KIPS Trans. Softw. and Data Eng.
Vol.12, No.8 pp.371~380
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.8.371

General Relation Extraction Using Probabilistic Crossover

Je-Seung Lee† · Jae-Hoon Kim††

ABSTRACT

Relation extraction is to extract relationships between named entities from text. Traditionally, relation extraction methods only extract relations between predetermined subject and object entities. However, in end-to-end relation extraction, all possible relations must be extracted by considering the positions of the subject and object for each pair of entities, and so this method uses time and resources inefficiently. To alleviate this problem, this paper proposes a method that sets directions based on the positions of the subject and object, and extracts relations according to the directions. The proposed method utilizes existing relation extraction data to generate direction labels indicating the direction in which the subject points to the object in the sentence, adds entity position tokens and entity type to sentences to predict the directions using a pre-trained language model (KLUE-RoBERTa-base, RoBERTa-base), and generates representations of subject and object entities through probabilistic crossover operation. Then, we make use of these representations to extract relations. Experimental results show that the proposed model performs about 3 ~ 4%p better than a method for predicting integrated labels. In addition, when learning Korean and English data using the proposed model, the performance was 1.7%p higher in English than in Korean due to the number of data and language disorder and the values of the parameters that produce the best performance were different. By excluding the number of directional cases, the proposed model can reduce the waste of resources in end-to-end relation extraction.

Keywords : Relation Extraction, Deep Learning, Pre-Trained Language Model, Probabilistic Crossover

확률적 교차 연산을 이용한 보편적 관계 추출

이 제 승† · 김 재 훈††

요 약

관계 추출은 텍스트로부터 개체(named entity) 사이의 관계를 추출하는 과정이다. 전통적으로 관계 추출 방법은 주어와 목적어가 미리 정해진 상태에서 관계만 추출한다. 그러나 종단형 관계 추출에서는 개체 쌍마다 주어와 목적어의 위치를 고려하여 가능한 모든 관계를 추출해야 하므로 이 방법은 시간과 자원을 비효율적으로 사용한다. 본 논문에서는 이러한 문제를 완화하기 위해 문장에서 주어와 목적어의 위치에 따른 방향을 설정하고, 정해진 방향에 따라 관계를 추출하는 방법을 제안한다. 제안하는 방법은 기존의 관계 추출 데이터를 활용하여 문장에서 주어가 목적어를 가리키는 방향을 나타내는 방향 표지를 새롭게 생성하고, 개체 위치 토큰과 개체 유형 정보를 문장에 추가하는 작업을 통해 사전학습 언어모델(KLUE-RoBERTa-base, RoBERTa-base)을 이용하여 방향을 예측한다. 그리고 확률적 교차 연산을 통해 주어와 목적어 개체의 표상을 생성한다. 이후 이러한 개체의 표상을 활용하여 관계를 추출한다. 실험 결과를 통해, 제안 모델이 하나로 통합된 라벨을 예측하는 것보다 3 ~ 4%p 정도 더 우수한 성능을 보여주었다. 또한, 제안 모델을 이용해 한국어 데이터와 영어 데이터를 학습할 때, 데이터 수와 언어적 차이로 인해 한국어보다 영어에서 1.7%p 정도 더 높은 성능을 보여주었고, 최상의 성능을 내는 매개변수의 값이 다르게 나타나는 부분도 관찰할 수 있었다. 제안 모델은 방향에 따른 경우의 수를 제외함으로써 종단형 관계 추출에서 자원의 낭비를 줄일 수 있다.

키워드 : 관계 추출, 심층 학습, 사전학습 언어모델, 확률적 교차 연산

1. 서 론

정보 추출(information extraction)은 자연어 처리 분야에서 텍스트 내의 중요한 정보를 추출하는 과정을 뜻한다. 이는 텍스트 문서로부터 구조화된 데이터를 추출하여 이해하기 쉽고 유용한 형태로 변환하는 기술로, 텍스트에서 특정한 유형의 정보를 찾아내는 작업이다. 정보 추출에는 개체명 인식(named entity recognition), 관계 추출(relation extraction) 등이 포함되며, 개체명 인식은 텍스트에서 개체의 유형을 식별하는 작업이고, 관계 추출은 식별된 개체들 간의 관계를 추출하는 작업이다. 일반적으로 개체명 인식과 관계 추출 작업은 별도의 태스크로 분리되어 수행되지만, 종단형(end-to-end) 관계 추출은 개체명 인식과 관계 추출을 동시에 수행하는 작업이다. 먼저 일반적으로 관계 추출은 하나의 문장에서 미리 지정된 두 개체 간의 관계를 추출하는 작업을 의미하며,

※ 이 논문은 본 논문은 2023년 이제승의 한국해양대학교 컴퓨터공학과 석사논문인 "DiREx: 확률적 교차 연산을 이용한 문장 수준의 한국어 관계 추출"을 확장한 것임.
† 비 회 원 : 한국해양대학교 컴퓨터공학과 석사
†† 종신회원 : 한국해양대학교 컴퓨터공학과 및 해양인공지능융합전공 교수
Manuscript Received : May 23, 2023
First Revision : July 5, 2023
Accepted : July 19, 2023
* Corresponding Author : Jae-Hoon Kim (jhoon@kmou.ac.kr)

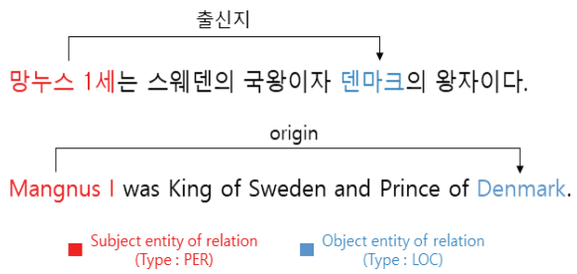

Fig. 1. An Example of Relation Extraction

관계 추출의 예시는 Fig. 1과 같다. 이는 문장에서 개체들 간의 관계를 추론하고, 그 결과로 구조화된 관계를 추출하는 것을 목표로 한다.

Fig. 1의 "망누스 1세는 스웨덴의 국왕이자 덴마크의 왕자이다." 에서 '망누스 1세'의 출신지가 '덴마크'임을 알 수 있다. 또한, 방향이란 주어 개체(head)에서 목적어 개체(tail)로 가는 방향을 의미하며, 문장에서 목적어가 주어보다 오른쪽에 위치할 경우 방향은 오른쪽이며, 반대의 경우 왼쪽이다. 따라서 Fig. 1에서의 방향은 오른쪽이며, 따라서 방향을 정하는 것은 어떤 개체가 주어와 목적어인지를 정하는 것과 같은 의미이다.

