사전 기반 자질과 동적 마스킹을 이용한 ELECTRA 기반 개체명 인식

김정욱○, 황태선, 김봉수, 이새벽,

와이즈넛
{jwkim, taesunwhang, usgnob, saebyeok}@wisenut.co.kr

Named Entity Recognition based on ELECTRA with Dictionary Features and Dynamic Masking

Jungwook Kim○, TaesunWhang, Bongsu Kim, Saebyeok Lee
Wisenut Inc.

요 약

개체명 인식이란, 문장에서 인명, 지명, 기관명, 날짜, 시간 등의 고유한 의미의 단어를 찾아서 미리 정의된 레이블로 부착하는 것이다. 일부 단어는 문맥에 따라서 인명 혹은 기관 등 다양한 개체명을 가질 수 있다. 이로 인해, 개체명에 대한 중의성을 가지고 있는 단어는 개체명 인식 성능에 영향을 준다. 본 논문에서는 개체명에 대한 중의성을 최소화하기 위해 사전을 구축하여 ELECTRA 기반 모델에 적용하는 학습 방법을 제안한다. 또한, 개체명 인식 데이터의 일반화를 개선시키기 위해 동적 마스킹을 이용한 데이터 증강 기법을 적용하여 실험하였다. 실험 결과, 사전 기반 모델에서 92.81 %로 성능을 보였고 데이터 증강 기법을 적용한 모델은 93.17 %로 높은 성능을 보였다. 사전 기반 모델에서 추가적으로 데이터 증강 기법을 적용한 모델은 92.97 %의 성능을 보였다.

주제어: ELECTRA, 개체명 인식, 사전 기반 모델, 데이터 증강

1. 서론

개체명 인식(Named Entity Recognition)은 주어진 문장에서 인명, 지명, 기관명, 날짜, 시간 등의 고유한 의미의 명명된 개체를 찾아서 미리 정의된 레이블로 부착하는 것이다. 개체명의 인식이 어려운 이유는 새로운 개체명이 꾸준히 만들어 지고 있으며, 문맥에 따라 다른 개체명으로 해석될 수 있는 중의성을 가지고 있기 때문이다. 즉, 개체명을 구성하는 단어만으로는 개체명의 유형을 결정할 수 없고, 문맥 지식을 활용해야 하는 경우가 많이 발생한다.[1] 예를 들어, “조지 워싱턴은 미국의 초대 대통령이다”, “워싱턴에서 한·미 정상회담을 하였다”, “워싱턴은 콜로라도를 상대로 3:0으로 승리했다.”와 같이 세 문장에서 “워싱턴”은 각각 인명, 지명, 기관명으로 서로 다르게 해석된다. 이런 문제들을 해결하기 위해 [1]은 중의성을 가진 다양한 자질을 생성하기 위한 자질 생성 방법을 제안하였고 [2]는 개체명 구성 원리를 이용하고 주변 문맥을 참고하는 기계학습 방법을 제안했다. 또한, BERT[3]와 같은 사전학습모델(PLM: Pre-trained language model)이 등장하면서 문맥에 따른 단어의 중의성을 일부 해소해주었다.

최근 개체명 인식은 대용량 텍스트 분석, 챗봇 등에서 많이 사용되고 있다. 특히, 자동화된 서비스를 제공하는 입장에서 제품명, 상품명, 서비스명 등 주기적으로 변화하는 개체명에 대해 추가 학습없이 인식하는 것에 대한 요구가 증가하고 있다. PLM의 등장 이후 전이학습 기반으로 다운스트림에서 개체명인식을 학습하기 때문에 개체명인식기의 성능이 많이 향상되었고, 학습데이터에 없는 개체명도 기존의 학습셋만으로 학습 할 때 보다 더 잘 인식할 수 있다.

