한국어 상호참조해결을 위한 BERT 기반 데이터 증강 기법

김기훈○*, 이창기*, 류지희**, 임준호**

강원대학교 컴퓨터과학과*, 한국전자통신 연구원**
[rlarlgnsu, leeck]@kangwon.ac.kr, [chrisjihee, joonho.lim]@etri.re.kr

BERT-based Data Augmentation Techniques for Korean Coreference Resolution

Kihun Kim*, Changki Lee*, Jihee Ryu**, Joonho Lim**
Kangwon National University*, Electronics and Telecommunications Research Institute**

요 약

상호참조해결은 문서 내에 등장하는 모든 멘션 중에서 같은 의미를 갖는 대상(개체)들을 하나의 집합으로 묶어주는 자연어처리 태스크이다. 한국어 상호참조해결의 학습 데이터는 영어권에 비해 적은 양이다. 데이터 증강 기법은 부족한 학습 데이터를 증강하여 기계학습 기반 모델의 성능을 향상시킬 수 있는 방법 중 하나이며, 주로 규칙 기반 데이터 증강 기법이 연구되고 있다. 그러나 규칙 기반으로 데이터를 증강하게 될 경우 규칙 조건을 만족하지 못했을 때 데이터 증강이 힘들다는 문제점과 임의로 단어를 변경 혹은 삭제하는 과정에서 문맥에 영향을 주는 문제점이 발생할 수 있다. 따라서 본 논문에서는 BERT의 MLM(Masked Language Model)을 이용하여 기존 규칙기반 데이터 증강 기법의 문제점을 해결하고 한국어 상호참조해결 데이터를 증강하는 방법을 소개한다. 실험 결과, ETRI 질의응답 도메인 상호참조해결 데이터에서 CoNLL F1 1.39% (TEST) 성능 향상을 보였다.

주제어: BERT, 데이터 증강, 상호참조해결, 딥러닝

1. 서론

상호참조해결(coreference resolution)은 문서 내에 등장하는 모든 명사(구)와 대명사(구) 중에서 같은 의미를 갖는 명사(구)와 대명사(구)를 하나의 집합으로 묶어주는 자연어처리 태스크이다. 여기에서 집합으로 묶일 수 있는 명사(구)와 대명사(구)를 멘션(mention)이라 부르며, 멘션들이 하나로 묶여있는 집합을 개체(entity)라고 한다. 상호참조해결은 기계번역, 질의 응답, 정보 추출 등 자연어처리 응용 태스크에서 문서에 등장하는 명사(구)와 대명사(구)의 관계를 올바르게 연결하여 대상에 대한 정보를 명확하게 전달하고 문맥을 이해하는데 중요한 역할을 한다.

최근 한국어 상호참조해결은 기계학습을 이용한 연구[1,2]가 활발히 진행되고 있다. 이러한 기계학습 방식의 상호참조해결 모델을 학습하기 위해서는 단어 사이의 참조관계 정보가 있는 문서 단위의 학습 데이터가 필수적이다. [1,2]에서는 수동으로 상호참조 관계를 주석한 ETRI 질의응답 도메인 상호참조해결 데이터 셋으로 학습과 평가를 진행하였다. 이 데이터는 외국어 상호참조해결 데이터[3]에 비하여 총 멘션 개수와 엔티티 개수가 부족하다([표 1] 참고). 영어권에 비하여 적은 양의 학습 데이터는 상호참조해결 모델의 성능 저하의 원인이 될 수 있다. 양질의 상호참조해결 데이터를 추가하면 성능 향상에 도움이 될 것이라 기대할 수 있다. 하지만 상호참조해결 데이터를 수동으로 구축하는 것은 상호참조에 대한 지식이 필수적이고 문서 내에 등장하는 모든 단어들의 관계를 파악해야 하기 때문에 많은 시간과 비용이 필요하다.

