BIT 표기법을 활용한 한국어 개체명 인식

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

Korean Named Entity Recognition Using BIT Representation

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

요 약

개체명 인식이란 주어진 문서에서 개체명의 범위를 찾고 개체명을 분류하는 것이다. 최근 많은 연구는 신경망 모델을 이용하며 하나 이상의 단어로 구성된 개체명을 BIO 표기법으로 표현한다. BIO 표기법은 개체명이 시작되는 단어의 표지에 B(Beginning)-를 붙이고, 개체명에 포함된 그 외의 단어의 표지에는 I(Inside)-를 붙이며, 개체명과 개체명 사이의 모든 단어의 표지를 O로 간주하는 방법이다. BIO 표기법으로 표현된 말뭉치는 O 표지가 90% 이상을 차지하므로 O 표지에 대한 혼잡도가 높아지는 문제와 불균형 학습 문제가 발생된다. 본 논문에서는 BIO 표기법 대신에 BIT 표기법을 제안한다. BIT 표기법이란 BIO 표기법에서 O 표지를 T(Tag) 표지로 변환하는 방법이며 본 논문에서 T 표지는 품사 표지를 나타낸다. 실험을 통해서 BIT 표기법이 거의 모든 경우에 성능이 향상됨을 확인할 수 있었다.

주제어: BIT 표기법, 개체명 인식, 자소 표상, Bi-LSTM/CRF

1. 서론

개체명이란 문서에서 나타나는 인명, 지명, 조직명, 시간, 날짜, 화폐 등 고유한 의미를 가지는 단어를 말한다. 개체명은 자연언어 처리의 응용 분야인 정보 검색에서 주요 검색대상이 되며 정보 추출에서는 추출하는 정보를 구성하는 요소가 된다[1]. 개체명 인식이란 주어진 문서에서 개체명의 범위를 찾고 개체명의 범주를 결정하는 것이며 문서 요약, 질의응답, 기계 번역, 챗봇과 같은 자연언어처리에 두루 응용된다. 개체명 인식은 단어로 구성된 개체명이 문맥에 따라 다른 개체명으로 해석될 수 있는 중의성 문제와 시간의 흐름에 따라 새롭게 생성되는 미등록어(out-of-vocabulary) 문제 등을 가지고 있다[2]. 중의성 문제는 기계학습 방법으로 문맥 패턴을 파악하여 어느 정도 해결할 수 있다[3]. 미등록어 문제는 말뭉치 확장과 단어 표상 확장 방법으로 해결하는 방법이 제안되었다[4,5].

기존의 한국어 개체명 인식 연구는 BIO 표기법[6]을 통해서 개체명의 표지를 결정하는 순차 표지 부착(sequence labeling) 방식으로 접근하여 어느 정도의 성능을 보이고 있다[3]. BIO 표기법은 개체명이 시작되는 단어의 표지에 B(Beginning)-를 붙이고, 개체명에 포함된 그 외의 단어의 표지에는 I(Inside)-를 붙이고, 개체명 사이의 모든 단어의 표지를 O(Outside)로 간주하는 방법이다. 개체명 인식 방법으로 제안된 거의 모든 연구에서 이 방법을 사용하고 있다. 최근에는 양방향 순환신경망(Bi-directional Recurrent Neural Networks, Bi-RNNs)과 조건부 랜덤 필드(Conditional Random Fields, CRFs)를 결합한 방식으로 많은 성능의 개선이 있었다[7]. 그러나 BIO 표기법을 사용할 경우, O 표지가 전체 말뭉치의 90% 이상을 차지하고 있어서 다른 표지에 비해 혼잡도(perplexity)가 매우 높다는 문제와 불균형 학습(imbalance learning) 문제[8]를 가지고 있다.