본 논문은 ELECTRA로 학습된 PLM을 이용하여 개체명인식 파인튜닝(fine-tuning)에서 개체명 사전을 추출하고 이를 자질로 추가해서 학습하는 방법과 사전학습 모델에서 사용하는 masked language model(MLM)처럼 입력데이터의 일부를 동적 마스킹을 적용하였고, 데이터 증강(augmentation)의 효과가 있는지 실험하였다. 그 결과 기존 연구들 보다 더 좋은 성능을 보여주었고, 최종 모델은 학습 이후 사전을 관리하는 것으로 추가 학습 없이 새로운 개체명도 잘 인식할 수 있었다.

2. 관련 연구

최근 대용량 코퍼스를 이용하여 masked language model(MLM)을 기반으로 한 언어 모델들[3, 4, 5, 6]이 다양한 자연어 다운스트림 태스크에서 뛰어난 성능을 보여주어 개체명 인식에서 또한 좋은 성능을 보여줬다. 대표적으로 BERT[3]는 self-attention[7]을 기반으로 하고 Transformer를 인코더로 사용하여 MLM을 통해 학습하는 언어모델이다. [8]는 BERT를 활용하여 개체명 인식을 수행했다. RoBERTa[4]는 동적 마스킹(dynamic masking LM)과 큰 배치 사이즈로 BERT 모델 구조를 더 견고하게 학습할 수 있는 방법을 제안했다. [9]은 RoBERTa 모델을 적

용하고 형태소 단위의 토크나이저와 BPE(Byte Pair Encoding) 단위의 토크나이저를 결합하여 미등록어에 강건한 하이브리드 방식의 토크나이즈 방식을 제안하여 기존 BERT 모델의 성능을 향상시켰다. ALBERT[5]는 큰 모델을 효과적으로 학습하기 위해 모델 크기를 줄이는 방법과 문장 순서 예측하는 태스크를 제안하여 성능을 향상시켰다. [10]은 ALBERT 모델을 활용하여 모델의 파라미터 수를 줄인 AL-RoBERTa를 제안하고 여러 다운스트림 태스크에 적용하였다. ELECTRA[6]는 기존 BERT 모델에서 학습 효율성을 높이기 위해 N개의 입력 토큰 전체에서 loss를 계산하기 위한 방법을 제안하였다. [11]은 ELECTRA 모델을 활용하여 개체명 인식기를 포함한 자연어 다운스트림 태스크를 수행하였다. [12]는 ELECTRA와 CRF(Condition Random Fields)의 단점을 보완한 LAN(Label Attention Network)을 활용한 개체명 인식 모델을 제안했다.

본 논문에서는 ELECTRA 모델에 사전 자질을 추가하고 데이터 증강 기법으로 동적 마스킹을 파인 튜닝에 적용하는 개체명 인식 모델을 제안한다.

3. 개체명 인식 모델

본 논문에서 제안하는 개체명 인식 모델은 ELECTRA의 Transformer Encoder의 구조를 그대로 사용하였다. 다만, 입력 단위가 문장이기 때문에, 세그먼트 임베딩은 사용하지 않고, 입력으로 사전 자질을 추가로 넣어 주고 출력은 토큰 단위의 분류(token classification)를 할 수 있도록 Dense 레이어를 추가하였다.

3.1 입력 층

토큰 임베딩은 문장의 처음과 끝에 각 [CLS], [SEP]을 부착하고 문장을 형태소 분석 후 하위 토큰으로 나눈 뒤 해당 토큰의 아이디를 부착한다. 세그먼트 임베딩은 토큰의 길이만큼 1로 채워준다. 사전 임베딩은 미리 만든 학습 사전을 통해 토큰이 해당하는 B-I 레이블 태깅으로 구성하고 레이블에 해당하지 않는 토큰은 O으로 태깅한다.

3.2 사전 구축 및 사전 자질 입력

개체명 사전은 학습 데이터에서 단어의 길이가 2 이상이면서 개체명이 부착된 단어 중 2개 이상의 레이블을 가지는 중의적인 단어를 제거 후, 출현 횟수로 정렬 하여 상위 단어들을 추출하였다. 학습할 때 자질로 사용하는 학습 사전과 평가할 때 자질로 사용하는 평가 사전으로 구성하였다. 그림 1은 사전의 예시이다.

