Bidirectional LSTM-CRF 앙상블을 이용한 공간 개체 추출

민태홍○, 이재성
충북대학교, 충북대학교
mintaehong@cbnu.ac.kr, jasonlee@cbnu.ac.kr

Spatial Entities Extraction using Bidirectional LSTM-CRF Ensemble

Tae Hong Min○, Jae Sung Lee
Chungbuk National University

요 약

공간 정보 추출은 대량의 텍스트 문서에서 자연어로 표현된 공간 관련 개체 및 관계를 추출하는 것으로 질의응답 시스템, 챗봇 시스템, 네비게이션 시스템 등에서 활용될 수 있다. 본 연구는 한국어에 나타나 있는 공간 개체들을 효과적으로 추출하기 위한 앙상블 기법이 적용된 Bidirectional LSTM-CRF 모델을 소개한다. 한국어 공간 정보 말뭉치를 이용하여 실험한 결과, 기존 모델보다 매크로 평균이 향상되어 전반적인 공간 관계 추출에 유용할 것으로 기대한다.

주제어: 공간 정보, 정보 추출, 딥러닝, 앙상블

1. 서론

대량의 텍스트가 컴퓨터 시스템에 기록되고 있고, 특히 웹이나 모바일 시스템을 통해 수집되는 텍스트의 양이 급속히 증가함에 따라, 이를 분석하거나 유의미한 정보를 추출하는 기술 또한 발전해 왔다. 공간 정보 추출은 정보 추출의 한 종류로 공간 개체와 그들 사이의 관계를 연결시켜주는 공간 관계를 추출하는 기술이다. 이는 질의응답 시스템, 챗봇 시스템, 네비게이션 시스템 등 공간 정보 추출과 공간 추론을 해야 하는 시스템에 활용될 수 있다.

공간 정보 표기법은 국제 표준인 ISO-Space[1]로 제정되었다. 이 표준에 따르면 공간 정보는 공간 정보 개체와 관계로 이루어져 있으며, 개체는 Place(장소), Path(경로), Spatial Entity(공간 안에 존재하는 개체), Spatial Signal(정적인 관계 어휘), Motion(동적인 관계 어휘), Motion Signal(이동을 설명하는 어휘), Measure(개체 측정치)로 총 7개 이며, 관계를 표현하는 링크는 Qualitative Spatial Link(개체간의 상대적인 위치), Orientation Link(개체간의 위상 정보), Movement Link(개체의 움직임 혹은 상태), Measurement Link(개체 측정치의 관계) 총 4개의 링크로 이루어져 있다. 국제적으로 이 표준에 기반을 두어 공간 정보를 추출하는 연구들이 진행되어 왔다[2,3]. 국내의 공간 정보 연구는 영어권에서 연구된 내용을 한국어의 특성에 맞게 변형을 하고 보완한 연구로 진행이 되었다[4].

본 논문에서는 공간 정보 개체 추출의 성능 향상을 위하여 앙상블 기법이 적용된 bidirectional LSTM-CRF을 제안한다. 또 이 모델을 다양한 단어 임베딩에 적용하여 평가하고 그 결과를 제시한다.

2. 모델 소개

bidirectional LSTM-CRF 모델은 개체명 인식(Named Entity Recognition) 등의 연구에서 매우 좋은 성능을 보인다[5]. 공간 정보 추출도 sequence labeling 문제로 볼 수 있으므로 본 논문에서는 해당 모델을 사용하였다.

2.1 Bidirectional LSTM-CRF

일반적으로 사용되는 bidirectional LSTM-CRF의 모델을 공간 정보 추출에 적용한 구조는 그림 1과 같다[5]. 그림에 나타난 bidirectional LSTM-CRF는 각각 단어 표상을 입력으로 받는다. 이를 오른쪽 방향으로 분석하는 R층, 왼쪽 방향으로 분석하는 L층 총 2개의 LSTM Layer을 통해 나온 데이터를 C로서 합친다. 이 C를 CRF(Conditional Random Field)를 이용하여 각각 태그의 점수를 벡터로 계산하고, 이를 최대로 하는 태그를 선정한다.

