BERT 모델을 이용한 한국어 공간 개체 추출

신형진○, 육대범, 이재성
충북대학교
{hjshin, daebum1994, jasonlee}@cbnu.ac.kr

Korean Spatial Elements Extraction using BERT

Hyeong Jin Shin○, Dae Bum Yuk, Jae Sung Lee
Chungbuk National University

요 약

텍스트에서 공간 정보를 추출하기 위해 그동안 통계 및 확률 기반 방법, 심층학습 방법 등이 연구되어 왔다. 본 연구에서는 최근 자연언어처리에서 우수한 성능을 보이고 있는 BERT 모델을 적용하여 공간 개체 정보를 추출한다. 공간 개체 추출은 공간 관계에 관련된 속성 추출을 함께 고려한 결합(joint) 모델로 구성하였으며, 한국어를 대상으로 기학습된 BERT 언어모델인 korBERT를 이용하였다. 실험결과, 기존의 방법들에 비해 1.9% 포인트 이상 증가한 성능을 보였다.

주제어: 한국어 공간 개체 정보 추출, BERT, korBERT, 결합모델, 언어모델

1. 서론

공간 정보 추출은 문장에 포함된 어휘들이 갖고있는 공간의 정적, 동적 관계를 추출하는 연구이다. 이를 위해 해당되는 공간 관련 개체를 식별하는 것이 선행되어야 한다.

공간 개체 추출은 규칙 또는 사전을 기반으로 정보를 추출하거나, 미리 구축된 학습 말뭉치를 활용하여 통계 및 확률 값을 기반으로 정보를 추출한다. 규칙 기반 정보 추출은 WordNet, PropBank 등 미리 구축된 어휘 사전의 정보에 맞는 공간 개체로 분류하는 방법으로 연구했고[1,2], 통계 및 확률 기반의 정보 추출은 준비된 말뭉치에서 연속으로 출현하는 빈도나 특정 개체로 쓰인 확률을 이용해 CRFs, SVM 등의 기계학습에 적용하여 연구했다[3-6].

공간 정보를 추출하는 문제는 주어진 문장을 구성하는 형태소 사이의 문맥 정보를 고려하여 순서대로 주석하는 순차 레이블링 문제라고 볼 수 있다. Kim 등(2016)의 연구에서는 당시 순차 레이블링 문제에 가장 우수한 성능을 보였던 CRFs 모델을 사용해 정보를 추출했고[7], 민태홍 등(2017)의 연구에서는 LSTM을 활용한 심층학습 모델과 CRFs 모델을 결합하여 공간 개체 정보를 추출했다[8].

최근 자연언어처리 연구에서는 대량의 데이터로 사전 훈련된 모델을 정밀 조정(fine-tuning)하여 목적에 따라 재훈련시켜 활용할 수 있는 BERT 언어 모델이 각광을 받고 있다[9]. 이 모델은 인코더와 디코더로 구성된 트랜스포머 모델을 기반으로 하며, 입력은 단어를 단어 조각(word piece)으로 분리한 후 단어 조각의 사전 번호와 문장에서의 위치 번호(position)를 결합하여 사용한다. 그 결과 언어 이해(GLUE), 질의응답(SQuAd) 등 다양한 태스크에서 우수한 결과를 얻었다[10, 11].

전자통신 연구소에서는 이 BERT 언어모델을 한국어에 맞게 사전 학습한 한국어 BERT 언어모델을 개발했다[12]. 본 연구에서는 이를 활용하여 한국어 공간 개체 정보를 추출한다. 그 결과, 기존의 한국어 공간 개체 정보 추출 연구들에 비해 우수한 성능을 얻었다.

2. 한국어 BERT 언어모델

BERT 모델은 영어, 중국어, 다국어 버전이 존재하며, 다국어 버전에서 한국어를 포함하고 있다. 한국어는 형태소 사이의 조합이 다양하기 때문에, 어절 단위로 학습하는 경우 더 많은 학습 데이터를 요구한다. 그에 따라, 다양한 국가의 언어에 대응하기 위해 어절 단위로 학습시킨 다국어 버전 외에 한국어에 적합한 형태소 단위로 학습시킨 모델을 개발할 필요가 있다.