모델의 전체 구조도는 그림 2와 같다. 입력층에서 기본적인 입력으로 사용하는 토큰 아이디 T, 세그먼트 S, 포지션 P와 추가적으로 사전 아이디 D를 사용한다. D 또한 다른 임베딩(T, S, P)과 마찬가지로 더해 준 다음, 아래 식 (1)과 같이 정규화와 드롭아웃을 거친다.

$$E = Dropout\big(LayerNorm(T + S + P + D)\big) \quad (1)$$

임베딩의 크기는 최대 길이 문자열로 모두 동일하다. ELCTRA에서 입력층으로부터 임베딩 값들을 받고 동적 마스킹 기법을 이용하여 매 학습마다 일정 확률로 새로운 마스킹을 하면서 학습을 한다. 마지막으로 입력 임베딩과 관련이 높은 레이블 값으로 출력한다.

PS	박지성
PS	문성민
PS	박찬호
OG	스포츠서울닷컴
OG	맨체스터 유나이티드
PS	신지애
PS	김주성
OG	K리그
. . .	

그림 1. 사전 예시

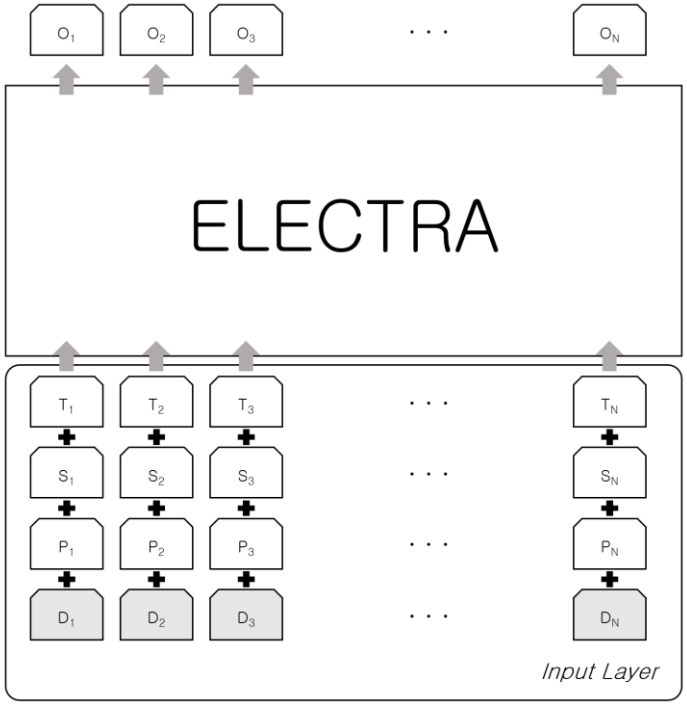

그림 2. 개체명 인식 모델 전체 구조도

3.3 데이터 증강 기법

본 논문에서 제안하는 데이터 증강 기법은 RoBERTa에서 제안한 동적 마스킹(dynamic masking LM)을 활용하였다. 동적 마스킹은 매 학습 단계마다 임의의 확률로 마스킹 변환을 새로 적용한 방법으로, 더 많은 학습 데이터로 더 많은 단계(step)를 학습하기 위해서는 동적 마스킹이 중요하다. RoBERTa 및 ELECTRA에서는 사전 학습을 위해 동적 마스킹을 사용하였지만 본 논문에서는 개체명 인식에 활용하여 데이터 증강에 활용했다. 마스킹 된 토큰을 맞추는 문제가 아닌, 학습 Step이 증가함에 따라 한정된 학습셋에서 더 다양한 데이터로 학습시키기 위한

목적이다. 마스킹 대상은 [CLS], [SEP] 토큰과 레이블이 부착된 토큰을 제외한 토큰들을 마스킹 후보 토큰으로 설정하였다.