데이터 증강 기법은 부족한 데이터를 늘릴 수 있는 효과적인 방법이다. [4]의 연구는 상호참조하는 대명사 고유명사 위치 변경과 괄호 제거 규칙을 사용하여 한국어 상호참조해결 데이터를 증강하는 방법을 제안하였으며, [5]의 연구에서 제안된 Easy Data Augmentation(EDA)는 분류 문제에서 영어권 학습 데이터를 단어 삭제, 단어 삽입, 위치 변경, 동의어 치환과 같은 간단한 규칙으로 자동으로 데이터를 증강하는 방법이다. 이 둘의 연구는 자동으로 데이터를 증강할 수 있지만 몇 가지 문제점이 있다. [4]의 연구는 증강할 문서가 괄호, 참조하는 대명사와 고유명사가 존재해야 한다는 조건을 만족해야 하고, [5]의 연구는 WordNet과 같은 한국어 어휘 의미 사전이 반드시 필요하며, 임의 삽입, 동의어 치환 과정에서 문서의 문맥을 보존할 수 없어 참조관계에 영향을 미칠 수 있다.

따라서 본 논문에서는 [6]의 BERT 기반 데이터 증강 기법을 상호참조해결 데이터에 적용하여 모든 데이터에 대해 증강할 수 있으며, 상호참조해결 대상이 아닌 단어에 대하여 MLM(masked language model)[7]을 이용하여 데이터의 문맥에 영향을 끼치지 않는 단어로 치환하는 데이터 증강 방법을 제안한다.

표 1. 각 언어별 상호참조해결 데이터 비교 (개수).

DATA	문서 수	엔티티 수	멘션 수
영어[3]	2,802	35,142	155,558

중국어[3]	1,810	28,256	102,853
한국어	2,819	10,069	30,967

2. 관련 연구

최근 한국어 상호참조해결은 모든 명사와 대명사 스팬을 멘션 후보로 지정하여 스팬 표현을 사용한 스코어링 방식의 End-to-end방식의 기계 학습 모델[1,2]이 연구되었다. 본 논문에서는 [2]의 연구인 BERT 기반 End-to-end 상호참조해결 모델을 베이스라인으로 선정했다.

[4]는 서로 참조하는 대명사와 고유명사의 위치를 교체하는 방식과 문서에 등장하는 괄호 부분을 제거하는 방식을 사용하여 부족한 상호참조해결 데이터를 증강한다. 다른 단어로 교체하거나 임의의 단어를 삽입하는 방식이 아니기 때문에 문맥에 영향을 미치지 않고 안전하고 적은 비용으로 데이터 증강이 가능하지만, 문서에 서로 참조하는 대명사 고유명사 멘션이 존재하거나 소괄호가 있어야 한다는 제약조건이 있다.

[5]는 분류 문제에서 제안된 규칙 기반의 데이터 증강 방법으로, 임의 단어 삭제, 임의 단어 삽입, 동의어 치환, 단어 위치 교환 방법을 사용하여 데이터를 증강한다. 감성 분석과 같은 분류문제에 적용하여 성능을 높였다.

[6]은 WordNet과 같은 어휘 사전이 필요한 EDA를 한국어에 적용하기 위하여 BERT 기반의 MLM을 이용하여 한국어 의미역 결정 데이터를 증강한다. 그 결과 EDA의 RD를 적용한 증강 데이터 보다 좋은 성능을 보였다.

3. 기존 규칙기반 데이터 증강 방법의 문제점

대명사 고유명사 교체, 괄호 제거 방식[4] : 이 방식은 간단한 규칙으로 문맥에 영향을 미치지 않도록 한국어 상호참조해결 데이터를 증강하는 방식으로, 구현이 쉽고 자동으로 상호참조해결 데이터를 증강할 수 있다는 장점이 있지만 문서에 서로 참조하는 대명사 고유명사 멘션이 존재하거나 소괄호가 있어야 한다는 제약조건이 있다. 이러한 제약조건 때문에 모든 문서에 대하여 데이터 증강이 불가능하다. 실제로 ETRI 질의응답 도메인 상호참조해결 데이터 셋의 학습 문서 2,819개 중에서 대명사 고유명사 교체 방식과 소괄호 제거 방식으로 각각 306문서, 815문서 만 증강이 가능했다. 이는 더 다양한 데이터를 학습할 수 있게 만들어주는 데이터 증강 기법의 목적을 제대로 만족시킬 수 없음을 의미한다.

