2019 한국컴퓨터종합학술대회
06.26. ~ 28.

Bi-LSTM/CRF 모델을 이용한 한국어 구묶음

남궁영†○, 김창현, 천민아†, 박호민†, 윤호†, 최민석†, 김재균†, 김재훈†*

Korean Chunking using Bi-LSTM/CRF

E-mail: young_ng@kmou.ac.kr

국립한국해양대학교
KOREA MARITIME AND OCEAN UNIVERSITY

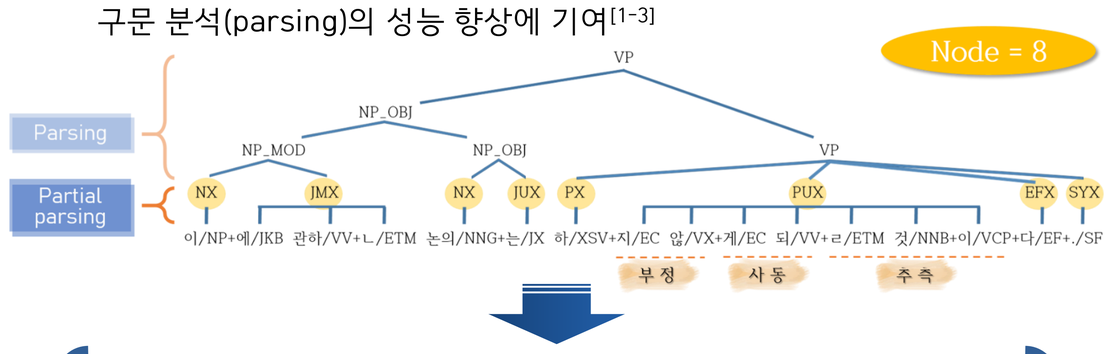
[감사의 글]
이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)과 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068187, 한국어 정보처리 원천 기술 연구 개발)

Ⅰ. 초 록

자연언어 처리에서 구묶음(Chunking)은 구문 분석 이전에 수행되는 전처리 단계로, 문장 내에서 단일한 기능을 수행하는 형태소들을 하나의 말덩이(Chunk)로 묶어 구문 분석의 성능 향상에 기여하는 역할을 한다. 본 논문에서는 한국어 문장 내의 모든 구성 성분에 대한 구묶음을 수행하기 위해 기구축된 구묶음 말뭉치와 sequence labeling에 우수한 성능을 보이고 있는 Bi-LSTM/CRF 모델을 이용한다. 또한, 말덩이 경계의 인식 정도 및 표지 부착의 적합성 여부에 따라 다양하게 평가를 진행하였다. 시스템의 성능 평가 결과 말덩이의 경계와 표지를 모두 정확히 예측한 경우 F1 점수가 97.02%로 측정되었다.

Ⅱ. 연구배경 및 목적

❖ 한국어 구묶음
- 문장 내에서 단일한 기능을 수행하는 형태소들을 하나의 말덩이(Chunk)로 묶어 구문 분석(parsing)의 성능 향상에 기여[1-3]



- 심층 학습 모델 중 하나인 Bi-LSTM/CRF 모델을 적용하여 한국어 문장 내의 모든 구성 요소에 대해 구묶음을 수행
- 차후 연구에서 의존 구문 분석에 본 논문의 결과를 이용

Ⅲ. 심층 학습을 이용한 한국어 구묶음

구묶음 (Chunking)

❖ 말덩이 (chunk)[3-4]
- 구묶음의 기본 단위
- 선행 연구에서 정의한 사항을 활용하여 본 연구 수행 (표 1)

표 1. 말덩이의 종류 및 표지

내용어 말덩이	체언구(NX), 본용언구(PX), 지정사구(CX), 부사구(AX), 관형사구(MX), 독립어구(IX)
기능어 말덩이	보조용언구(PUX), 격조사구(JKX), 관형격조사구(JMX), 보조사구(JUX), 접속조사구(JCK), 선어말어미구(EPX), 연결어미구(ECX), 전성어미구(ETX), 종결어미구(EFX), 호격조사구(JVX), 문장부호구(SYX), 분석불능구(NAX)

❖ 구묶음 (chunking)
- 말덩이 인식 및 표지 부착
- 말덩이의 정의 및 그 표지를 기준으로 구묶음에 대한 sequence labeling 수행

Sequence labeling

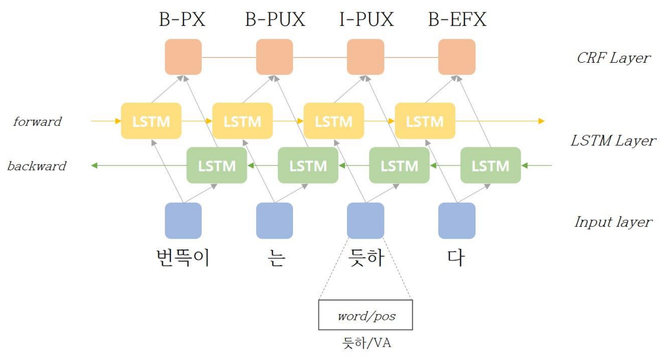


그림 1. Bi-LSTM/CRF 모델

❖ Bi-LSTM/CRF[5]
- 기존의 순환신경망(RNN)에서 장기 의존성을 보완하기 위해 고안된 장단기 기억 계층(LSTM)을 양방향으로 이용
- 출력계층에서 CRF를 통해 가장 적합한 말덩이 표지 선택
- sequence labeling에서 뛰어난 성능을 보임