원문	<DT>올해</DT> 각각 친정팀으로 돌아온 뒤에는 <OG>부산</OG>의 프로스포츠 바람을 주도하고 있다.
토큰화	['[CLS]', '올해/NNG_', '각각/MAG_', '친', '정/NNG_', '팀/NNG_', '으로/JKB_', '돌아오/VV_', 'ㄴ/ETM_', '뒤/NNG_', '에/JKB_', '는/JX_', '부산/NNP_', '의/JKG_', '프로/NNG_', '스포츠/NNG_', '바람/NNG_', '을/JKO_', '주도/NNG_', '하/XSV_', '고/EC_', '있/VV_', '다/EF_', './SF_', '[SEP]']
토큰 교체	['[CLS]', '올해/NNG_', '각각/MAG_', '친', '정/NNG_', '팀/NNG_', '으로/JKB_', '[MASK]', 'ㄴ/ETM_', '뒤/NNG_', '에/JKB_', '는/JX_', '부산/NNP_', '의/JKG_', '프로/NNG_', '스포츠/NNG_', '[MASK]', '을/JKO_', '주도/NNG_', '하/XSV_', '고/EC_', '있/VV_', '다/EF_', './SF_', '[SEP]']
토큰 교체	['[CLS]', '올해/NNG_', '각각/MAG_', '친', '정/NNG_', '팀/NNG_', '으로/JKB_', '돌아오/VV_', 'ㄴ/ETM_', '뒤/NNG_', '에/JKB_', '는/JX_', '부산/NNP_', '[MASK]', '프로/NNG_', '스포츠/NNG_', '바람/NNG_', '을/JKO_', '주도/NNG_', '[MASK]', '고/EC_', '있/VV_', '다/EF_', './SF_', '[SEP]']
...	

그림 3. 데이터 증강 예시

그림 3에서 "올해"와 "부산"의 단어는 레이블이 부착된 토큰이기 때문에 제외하고 나머지 토큰은 마스킹 대상이다. 매 학습 단계마다 해당 토큰들은 "[MASK]" 토큰으로 교체되어 학습하게 된다.

4. 실험 및 결과

4.1 실험 데이터

한국어 ELECTRA 사전 학습 모델[11]은 4.5B 토큰들로 구성된 신문기사, 방송, 웹 문서를 사용하고 형태소 분석과 함께 워드피스 토큰화(word piece tokenizer)를 통해 만들었다. 본 논문에서는 small 모델을 사용한다. 개체명 인식 학습과 실험 데이터는 엑소브레인 언어 분석 말뭉치 중 개체명 인식 말뭉치를 사용하였다. 총 10,000 문장 중 8,000 문장은 학습 데이터로, 2,000 문장은 평가 데이터로 구성하였다.

4.2 실험 결과

표 1과 표2는 학습 사전과 평가 사전에 따른 실험 결과이다. 학습 사전 및 평가 사전은 정렬된 개체명-단어 쌍에서 상위 %로 구성하였다. 배치 사이즈는 32, 학습 단계(step)는 500으로 학습하였다.

표 1에서 사전 임베딩을 추가하지 않은 기존 모델[11]은 92.13 %의 성능을 보였다. 사전 임베딩을 추가한 모델에서는 상위 단어 50 %의 사전으로 학습한 모델이 92.81 %로 성능이 높았고 오히려 사전의 크기가 클수록 성능이 떨어지는 것으로 나타났다.

표 2는 50 %의 학습 사전으로 학습을 하고 평가 사전을 다르게 구성하여 실험을 한 결과이다. 학습과 동일하게 50 % 사전을 사용한 모델이 성능이 가장 높았다.

학습 사전	단어 수	F1 Score
0%	0	92.13 [11]
25%	1,569	92.31
50%	3,138	**92.81**
75%	4,707	92.03
100%	6,276	89.96

표 1. 학습 사전에 따른 실험 결과

학습 사전	평가 사전	단어 수	F1 Score
50%	0%	0	92.58
	25%	1,569	92.69
	50%	3,138	**92.81**
	75%	4,707	92.79
	100%	6,276	92.78