이러한 문제를 해결하기 위해서 본 논문에서는 기존의 BIO 표기법 대신에 BIT 표기법을 제안한다. BIT 표기법이란 BIO 표기법에서 O 표지를 T(Tag) 표지로 변환하는 방법이며 본 논문에서 T 표지는 품사 표지를 나타낸다. 실험을 통해서 BIT 모델이 BIO 모델보다 우수한 성능을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서는 BIT 표기법을 기술한다. 4장에서는 실험을 통해서 두 표기법의 성능을 비교하고 분석한다. 마지막으로 5장에서 본 논문의 결론 및 향후 연구 방향을 기술한다.

2. 관련 연구

이 장에서는 심층학습을 이용한 한국어 개체명 인식에 대해서 간단히 기술하고 실험에서 사용되는 다양한 입력 표상의 생성 방법에 대해서 간단히 소개한다.

2.1 심층학습을 이용한 한국어 개체명 인식

기존의 한국어 개체명 인식은 규칙 기반의 모델이나 기계학습 모델[9]이 주로 사용되었다. 최근에는 개체명 인식의 많은 연구가 심층학습 모델을 이용하고 있다. 심층학습 모델 중 순환신경망을 사용하는 방법으로는 장단기 기억(LSTM) 모델에 조건부 랜덤 필드(CRFs) 층을 쌓은 모델을 이용하여 성능 향상을 보였다[10]. 또한 순환신경망과 합성곱신경망(Convolution Neural Network, CNN)을 결합한 방법[11]도 제안되었으며, 순환 신경망의 문제점 중 하나인 장기 의존성 문제를 보완하기 위해 주의 집중 방법(attention mechanism)을 사용하는 방법도

제시되었다[12].

뿐만 아니라 모델의 입력에 해당하는 단어 표상을 확장하여 개체명 인식 모델의 성능을 향상하는 방법도 연구되었다. [13]에서는 순환 신경망의 입력에 해당하는 단어 표상에 사전 학습된 단어 표상, 사전 학습된 품사 표상, 단어를 이루는 음절 단위 표상을 차례로 결합하여 성능 향상을 보였다. 최근에는 말뭉치로부터 접사 자질을 추론하여 단어와 음절 단위 정보에 결합하여 단어 표상을 확장하는 방법도 제안되었다[14].

2.2 입력 표상 생성 방법들

모델의 입력이 되는 단어를 표현하는 방법은 꾸준히 연구되어 왔으며, 최근에는 신경망을 통해 입력 단어를 표현하는 다양한 방법이 제시되었다. Word2Vec[15]은 신경망을 이용하여 주변 단어와 중심 단어 쌍을 통해 입력 단어를 표현하는 방법이다. Glove[16] 학습 말뭉치에서 동시에 등장한 단어의 빈도수를 말뭉치 전체의 단어 수로 나눈 동시 등장 확률을 통해 단어를 표현하는 방법이다. fastText[17]의 경우 주변 단어뿐만 아니라 단어의 부분 정보(subword)를 이용해서 입력 단어를 표현하는 방식으로, 기존의 단어 표상 방법보다 미등록어 문제에 좀 더 유연한 결과를 보였다.

이후 컴퓨터 계산 능력의 향상과 함께 문맥 전체를 고려한 단어 표현 방법도 제안되었다. [18]에서는 문맥에 따라 같은 단어의 표현 방법을 달리 하기 위해 양방향 언어 모델(bidirectional language model)을 사용한 ELMo를 제시하였다. [19]에서는 문장 안에 있는 단어 중 일부를 무작위로 가려 이를 예측하는 모델인 MLM(Masked Language Model)을 이용한 BERT를 제시하였다.