기존의 관계 추출 모델은 방향이 미리 주어지기 때문에 성능이 높을 수 있으나, 종단형 관계 추출과 같은 작업에서 모든 방향에 대한 경우의 수만큼 같은 문장을 여러 번 표상화(embedding)해야 한다는 문제점이 있다. 매개변수의 수가 방대한 사전학습 언어모델로 경우의 그 수만큼의 문장을 표상화하면 자원의 효율성과 학습 및 추론 속도가 저하된다. 본 논문에서는 방향까지 고려하여 두 개체의 관계와 주어, 목적어를 나타내는 모델을 제안함으로써, 이러한 문제를 완화하고자 한다. 이 모델은 기존의 관계 추출 모델과 달리 방향에 따라 여러 번의 표상화를 할 필요 없이, 한 번의 표상화로 두 개체 간의 관계와 방향을 동시에 판별할 수 있다.

Fig. 2는 여러 개의 개체가 있는 문장에서 관계를 추출하기 위한 문장을 만들어낸 예시이다. 'with subject, object'는 기존의 관계 추출 방법으로 각 개체에 주어를 나타내는 위치 토큰인 '[s]', '[/s]'와 목적어를 나타내는 위치 토큰인 '[o]', '[/o]'을 사용하고, 'without subject, object'는 제안 모델의 입력 문장으로 주어, 목적어가 없이 같은 위치 토큰인 '[e]', '[/e]'을 사용한다. 이는 방향에 대한 경우의 수를 제외할 수 있어 기존 방법보다 생성되는 문장의 수가 절반이 되고, 이에 따라 자원의 효율성을 향상시키며 학습 및 추론 속도를 개선할 수 있다. 제안 모델은 생성된 문장을 기반으로 방향을 예측하고, 각 방향의 확률을 이용하여 확률적 교차 연산(probabilistic crossover)을 수행하여 주어와 목적어 개체의 표상을 생성하고, 이를 통해 관계를 추출한다.

Fig. 3은 예시 문장을 제안 모델로 추출한 결과를 개략적으로 나타낸 그림이다. 방향이 없이 두 개체만 주어진 문장을 통해 방향 표지와 관계 표지를 예측하여 Fig. 1과 같은 결과를 얻을 수 있다. 즉, 방향이 'right'이므로 주어는 '망누스 1세',

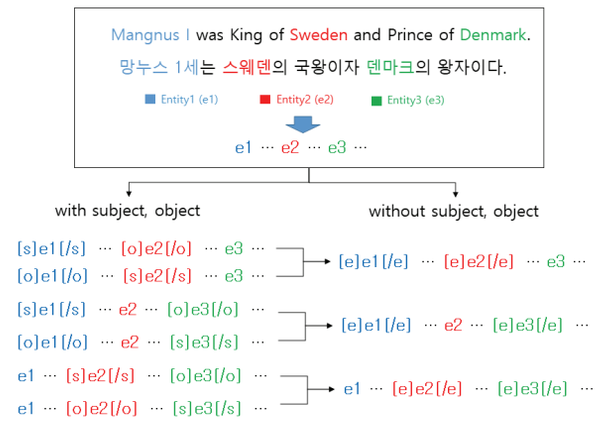

Fig. 2. Sentence Encoding Scheme for End-to-end RE Without Directions

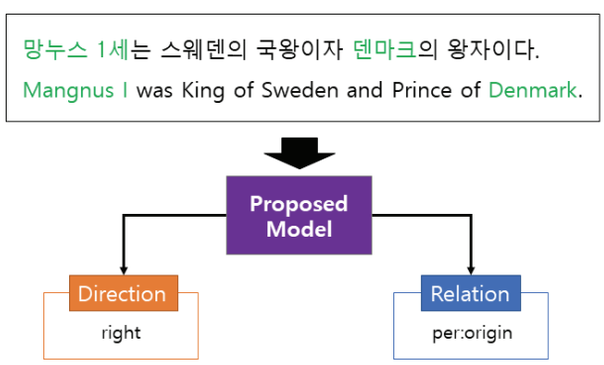

Fig. 3. The Expected Output of the Proposed Model

목적어는 '덴마크'라는 결과를 얻고, 관계 'per:origin'이 추출된다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 본 논문에서 제안하는 모델을 설명한다. 4장에서는 실험 결과를 분석하며, 마지막으로 5장에서는 본 논문의 결론을 도출하고, 향후 연구의 방향에 대해 다룬다.

2. 관련 연구

2.1 사전학습 언어모델

사전학습 언어모델(pre-trained language model)은 자연어 처리(natural language processing) 분야의 다양한 작업에서 범용적으로 사용되는 모델로, 대용량의 텍스트 데이터를 학습하여 단어 간의 의미적 유사성이나 문맥 정보를 단어나 문장을 벡터 형태로 표현한다. 단어를 벡터로 표현한 것을 단어 표상(word embedding)이라고 하는데, 심층 학습 등장 이후 초기 단계에서는 각 단어를 사전에 정의한 벡터로 나타내었으나[1,2], 이는 동형이의어를 같은 벡터로 나타내어 문맥 정보를 반영하지 못하는 단점이 있었다. 이후 등장한 BERT (Bidirectional Encoder Representations from Transformers) [3]에서는 자가주의집중(self-attention)을 통해 단어 표상에

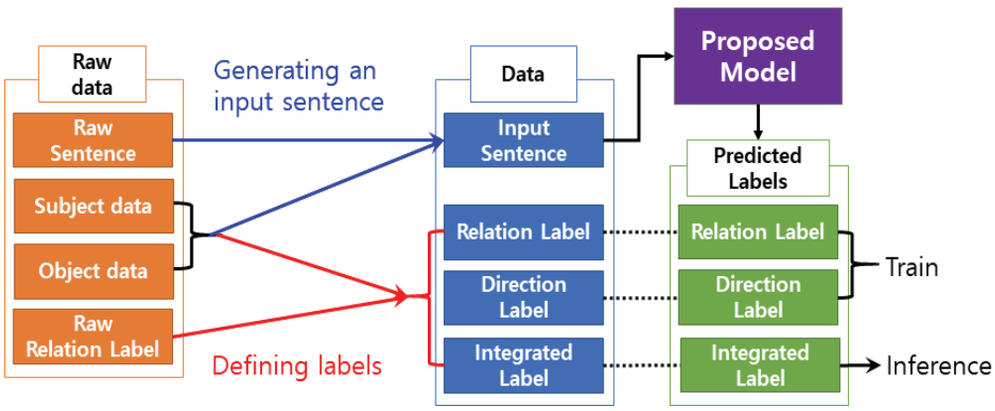

Fig. 4. The Overall Process of the Proposed Relation Extraction System

복잡한 문맥 정보가 반영되었고, 대용량의 말뭉치를 이용하여 범용적인 언어 이해 능력을 갖추게 되었다. 또한 추가적인 특정 작업에 대한 미세조정(fine-tuning)을 통해 자연어처리 대부분의 분야에서 이전보다 높은 성능을 보여주었다. 이후 RoBERTa[4], ELECTRA[5], ALBERT[6], BART[7] 등 기존 BERT에서 변형된 다양한 사전학습 언어모델이 발표되었다.

