Journal of Digital Convergence
Vol. 17. No. 12, pp. 243-248, 2019

ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.12.243

다중작업학습 기법을 적용한 Bi-LSTM 개체명 인식 시스템 성능 비교 분석

김경민[1], 한승규[1], 오동석[2], 임희석[3*]
[1]고려대학교 컴퓨터학과 석사과정, [2]고려대학교 컴퓨터학과 박사과정, [3]고려대학교 컴퓨터학과 교수

Performance Comparison Analysis on Named Entity Recognition system with Bi-LSTM based Multi-task Learning

GyeongMin Kim[1], Seunggnyu Han[1], Dongsuk Oh[2], HeuiSeok Lim[3*]
[1]Master student, Department of Computer Science and Engineering, Korea University
[2]Ph.D. student, Department of Computer Science and Engineering, Korea University
[3]Professor, Department of Computer Science and Engineering, Korea University

요 약 다중작업학습(Multi-Task Learning, MTL) 기법은 하나의 신경망을 통해 다양한 작업을 동시에 수행하고 각 작업 간에 상호적으로 영향을 미치면서 학습하는 방식을 말한다. 본 연구에서는 전통문화 말뭉치를 직접 구축 및 학습데이터로 활용하여 다중작업학습 기법을 적용한 개체명 인식 모델에 대해 성능 비교 분석을 진행한다. 학습 과정에서 각각의 품사 태깅(Part-of-Speech tagging, POS-tagging) 과 개체명 인식(Named Entity Recognition, NER) 학습 파라미터에 대해 Bi-LSTM 계층을 통과시킨 후 각각의 Bi-LSTM을 계층을 통해 최종적으로 두 loss의 joint loss를 구한다. 결과적으로, Bi-LSTM 모델을 활용하여 단일 Bi-LSTM 모델보다 MTL 기법을 적용한 모델에서 1.1%~4.6%의 성능 향상이 있음을 보인다.

주제어 : 딥러닝, 다중작업학습, 품사 태깅, 개체명 인식, 전통문화

Abstract Multi-Task Learning(MTL) is a training method that trains a single neural network with multiple tasks influences each other. In this paper, we compare performance of MTL Named entity recognition(NER) model trained with Korean traditional culture corpus and other NER model. In training process, each Bi-LSTM layer of Part of speech tagging(POS-tagging) and NER are propagated from a Bi-LSTM layer to obtain the joint loss. As a result, the MTL based Bi-LSTM model shows 1.1%~4.6% performance improvement compared to single Bi-LSTM models.

Key Words : Deep Learning, Multi-task Learning, Part of speech tagging, Named entity Recognition, Traditional culture

1. 서론

일반적인 딥러닝 학습 방식으로는 하나의 모델 학습을 통해 단일 모델에 대한 성능 결과를 예측하거나 다양한 모델에 앙상블 기법을 적용하여 성능 향상을 보이는 다양한 기법들이 존재한다. 다중작업학습 기법이란 한 신경망이 여러 작업(task)을 동시에 수행할 수 있도록 각 작업이 서로 다른 작업에 영향을 주면서 학습한다. 작업 A

*This research is supported by Ministry of Culture, Sport and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research&Development Program 2017. (No. R2017030045).
*Corresponding Author : HeuiSeok Lim(limhseok@korea.ac.kr)
Received October 28, 2019 Revised November 25, 2019
Accepted December 20, 2019 Published December 28, 2019

에서 작업 B로 일련의 순차적인 과정을 거치는 전이 학습(transfer learning) 방식과는 다르게 이 기법은 다양한 정보를 동시에 학습시킨다. 따라서 독립된 신경망으로 훈련시키는 기존 방법보다 각 작업의 정보를 공유하면서 학습하는 방식은 모든 데이터에 태그가 붙어있지 않아도 학습이 가능하고, 하나의 모델로 동일한 파라미터를 공유한다는 점에서 기존 방식과 비교하여 장점을 갖는다[1].

품사 태깅(POS-tagging)과 개체명 인식(NER)은 텍스트의 고유한 정보를 얻기 위한 기존 순차적 레이블링(Sequence Labeling) 연구의 하위 작업이다. 품사와 개체명은 동일한 토큰 단위의 정보로부터 해당 토큰에 대한 각각의 정답 값으로, 학습을 위한 딥러닝 모델의 입력으로 활용된다.