3. BIT 표기법을 활용한 개체명 인식

BIO 표기법은 기존의 IOB 형식(Inside, Outside, beginning)[6]에서 출발하여 현재는 BIO 표기법으로 통용되고 있다. BIO 표기법은 개체명이 시작되는 단어의 표지에 B(Beginning)-를 붙이고, 개체명에 포함된 그 외의 단어의 표지에는 I(Inside)-를 붙이고, 개체명과 개체명 사이의 다른 모든 단어의 표지를 O(Outside)로 간주하는 방법이다. BIO 표기법은 개체명 인식뿐 아니라 구묶음과 같은 순차 표지 부착에 두루 사용된다. 그러나 표 1[1]에서 보는 바와 같이 O 표지가 전체 말뭉치에서 차지하는 비율이 92.7%에 달한다. 이처럼 O 표지가 차지하는 비중이 너무 높아서 혼잡도가 높아질 뿐 아니라 불균형학습 문제[8]가 발생된다.

표 1. 표지별 빈도수와 전체에서 차지하는 비율

개체명 표지	빈도수	전체 말뭉치에 대한 표지 비율
O	152,223	92.70%
B_PS	3,680	2.24%
I_DT	2,043	1.24%
B_LC	1,951	1.18%
B_DT	1,561	0.95%
B_OG	1,518	0.92%
I_PS	389	0.23%
B_TI	297	0.18%
I_OG	291	0.17%
I_TI	174	0.10%
I_LC	81	0.04%
계	164,208	100.00%

이러한 문제를 다소 완화시키기 위해서 본 논문에서는 BIO 표기법 대신 BIT 표기법을 사용할 것을 제안한다. BIT 표기법은 BIO 표기법에서 B(Beginning) 표지, I(Inside) 표지는 그대로 사용하지만, O(Outside) 표지를 T(Tag) 표지로 바꾼 표기법이며, 본 논문에서 T 표지로는 품사 표지를 사용한다. 즉, 한국어 개체명 인식의 입력 단위는 형태소이므로 형태소의 개체명 인식 결과를 O 표지 대신 형태소의 품사 표지가 출력되도록 한다. 그림 1은 BIO 표기법과 BIT 표기법을 비교하여 보이고 있다. 그림 2는 한국어 개체명 말뭉치를 BIT 표기법으로 표기했을 때 각 표지별 분포를 그래프로 표현한 것이며, 표 1은 BIO 표기법을 이용했을 때 각 표지별 빈도수 및 전체에서 차지하는 비율을 나타낸 것이다. 그림2와 표 1에서 볼 수 있듯이 BIT 표기법은 BIO 표기법처럼 어느 한 표지에만 집중적으로 분포되지 않고 표지별로 골고루 분포됨을 알 수 있다.

입력 형태소	부산	에서	열리	ㄴ다
BIO 표기법	B_LC	O	O	O
BIT 표기법	B_LC	JKB	VV	EF

그림 1. BIO 표기법과 BIT 표기법 비교

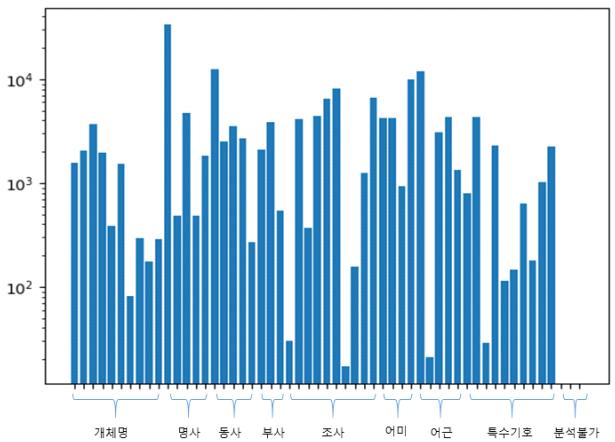

그림 2. BIT 표기법 범주별 빈도수 분포(Log Scale)

1) 2016년 국어 정보 처리 시스템 경진 대회에서 배포한 개체명 인식 말뭉치의 개체명 표지를 사용한다.

4. 실험

실험을 위해서 개체명 인식에 가장 널리 사용되는 Bi-LSTM/CRF 모델을 구현하였으며, 그림 3은 본 논문에서 구현한 Bi-LSTM/CRF 모델의 구성도이다. 이 장에서는 구현된 개체명 인식 시스템의 입력 자질에 대해서 간단하게 기술하고 실험 환경 및 결과를 구체적으로 기술한다.