2.2 관계 추출

관계 추출은 문장, 문서, 대화에서 텍스트를 이해하고 이해한 정보를 바탕으로 개체 간의 관계를 예측하는 작업으로, 사전학습 언어모델의 등장으로 복잡한 문맥 정보를 반영하게 되면서 더 정확하고 의미론적으로 일관된 관계 추출이 가능하게 되었다. 문장 수준의 관계 추출에서는 사전학습 언어모델로 개체에 해당하는 단어의 표상만을 이용한 모델[8]이 사전학습 언어모델 등장 이전의 관계 추출 연구보다 뛰어난 성능을 보여주었다. 하지만 이는 사전학습 언어모델이 개체를 인식할 수 없다는 단점이 있었다. 이를 해결하기 위해 입력 문장에서 개체 위치 토큰을 추가하여 사전학습 언어모델에 관계를 추출할 개체에 대한 정보를 제공하여 성능을 높인 연구[9]가 있었고, 한국어에서 개체 위치 토큰 안의 개체에 개체의 어미를 결합한 연구[10]에서는 한국어에서도 개체 위치 토큰이 관계 추출에 효과적임을 입증하였다. 이후 개체 유형을 추가 정보로 이용하는 연구가 진행되었다. 개체 유형을 문장에 추가하여 성능을 향상시킨 TEM(Typed Entity Marker)를 제안한 연구[11]가 있었고, 개체 유형에 따라 가능한 관계만을 다른 분류기로 분류하는 모델[12] 또한 개체 유형 정보로 성능을 향상시켰다. 또한, 개체 유형 정보와 그래프를 한국어에서 사용한 연구[13]에서 한국어에서 효과적으로 개체 유형 정보를 이용하는 방법을 제시하였고, 그래프를 통해 성능을 더욱 높일 수 있음을 보여주었다. 종단형 문장 수준의 관계 추출 연구에서는 서론에서 언급한 많은 문장을 표상화해야 하는 문제를 해결하기 위한 연구들[14,15]이 있었지만, 모두 주어와 목적어를 정한 상태에서 관계를 추출한다는 한계점이 있다.

3. 확률적 교차 연산을 이용한 보편적 관계 추출

본 논문에서는 방향이 정해진 기존 관계추출 모델의 문제점을 보완하는 모델을 제안한다. 제안 모델은 방향과 관계를 하나의 모델에서 예측한다. 관계 예측 시 방향의 정보를 반영하기 위한 두 가지 연산인 교차 연산(crossover)과 확률적 교차 연산(probabilistic crossover)을 제안하며, 이를 통해 주어와 목적어 개체의 표상을 만들어낸다. Fig. 4는 본 논문에서 제안된 모델을 이용한 관계추출 시스템의 전체적인 구조이다. 말뭉치에서 제공하는 원시 데이터(raw data)는 원시 문장(raw sentence), 주어(subject)와 목적어(object) 정보, 원시 관계 표지(raw relation label)가 제공되는데, 표지 정의 단계(3.1절 참조; Fig. 4에서 Defining labels)에서는 주어와 목적어 정보 및 원시 관계 표지를 이용하여 관계, 방향, 통합 표지를 만들어내며, 입력 문장 생성단계(3.2절 참조; Fig. 4에서 Generating an input sentence)에서는 원시 문장과 주어, 목적어 정보를 이용하여 몇 가지 토큰을 추가하여 입력 문장을 만들어낸다. 방향 및 관계 추출 단계(3.3절 참조; Fig. 4에서 Proposed Model)는 제안 모델을 이용하는 과정으로, 입력 문장을 통해 방향과 관계를 예측한다. 예측된 결과를 정답 표지와 비교하여 학습하고, 추론 결과를 평가한다.

3.1 표지 정의

이 절에서는 모델이 예측하는 관계와 방향에 대한 표지 및 관계 표지와 방향 표지를 하나로 합친 통합 표지를 정의한다. 입력 데이터에서 방향 정보가 사라지면서 기존의 관계 표지 중 몇 개가 통합되고, 주어와 목적어의 위치에 따라 방향 표지를 생성하며, 이 두 표지를 합쳐 통합 표지가 정의된다.

1) 관계 표지

본 논문의 실험 말뭉치인 KLUE-RE[16]와 TACRED[17]에서 제공하는 관계 추출 표지 중에서 자식을 나타내는 'per:children' 관계의 방향을 바꾸면 부모를 나타내는 'per:parents' 관계가 된다. 이처럼 방향을 바꾸었을 때, 다른 관계를 나타내는 관계

Table 1. Labels Converted into Other Labels by Interchanged Direction

Label	Interchanged direction label
org:members	org:member_of
per:parents	per:children

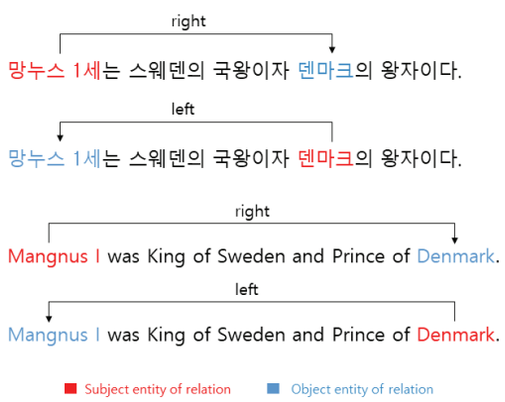

Fig. 5. Examples of Direction Labels

표지가 존재하는데, 이러한 표지들은 방향이 주어지지 않으면 두 개의 관계가 모두 성립한다. 이러한 경우 가장 확률이 높은 하나의 답만을 예측하는 분류기(classifier)의 특성상 학습에 혼선이 생길 수 있어 하나의 관계로 표현한다. 따라서 Table 1과 같이 기존의 'org:member_of' 표지는 'org:members'로, 'per:children' 표지는 'per:parents'로 대체된다. 따라서 'org:member_of', 'org:mem- bers' 표지는 모두 'org:members'인 하나의 표지로 표현되고, 'per:children'과 'per:parents'는 모두 'per:parents'인 표지로 표현된다. 이를 통해 KLUE-RE, TACRED에서의 관계 표지는 기존의 표지보다 2개 적은 28개, 40개 종류의 관계를 가진다.

2) 방향 표지

개체는 문장 안에서 위치 순서를 가지고 있으며, 서론에서 언급한 바와 같이 방향은 주어에서 목적어 방향을 뜻한다. 방향 표지는 문장에서 주어가 목적어를 가리키는 방향이 오른쪽이면 'right'를, 왼쪽이면 'left'를 부여한다. Fig. 5는 주어와 목적어의 위치에 따른 방향을 나타낸 그림이다. 방향을 바꾼 표지로 통일되는 표지(Table 1 참조)를 가진 문장은 방향 표지 또한 뒤집어준다('right'→'left', 'left'→'right').