한국학중앙연구원 디지털 인문학[1]은 '한국의 기록문화 유산' 중 스토리텔링 자원으로 활용 가치가 높은 분야를 선정하여 대표적인 기록물들에 대해 디지털 콘텐츠 정보화하여 구축해 놓은 사이트이다. 해당 콘텐츠에는 기록물에 대한 역사적 및 지리적 배경, 인물, 관련 내용 등에 대한 정보를 담은 한국어 전통문화 데이터가 존재한다. 본 연구에서는 해당 웹사이트 데이터를 추출하고, 추출한 데이터를 학습 말뭉치로 구축한 방법을 기술한다. 또한, 구축한 말뭉치를 다중작업학습 기법을 통해 BI-LSTM 모델에 실험한 결과 및 기존 다양한 모델에 적용한 실험 결과에 대해 비교 분석하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 MTL, 딥러닝 모델 및 임베딩 기법에 관한 관련 연구를 설명하고 3장에서는 다중학습기법을 활용한 pos-ner multitask learning 연구 모델을 설명한다. 4장에서는 말뭉치 구축 방법 및 통계에 대한 설명과 5장에서는 비교 분석 결과, 6장은 결론 및 향후 연구에 대해 설명한다.

2. 관련 연구

Multitask-learning의 대표적인 방법에는 각 작업이 공유하는 은닉 계층과 작업에 최적화된 층을 갖는 형태를 취함으로써 over-fitting을 방지시키는 Hard Parameter Sharing 방법과 각 작업이 독립적인 층을 가지면서 각 계층의 Loss를 최소화시키는 Soft Parameter Sharing 두 가지 방법이 존재한다[2]. 최근 컴퓨터 비전 분야에서 각 작업이 공유된 pre-trained CNN (Convolutioal Neural Network, CNN)을 활용하여 fine-tuning 하는 방식의 네트워크 모델과 각 작업에 따라 서로 다른 은닉 계층의 손실 값을 공유함으로써 그 값을 근사시키는 기법을 활용한 방법[3, 4]은 MTL 기법을 통해 각 작업에서 우수한 성능을 보일 수 있었다. 자연언어 처리 방식 중 연속적 데이터 처리를 위해 사용되는 순환신경망(Recurrent Neural networks, RNN) 기반의 Long Short term memory (LSTM) 방식은 자연어 처리 분야뿐만 아니라 최근 의학, 보안 등 다양한 분야에 적용되어 많은 성과를 이뤄내고 있다[12-15]. 그 중, 문장의 양방향 정보를 고려하는 Bi-LSTM[16,17]은 기존 LSTM의 정보를 앞, 뒤 모든 방향으로 고려하는 기법으로 Word2vec[5], Glove[6], Fasttext[7] 등의 토큰 임베딩을 입력 벡터로 활용하여 모델을 학습시키는 방법이다.

모델 생성에 있어 아키텍처를 구성하는 방법과 컴퓨터가 이해할 수 있는 체계적인 지식적 표현 방법은 어려우면서도 상당히 중요하다. 언어 표현방식 중 하나인 one-hot encoding은 언어의 의미 간 관계 파악이 어렵고 단어의 수가 곧 벡터의 차원으로 표현되어 리소스 낭비가 심하다는 단점이 있다. 최근 다양한 연구 모델에 입력으로 활용되는 Word2vec, Glove, Fasttext 등의 임베딩 기법들은 단어를 밀집된 고차원의 특정 벡터로 맵핑시키는 방식으로 단어 간의 관계 파악을 가능하게 할 수 있으나 문맥을 파악하기 어렵다는 한계가 존재한다. 문장 임베딩 기법인 언어 모델 임베딩(Embedding from Language Model, ELMO)[8]은 전이 학습 기법을 자연어 처리에 적용한 기법으로 훈련된 Bi-LSTM 신경망을 사전학습(pre-training)을 통해 각 하위 작업에 적용하였다. 이를 통해 같은 단어가 문맥에 따라 다른 벡터로 표현 가능하게 되었으며, Transformer 인코더 기반의 모델인 BERT(Bidirectional Encoder Representations from Transformers, BERT)[9] 는 한 단계 더 나아가 단순히 forward와 backward를 합하는(concatenate) 방식을 취했던 기존 ELMO 방식에 random masking과 next sentence prediction 기법을 적용하여 언어를 완전한 양방향으로 이해할 수 있는 언어적 표현(Language Representation)방식으로 체계화한다. 최근 BERT의 문맥적 위치 정보를 고려하지 않는다는 특징을 개선한 XLNET[10] 모델과 추가 하이퍼파라미터 튜닝 및 Dynamic Masking 기법을 활용한 RoBERTa[11] 모델이 있다. 그러나 이러한 문장 임베딩 기법은 아직 한국어 특성을 반영하지 못하거나, 개인 리소스의 한계로 연구에 다소 어려운 점이 존재하기도 한다.