Easy Data Augmentation(EDA) 방식[5] : 이 방식은 단어 임의 제거(RD - Random Deletion), 동의어 교체(SR - Synonym Replacement), 임의 단어 위치 교환(RS - Random Swap), 임의 단어 삽입(RI : Random Insertion) 규칙을 사용하여 데이터를 증강한다. EDA 방식을 한국어 상호참조해결 데이터에 그대로 적용할 경우, 상호참조 대상인 멘션을 삭제 혹은 교체할 수 있고, 이는 문서의 문맥을 바꾸어 상호참조 관계에 직접적인 영향을 미칠 수 있다. 또한, 동의어 교체 방식은 어휘 의미 사전에 기반하여 단어를 교체하기 때문에 어휘 중의성이 있는 경우 문맥에 영향을 미치거나 비문이 되는 문제가 발생한다. 이러한 문제들은 데이터 증강 시 오류 데이터를 만들어 내며, 상호참조해결 모델의 학습에 악영향을 끼칠 수 있다. [표 2]는 EDA를 적용하여 기존 상호참조해결 데이터를 증강할 때 발생한 문제점들을 보여준다.

표 2. 상호참조해결 데이터에 EDA 적용시 오류 예제

원문	오스트리아 출생으로 **[적혈구가 다른 사람의 혈청에 의하여 응집되는 것]**을 발견하고, **[이]**에 의해 A식 혈액을 분류하였다.
임의 단어 제거(RD)	오스트리아 출생으로 **[적혈구가 다른 사람의 혈청에 의하여 응집되는 것]**을 발견하고, **[이]**에 의해 A식 혈액을 분류하였다.
원문	**[패스트푸드의 총칭]**으로 나트륨, 당, 지방 등의 성분이 일정 기준 이상 들어있어 **[비만을 초래하는 것]**은?
동의어 교체(SR)	**[패스트푸드의 총칭]**으로 나트륨, 연합, 지방 불의 성분이 일정 기준 이상 들어있어 **[비만을 초래하는 것]**은?
원문	**[삼성]**은 2017년 2분기 영업이익 1위를 지켜온 애플을 제쳤다. **[세계에서 가장 돈을 많이 번 제조 기업]**이다.
임의 단어 위치 교환(RS)	**[삼성]**은 2017년 애플을 영업이익 1위를 지켜온 2분기 제쳤다. **[세계에서 가장 돈을 많이 번 제조 기업]**이다.
원문	**[세종]**은 **[조선의 제4대 국왕]**이며 **[언어학자]**이다. **[그]**는 근정전에서 훈민정음을 반포했다.
임의 단어 삽입(RI)	**[세종]**은 **[조선의 제4대 국왕]**이며 **[언어학자]**이다. **[그]**는 조선의 근정전에서 훈민정음을 반포했다.

[표 2]의 임의 단어 제거(RD) 예제를 보면, '출생으로' 단어가 삭제가 되어 오스트리아가 [적혈구가 다른 사람의 혈청에 의하여 응집되는 것]을 수식하는 것처럼 보이게 된다. 또한 대명사 [이]를 삭제하여 상호참조 관계에 영향을 미친다. 동의어 교체(SR)예제에서는 '당'을 당질의 의미가 아닌 '연합'으로 바꾸거나, '~등'을 등불의 '불'로 교체하는 오류가 발생한다. 임의 단어 위치 교환과 같은 경우 예제와 같이 문법적으로 맞지 않은 비문을 생성할 확률이 높다. 임의 단어 삽입(RI)은 새로운 명사가 삽입이 될 경우 상호참조 관계에 영향을 미칠 수 있다. 이러한 오류들은 상호참조해결 학습 과정에서 악영향을 끼칠 수 있다.