3) 통합 표지

모델의 성능을 측정할 때는 앞서 정의한 관계 표지와 방향 표지를 모두 맞게 예측해야 한다. 따라서 이 두 표지를 모두 맞게 예측했는지 여부를 측정하기 위해 관계 표지와 방향 표지를 합친 통합 표지를 생성한다. 예를 들어 관계 표지가 'per:title'이고, 방향 표지가 'left'인 문장은 'per:title__left'라는 통합 표지를 가지게 된다. 여기서 Table 2의 관계 표지들과 같이 어떤 방향이든 성립되는 표지들은 양방향 관계 표지라고 정의한다. 양방향 관계 표지는 위의 방향 표지를 붙이는 규칙을 따르지 않는다. 또한, 본 논문에서 'no_relation' 표지는 양방향 관계 표지라고 가정한다. KLUE-RE 말뭉치에서의 통합 표지는 Table 2의 5개 표지가 방향표지 없이 하나의 표지로 표현되어 총 51개가 되고, TACRED 말뭉치에서의 통합 표지는 Table 2에서 존재하지 않는 표지인'per:colleagues'를 제외한 4개 표지가 하나의 표지로 표현되어 총 76개로 이루어진다.

Table 2. Bi-directional Relation Labels

Label
No_Relation
Per:Other_Family
Per:Colleagues
Per:Siblings
Per:Spouse

3.2 입력 문장 생성

이 절에서는 사전학습 언어모델이 개체 정보를 학습하도록 개체 정보를 문장에 추가하는 과정을 기술한다. 이전 연구에서는 개체 위치 토큰을 사용하여 사전학습 언어모델에 개체 위치 정보를 추가하는 방법이 제안되었다[10]. 이후, TEM을 제안한 연구[11]에서는 이에 개체 유형을 추가하여 성능을 더욱 향상시켰다. 본 논문에서는 이러한 연구들을 바탕으로 개체 유형과 토큰을 이용하는 방법을 제시한 연구[12]의 문장 생성 방법을 이용하여 방향이 주어지지 않은 문장에서 개체 위치 토큰과 개체 유형을 문장에 삽입하는 과정을 설명한다.

1) 개체 위치 토큰 추가

기존 연구들에서는 개체 위치 토큰에 주어와 목적어 정보, 즉 방향 정보가 포함된 상태에서 관계를 추출하는 방식으로 작동한다. 그러나, 본 논문에서는 방향 예측이 필요하기 때문에, 문장의 개체 위치 토큰에 방향 정보를 포함시키지 않는다. 이에 따라 각 개체는 주어와 목적어 여부가 정해지지 않았기 때문에, 각 개체는 같은 위치 토큰을 사용한다. 따라서 Fig. 6과 같이 동일한 위치 토큰 '[e]'와 '[/e]'를 추가한다. 본 논문에서 '[e]'와 '[/e]'를 개체 위치 토큰이라고 정의하며, '[e]'를 개체 시작 토큰, '[/e]'를 개체 끝 토큰이라고 정의한다.

2) 개체 유형 추가

문장에 개체 유형을 추가하여 사전학습 언어모델이 개체 유형 정보도 함께 집중(attention)하도록 한다. 제공되는 개체 유형은 KLUE-RE말뭉치의 경우 영어 약자로 되어있으므로, Table 3과 같이 한국어 단어로 변환한다. TACRED는 'PERSON',

[e]망누스 1세[/e] 는 스웨덴의 국왕이자 [e]덴마크[/e]의 왕자이다.
[e]Mangnus I[/e] was King of Sweden and Prince of [e]Denmark[/e].

Fig. 6. An Example of a Sentence with Entity Position Tokens

Table 3. KLUE-RE's Entity Types Translated to Korean

Type	Meaning	Korean
PER	person	인물
ORG	organization	기관
LOC	location	장소
POH	other proper nouns	명사
DAT	date and time	날짜
NOH	other numerals	수량

[e]망누스 1세[/e] [t]인물[/t] 는 스웨덴의 국왕이자 [e]덴마크[/e] [t]장소[/t] 의 왕자이다.
[e]Mangnus I[/e] [t]person[/t] was King of Sweden and Prince of [e]Denmark[/e] [t]location[/t].

Fig. 7. An Example of a Sentence with Entity Types

'ORGANIZATION' 등 약자로 되어있지 않기 때문에 소문자로만 변환한다. 단어로 변환된 개체 유형은 사전학습 언어모델에서 사전학습된 정보에 의해 개체의 유형을 판별할 수 있게 된다[12].

이후 변환된 개체 유형 스팬(span)의 시작과 끝 위치를 의미하는 토큰을 추가한다. 이 때, 개체 위치 토큰과 같이 이전에 생성된 두 개의 개체 유형은 주어, 목적어가 없기 때문에 같은 토큰인 '[t]', '[/t]'를 추가한다. 본 논문에서 '[t]', '[/t]'를 개체 유형 위치 토큰이라고 정의하며, '[t]'를 개체 유형 시작 토큰, '[/t]'를 개체 유형 끝 토큰이라고 정의한다. 개체 유형 위치 토큰이 추가된 개체 유형은 문장에서 개체 끝 토큰인 '[/e]'의 뒤에 추가한다. Fig. 7은 최종 생성된 예시 문장이다.

3.3 방향 및 관계 추출

이 절에서는 3.2절에서 생성된 문장을 입력으로 하는 제안 모델에 대해 설명한다. 전체적인 과정은 사전학습 언어모델로 표상화하여 개체의 표상을 얻고, 개체의 표상을 토대로 방향을 예측한 다음 개체의 표상과 방향에 대한 확률을 이용해 확률적 교차 연산을 수행하여 최종 관계를 추출한다. 이 과정

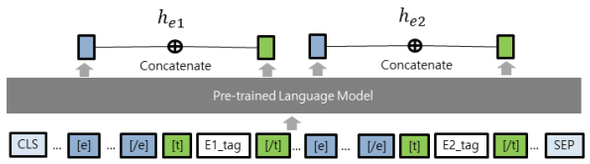

Fig. 9. The Sentence Embedding Step

을 1) 개체 표상 생성, 2) 방향 예측, 3) 관계 추출로 나누어 자세히 기술한다. 제안 모델의 구조는 Fig. 8과 같다.

1) 개체 표상 생성

이 단계에서는 개체의 표상을 생성한다. 이를 위해 먼저 문장을 토큰화한 후 사전학습 언어모델로 표상화한다. 개체의 표상은 개체 시작 토큰인 '[e]'와 개체 유형 끝 토큰인 '[/t]'로 생성하는데, 두 토큰을 표상화한 벡터를 연접 연산(concatenate)한 벡터를 개체 표상이라고 하며, 문장에서 앞에 위치한 개체의 표상을 h_{e1}, 뒤에 위치한 개체의 표상을 h_{e2}라고 정의한다. 개체 표상 생성 단계를 그림으로 나타내면 Fig. 9와 같다.