1) http://dh.aks.ac.kr/

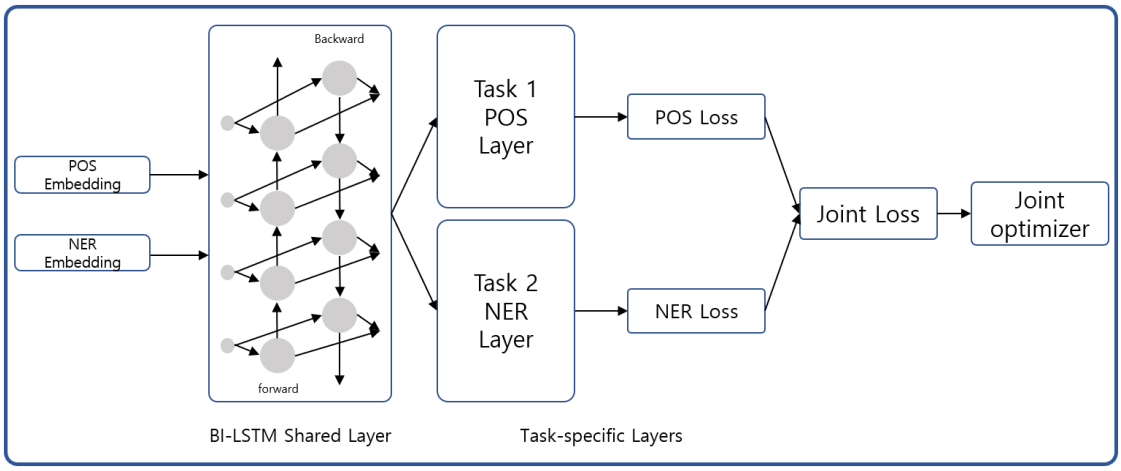


Fig. 1. Hard parameter sharing architecture for multi-task Learning.

본 연구에서는 단어 임베딩 수준에서 전통문화 데이터를 활용하여 다중작업학습 기법에 적용한 모델과 기존 다양한 모델의 성능 비교 분석을 진행한다.

3. POS-NER Multitask-Learning

자연언어처리에는 pos, ner, dependency parsing 등의 다양한 작업이 존재하는데 일반적인 작업은 해당 작업의 목적에 맞는 아키텍처를 구성하지만 MTL은 각 작업에 상호 간 영향을 미칠 수 있도록 모델링하여 전반적인 작업 성능을 향상시킨다. 본 연구에서는 pos와 ner 두 작업에 대해 MTL에서 가장 활용도가 높은 모델링 방법 중 하나인 Hard parameter sharing을 적용하여 실험을 진행한다.

Fig. 1은 본 연구에서 사용한 MTL 모델에 대해 나타낸다. 이는 은닉 계층을 공유하고 각 작업에 따라 추가적인 계층을 통과시키는 기법으로, 은닉 계층으로 Bi-LSTM 모델을 활용하여 pos와 ner을 각각 통과시키고 각 작업에 최적화된 Bi-LSTM 계층을 다시 한번 통과시켜 두 Loss를 통해 joint loss를 계산하는 과정을 진행한다. Fig. 2에서의 주황색 실선은 joint loss를 나타내고 파랑색 실선은 ner loss를 나타내며 joint 모델의 loss가 더 빠르게 떨어지는 것을 확인할 수 있다.

성능 비교 실험을 위해 입력으로는 형태소 단위의 토큰을 사용했으며 모든 모델에 같은 하이퍼파라미터를 사용한다. Table 1은 학습에 사용한 하이퍼파라미터를 나타낸다. 전체 학습 epoch은 20, dropout 비율은 0.8로 128의 배치사이즈로 학습을 진행한다. 신경망 학습 최적

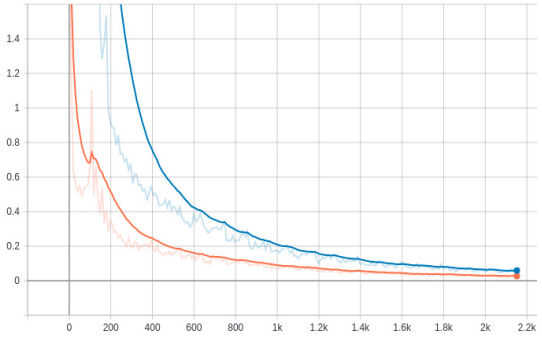


Fig. 2. Loss comparison. There are two losses which orange color is joint loss and blue color is ner loss.