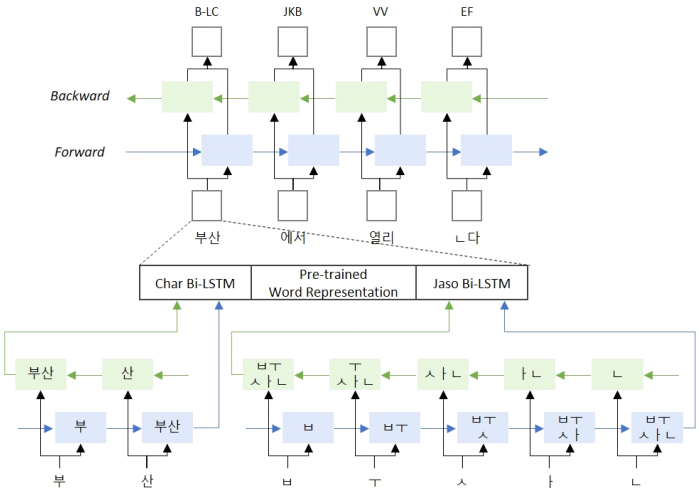

그림 3. 확장된 단어 표상에 대한 Bi-LSTM/CRF 모델

4.1. 실험을 위한 자질 종류

기존의 연구에서는 단어 표상에 음절 정보를 추가하여 단어 표상을 확장하였다[20]. 본 논문에서는 음절뿐 아니라 자소 정보를 추가하여 단어 표상을 확장하였다. 그림 4은 모음 'ㅢ'에 대한 개체명 분포도이며 OG와 PS 표지에서만 높은 분포를 가진다. 따라서 모음 'ㅢ'를 포함하는 개체명은 OG와 PS일 확률이 높다는 사실을 알 수 있다.

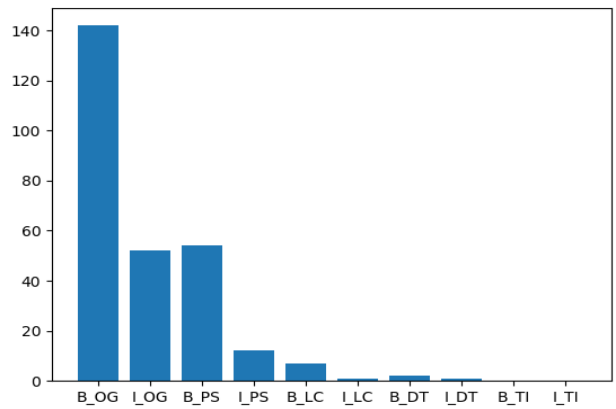

그림 4. 모음 'ㅢ'에 대한 개체명 분포도

그림 4에 알 수 있듯이 본 논문에서는 단어 표상뿐 아니라 음절 자질과 자소 자질을 추가하여 입력 표상을 확장하였다.

4.2 실험 환경

성능 평가를 위해 제안한 모델은 Tensorflow로 구현되었다[2]. 개체명 인식 말뭉치로는 2016년 국어 정보 처리 시스템 경진 대회[3]에서 배포한 개체명 인식 말뭉치를 사용하였다. 말뭉치 정보는 표 2와 같으며 성능 평가는 개체명 표지에 대한 F1-점수를 사용하였다. 단어 표상은 세종말뭉치를 기준으로 학습하였으며 BERT 모델은 ETRI에서 공개한 KorBERT[4]를 사용하였다.