2) 방향 예측

개체의 표상인 h_{e1}, h_{e2}를 연접 연산을 통해 연결하여 하나의 벡터로 만들어 선형층(linear layer)을 지나 왼쪽 방향일 확률인 P_l, 오른쪽 방향일 확률인 P_r를 얻는다. 손실값 L_{dir}의 계산에는 cross-entropy 손실 함수를 이용하고, 양방향 관계 표지의 경우 L_{dir}을 0으로 고정하여 방향의 학습에 반영되지 않도록 한다. Fig. 10에 방향 예측 단계를 그림으로 표현하였다.

3) 관계 추출

기존의 관계 추출 모델은 주어 개체 표상과 목적어 개체 표상을 연결해서 이용하여 벡터에서 주어와 목적어 표상의 위치를 고정하며, 이를 통해 주어와 목적어의 자질을 따로 학습하

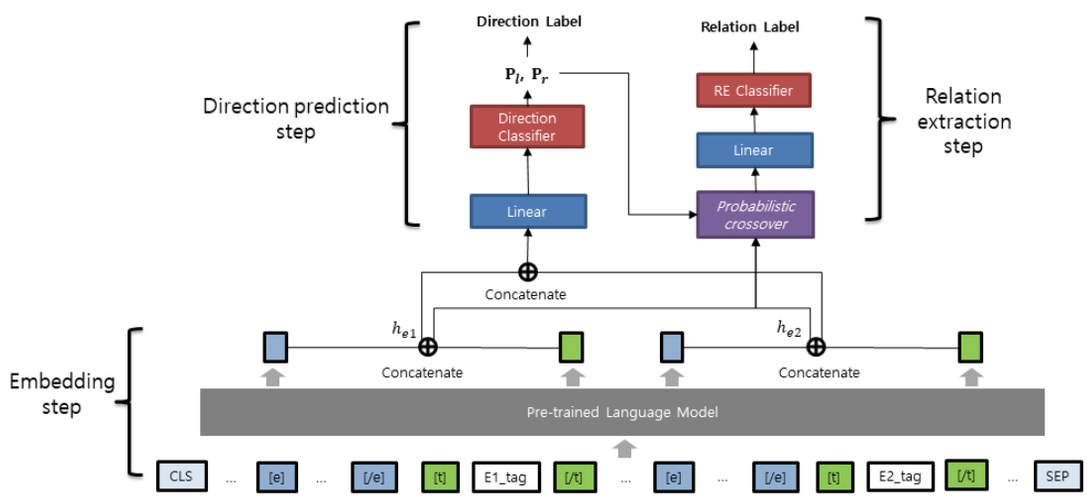

Fig. 8. The Overall Structure of the Proposed Model

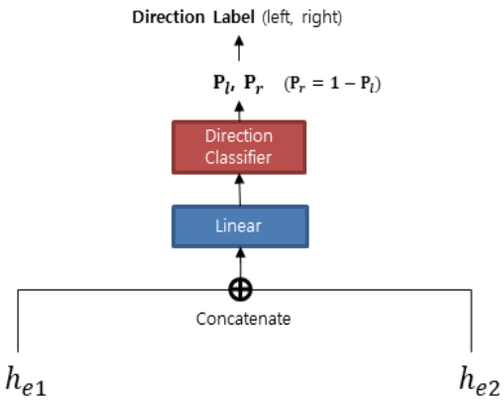

Fig. 10. The Direction Prediction Step

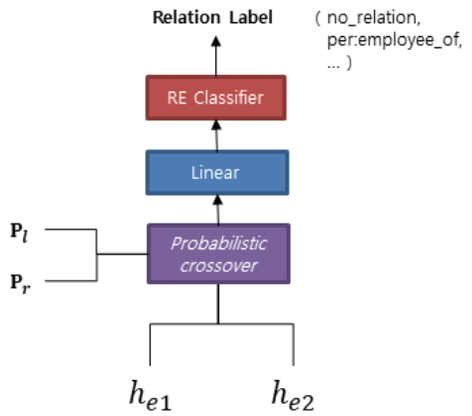

Fig. 11. The Relation Extraction Step

게 된다. 본 논문에서는 기존의 관계 추출 방법을 최대한 반영하기 위한 간단한 방법으로 교차 연산(crossover)을 제안하고, 교차 연산의 단점을 보완한 확률적 교차 연산(probabilistic crossover)을 제안한다. 이를 통해 주어와 목적어 표상을 생성하고, 주어와 목적어의 자질을 효과적으로 학습하는 방법을 제시한다.

교차 연산(crossover) : 이 연산은 예측된 방향에 따라 개체를 연결하는 순서를 정한다. 두 개체 표상을 연결할 때, 주어는 왼쪽, 목적어는 오른쪽에 위치시킨다. 따라서 P_r이 P_l보다 클 경우 h_{e1}과 h_{e2}를 순서대로 연결한다. 반대로 P_l이 P_r보다 크다면 h_{e2}과 h_{e1}를 순서대로 연결한다. 예를 들어 방향이 왼쪽으로 예측되었다면, 오른쪽의 개체가 주어이므로 h_{e2}가 왼쪽에 위치한다. 따라서 h_{e2}, h_{e1} 순서로 연결하게 된다. 이를 통해 연산 결과 벡터의 왼쪽에는 주어일 확률이 높은 개체 표상이, 오른쪽에는 목적어일 확률이 높은 개체 표상이 위치하게 된다. Equation (1)에 교차 연산을 수식으로 나타내었다.

$$crossover(h_{e1}, h_{e2}, P_r, P_l) = \begin{cases} concat(h_{e1}, h_{e2}) & \text{if } P_r > P_l \\ concat(h_{e2}, h_{e1}) & otherwise \end{cases} \quad (1)$$

교차 연산은 방향 확률을 바탕으로 주어와 목적어 개체를 만들어내는 간단한 방법이지만, 미분 불가능한 연산이기 때문에 역전파(backward) 과정에서 교차 연산 이후의 손실값이 교차 연산 이전으로 전파되지 않는 단점이 있다. 또한, 순전파(forward) 과정에서 예측된 방향에 따라 위치를 고정하기 때문에 방향 예측의 오류가 이후 단계에 그대로 전파되는 단점이 있다. 이 문제를 완화하기 위하여 확률적 교차 연산을 제안한다.