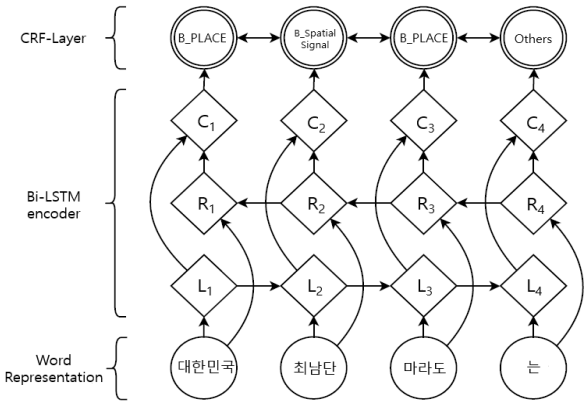

그림 1. bidirectional LSTM-CRF 모델

2.2 공간 정보 추출을 위한 단어 표상 확장

단어 표상(word representation)이란 각각 단어에 관하여 정보를 표기하는 방법이다. 본 연구는 워드 임베딩 벡터(100), 형태소 정보(46), 개체명 인식(35), 기분석 사전(15) 총 196차원의 벡터로 단어를 표현하였다.

(1) 워드 임베딩 벡터(word embedding vector)

워드 임베딩 벡터는 단어 의미 자체를 특정 차원인

벡터로 표현하는 것을 의미한다. 본 실험은 Word2Vec의 CBOW(Continuous Bag of Words) 및 Skip-gram[6]과 Stanford의 GloVe[7], Facebook의 fastText[8] 총 4가지 워드 임베딩 벡터 모델을 이용하여 100차원의 워드 벡터를 만들어 비교 실험하였다.

(2) 형태소 정보

형태소 정보는 [9]에서 규정한 45개와 문장 시작, 띄어쓰기, 문장 끝(<SOS>, <BLN>, <EOS>)를 1개의 태그로 총 46개를 one-hot 벡터로 표현하였다.

(3) 개체명 인식(Named Entity Recognition)

개체명 인식의 정보는 [10]에서 규정한 대분류들을 사용하였으며, LC, AF, QT는 소분류 까지 사용하여 총 35개의 태그를 one-hot 벡터로 표현하였다.

(4) 기분석 사전

기분석 사전이란 학습데이터에서 동일한 태그로 주석된 어휘에 대하여, 해당 정보를 주는 것이다. 본 연구에서는 동일한 태그로 3번 이상 주석 된 어휘를 대상으로 공간 정보 개체의 개수인 15차원 one-hot 벡터로 표현하였다.

2.3 앙상블 모델

신경망 모델에서 성능을 향상시키기 위한 연구 중의 하나로 앙상블 알고리즘을 적용한 연구들이 있다[11]. 기본적인 앙상블 모델의 구조는 그림 2와 같다.

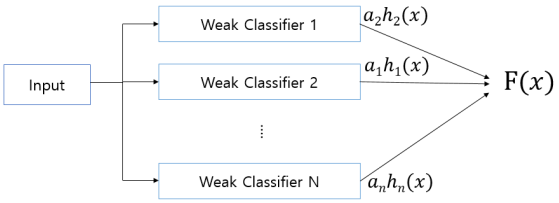

그림 2. 앙상블 모델의 구조

앙상블 기법은 단일 모델을 여러 개 학습시킨 후 그 모델들의 결과 $h(x)$에 가중치 a를 곱한 값들을 더하여 결과 $F(x)$를 도출한다. 이를 표현하는 수식은 다음과 같다.

$$F(x) = \sum_{i=0}^{n} (a_i h_i(x))$$

동일한 모델이라도 초기의 신경망 가중치를 랜덤으로 지정하는 점, dropout 기법 등 랜덤 수치가 적용되는 점이 있기에 학습되는 결과가 조금씩 다르다. 이를 통해 동일한 모델을 여러 개 (현 실험에서는 5개) 학습시킨 후 이들 값을 입력받아, 최종적인 분류를 하는 모델을 만들어 성능을 향상시켰다. 앙상블 기법을 적용시킨 bidirectional LSTM-CRF모델은 그림 3과 같다.