이러한 한국어의 특성에 맞추기 위해 전자통신연구소(ETRI)에서는 신문기사 및 백과사전 등 약 23GB의 대용량 텍스트를 이용하여 학습시킨 한국어 BERT 언어 모델을 공개했다. 이 모델은 30,349개의 단어 조각(word-piece)으로 학습되었으며, BERT 다국어 모델과 비교했을 때 의미역 인식 분야에서 약 3.9% 포인트 더 나은 성능을 얻어낸 모델이다.

3. 공간 개체 정보 추출 모델

공간 관계 정보는 문장 내에 포함된 어휘 사이에 존재하는 공간적 연관성을 의미한다. 예를 들어, "*창고에 자전거가 있다.*"라는 문장이 있다면, '*자전거*'는 '*창고*', '*에 있다*'는 의미를 가지고 있을 것이다. 다만, 여기서 '*에 있다*'는 어절로 분리되어 있기 때문에 '*에*'에 의미를 부여하고, 이 의미를 컴퓨터가 이해할 수 있도록 '*자전거*', '*창고*', '*에*', '*있다*'가 어떤 개체로 사용되었는지 식별해 줄 필요가 있다.

공간 개체 정보 추출은 '*자전거*', '*창고*', '*에*' 등의 어휘에 대해 개체를 식별하는 단계이다. 위의 예시처럼 공간 개체는 형태소 단위로 주석 되므로, 입력 문장을 임베딩하기 전에 형태소 단위로 분석할 필요가 있다. 그를 위해, 전자통신연구소(ETRI)에서 제공하는 형태소 분석기를 사용했다.

각각의 형태소들은 토큰화(tokenizer) 단계를 거쳐 단어 조각으로 분리되고, 분리된 조각들은 임베딩 단계를 거쳐 BERT 모델의 입력으로 사용된다.

BERT 모델의 사전 학습된 트랜스포머 기반의 인코더-디코더 모델을 정밀 조정(fine-tuning)하여 한국어 공간 개체 정보 추출에 맞게 학습시켰다. 출력은 Logits 분류기를 연결하여 선택하도록 했다. 모델의 구성은 아래 그림 1과 같이 구성했다.

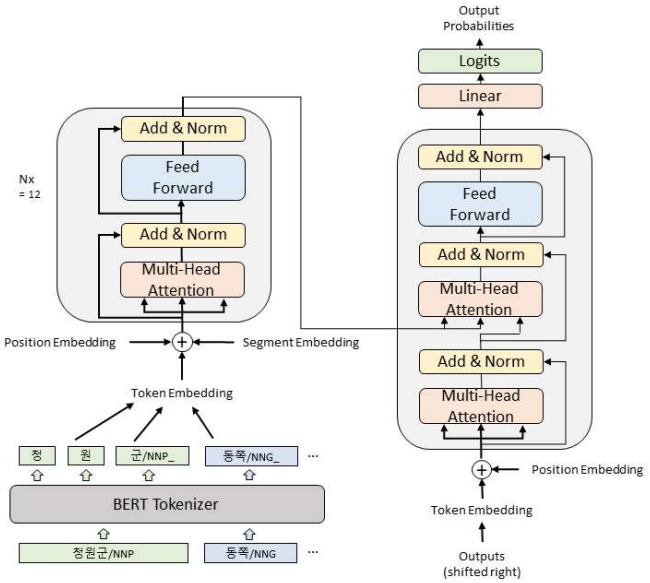

그림 1. BERT 기반 공간 개체 정보 추출 모델

출력은 ISOspace 기준으로 'PLACE', 'MOTION' 등 7가지 태그가 확률에 따라 결정된다[13]. 본 연구팀은 보다 나은 성능을 얻기 위해 그림 2와 같은 결합 모델 형태로 구성했다[14]. 관계 속성은 공간 개체가 공간 관계(방향, 이동 등)에서 어떤 역할을 하는지 나타낸다[2].

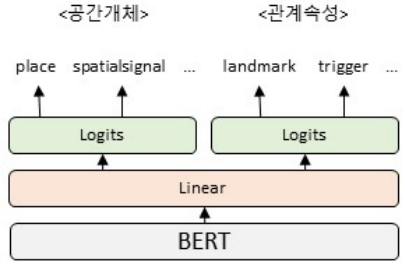

그림 2. 공간 개체 및 관계 속성 결합 모델

4. 실험 및 평가

실험은 전자통신연구소에서 연구용으로 배포하는 한국어 공간 정보 말뭉치 v2.1(ISOspace 기준 주석)을 이용해 진행하였다[12]. 상세한 구성은 아래 표 1과 같다.