확률적 교차 연산(probabilistic crossover) : 교차 연산은 확률이 높은 개체를 그대로 주어, 목적어 개체에 반영하였지만, 확률적 교차 연산은 개체를 확률값만큼만 반영시킨다. 먼저 주어 개체 표상인 h_{sbj}는 Equation (2)와 같이 앞 개체의 표상인 h_{e1}에 오른쪽 방향일 확률, 즉 앞 개체가 주어일 확률 P_r를 곱한 벡터와 뒤 개체의 표상인 h_{e2}에 왼쪽 방향일 확률, 즉 뒤 개체가 주어일 확률 P_l을 곱한 벡터를 더하여 만들어낸다. 이를 통해 h_{sbj}는 각 개체가 주어일 확률만큼 반영된다. 같은 방법으로 목적어 개체 표상인 h_{obj}는 Equation (3)과 같이 h_{e1}와 h_{e1}이 목적어일 확률인 P_l를 곱한 벡터와 h_{e2}와 h_{e2}가 목적어일 확률인 P_r을 곱한 벡터를 더한다. h_{sbj}와 같이 h_{obj}에는 각 개체가 목적어일 확률만큼 반영된다. 이후 h_{sbj}과 h_{obj}를 연결한다. Equation (4)에 확률적 교차 연산을 수식으로 나타내었다.

$$h_{sbj} = P_r \cdot h_{e1} + P_l \cdot h_{e2} \quad (2)$$

$$h_{obj} = P_l \cdot h_{e1} + P_r \cdot h_{e2} \quad (3)$$

$$probablistic\_crossover(h_{e1}, h_{e2}, P_r, P_l) = concat(h_{sbj}, h_{obj}) \quad (4)$$

확률적 교차 연산은 개체를 확률만큼 주어, 목적어 개체에 반영하기 때문에, 순전파 과정에서 방향 예측의 오류가 교차 연산보다 덜 전파될 수 있다. 또한, 미분 가능하기 때문에 역전파 과정에서 이후 관계 예측의 오류가 방향 예측단계까지 전파된다. 따라서 방향을 학습할 때, 관계 예측의 오류가 영향을 주게 된다.

이후 단계는 생성된 은닉 벡터를 관계 표지의 개수만큼 사상시킨다. Fig. 11에 관계 추출 단계를 그림으로 나타내었다. 관계 추출 데이터의 특성상 불균형 데이터의 학습에 유리한 CB(Class-Balanced) focal loss[18]으로 손실값 L_{re}를 계산하며, 최종 손실값 L은 관계 추출시 손실값인 L_{re}와 방향 예측시 손실값인 L_{dir}을 α값에 따라 반영하여 더한다. α값은 최종 손실값에 방향과 관계 손실값을 반영하는 비율이며, Equation (5)에 최종 손실값 L의 식을 나타내었다.

$$L = \alpha L_{re} + (1-\alpha) L_{dir} \quad (5)$$

4. 실험 및 평가

본 장에서는 실험 환경 및 실험 척도를 설명하고, 비교 실험 모델을 소개하며, 실험 결과를 분석하고 토의한다. 실험은 한국어, 영어 관계 추출 말뭉치에 대해 진행하였고, 기존 방향

Table 4. The Statistics of Relation Extraction Corpus

Corpus	KLUE-RE	TACRED
Training	32,470	68,124
Validation	-	22,631
Test	7,765	15,509
Total	40,235	106,264

이 정해진 관계 추출 말뭉치에서 주어와 목적어를 바꾼 문장을 추가하여 학습한 모델과의 비교 실험을 진행한다. 평가는 통합 표지, 방향 표지, 관계 표지에 대한 성능을 실험 모델별로 평가하며, 통합 표지와 관계 표지는 미시 F1-점수(Micro F1-score)를 이용하고, 방향 표지의 평가는 거시 F1-점수(Macro F1-score)로 평가한다. 통합 표지와 관계 표지는 일반적인 관계 추출과 같이 불균형 데이터이므로 KLUE-RE[16], TACRED[17]에서 평가 지표로 사용하는 미시 F1-점수를 통해 불균형 데이터에 대한 예측 결과를 효과적으로 측정할 수 있고, 방향 표지는 불균형을 크게 이루지 않기 때문에 거시 F1-점수를 이용하였다.

4.1 실험 환경

한국어의 관계 추출 실험을 위해서 KLUE-RE 말뭉치를 사용하였다. 학습에는 KLUE-RE의 학습(training) 말뭉치를 사용하였고, KLUE에서 평가 말뭉치를 공개하지 않기 때문에 개발(development) 말뭉치를 평가 말뭉치로 사용한다. 또한, 영어 실험을 위한 말뭉치는 TACRED 말뭉치를 사용하였다. Table 4는 각 말뭉치에서 문장의 개수이다.

Table 5는 문장에서 각 말뭉치마다 실험한 매개변수들이다. PLM(Pre-trained Language Model)은 사전학습 언어모델을 뜻하고, KLUE-RE는 한국어 말뭉치 위주로 학습한 KLUE--RoBERTa-base를, TACRED는 영어 말뭉치 위주로 학습한 RoBERTa-base를 이용하였다. 또한, 학습 말뭉치가 비교적 많은 TACRED는 15번의 epoch와 1e-5의 학습률로 학습하였다.

4.2 비교 실험 모델

이 절에서는 제안 모델과의 비교 실험을 진행할 기준(base) 모델과 non-crossover 모델을 소개한다. 먼저 기준 모델은 방향이 정해진 상태에서 관계를 추출하는 기존의 방식을 이용한 모델이나, 제안 모델과의 비교를 위해 방향을 뒤바꾼 데이터를 추가하여 학습하였다. 기준 모델은 관계와 방향 표지를 따로 나누지 않은 통합 표지를 학습하는 모델로, 사전학습 언어모델에서 문맥 전체 정보를 담고 있다고 알려진 CLS토큰을 이용한 base(CLS) 모델과 h_{e1}과 h_{e2}를 이용한 base(entity) 모델 2가지로 나뉜다. 기준 모델의 손실 함수는 CB focal loss를 사용한다. non-crossover 모델은 제안 모델과 같이 방향과 관계를 나누어 예측하는 모델이다.

base(CLS) : 이 모델은 일반적인 문서 분류에서 이용하는 CLS토큰의 표상을 이용한다. Fig. 12는 이 모델의 구조이다.