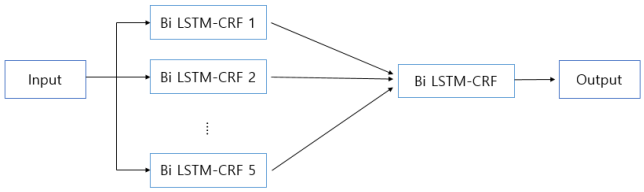

그림 3. 앙상블 기법을 적용시킨 최종 모델

3. 실험 및 평가

본 실험은 한국어 공간 정보 주석 말뭉치 v1.0을 이용하였으며, 각각의 개체 개수는 아래 표 1과 같다[12].

표 1. 한국어 공간 정보 주석 말뭉치 개체 개수

개체	개수	개체	개수
문장	1654	Spatial Entity	390
Place	5049	Spatial Signal	1171
Path	275	Motion	261
Measure	235	Motion Signal	245

실험은 5-fold test로 진행하였으며, 한국어 앙상블 기법을 적용한 모델들과 적용하지 않은 단일 모델들로 나누고, 각 모델에 대해 4가지의 워드 임베딩 모델을 적용하여 실험하였다. 그 결과는 표 2와 같다.

표 2. 공간 정보 개체 추출 성능 결과(%)

모델	워드 임베딩	정확률	재현율	F1
단일	CBOW	79.9	85.1	82.4
	Skip-gram	79.4	85.2	82.2
	GloVe	79.2	85.0	82.0
	fastText	81.7	86.7	84.1
앙상블	CBOW	81.3	87.3	84.1
	Skip-gram	81.3	86.7	83.9
	GloVe	81.2	88.2	84.5
	fastText	**82.8**	**88.4**	**85.5**

표 2에서 보듯이 fastText를 워드 벡터로 사용하였을 경우 성능이 단일 모델이나 앙상블 모델에서 가장 좋게 나왔으며, 앙상블 기법을 적용하였을 때 성능이 더 우수하다.

한국어 공간 정보 개체 추출 시스템으로서 기존 연구 중 가장 성능이 좋은 것은 CRF 모델을 이용한 것이다[4]. 이 모델에서는 형태소 원형, 형태소 품사, 어절 띄어쓰기 정보, 형태소의 의미 부류, 개체명 인식 정보, 의존 구문 레이블, 의존 구문 head 레이블, 의존 구문 head 레이블의 형태소, 워드 클러스터 정보, 의존 구문 head의 워드 클러스터 정보, 총 10가지 자질을 각각 개체마다 조금씩 다르게 적용하여 학습을 진행하였다. 그리고 성능을 높이기 위해 앙상블 모델을 사용하여 각각

결과를 합친 후, 유효한 태그를 분별하기 위해 태그 벡터를 사용하였다. 기존의 CRF 기반 공간정보 추출 모델과 본 연구에서 최고 성능을 보인 모델의 성능 비교는 표 3와 같다.

표 3. 기존 모델[4]과 제안 모델 비교(%)

개체	정확률		재현율		F1	
	기존 모델	제안 모델	기존 모델	제안 모델	기존 모델	제안 모델
Place	96.1	90.0	95.8	91.0	96.0	90.5
Path	55.2	61.3	53.9	75.2	54.5	67.2
S.Entity	32.6	42.5	44.4	68.8	37.6	52.3
Motion	54.4	59.4	70.9	74.1	61.6	65.8
M.Signal	55.6	49.2	69.4	74.2	61.7	58.2
S.Signal	89.2	80.1	83.6	85.0	86.3	82.4
Measure	95.1	90.4	90.6	95.0	92.8	92.5
마이크로 평균	**86.1**	**82.8**	**88.0**	**88.4**	**87.0**	**85.5**
매크로 평균	**68.3**	**67.6**	**72.7**	**80.5**	**70.1**	**72.7**