화를 위한 알고리즘으로는 rmsprop 알고리즘을 사용하고 0.01 learning rate부터 매 학습 시 0.9 decay로 학습을 진행한다.

Table 1. training hyperparameter

Parameter	
Epoch	20
dropout	0.8
batch_size	128
learning rate method	rmsprop
learning rate	0.01
learning rate decay	0.9

4. 말뭉치 구축

한국학중앙연구원 디지털 인문학 웹사이트는 한국의 대표적인 기록물들을 디지털 콘텐츠로 제작해놓은 사이트이다. 본 연구에서는 한국의 기록 문화유산과 관련한 백과사전 기사 중 문맥 기사의 개요와 내용을 크롤링하여 각 토큰에 태그하는 작업을 진행하였으며 기구축한 기획 및 중심기사의 데이터와 통합하는 과정을 순차적으로 진행했다.

Table 2. Category and Tag ratio in Korean culture Corpus

Category	Count	Frequency	Percentage
B	21699	6.96%	-
B_PS(Person)	8757	2.81%	40.36%
B_LC(Location)	6183	1.98%	28.49%
B_DT(Date)	4403	1.41%	20.29%
B_OG(Organization)	1996	0.64%	9.20%
B_TI(Time)	360	0.12%	1.66%
I	12909	4.14%	-
I_DT	4651	1.49%	36.03%
I_LC	3765	1.21%	29.17%
I_PS	3068	0.98%	23.77%
I_OG	1135	0.36%	8.79%
I_TI	290	0.09%	2.25%

코모란 형태소 분석기[2]를 사용하여 각 문장을 형태소 단위로 분리하여 품사를 태깅하였고, 인물(PS), 장소(LC), 기관(OG), 시간(TI), 날짜(DT) 총 5개 클래스에 대해 BIO(Begin, Inside, Outside) 태깅 기법을 사용해 각각의 토큰에 태그하는 작업을 진행했다. 문장을 구분하기 위한 구분자로는 '.'(구두점)과 공백을 사용한다. 결과적으로 전체 1731개의 기사로부터 21189개의 문장과 약 31만 형태소 단위의 데이터를 추출하였으며 말뭉치는 8:1:1로 분리하여 각각 Train, Test, Evaluation에 사용하였다. 말뭉치에 대한 자세한 정보는 Table 3에 나타내었다.

Table 3. Data details

Dataset	Data type	Sentences	Tokens
Culture dataset	Train	16951	249479
	Test	2119	31185
	Evaluation	2119	31185

2) https://www.shineware.co.kr/products/komoran/

Table 2는 전체 카테고리 중 B 태그와 I 태그로 태깅된 비율을 나타낸다. B 태그는 전체 태그 중 6.96% 비율을 차지하며 가장 빈도수가 높은 데이터는 B_PS로 2.81%의 태그 비율을 가지고 동종 내 비율(Percentage) 40.36%를 나타낸다. 가장 낮은 데이터는 B_TI로 태그 비율 0.12%로 B 태그 내에서 1.66%를 차지한다. I 태그의 경우 전체 태그 중 4.14% 비율을 차지하고 가장 빈도수가 높은 데이터는 I_DT 태그로 1.49% 태그 비율을 가졌으며 동종 내 비율 36.03%를 나타낸다. B 태그에서 PS가 비율이 가장 높은 반면, I 태그의 PS가 상대적으로 낮은 빈도수를 나타낸 것은 PS가 DT, LC과 비교하여 단일 태그된 개체명이 많이 등장했다는 것을 알 수 있다.

5. 비교 분석 결과

본 연구에서의 실험은 총 7가지로 Table 4를 통해 확인할 수 있다. 성능 평가는 개체명 인식 기술에서 가장 널리 사용되는 정량적 평가 방식인 F-score를 사용한다. (1~3)의 실험은 기존 Bi-LSTM, Bi-LSTM-CRF 기반의 단일 모델에 입력 자질을 조금씩 다르게 진행하였고 (4~5) 실험은 MTL 기반 Bi-LSTM에 어절, 형태소 단위 자질을 실험하였다. (6~7)실험은 단일 Bi-LSTM 모델에 CNN 모델에 문자 단위의 입력 자질을 활용하였다.

