Bi-LSTM/CRF 모델을 이용한 한국어 구묶음

남궁영[†○], 김창현[‡], 천민아[†], 박호민[†], 윤호[†], 최민석[†], 김재균[†], 김재훈[†*]
한국해양대학교[†], 한국전자통신연구원[‡]
young_ng@kmou.ac.kr, chkim@etri.re.kr, minah2018@kmou.ac.kr, homin2006@hanmail.net, 4168615@naver.com, ehdgus5136@naver.com, jgk20000@naver.com, jhoon@kmou.ac.kr

Korean Chunking using Bi-LSTM/CRF

Young Namgoong[†○], Chang-Hyun Kim[‡], Min-Ah Cheon[†], Ho-Min Park[†], Ho Yoon[†], Min-Seok Choi[†], Jae-Kyun Kim[†], Jae-Hoon Kim[†*]
Korea Maritime and Ocean University[†], Electronics and Telecommunications Research Institute[‡]

요 약

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.

1. 서 론

자연언어 처리에 있어 구묶음(Chunking) 또는 부분 구문분석(Partial parsing)은 구문 분석 이전에 수행 되는 전처리 단계이다[1]. 이는 문장 내에서 구문적으로 단일한 역할을 수행하는 형태소들을 하나의 말덩이(Chunk)로 묶어 구문 분석의 입력 성분 수를 줄이고 분석 결과의 모호성을 해소하는 등 구문 분석의 문제들을 완화하는 기능을 한다[2, 3].

구묶음(Chunking)은 형태소 분석, 개체명 인식과 함께 sequence labeling으로 문제를 해결하는 대표적인 분야이다. 이는 크게 말덩이를 인식하는 단계와 이에 해당하는 말덩이 표지를 부착하는 단계로 이루어진다[2]. 최근에는 자연언어 처리에 심층 학습 기법을 적용하면서 sequence labeling을 위한 자질 추출에 드는 노력을 경감하면서도 우수한 결과를 보이고 있다.

본 논문에서는 sequence labeling에 심층 학습 모델 중 하나인 Bi-LSTM/CRF 모델을 적용하여 한국어 문장 내의 모든 구성 요소에 대해 구묶음을 수행한다. 2장에서는 한국어 구묶음에 관한 기존의 연구들을 소개하고 sequence labeling에 사용되는 대표적인 모델을 설명한다. 3장에서는 학습 및 평가에 사용된 구묶음 말뭉치에 대해 설명하고, 4장에서 실험 및 평가, 5장에서는 결론 및 향후 연구 방향에 대해 기술한다.

2. 관련 연구

2.1 구묶음(Chunking)

한국어 구묶음에 대한 연구는 그 중요성에 비해 많은 연구가 활발히 이루어지지는 않았지만, 2000년대 초반까지 꾸준히 진행되어 왔다. [3]에서는 한국어 처리에 있어 부분 구문분석의 역할 및 필요성을 제시하고 이에 대한 방법론 및 다양한 응용 분야에 대해 기술하였다. [4]에서는 기본구를 인식하기 위해 기계학습을 적용하였으며, [5]에서는 규칙 기반의 한국어 부분 구문분석기를 구현하고 이를 포함한 구문분석과 기존의 방식을 비교하였다. [6]에서는 기존에 명사열 위주로 수행되었던 구묶음을 비롯하여 보조 용언과 의존 명사에 대한 구묶음 방식을 제안하고, 이를 활용한 구문 분석에서 의존 관계를 보다 정확하게 추출할 수 있음을 보였다. 최근에 연구된 [7]에서는 구문 분석을 위한 전처리 단계로서 한국어 문장의 모든 구성 성분에 대해 구묶음을 수행하기 위한 말덩이의 기준 및 그 표지를 제시하였다.

본 논문에서는 [7]에서 연구된 말덩이의 정의 및 그 표지를 보완 및 활용하여 형태소 분석된 문장에 대해 말덩이 표지를 부여하고, 이렇게 구축된 말뭉치를 이용하여 구묶음에 대한 sequence labeling을 수행한다. 본 논문에서 사용한 말덩이 표지는 표 1과 같다.

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)

2.2 Sequence Labeling

본 논문에서는 말덩이 인식을 위해 sequence labeling에서 가장 뛰어난 성능을 보이고 있는 Bi-LSTM/CRF 모델[8]을 이용한다. 이는 그림 1에서와 같이 기존의 순환신경망(Recurrent Neural Network)에서 장기 의존성을 보

완하기 위해 고안된 장단기 기억(LSTM) 계층을 양방향으로 이용하고, 출력 계층에서 Conditional Random Field(CRF)를 통해 가장 적합한 말덩이 표지를 선택하는 방식으로 sequence labeling에 최근 전형적으로 이용되는 모델이다.