표 3을 보면 Path, Spatial Entity, Motion의 추출 성능이 기존 모델보다 오른 것을 볼 수 있다. 특히 Path와 Spatial Entity는 각각 12.7%포인트, 14.7%포인트만큼 큰 폭으로 추출 성능이 향상되었다. Spatial Entity는 이동하는 개체 혹은 이동의 가능성이 있는 개체에 대한 태그이기에 태그들 간의 관계를 형성하는 경우에만 추출한다. 그렇기에 Spatial Entity의 성능이 오른 점은 차후 공간 정보 관계 추출에 유용할 것으로 기대한다.

전체적인 성능으로는 제안 모델의 F1값은 마이크로 평균은 85.5%이며, 매크로 평균은 72.7%이다. 기존 모델과 비교해 볼 때, 마이크로 평균은 1.5%포인트 하락하였으며, 매크로 평균은 2.6%포인트 상승하였다. 비록 마이크로 평균은 하락하였지만, 매크로 평균이 상승하여, 학습 데이터의 데이터 편중의 문제를 완화하였음을 보여준다.(한국어 공간 정보 말뭉치 특성상 Place 태그의 빈도가 다른 태그들의 빈도보다 높아 데이터가 편중되어 있다. 그 결과 Place 태그 추출의 성능은 좋게 나오는 반면 다른 태그 추출 성능은 낮은 것을 볼 수 있다. 특히 Place, Path, Spatial Entity는 주로 명사이기에 이를 구분하는 것이 어렵다는 점이 [4]의 성능 저하의 원인 중 하나다.) 본 연구에서 Path, Spatial Entity, Motion의 성능이 전반적으로 상승하여 데이터 편중 문제를 어느 정도 보완한 것으로 판단할 수 있다.

4. 결론 및 향후 연구계획

본 연구는 앙상블 기법이 적용된 bidirectional LSTM-CRF을 사용하여 공간정보를 추출하는 방법을 제안하였다. 다양한 단어 임베딩을 사용하여 실험한 결과, fastText 임베딩을 이용하고, 앙상블 기법을 사용한 모델이 가장 우수하였다. 기존 CRF 모델의 성능과 비교하여 보면, 마이크로 평균은 하락하였지만, 매크로 평균이 상승하여, 학습데이터의 데이터 편중의 문제를 완화하였다고 볼 수 있다.

향후 연구로는 보다 다양한 파라미터를 사용하여 앙상블 모델을 확장하고 이를 딥러닝 기반의 공간 관계 추출 시스템과 통합하여 전체 공간 정보 추출의 성능을 향상시킬 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [2013-0-00131,(엑소브레인-1세부) 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

참고문헌

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

[2] Pustejovsky, James, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum, “SemEval-2015 task 8: SpaceEval,” SemEval 2015, 2015.

[3] Kolomiyets, Oleksandr, Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens, “SemEval-2013 task 3: spatial role labeling,” SemEval 2013, 2013.

[4] Kim, Bogyum and Jae Sung Lee. “Extracting spatial entities and relations in Korean text,” COLING. 2016.

[5] Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. “Neural architectures for named entity recognition,” arXiv preprint arXiv:1603.01360, 2016.

[6] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[7] “GloVe: Global Vectors for Word Representation,” https://nlp.stanford.edu/projects/glove

[8] Bojanowski, Piotr, Edouard Grave, and Armand Joulin, Tomas Mikolov, “fastText,” https://research.fb.com/fasttext/

[9] 국립국어원, 21세기 세종계획 최종성과물(2011년 12월 수정판), 2011.

[10] TTAK0.KO-10.0852:2015, 개체명 태그세트 및 태깅 말뭉치.

[11] 권순재, 허윤석, 이건철, 임지수, 최호정, 서정연, “가중 투표 기반의 앙상블 기법을 이용한 한국어 개체명 인식기”, 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp.333-336, 2016.

[12] 충북대학교 언어지식공학 연구실, “한국어 공간 정보 주석 가이드라인”, 2016.