표 1. 한국어 공간 정보 주석 말뭉치 개체 구성

개체	개수	개체	개수
문장	2,826	Spatial Entity	1,417
Place	6,730	Spatial Signal	2,104
Path	566	Motion	701
Measure	526	Motion Signal	835

평가는 IOB 기반 태그들의 일치 여부를 F1-score로 측정했으며, 5배수 교차검증으로 평가한 후 결과의 평균으로 기입하였다. 학습된 모델을 이용해 각각의 개체에 대한 F1-score를 측정한 결과 아래 표 2와 같은 성능을 얻었다.

표 2. 공간 개체 정보 추출 결과

개체	정확률	재현율	F1
Place	90.6	92.5	91.6
S.Entity	81.8	80.0	80.6
S.Signal	87.5	88.1	87.7
Motion	82.5	84.7	83.5
M.Signal	82.8	88.2	85.4
Path	82.3	83.6	82.9
Measure	95.3	93.2	94.2
마이크로평균	**89.5**	**88.3**	**88.9**
매크로 평균	**86.1**	**87.1**	**86.6**

이 성능은 표 3과 같이 CRFs 기반 모델, LSTM-CRFs 모델의 마이크로 평균과 비교해 본 결과 본 연구가 가장 우수한 성능을 보임을 확인할 수 있었다.

표 3. 기존 연구와 비교

모델	정확률	재현율	F1
통계 기반 CRFs[7]	85.8	86.2	86.0
Bi-LSTM CRFs[8]	86.1	88.0	87.0
BERT joint(제안 모델)	**89.5**	**88.3**	**88.9**

5. 결론

본 연구에서는 BERT 모델을 활용하여 공간 개체 정보를 추출한 결과를 확인하고 기존 연구와 비교했다. 대량의 데이터를 학습시켜 얻어낸 BERT 언어 모델과 관계 속성과 결합한 모델로 사용한 결과 기존 연구보다 나은 성능을 얻어낼 수 있었다. 차후 연구에서는 본 연구를 통해 얻어낸 모델을 활용하여 공간 관계 정보를 추출하는 연구에 적용할 것이다.

감사의 글

이 논문은 2019년도 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] Salaberri, H., O. Arregi, B. Zapirain, "IXAGroupEHUSpaceEval: (X-Space) A WordNet-based Approach Towards the Automatic Recognition of Spatial Information Following the ISO-Space Annotation Scheme," in Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 856-861, 2015.

[2] Pustejovsky, J., P. Kordjamshidi, M. F. Moens, A. Levine, S. Dworman, Z. Yocum, "SemEval-2015 Task 8: SpaceEval," in Proceedings of the 9th International Workshop on Semantic Evaluation(SemEval 2015), pp. 884-894, 2015.

[3] Lafferty, John; Mccallum, Andrew; Pereira, Fernando CN, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data,", 2001.

[4] Kordjamshidi, P., M. V. Otterlo, M. F. Moens, "Spatial Role Labeling: Task Definition and Annotation Scheme," in Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), ELRA, pp. 413-420, 2010.

[5] Nichols, E., F. Botros, "SpRL-CWW: Spatial Relation Classification with Independent Multi-class Models," in Proceedings of the 9th International Workshop on Semantic Evaluation, pp. 895-901, 2015.

[6] Bastianelli, E., D. Croce, D. Nardi, R. Basili, "UNITO-HMM-TK: Structured Kernel-based Learning for Spatial Role Labeling,"in Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), 2, pp. 573-579, 2013.

[7] 김보겸, "한국어 텍스트의 공간 정보 자동 추출", 충북대학교 박사학위 논문, 2016.

[8] 민태홍, 이재성, "Bidirectional LSTM-CRF 앙상블을 이용한 공간 개체 추출", 제 29회 한글 및 한국어 정보처리 학술대회 논문집, pp.133-136, 2017.

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

[10] Wang, Alex, et al., "Glue: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint arXiv:1804.07461, 2018.

[11] Rajpurkar, Pranav, et al., "Squad: 100,000+ questions for machine comprehension of text," arXiv preprint arXiv:1606.05250, 2016.

[12] ETRI, "공공 인공지능 오픈 API·DATA 서비스 포털", Last modified June 10, 2019, http://aiopen.etri.re.kr/service_dataset.php

[13] ISO 24617-7:2014, language resource management - part 7: Spatial information (ISOspace).

[14] Xue, Kui, et al., "Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text," arXiv preprint arXiv:1908.07721, 2019.