따라서 본 논문에서는 비문을 생성하는 임의 단어 위치 교환(RS)과 상호참조 관계에 영향을 미치는 임의 단어 삽입(RI)을 상호참조해결 데이터 증강에서 제외하였으며, 임의 단어 제거(RD)와 동의어 교체(SR)방식은 멘션을 제외한 멘션의 주변 단어에만 데이터 증강을 적용하

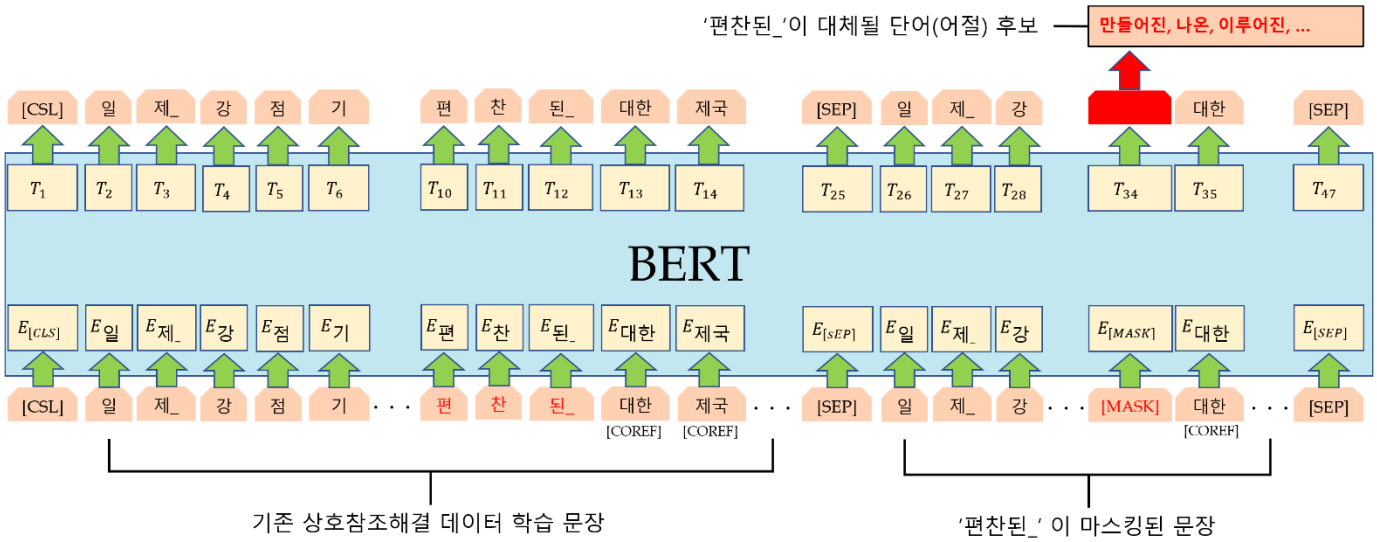

그림 1. BERT 기반 상호참조해결 데이터 증강 기법(BERT_SR4COREF) 모델 구조.

여 실험한다. 5절에서는 이 두 방법(EDA RD, EDA SR)과 BERT 기반 상호참조해결 데이터 증강 기법(BERT_SR4COREF)을 비교한다.