표 2. 학습 말뭉치 문장, 형태소, 개체명 정보

말뭉치	문장수	형태소개수	개체명수
학습말뭉치	3,408	131692	7,233
개발말뭉치	425	16,353	891
평가말뭉치	426	16,163	883

단어 표상별 매개변수는 표 3과 같다. 표 3에서 Word2Vec, GloVe, fastText와 같은 경우 모델마다 기준값이 달라서 min_count를 5, window_size를 10, iteration을 50으로 동일하게 설정하고, 학습을 진행하였다. ELMo의 경우, 한국어의 특성에 맞게 max_characters_per_token을 40으로 변경한 뒤, 학습하였다. KorBERT에 경우, 기존의 모델을 사용하고 BERT층 위에 CRF층을 쌓아서 출력되게 하였다. 모델의 경우 표 4와 같이 매개변수를 두고, Bi-LSTM/CRF, char LSTM + Bi-LSTM/CRF, jaso LSTM + Bi-LSTM/CRF, char LSTM + jaso LSTM + Bi-LSTM/CRF의 순서로 모델을 구성하여 평가하였다.

표 3. 단어 표상별 매개변수

단어 표상	매개변수
Xavier Uniform Initializer[21]	Demension = 300
Word2Vec GloVe fastText	Demension = 300 **min_count = 5** **window_size = 10** **iter = 50**
ELMo	lstm = 4096 char_cnn = 1024 n_highway = 2 **max_characters_per_token = 40**
BERT	num_hidden_layes = 12 num_attention_heads = 12 hidden_size = 768 hidden_act = "gelu"

표 4. 모델별 매개변수

시스템	매개변수
Bi-LSTM/CRF	dropout = 0.5 batch_size = 20 lstm_size = 500
char LSTM, jaso LSTM	output dim = 100 lstm_size = 25

2) https://tensorflow.org
3) https://ithub.korean.go.kr/user/contest/contestIntroView.do
4) http://aiopen.etri.re.kr/service_dataset.php

표 5. 각 모델 및 단어 표상별 성능 비교 (F1-점수)

입력 자질 / 표상 생성 방법	Morph		Morph + Syllable LSTM		Morph + Jaso LSTM		Morph + Syllable LSTM + Jaso LSTM	
	BIO	BIT	BIO	BIT	BIO	BIT	BIO	BIT
Xavier	84.45	**87.40**	85.21	**87.57**	83.44	**84.36**	86.03	**86.45**
Word2Vec	83.36	**84.76**	87.49	**87.79**	87.41	**87.54**	88.15	**88.31**
GloVe	86.68	**87.21**	86.88	**88.34**	87.31	**87.57**	87.11	**88.69**
fastText	86.86	**88.48**	86.37	**88.50**	87.92	**88.58**	88.05	**88.61**
ELMo	88.64	**89.52**	N/A	N/A	N/A	N/A	N/A	N/A
BERT	92.06	**92.24**	N/A	N/A	N/A	N/A	N/A	N/A

4.3 실험 결과

표 5는 BIO 표기법과 BIT 표기법에 대한 한국어 개체명 인식의 성능 평가 결과이다. 모든 입력 자질과 모든 표상 생성 방법에서 BIT 표기법이 좋은 성능을 보였다. 또한 음절 자질뿐 아니라 자소 자질도 한국어 개체명 인식을 위한 유용한 자질임을 확인할 수 있었다. 표상 생성 방법은 Xavier, Word2Vec, fastText, GloVe, ELMo, BERT 순으로 성능이 향상되었다. 그 BERT가 가장 높은 성능을 보였는데 이는 단어 표상 생성의 위한 학습 말뭉치의 크기도 큰 영향을 준 것으로 파악된다[5].

5. 결론

본 논문에서는 개체명 인식에 널리 사용되는 BIO 표기법 대신 BIT 표기법을 제안한다. BIT 표기법은 O 표지에 편중된 단어의 분포를 골고루 분산함으로써 혼잡도를 줄이고, 모든 기계학습에서 문제가 되는 불균형 학습 문제가 어느 정도 완화됨을 실험을 통해서 확인할 수 있었다. BIT 표기법은 모든 입력 자질에 대해서도 BIO 표기법에 비해 우수한 성능을 보였으며, 또한 모든 표상 생성 방법에 대해서도 더 좋은 성능을 보였다. 또한 한국어 개체명 인식에 있어서 자소 자질이 성능을 개선하는데 도움이 됨을 알 수 있었다. 향후에는 영어와 같은 다른 언어에서도 BIT 표기법이 유용한지를 살펴볼 것이며, 구묶음과 같은 다른 순차 표지 부착 문제에도 적용할 것이다.