Table 5. Hyper-parameters of Relation Extraction Corpus

Hyper parameter	KLUE-RE	TACRED
PLM	KLUE-RoBERTa-base	RoBERTa-base
Epoch	10	15
learning rate	2e-5	1e-5
weight decay	0.01	
Optimizer	AdamW	

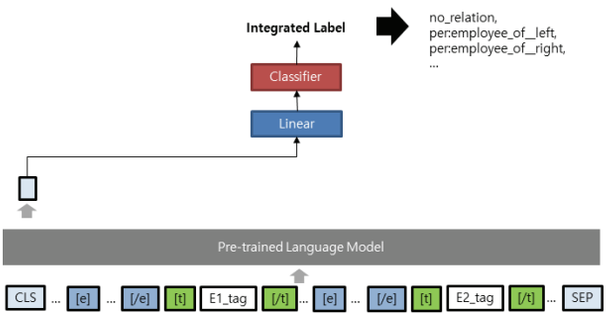

Fig. 12. The Structure of the Baseline (CLS) Model

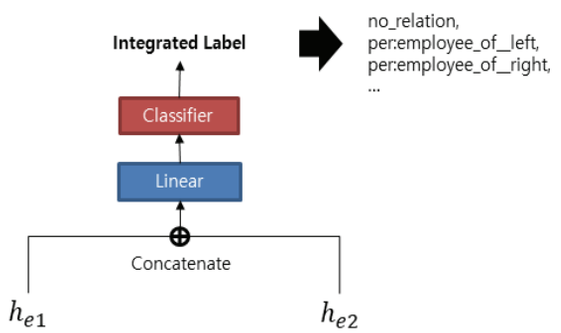

Fig. 13. The Structure of the Baseline (entity) Model

base(entity) : 이 모델은 각 개체의 표상을 나타내는 벡터인 h_{e1}과 h_{e2}를 순서대로 연접 연산한 벡터로 통합 표지를 예측하며, base(entity) 모델의 구조는 Fig. 13과 같다.

non-crossover : 이 모델은 관계 표지와 방향 표지를 분리하여 학습하는 기본 모델이다. 각 개체의 표상인 h_{e1}과 h_{e2}를 순서대로 연접 연산한 벡터로 관계와 방향 표지를 예측하며, 제안 모델과 같이 관계 예측의 손실 함수에는 CB Focal loss를, 방향 예측의 손실 함수는 cross-entropy loss를 사용하였다. 모델의 구조는 Fig. 14와 같다.

4.3 실험 결과 및 분석

이 절에서는 제안 모델과 비교 실험 모델의 평가를 진행하고 분석한다. 실험 모델별 성능은 Table 6과 같고, 통합 표지와 관계 표지의 평가는 KLUE-RE, TACRED의 평가 방식에 따라 'no_relation'을 제외한 미시 F1-점수를 이용하여 평가한 결과이며, 방향 표지는 거시 F1-점수를 이용하여 평가하였다.

실험 결과, 관계와 방향 표지를 나누어 학습한 모델들이 방

Table 6. The Result of Comparison by Experimental Model

Model	KLUE-RE			TACRED		
	Integrated label	Relation label	Direction label	Integrated label	Relation label	Direction label
Models Learned Integrated Labels						
base(CLS)	57.64	59.06	76.73	60.44	65.06	73.33
base(entity)	60.46	62.72	78.05	60.46	65.16	72.46
Models Learned Direction Labels and Relation Labels						
non-crossover	60.96	62.81	87.02	61.29	66.62	88.48
crossover	60.61	62.80	87.02	61.16	66.06	88.59
probabilistic crossover	61.70	64.00	87.55	63.41	68.77	89.93

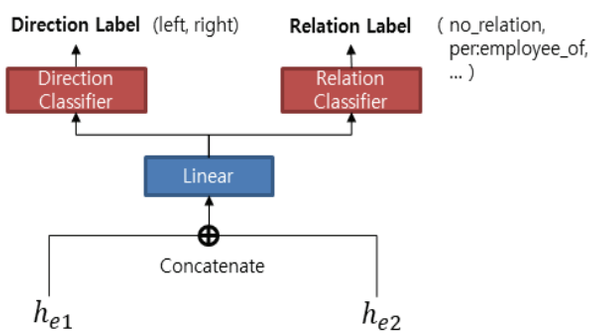

Fig. 14. The Structure of the Non-crossover Model

향을 정해 놓고 예측한 모델보다 더 높은 성능을 보여주어 방향과 관계를 한 번에 학습하여 추론하는 것 보다 나누어 학습 및 추론하는 방법이 더욱 효과적임을 알 수 있다. 관계에 따라서 방향이 정해질 수 있고, 반대로 방향에 따라 관계가 달라질 수 있기 때문에 관계와 방향은 서로 독립적인 자질이 될 수 없는 것은 자명한 사실이다. 실험 결과에서 통합 표지의 성능이 높은 모델이 관계, 방향 표지 예측률 또한 높은 결과를 보여준 부분에서 이러한 특징이 반영된 것으로 보인다. 그러나 관계와 방향 표지를 나누어 학습한 모델이 더 높은 성능을 보여 준 것에서 어느 정도는 관계와 방향이 독립적인 자질을 가지고 있음을 시사한다.

또한 방향과 관계를 나누어 학습하는 방법에서 주어와 목적어 개체의 표상을 만들어내지 않는 non-crossover방식이 crossover보다 더 좋은 성능을 보였다. crossover모델은 확률적 교차 연산과 같이 예측한 방향을 이용해 주어와 목적어 표상을 만들어 내지만, 개체의 표상을 그대로 이용하였기 때문에 순전파(forward) 단계에서 crossover모델의 방향 예측 오류가 더 크게 전파되고, 또한 역전파(backward) 단계에서 관계 표지 예측의 손실값이 방향 예측 단계까지 전파되지 않기 때문에 성능이 더 낮은 것으로 추측된다. 그러나 확률적 교차 연산은 위에서 언급한 crossover의 단점을 보완하여 고안한 모델로, 개체의 표상을 주어와 목적어에 확률만큼 반영하여 순전파 단계에서 방향 예측 오류의 전파를 완화하였고, 역전파 단계에서 관계 예측에서의 손실값을 방향 예측 단계까지 전달하여 효과적으로 관계 표지와 방향 표지를 함께 학습할 수 있었다. 이를 통해 확률적 교차 연산을 이용한 모델이 가장 높은 성능을 보여주었다.

Table 7. The Result of Comparison by α

α	KLUE-RE	TACRED
0.4	57.98	62.81
0.5	59.31	**63.41**
0.6	59.67	62.69
0.7	59.83	62.31
0.8	60.14	62.18
0.9	**61.70**	61.90
0.95	60.43	61.55

언어별 실험 결과를 비교해 보면, 한국어보다 영어 말뭉치를 학습 및 평가한 모델이 더 높은 성능을 보여주었다. 이러한 결과는 데이터 개수의 차이와 언어적 차이에서 기인하는 것으로 보이는데, 학습하는 영어 문장의 수가 한국어보다 2배 이상 많기 때문에(Table 4 참조) 데이터가 많을수록 유리한 심층학습 특성상 영어의 성능이 더 높게 나타난 것으로 보인다. 또한, 한국어에서 개체 위치 토큰을 적용한 연구[10]에 따르면, 한국어에서는 영어와 달리 문장에서 개체의 역할이 조사(postposition)까지 포함하여야 결정될 수 있으므로, 개체 위치 토큰 사이의 개체 단어를 조사까지 포함하여 관계를 추출하는 방식이 효과적임을 보였다. 본 실험 결과 또한 이러한 한국어와 영어의 언어적인 차이에서 비롯된 것으로 보이며, 조사를 포함하는 방법을 이용하는 등의 추가 연구를 통해 한국어 관계 추출의 성능을 향상할 수 있을 것으로 예상된다. 마지막으로 제안 모델에서 관계와 방향의 손실값의 비율을 정하는 α값에 따른 성능이 언어별로 다소 차이가 있었는데, Table 7에 α값에 따른 성능 차이를 나타내었다. 성능은 모두 확률적 교차 연산을 이용한 모델을 이용하였고, 통합 표지의 미시 F1-점수로 측정하였다.