4. BERT기반 데이터 증강방법

본 논문에서 제안하는 BERT 기반 상호참조해결 데이터 증강 기법(BERT_SR4COREF)의 모델 구조는 [그림 1]과 같다. 상호참조해결 데이터는 문서 단위로 되어있지만 BERT의 512토큰 입력 제한으로 인해 문장 단위로 실행한 후 동일한 문서의 결과 문장들을 다시 합쳐서 문서 단위로 만든다. 기존의 상호참조해결 데이터 문장과 기존 문장에 데이터 증강을 위하여 마스킹이 적용된 문장은 [SEP]토큰으로 구분되어 사전학습된 BERT 모델의 입력으로 들어간다. 각 문장은 BPE가 적용되며, 교체시켜줄 단어를 어절 단위로 마스킹한다. 이때, 상호참조 관계를 가지고 있는 멘션의 중심어(head)는 마스킹 대상에서 제외된다. 마스킹 비율(α)은 하이퍼 파라미터로 본 논문에서는 $\alpha = 0.1$ 값을 사용한다. BERT 모델은 MLM(masked language model) 태스크로 사전학습이 되었기 때문에 마스킹된 어절 자리의 문맥에 어울리는 단어의 확률이 높게 나오며, 기존의 단어를 제외하고 확률이 가장 높은 단어를 선택하여 출력한다. [그림 1]의 예제에서는 '편찬된'이라는 어절 대신 '만들어진, 나온, 이루어진' 등의 대체어 중 가장 높은 확률을 가진 단어가 출력이 된다. [표 3]은 BERT_SR4COREF를 사용하여 한국어 상호참조해결 데이터를 증강한 결과 예제를 보여준다. 대괄호([])로 묶여있는 부분은 멘션이다.

표 3. BERT_SR4COREF을 사용한 데이터 증강 예제.

원문 1	**[이 책]**은 **[임진왜란의 원인, 전황 등을 기록한 책]**으로 **[선조 때 영의정을 지낸 유성룡이 벼슬에서 물러나 낙향해 있을 때 집필한 것]**이다.
BERT_SR4COREF 1	**[이 책]**은 **[임진왜란의 원인, 전황 등을 적은 책]**으로 **[선조 때 영의정을 한 유성룡이 벼슬에서 나와 낙향해 있을 때 쓴 것]**이다.
BERT_SR4COREF 2	**[이 책]**은 **[임진왜란의 원인, 전황 등을 밝힌 책]**으로 **[선조 때 영의정을 했던 유성룡이 벼슬에서 내려 낙향해 있을 때 기록한 것]**이다.
원문 2	한-이탈리아 양국은 또 **[[한국]의 대구시가 추진 중인 `밀라노 프로젝트']**와 관련, 양국간 섬유기술교류 및 공동연구를 위한 협약도 체결할 예정이다.
BERT_SR4COREF 1	한-이탈리아 정부는 또 **[[한국]의 대구시가 추진 하는 `밀라노 프로젝트']**와 관련, 양국간 섬유기술교류 및 공동연구를 지원하는 협약도 체결할 예정이다.
BERT_SR4COREF 2	한-이탈리아 정상은 또 **[[한국]의 대구시가 추진 준비중인`밀라노 프로젝트']**와 관련, 양국간 섬유기술교류 및 공동연구를 통한 협약도 체결할 예정이다.
원문 3	**[조선 시기 마지막 법전]**으로 알려진 **[이 법전]**은 **[무엇]**일까?
BERT_SR4COREF 1	**[조선 시기 마지막 법전]**으로 전해진 **[이 법전]**은 **[무엇]**일까?
BERT_SR4COREF 2	**[조선 시기 마지막 법전]**으로 공개된 **[이 법전]**은 무엇일까?
원문 4	**[김 대통령]**의 연설에 앞서 **[[페터 게트겐스] 총장]**은 **[[아시아 민주주의의 상징인] 김 대통령]**의 베를린 방문을 환영한다는 내용의 환영사를 한 뒤 **[김 대통령]**에게 `자유의 메달'을 증정했다.
BERT_SR4COREF 1	**[김 대통령]**의 연설에 이어 **[[페터 게트겐스] 총장]**은 **[아시아 민주주의의 중심 김 대통령]**의 베를린 방문을 기대한다는 내용의 환영사를 한 후 **[김 대통령]**에게 `자유의 메달'을 증정했다.
BERT_SR4	**[김 대통령]**의 연설에 대해 **[[페터 게트**

COREF 2	**겐스] 총장]**은 **[아시아 민주주의의 중심으로 김 대통령]**의 베를린 방문을 지지하는 내용의 환영사를 한 직후 **[김 대통령]**에게 `자유의 메달'을 증정했다.

