모델 성능 (0.700)에 비해 눈에 띄는 상승은 아니다. KoELECTRA Small을 기반으로 하는 사실 확인 분류기는 본 논문에서 소개하는 data augmentation 방법을 개별적으로 적용할 때는 모델 성능을 크게 개선하지 못하였지만, 모든 방법을 추가한 Total의 경우 기본 학습 자료보다 accuracy가 0.03만큼 증가하여 data augmentation의 효과가 관찰되었다.

사실 확인 분류기로 KoELECTRA Base 모델을 사용할 경우 data augmentation의 효과가 더 뚜렷하게 나타난다.

KoELECTRA small 모델에서는 FEVER 자료를 기본 학습 자료에 추가하는 것이 제일 높은 성능을 기록한 반면, KoELECTRA Base는 E2C와 C2E의 data augmentation 기법들이 우세를 보였다. FEVER의 예제보다 E2C나 C2E의 예제가 더 길고 복잡해서, Base 모델에서의 성능이 더 높게 기록된 것으로 사료된다. 모든 방법을 다 추가한 Total은 accuracy가 0.744로 가장 높게 나왔는데, 이는 기본 학습 자료를 기반으로 한 모델의 accuracy보다 0.03 만큼 증가한 수치이다.

4.3. C2E 관련 추가 실험

본 논문에서는 KoELECTRA Base에서 제일 높은 accuracy와 f1을 기록한 C2E 기법과 관련된 추가 실험을 KoELECTRA Base를 사실 확인 분류기로 사용하여 진행하였다.

표 9. Not Entailment 주장 구성 방법에 따른 성능 비교

	$C \& C'$	C'
Accuracy	**0.739**	0.738
Recall	0.739	**0.745**
Precision	**0.868**	0.862
F1	0.798	**0.799**

기존 C2E 방식은 Not Entailment 주장의 절반을 정답 A 기반으로 한 주장 C, 나머지 절반은 틀린 답 A'를 기반으로 한 C'로 구성한다 ($C \& C'$). 표 9.에서 Not Entailment 주장을 전부 C'로 구성하였을 때 (C') 모델 성능의 변화를 확인해보는데, C' 기반 모델 성능이 $C \& C'$ 기반 모델 성능보다 높은 recall과 낮은 precision을 기록함을 확인하였다. C' 이 $C \& C'$보다 더 좁은 범위의 주장을 포함하기에, C' 기반 사실 확인 분류기는 Entailment로 추측값을 내놓는 threshold가 기존 $C \& C'$ 기반 분류기보다 낮아졌을 것이다.

표 10. HyperCLOVA 모델별 학습 자료의 구성

	HyperCLOVA 13B	HyperCLOVA 39B
Entailment %	50.0%	31.3%
# Samples	13,124	13,124

본 논문은 HyperCLOVA 모델의 크기가 생성된 학습 자료의 품질에 미치는 영향을 연구하기 위해 네이버에서 자체 학습한 HyperCLOVA 39B 모델과 HyperCLOVA 13B 모델을 각각 이용해 학습 자료를 구성해보았다. 우선, HyperCLOVA 13B를 이용하여 JEOPARDY 자료를 seed로 하는 사실 확인 예제 30,000개를 생성한 후, Entailment : Not Entailment 예제의 비율이 1:1이 되도록 예제의 크기를 13,124개로 줄여 HyperCLOVA 13B 모델 기반의 학습 자료를 구성하였다. HyperCLOVA 39B 모델로는 앞서 생성한 13,124개 예시의 seed가 되는 JEOPARDY의 질의응답 예제를 동일하게 사용하여 13,124개의 예제를 만들었다. 그 결과, HyperCLOVA 39B가 생성한 자료는 Entailment : Not Entailment 비율이 맞춰지지 않았다.

표 11. HyperCLOVA 크기에 따른 분류기 성능 비교

	HyperCLOVA 13B	HyperCLOVA 39B
Accuracy	0.730	**0.741**
Recall	0.694	**0.733**
Precision	**0.896**	0.876
F1	0.782	**0.798**

HyperCLOVA 39B가 C2E 방식으로 생성한 학습 자료가 HyperCLOVA 13B 기반 학습 자료보다 여러 평가 지표에서 더 높은 성능을 기록하였음을 확인할 수 있는데, 이는 생성 모델의 크기가 클수록 고품질의 학습 자료 샘플이 생성됨을 의미한다.