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).

5) BERT의 학습말뭉치 크기는 23GB, 47억 형태소이고, 그 외의 모든 방법은 세종말뭉치를 학습말뭉치로 사용했으며 그 크기는 79.3MB, 1558만 형태소이다.

참고문헌

[1] 이경희, 이주호, 최명석, 김길창, “한국어 문서에서 개체명 인식에 관한 연구”, 제12회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 292-299, 2000.

[2] 김재훈, 김형철, 최윤수, “기계학습 기반 개체명 인식을 위한 사전 자질 생성”, 정보관리연구, 제41권, 제2호, pp. 31-46, 2010.

[3] 이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현진, 왕지현, 장명길 “Conditional Random Fields를 이용한 세부 분류 개체명 인식”, 제18회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 268-272, 2006.

[4] 노경목 김창현, 천민아, 박호민, 윤호, 김재균, 김재훈, “개체명 사전 기반의 반자동 말뭉치 구축 도구”, 제29회 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 309-313, 2017.

[5] J. Chiu and E. Nichols, “Named entity recognition with bidirectional LSTM-CNNs”, Transactions of the Association for Computational Linguistics, vol. 4, pp. 357-370, 2016.

[6] C. LEE, “LSTM-CRF models for named entity recognition”, IEICE Transactions on Information and Systems, vol. E100-D, no. 4, pp. 882-887, 2017.

[7] L. M. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning“, Proceedings of the Third ACL Workshop on Very Large Corpora, Association for Computational Linguistics, 1995.

[8] H. He and E. A. Garcia, “Learning from imbalanced data”, IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009.

[9] C. LEE, P.-M. Ryu, and H. Kim, “Named entity recognition using a modified Pegasos algorithm”, Proceedings of the 20th ACM International Confer

ence on Information and Knowledge Management, pp. 2337-2340, 2011.
[10] Z. Huang, W. Xu, K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging”, arXiv:1508.01991, 2015.
[11] 조형미, 김종구, 권홍석, 이종혁, “순환 신경망과 합성곱 신경망을 이용한 개체명 인식”, 한국정보과학회 학술발표논문집, pp. 636-638, 2017.
[12] 박성식, 김학수, “주의 집중 방법을 이용한 개체명 인식”, 한국정보과학회 학술발표논문집, pp. 678-680, 2018.
[13] 유홍연, 고영중, “품사 임베딩과 음절 단위 개체명 분포 기반의 Bidirectional LSTM CRFs를 이용한 개체명 인식”, 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp. 105-110, 2016.
[14] 정예원, 이종혁, “Bidirectional LSTM-CRF 기반 한국어 개체명 인식을 위한 접사 자질을 이용한 단어 표상 확장”, 한국정보과학회 학술발표논문집, pp. 611-613, 2019.
[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality”, Proceedings of Advances in Neural Information Processing Systems, vol. 26, pp. 3111-3119, 2013.
[16] P. Jeffrey, R. Socher, and C. D. Manning, “GloVe: Global vectors for word repersentation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543, 2014.
[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information", Transactions of the Association for Computational Linguistics, vol. 5, pp.135-146, 2017.
[18] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations”, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 2227-2237, 2018.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171-4186, 2019.
[20] 유홍연, 고영중, “Bidirectional LSTM CRF 기반의 개체명 인식을 위한 단어 표상의 확장”, 정보과학회 논문지, 제44권, 제3호, pp. 306-313, 2017.
[21] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks”, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249-256, 2010.