5. 실험 및 결과

본 논문에서 제안한 BERT 기반 상호참조해결 데이터 증강 방법을 실험하기 위하여 ETRI 질의응답 도메인 상호참조해결 데이터 셋을 이용하였으며, 학습(Train) 데이터 2,819 문서, 개발(Dev) 데이터 645 문서, 평가(Test) 데이터 571 문서로 구성된다. [2]의 BERT 기반 End-to-end 상호참조해결 모델을 baseline으로 선정하여 성능을 측정하였다. 성능 측정은 MUC, B^3, CEAF-e, CoNLL F1[8]으로 진행하였으며, 중심어 경계(head boundary)를 기준으로 성능을 측정하였다. 데이터 증강 실험에 사용된 파라미터는 마스킹 비율 $\alpha = 0.1$을 사용했으며, 기존 데이터와 증강 데이터 비율 γ은 2:1(기존데이터 문서 수 : 증강데이터 문서 수)를 사용하였다. 실험에 사용된 GPU는 TITAN RTX 24GB이다.

표 4. 데이터 증강 방법에 따른 상호참조해결 성능 비교 (%).

DEV				
Model	**MUC**	**B^3**	**CEAFe**	**Dev F1**
Baseline[2]	72.06	69.68	70.72	70.78
대명사 고유명사 교체 + 소괄호 제거[4]	72.67	70.02	70.77	71.21
EDA SR[5]	72.41	70.16	71.39	71.32
EDA RD[5]	**73.18**	70.43	70.79	71.47
EDA SR + EDA RD [5]	72.78	70.37	71.48	71.54
BERT_SR4COREF (Ours)	72.88	**70.76**	**71.83**	**71.82**
TEST				
Model	**MUC**	**B^3**	**CEAFe**	**Test F1**
Baseline[2]	69.69	67.17	68.47	68.44
대명사 고유명사 교체 + 소괄호 제거[4]	71.14	68.44	69.21	69.60
EDA SR[5]	70.52	68.08	69.20	69.27
EDA RD[5]	70.66	67.90	69.10	69.22
EDA SR + EDA RD [5]	71.08	68.41	**69.59**	69.69
BERT_SR4COREF (Ours)	**71.23**	**68.68**	**69.59**	**69.83**

[표 4]는 데이터 증강 방법에 따른 상호참조해결 성능을 비교 결과이다. 각 데이터 증강 방식으로 증강한 데이터에 기존 데이터를 합하여 학습을 진행한 결과이다. 본 논문에서 제안한 BERT_SR4COREF 방식으로 증강한 데이터를 실험하였을 때, 기존 데이터로 실험했을 때보다 개발 셋(DEV)에서 CoNLL F1 1.04% 증가한 71.82%의 성능을 보였다. 이는 [4]보다 0.61%, [5]의 SR 방법보다 0.5%, [5]의 RD 방법보다 0.35%, [5]의 RD와 SR을 동시에 적용했을 때보다 0.28% 높은 성능이다. 평가 셋(TEST)에서는 1.39% 증가한 69.83%의 성능을 보였으며, [4]보다 0.23%, [5]의 SR 방법보다 0.56%, [5]의 RD 방법보다 0.61%, [5]의 RD와 SR을 동시에 적용했을 때보다 0.14% 높은 성능이다.