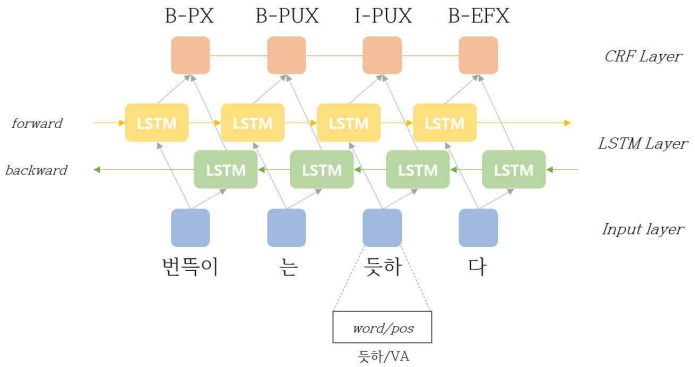

그림 1. Bi-LSTM/CRF 모델

인식한 말덩이의 표지를 표기하기 위한 방법으로는 그림 1의 최종 출력에서 표기한 바와 같이 sequence labeling에 보편적으로 사용되는 IOB 형식[9]을 이용한다. 이 때 구묶음 작업에 있어서 문장 내에 있는 형태소는 모두 임의의 말덩이로 구묶음 되므로, 하나의 형태소는 반드시 하나의 말덩이 표지를 가지게 된다[7]. 따라서 어떤 형태소든 말덩이에 속하지 않는 예외가 없으므로 흔히 사용되는 IOB 형식 중 'O' 태그는 사용하지 않는다.

3. 구묶음 말뭉치 (Chunked Corpus)

본 연구는 세종 형태 분석 말뭉치[10]에 [7]에서 정의한 기준 및 그 표지를 중심으로 구묶음을 수행한 말뭉치를 사용한다. 따라서 입력이 되는 문장에 부착되어 있는 품사의 종류는 총 45개이며, 구묶음을 통해 부착되는 말덩이 표지의 개수는 [7]의 내용을 바탕으로 좀 더 보완하여 총 18개의 범주를 적용하였다. 이때 세종 품사 태그의 추정 및 분석불능범주(NF, NV, NA)는 'NAX'라는 표지를 이용하여 구묶음 하였다.

말덩이 표지가 부착된 구묶음 말뭉치 구축은 [11]의 언어 정보 부착 시스템을 활용하여 반자동 형식으로 수행하였으며, 이를 그림 2에서와 같이 CoNLL 형식으로 정리하여 시스템의 학습 및 평가에 이용한다. 이 때, 'text'는 원문이며, 각 열은 왼쪽부터 차례대로 순번, 형태소, 품사, 띄어쓰기 여부, 말덩이 표지를 나타낸다.

```
# 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

4. 실험 및 평가

구묶음은 형태소 분석된 문장에 대해 수행되며, 같은 형태소라도 품사에 따라 구묶음 결과가 달라지는 등 품사 태그에 영향을 많이 받게 된다. 따라서 모델의 입력은 2.2장의 그림 1과 같이 각 형태소와 그에 상응하는 품사를 결합한 것을 한 단위로 임베딩하여 사용한다.

실험에 사용한 모델의 활성 함수는 ReLU를 이용하고 학습기로는 RMSprop을 이용하였으며, 모델의 각종 hyper-parameter들은 실험적으로 조절해가며 평가에 사용하였다.

실험에 사용된 말뭉치는 3장에 기술한 구묶음 말뭉치이며, 이 중 13,113개를 실험에 사용하였다. 학습에 사용된 문장은 10,490개이고 평가에 사용된 문장은 1,312개, 검증에 사용된 문장은 1,311개이다. 실험에 사용한 문장 및 형태소 개수는 표 2와 같다.

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

평가 방법은 구묶음과 마찬가지로 개체의 경계를 찾고 해당 개체에 부착된 표지의 적합성을 판별하는 개체명 인식 시스템 평가에 이용되는 지표를 사용하였다. 개체명 인식 시스템을 평가하는 방법에는 대표적으로 MUC[12, 13]에 사용된 방법과 이를 기반으로 평가 방식에 따라 세분화 하여 측정한 SemEval[14]에 사용된 방법이 있다. 본 논문에서는 구묶음 시스템의 평가를 위해 [12-14]에 사용된 방법을 이용하여 표 3에 설명한 4가지 경우에 대해 정밀도, 재현율, F1 점수를 각각 측정하였다.

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

이를 바탕으로 각 경우에 따른 구묶음 시스템의 성능을 측정한 결과 말덩이의 경계와 표지가 모두 일치하는 경우의 F1 점수는 97.02%였으며, 전체 평가 방식에 대한 실험 결과는 표 4와 같다.

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

5. 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.

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)

참고문헌

[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. Hwang, H. Chung, S. Park, Y. Kwak, and H. Rim, "Improving the Performance of Korean Text Chunking by Machine Learning Approaches based on Feature Set Selection", Journal of KISS : Software and Applications, vol. 29, no. 9/10, pp. 654-668, 2002.

[5] K. Lee, J. Kim, “Implementing Korean Partial Parser based on Rules”, Korea Information Processing Society Transactions: Part B, vol. 10, no. 4, pp. 389-396, 2003.

[6] E. Park and D. Ra, "Processing Dependent Nouns Based on Chunking for Korean Syntactic Analysis", Korean Journal of Cognitive Science, vol. 17, no. 2, pp. 119-138, 2006.

[7] 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.

[8] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging”, arXiv preprint arXiv:1508.01991, 2015.

[9] L. Ramshaw and M. Marcus, “Text Chunking using Transformation-Based Learning”, arXiv:cmp-lg/9505040, 1995.

[10] The National Institute of the Korean Language, 21th Century Sejong Project Final Result, 2011.12 Revised Edition, 2011.

[11] K. Noh, C. Kim, M. Choi, H. Yoon, and J. Kim, “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.

[12] N. Chinchor and B. Sundheim, “MUC-5 Evaluation Metrics”, Proceedings of the 5th Message Understanding Conference, pp. 69-78, 1993.

[13] 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.

[14] 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.