표 2. 평가 사전에 따른 실험 결과

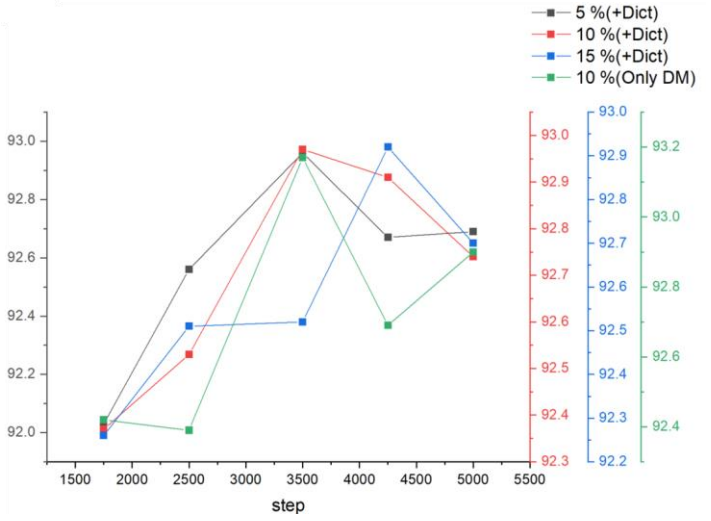

그림 4. 마스킹 확률과 데이터 증강에 따른 실험 결과

Model	F1 Score
BERT + CRF [8]	91.58
ALBERT + LSTM + CRF [10]	91.87
RoBERTa + LSTM + CRF [9]	91.94
ELECTRA + LAN [12]	92.78
ELECTRA small [11]	92.13
ELECTRA + Dictionary + DM (Ours)	**92.97**
ELECTRA + DM(Ours)	**93.17**

표 3. 기존 모델과 제안 모델의 성능 비교

목표 단어	입력 문장	예측 결과	
박지성길: LC	수원시와 화성시는 총연장 4.7㎞ 중 1.3㎞ 구간은 수원시, 3.4㎞ 구간은 화성시에 걸쳐 있는 박지성길의 명칭을 '동탄지성로'로 바꾸기로 합의했다고 밝혔다.	사전O	박지성길: LC
		사전X	박지성길: PS
세종: PS	세종(世宗, 1397년 5월 7일 ~ 1450년 3월 30일)은 조선의 제4대 국왕(재위 : 1418년 9월 9일 ~ 1450년 3월 30일)이다. 세종의 업적에 대한 존경의 의미를 담아 세종대왕(世宗大王)으로 일컬어진다.	사전O	세종: PS
		사전X	세종: OG
레오나르도 다빈치: PS	레오나르도 다빈치의 스케치인 비트루비우스적 인간은 인체 비례에 대한 상징처럼 여겨질 정도로 유명하다.	사전O	레오나르도 다빈치: PS
		사전X	-

표 4. 목표단어들의 사전 적용 전, 후의 정성적 평가

그림 4는 동적 마스킹을 적용하여 더 많은 학습 단계로 실험한 결과이다. 마찬가지로 배치사이즈는 32로 설정했다. 학습 단계가 많아지면 더 많은 데이터로 학습할 수 있기 때문에 학습 단계가 많아질수록 성능이 높아지는 것을 보였다. 3500 step에서 마스킹 비율 10 %로 설정했을 때 F1 score 92.97 %로 가장 높았다. 일정 단계부터는 과적합(overfitting)이 발생하여 성능이 다소 떨어진다. 사전과 데이터 증강 기법을 통해 학습한 모델은 기존 모델[11] 보다 0.84 % 가량 높은 성능을 보였고 데이터 증강 기법만 적용해서 실험한 결과 93.17 %로 더 높은 성능을 보였다.

표 3는 ETRI 데이터셋에 대해서 기존 모델과 제안 모델의 성능 비교이다. 본 논문에서 제안한 모델이 기존 모델과 비교하여 더 좋은 성능을 보였다.