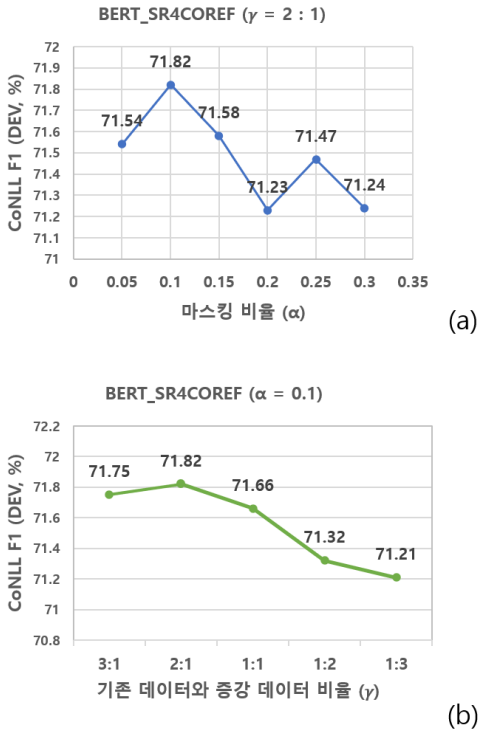

그림 2. BERT_SR4COREF의 마스킹 비율(α) 파라미터 튜닝(a), 기존 데이터와 증강 데이터 비율(γ) 실험(b).

[그림 2]는 BERT_SR4COREF 방식의 하이퍼 파라미터인 마스킹 비율 α와 기존 데이터와 증강 데이터 비율 γ의 실험 결과이다. 파라미터 튜닝은 개발 셋을 기준으로 진행하였으며 α = 0.1 (전체 데이터 중에 10%)이고, γ = 2 : 1 (증강 데이터 한 번 학습할 때 기존 데이터 두 번 학습)일 때 71.82%로 가장 높은 성능을 보였다.

6. 결론

본 논문에서는 기존 규칙기반의 데이터 증강 기법의 문제점을 보완하는 BERT_SR4COREF 방식의 한국어 상호참조해결 데이터 증강 기법을 제안하였다. BERT_SR4COREF방식은 모든 상호참조해결 데이터 문서에 적용할 수 있으며 BERT의 MLM을 사용하여 기존의 단어를 새로운 단어로 치환할 때 문맥에 영향을 주지 않는다. 실험 결과 증강된 데이터를 함께 학습한 결과 평가 셋에서 baseline보다 1.39%향상된 69.83%를 보였다. 이는 규칙기반 방법인 [4]보다 0.23%, [5]의 RD와 SR을 동시에 적용했을 때보다 0.14% 높은 수치이다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의

재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능진화형 Wise QA 플랫폼 기술 개발

참고문헌

[1] 김기훈, 박천음, 이창기, 김현기. “고차 추론을 이용한 한국어 End-to-end 신경망 기반 상호참조해결”, 한국정보과학회 학술발표논문집, 한국정보과학회 2019 한국컴퓨터종합학술대회 논문집, 2019.

[2] 김기훈, 박천음, 이창기, 김현기. “BERT 기반 End-to-end 신경망을 이용한 한국어 상호참조해결”, 제31회 한글 및 한국어 정보처리 학술대회 논문집, pp. 181-184. 2019.

[3] Pradhan, Sameer, et al. "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes." *Joint Conference on EMNLP and CoNLL-Shared Task*. 2012.

[4] 김기훈, 이창기, 임준호, 김현기. “단순 규칙을 이용한 한국어 상호참조해결 데이터 증강 방법”, 한국정보과학회 학술발표논문집, 한국정보과학회 2020 한국컴퓨터종합학술대회 논문집, pp.326-328, 2020.

[5] Wei, Jason, and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." *arXiv preprint arXiv:1901.11196,* 2019.

[6] 배장성, 이창기, 임준호, 김현기. “한국어 의미역 결정을 위한 BERT 기반 데이터 증축 기법”, 한국 정보과학회 학술발표논문집, 한국정보과학회 2020 한국컴퓨터종합학술대회 논문집, pp.335-337, 2020.

[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805,* 2018.

[8] Lee, Heeyoung, et al. "Deterministic coreference resolution based on entity-centric, precision-ranked rules." *Computational linguistics* 39.4 pp. 885-916, 2013.