표 12. C2E내 분류기의 학습 자료에 따른 성능 비교

	기본	KorNLI-MNLI
Accuracy	0.739	**0.743**
Recall	0.739	**0.748**
Precision	**0.868**	0.866
F1	0.798	**0.803**

C2E 방식의 pseudo label 생성기로 사용되는 사실 확인 분류기는 기본 학습 자료를 토대로 학습되어 있기에, 수작업으로 구성된 사실 확인 자료 없이는 C2E 방식을 시도하기 어렵다. 따라서 본 연구는 pseudo label 생성용 사실 확인 분류기를 KorNLI-MNLI에 fine tuning 시키어 C2E 방법을 적용해보았다. 그 결과, KorNLI-MNLI를 기반으로 하는 pseudo label 생성기가 더 높은 성능을 기록하는 학습 자료를 구성하여, 수작업으로 구성된 사실 확인 자료 없이도 HyperCLOVA의 Few Shot 예제만으로 C2E 방식을 도입할 수 있음을 확인하였다.

5. 결론

본 논문은 국문 사실 확인 과제를 위한 학습 자료 확장 방법을 제안한다. FEVER를 기계 번역으로 국문으로 변환하는 방법, 위키피디아 문서를 근거로 삼는 주장을 BART로 생성하는 방법, 그리고 Jeopardy 예제를 seed로 삼아 주장과 근거를 HyperCLOVA를 이용해 생성하는 방법을 시도해보았다. 그 결과, HyperCLOVA의 pre-trained 지식을 이용하는 C2E가 개별 방식 중 가장 높은 성능을 기록하였다. 더 나아가, C2E는 수작업으로 구성된 사실 확인 자료 없이 기존의 국문 자연어 추론 과제 학습 자료만으로도 양질의 사실 확인 학습 자료를 생성해 국문 사실 확인 자료 생성의 자동화 가능성을 확인하였다.

C2E 방식의 한계는 바로 생성된 예제들의 label이 pseudo-label이라는 점이다. 현재 방식으론 C2E 내부의 사실 확인 분류기가 mislabel하는 주장, 근거 쌍들을 noise로 처리하는데, pseudo-label 생성기의 정확도를 높이는 방법은 추가로 연구가 필요하다. 또한, C2E 방식이 생성하는 Entailment 예제들 중 여전히 주장과 근거 사이에 중복 어휘 출현이 나타나는 경우가 종종 있어, 이를 해소할 수 있는 방안 역시 향후 연구가 필요하다.

참고문헌

[1] J.Thorne, A.Vlachos , C.Christodoulopoulos , and A.Mittal. FEVER: a large-scale dataset for Fact Extraction and VERification. In NAACL-HLT, 2018.

[2] L. Pan, W. Chen, W. Xiong, M. Kan, and W. Wang. Zero-shot Fact Verification by Claim Generation. In ACL-IJCNLP, 2021.

[3] J. Th orne, A . Vlachos, O . Cocaras cu, C. Christodoulopoulos, and A. Mittal. The FEVER2.0 Shared Task. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 1-6, 2019.

[4] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, and A.Mittal. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In CoRR, 2021.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, B. Stoyanov, L.Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre -training for Natural L anguage G eneration, Trans lation, and Comprehension. In ACL, 2020.

[6] B. Kim, H. Kim, S. Lee, G. Lee, D. Kwak, D. Jeon, S. Park, S. Kim, S. Kim. D. Seo, H. Lee, M. Jeong, S. Lee, M. Kim, S. Ko, S. Kim, T. Park, J. Kim, S. Kang, N. Ryu, K. Yoo ,M. Chang, S. Suh, S. In, J. Park, K. Kim, H. Kim, J. Jeong, Y. Yeo, D. Ham, D. Park, M. Lee, J. Kang, I. Kang, J. Ha, W. Park, and N. Sung. What Changes Can Large-scale Language Models Bring? Intensiv e Study on NyperCLOVA: Billions-scale K orean Generativ e Pretrained Transformers. In EMNLP, 2021.

[7] K. Clark, M. Luong, Q. Le, and C. Manning. ELECTRA: Pre-training Tex t Encoders as Discriminators Rather Than Generators. In ICLR, 2020.

[8] J. Nørregaard, L . Derczynsk . D anFEVER: claim verification dataset for Danish. In Proceedings of NODALIDA, 2021.

[9] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiels. Adversarial NLI: A New Benchmark for Natural Language Understanding. In ACL, 2020.

[10] D. Demszky, K. Guu, and P. Liang. Transforming Question Answering Datasets Into Natural Language Inference. In CoRR, 2018.

[11] J. Park, S. Min, J, Kang, L. Zetttlemoyer, and H. Hajishirz i. FaVIQ: Fact Verification from Information-seeking Questions. In CoRR, 2021.

[12] [trexmatt], (2014, Jan. 11). 200,000+ Jeopardy! Questions in a JSO N file. [Online forum post]. Reddit.https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/

[13] J. Ham, Y. Choe, K. Park, I. Choi, and H. Soh. KorNLI and K or STS: New Benchmark Datasets for Korean Natural L anguage Understand ing. I n CoRR, 2020.