❖ BI 형식[6] 이용

구묶음 말뭉치 (chunked corpus)

❖ 형태소 분석된 말뭉치로 구묶음을 수행

❖ 모델의 golden answer가 되는 말덩이 표지(chunk tag)가 부착된 말뭉치 (언어 정보 부착 시스템[7]을 활용하여 반자동 형식으로 구축)

❖ ID / 형태소 / 품사 / SP / 말덩이 표지

```
# text = 아마 그런 사람은 없으리라 본다.
1       아마        MAG     1       B-AX
2       그런        MM      1       B-NX
3       사람        NNG     0       I-NX
4       은          JX      1       B-JUX
5       없          VA      0       B-PX
6       으리라      EC      1       B-PUX
7       보          VV      0       I-PUX
8       ㄴ다        EF      0       B-EFX
9       .           SF      0       B-SYX
```

그림 2. 구묶음 말뭉치 예시

Ⅳ. 실험 및 평가

❖ Model parameter
- activation func.: ReLU
- optimizer: RMSprop
- batch size, epoch: 32, 8

표 2. 성능 평가에 사용한 문장 및 형태소 개수

(단위: 개)

	문장 수	형태소 수
학습용 말뭉치	10,490	163,641
평가용 말뭉치	1,312	20,416
검증용 말뭉치	1,311	20,191
전체 말뭉치	13,113	204,248

❖ 평가방법
- MUC[8,9], SemEval[10]에 사용된 방법을 이용하여 표 3의 네 가지 경우에 대해 각각 정밀도, 재현율, F1 점수 측정

표 3. 성능 평가에 이용한 평가 방식

평가 방식	설 명
경계/표지 일치 (strict)	시스템이 예측한 말덩이의 **경계** 및 **표지**가 모두 정답과 일치하는 경우
경계 일치 (exact)	표지의 일치 여부와 관계 없이 말덩이의 **경계**를 잘 인식한 경우
부분 경계 일치 (partial)	표지의 일치 여부와 관계 없이 시스템이 예측한 말덩이의 경계와 정답의 **경계가 일부** 겹치는 경우
표지 일치 (type)	말덩이의 경계 일치 여부와 관계 없이 시스템이 예측한 **표지**가 일치하는 경우

❖ 실험 결과
- 말덩이의 경계와 표지가 모두 일치하는 경우의 F1 점수: 97.02%

표 4. 평가 방식에 따른 실험 결과

	경계/표지	경계	부분경계	표지
정밀도	97.26	97.69	97.69	97.54
재현율	96.78	97.21	97.21	97.07
F1	97.02	97.45	97.45	97.30

Ⅴ. 결론 및 향후 연구

본 논문은 기존의 한국어 구묶음에 심층 학습 기법 중 하나인 Bi-LSTM/CRF 모델을 적용하여 문장 내 모든 구성 성분에 대해 구묶음을 수행하였다. 실험 결과 말덩이의 경계와 해당 말덩이의 표지를 모두 정확히 찾은 경우의 F1 점수는 97.02%였다. 향후 각 말덩이의 종류에 따른 인식 성능을 실험하고, 그 결과에 따라 전체 시스템의 성능을 높이기 위한 방안에 대해 지속적인 연구를 진행할 예정이다.

참 고 문 헌

1. S. Abney, “Parsing by chunks”, Principle-based parsing, eds. Berwick, R. Abney, S. and Tenny, C., Kluwer Academic Publishers, 1991.
2. S. Abney, “Part-of-speech and partial parsing,” Corpus-Based methods in language and Speech Processing, eds. Young, S and Bloothooft, G., Kluwer Academic Publishers, pp. 118-173, 1996.
3. J. Kim, “Partial Parsing”, Korea Information Processing Society Review, vol. 7, no. 6, pp. 83-96, 2000.
4. Y. Namgoong, M. Cheon, H. Park, H. Yoon, M. Choi, and J. Kim, “Defining Chunks for Parsing in Korean”, Proceedings of the 30th Annual Conference on Human and Cognitive Language Technology, pp. 409-412, 2018.
5. Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging”, arXiv preprint arXiv:1508.01991, 2015.
6. L. Ramshaw and M. Marcus, “Text Chunking using Transformation-Based Learning”, arXiv:cmp-lg/9505040, 1995.
7. K. Noh et al., “LiAS: A linguistic information annotation system for linear structures of language based on incremental expansion of dictionary and machine learning”, Journal of the Korean Society of Marine Engineering, vol. 42, no. 7, pp. 580-586, 2018.
8. N. Chinchor and B. Sundheim, “MUC-5 Evaluation Metrics”, Proceedings of the 5th Message Understanding Conference, pp. 69-78, 1993.
9. N. Chinchor and P. Robinson, “Appendix E: MUC-7 Named Entity Task Definition (version 3.5)”, Proceedings of the 7th Message Understanding Conference, https://www.aclweb.org/anthology/M98-1028 (accessed 2019.05.03.) 1998.
10. Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo, “SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)”, Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics and the 7th International Workshop on Semantic Evaluation, pp. 341-350, 2013.