표 4는 사전 적용의 효과를 보기 위해서 파인 튜닝된 모델을 수정하지 않고, 사전에 없는 단어를 사전에 추가했을 경우 결과가 변경되는 것을 정성적으로 보여주기 위한 표이다. 첫 번째 예시에서 사전에 “박지성길”을 지명(LC)으로 추가하면 기존에 인명(PS)로 예측되는 결과가 지명(LC)으로 예측된다. 마찬가지로 두 번째 예시에서 “세종”이 기관명(OG)으로 예측되는 결과를 사전 추가를 통해 인명(PS)으로 예측되도록 한다. 세 번째 예시는 기존에 개체명 부착이 안되었던 “레오나르도 다빈치”라는 단어를 인명(PS)으로 사전에 추가 시 인명(PS)으로 예측된다.

5. 결론

본 논문에서는 사전 학습 모델로 한국어 형태소 기반으로 만든 ELECTRA를 활용하고 학습데이터를 통해 만든 사전을 추가적인 자질로 사용하여 학습하고 이후 재 학습 없이 사전을 통한 품질 제어가 가능한 것을 보였다. 특히, 데이터 증강 방법으로 동적 마스킹을 파인 튜닝에 적용함으로써 과적합에 강하고 학습이 더 잘 수렴되어 더 좋은 성능이 나오는 것을 확인하였다. 실험결과에서는 사전자질을 사용하는 것 보다, 동적 마스킹만을 사용하는 것이 더 좋은 성능을 갖지만, 사전 자질을 사용할 경우 학습 이후에도 일부 결과에 대한 컨트롤이 가능하다는 장점이 있다. 기존의 모델은[9, 10, 12] 모두 추가 레이어나 구조를 이용하여 성능을 높였지만, 본 논문에서는 사전 학습 모델을 이용하여 데이터와 입력만을 다뤄 더 좋은 성능 향상을 보였다. 또한 레이어나 구조를 추가하지 않았기 때문에 계산할 파라미터의 수의 차이가 거의 나지 않는다는 점과 LSTM-CRF처럼 병목 현상이 있는 레이어를 사용하지 않아 더 빠른 추론 서비스가 가능하다는 점에서 의미가 있다. 이후의 연구로는 다른 사전 구축 방법과 데이터 증강 기법을 적용하여 실험을 진행할 예정이다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 1711125985, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)과 2021년도 정부(산업통상자원부)의 재원으로 한국이스라엘산업연구개발재단의 지원을 받아 수행된 연구임 (금융 지식 그래프를 위한 다국어 자연어처리 기술 개발, 과제번호: 2018-35-169)

참고문헌

[1] 황이규, 이현숙, 정의석, 윤보현, 박상규, “개체명 구성 원리를 이용한 교사학습 기반의 한국어 개체명 인식”, 한글 및 한국어 정보처리 학술대회 논문집, pp.111-117, 2002.

[2] 김재훈, 김형철, 최윤수, "기계학습 기반 개체명 인식을 위한 사전 자질 생성", 정보관리연구, 제41 권, 제2호, pp.31-46, 2010.

[3] J. Devlin, M.-W. 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, Volume 1 (Long and Short Papers), pp. 4171-4186, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov,

"Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," International Conference on Learning Representations, 2020.
[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," International Conference on Learning Representations, 2020.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, pp. 5998-6008, 2017.
[8] 박광현, 나승훈, 신종훈, 김영길, "BERT를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정" 한국정보과학회 학술발표 논문집, pp. 584-586, 2019.
[9] 민진우, 나승훈, 신종훈, 김영길, "RoBERTa를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱", 한국정보과학회 학술발표 논문집, pp.407-409, 2019.
[10] 이영훈, 나승훈, 최윤수, 이혜우, 장두성, "ALBERT를 이용한 한국어 자연어처리: 감성분석, 개체명 인식, 기계독해", 한국정보과학회 한국소프트웨어종합학술 대회 논문집, pp. 332-334, 2020.
[11] 황태선, 김정욱, 이새벽, "한국어 ELECTRA 모델을 이용한 자연어처리 다운스트림 태스크", 한글 및 한국어 정보처리 학술대회 논문집, pp.279-282, 2020.
[12] 김홍진, 오신혁, 김학수, "ELECTRA와 Label Attention Network를 이용한 한국어 개체명 인식", 한글 및 한국어 정보처리 학술대회 논문집, pp.355-336, 2020.