한국어 말뭉치인 KLUE-RE에서는 α값이 0.9일 때 제일 좋은 성능을 보여주었고, 영어 말뭉치인 TACRED에서는 α값이 0.8일 때 최고 성능을 보여주었다. 이는 말뭉치마다 추출하는 관계가 다르고, 표지별 데이터의 분포와 데이터 수 등이 달라 이러한 차이가 생기며, 영어와 한국어가 언어의 성격이 달라 이러한 차이를 만들어낸 것으로 보인다. 또한, base모델보다

더 성능이 낮게 측정되는 α값도 존재하였다. 따라서 말뭉치에 따라 관계와 방향 손실값의 비율인 α값의 조절이 필수적이라는 것을 알 수 있다.

5. 결 론

관계 추출은 다양한 텍스트 소스에서 개체 간의 관계를 추출하여 방대한 데이터를 만들어낼 수 있어 정보 검색 및 자연어 처리 분야에서 매우 중요한 기술 중 하나이다. 관계 추출 모델은 학습과 추론에 소요되는 시간 및 자원이 적을수록 이점이 있지만, 문장 수준 관계 추출은 대부분의 연구가 방향이 정해진 상태로 관계를 추출하며, 종단형 관계 추출 모델에서 이러한 모델을 이용하면 방향에 따른 경우의 수만큼 사전학습 언어모델로 표상화하여야 하므로 많은 시간과 자원이 소요되는 단점이 있다.

본 논문에서는 이러한 점을 완화하기 위해 주어진 두 개체에 대해 관계와 관계의 방향을 예측하여 기존의 방식보다 사전학습 언어모델의 사용을 0.5배로 줄여 시간과 자원의 소요를 줄인 모델을 제안했다. 입력 문장에서 기존의 주어, 목적어에 따라 달라지는 개체 위치 토큰을 하나의 같은 토큰으로 적용하고, 이후 방향을 예측하고 방향에 대한 확률을 이용한 확률적 교차 연산을 제안하였고, 이를 통해 관계를 추출한다. 실험 결과 방향과 관계를 학습 및 추론하는 관계 추출 모델이 통합 표지를 학습한 모델보다 더 성능이 높아 방향과 관계를 따로 추출하는 방법의 효과성을 입증하였다. 또한 방향과 관계를 나누어 예측하는 실험 모델(non-crossover, crossover, probabilistic crossover) 중 확률적 교차 연산을 이용한 모델의 성능이 한국어뿐 아니라 영어에서 또한 가장 높았고, 이를 통해 제안 방법의 효과성을 확인하였으다. 그러나 영어보다 한국어에서의 성능이 1~2%정도 낮게 측정되었는데, 이는 개체 위치 토큰을 이용하였을 때 한국어에서 개체의 역할을 포착하는 데 더욱 어려움이 있는 것으로 판단되었고, 언어마다 최적의 α값이 다름을 관찰할 수 있었다.

추후 연구로는 한국어에서 개체에 조사를 포함하는 등의 방법을 통해 영어와의 성능 차이를 줄이고, 한국어와 영어에서 더욱 다양한 말뭉치에 적용하여 볼 필요가 있다. 또한, F1-점수 이외의 다른 평가 지표에 대해 실험하여 관계 추출에서 평가 지표별 차이를 상세히 분석하여 보고, 종단형 관계 추출 모델에 적용하여 기존 방법과의 성능 비교 및 소요되는 시간과 자원을 수치로써 비교 실험할 예정이다.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol.2, pp.3111-3119, 2013.

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, Vol.5, pp.135-146, 2017.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol.1, pp.4171-4186, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692, 2019.

[5] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," *Proceedings of International Conference on Learning Representations*, 2019.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, 2020, "ALBERT: A lite BERT for self-supervised learning of language representations," *Proceedings of International Conference on Learning Representations*.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7871-7880, 2020.

[8] H. Wang, M. Tan, M. Yu, S. Chang, D. Wang, K. Xu, X. Guo, and S. Potdar, "Extracting multiple-relations in one-pass with pre-trained transformers," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.1371-1377, 2019.

[9] L. B. Soares, N. Fitzgerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: distributional similarity for relation learning," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.2895-2905, 2019.

[10] Y. Hur, S. Son, M. Shim, J. Lim, and H. Lim, "K-EPIC: Entity-perceived context representation in Korean relation extraction," *Applied Sciences*, Vol.11, No.23, pp.11472, 2021.

[11] W. Zhou and M. Chen, "An improved baseline for sentence-level relation extraction," *arXiv:2102.01373*, 2021.

[12] S. Lyu and H. Chen, "Relation classification with entity type restriction," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp.390-395, 2021.

[13] J. Lee and J. Kim, "Korean relation extraction using pre-trained language model and GCN," *Proceedings of the 34th Annual Conference on Human and Cognitive Language Technology*, pp.379-384, 2022.

[14] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.50-61, 2021.

[15] D. Ye, Y. Lin, P. Li, and M. Sun, "Packed levitated marker for entity and relation extraction," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol.1, pp.4904-4917, 2022.

[16] S. Park et al., KLUE: Korean Language Understanding Evaluation, *arXiv:2105.09680*, 2021.

[17] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning, "Position aware attention and supervised data improve slot filling," In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp.35-45, 2017.

[18] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9268-9277, 2019.

이 제 승

https://orcid.org/0000-0002-4699-2928
e-mail : leeje1231@naver.com
2018년 한국해양대학교 컴퓨터공학과(학사)
2023년 한국해양대학교 컴퓨터공학과(석사)
관심분야 : Natural Language Processing, Relation Extraction, Information Extraction

김 재 훈

https://orcid.org/0000-0001-8655-2591
e-mail : jhoon@kmou.ac.kr
1986년 계명대학교 전계계산학과(학사)
1988년 한국과학기술원 전산학과(석사)
1996년 한국과학기술원 전산학과(박사)
1988년 ~ 1997년 한국전자통신연구원(선임)
2001년 ~ 2002년 Information Sciences Institute USC 방문연구원
2007년 ~ 2008년 Beckman Institute UIUC 방문연구원
1997년 ~ 현 재 한국해양대학교 컴퓨터공학과 및 해양인공지능융합전공 교수
관심분야 : Natural Language Processing, Information Retrieval, Corpus Linguistics, Sentiment Analysis