Table 4. Performance in models

No.	Model	Features	F1-score
1	Bi-LSTM	word	75.60%
2	Bi-LSTM	word, morpheme	76.48%
3	Bi-LSTM-CRF	word, morpheme	78.27%
4	MTL Bi-LSTM	word	79.36%
5	MTL Bi-LSTM	word, morpheme	80.20%
6	Bi-LSTM-CNN	character, word, morpheme	81.50%
7	Bi-LSTM-CNN-CRF	character, word, morpheme	82.50%

본 연구에서의 어절, 형태소 단위 자질을 활용한 MTL 모델의 실험 (4~5) 은 Bi-LSTM, Bi-LSTM-CRF 모델에 어절, 형태소 단위 자질로 실험한 (1~3) 보다 높은 성능을 보이고, Bi-LTM-CNN, Bi-LSTM-CNN-CRF 모델에 문자 단위 자질을 활용한 (6~7) 성능보다는 비교적 낮은 성능을 보였다. 결과적으로, MTL 모델은 단일 모델 혹은 CRF를 추가하여 실험한 성능보다 1.1%~4.6%의 성능 차이를 보였다.

6. 결론

본 연구는 다중작업 기법을 적용한 Bi-LSTM 기반 모델의 기존 단일 모델과의 성능 비교 분석을 진행하였다. 전통문화 말뭉치를 직접 구축하여 품사 태깅 정보 및 개체명 태그 정보를 학습데이터로 활용하여 MTL 기반 Bi-LSTM 모델이 기존 단일 모델보다 1.1%~4.6% 성능 차이를 보였다. 기존 단일 모델에서는 pos와 ner 각각 서로의 정보만을 가지고 모델을 생성했다면 MTL 기법은 상호 간 자질 정보를 공유함으로써 기존 단일 모델에서는 고려하지 못했던 단어정보까지 포함할 수 있었기 때문에 성능 향상이 가능했던 것으로 보인다. 풍부한 의미적 자질 표현을 위해 문자, 어절, 형태소 등의 정보를 활용한다면 개체명 인식에 있어서 단일 모델보다 더 높은 인식률 얻을 수 있을 것으로 기대한다.

REFERENCES

[1] S. Ruder. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

[2] R. Caruana. (1997). Multitask learning. *Machine learning, 28(1),* 41-75.

[3] M. Long & J. Wang. (2015). Learning multiple tasks with deep relationship networks. *arXiv preprint arXiv:1506.02117,* 2.

[4] Y. Zhang, Y. Wei & Q. Yang. (2018). Learning to multitask. *In Advances in Neural Information Processing Systems* (pp. 5771-5782).

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado & J. Dean. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems* (pp. 3111-3119).

[6] J. Pennington, R. Socher & C. Manning. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[7] P. Bojanowski, E. Grave, A. Joulin & T. Mikolov. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5,* 135-146.

[8] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365.*

[9] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[10] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov & Q. V. Le. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237.*

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen & V. Stoyanov. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[12] A. Lamurias, D. Sousa, L. A. Clarke & F. M. Couto. (2019). BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. *BMC bioinformatics, 20(1),* 10.

[13] C. Lyu, B. Chen, Y. Ren & D. Ji. (2017). Long short-term memory RNN for biomedical named entity recognition. *BMC bioinformatics, 18(1),* 462.

[14] A. R. Tuor, R. Baerwolf, N. Knowles, B. Hutchinson, N. Nichols & R. Jasper. (2018, June). Recurrent neural network language models for open vocabulary event-level cyber anomaly detection. *In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.*

[15] S. KP. (2019). RNNSecureNet: Recurrent neural networks for Cyber security use-cases. *arXiv preprint arXiv:1901.04281.*

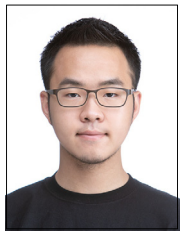
[16] G. Kim, K. Kim, J. Jo & H. Lim. (2018). Constructing for Korean Traditional culture Corpus and Development of Named Entity Recognition Model using Bi-LSTM-CNN-CRFs. *Journal of the Korea Convergence Society, 9(12),* 47-52.
DOI : 10.15207/jkcs.2018.9.12.047

[17] D. Lee, W. Yu & H. Lim. (2017). Bi-directional LSTM-CNN-CRF for Korean Named Entity Recognition System with Feature Augmentation. *Journal of the Korea Convergence Society, 8(12),* 55-62.
DOI : 10.15207/JKCS.2017.8.12.055

김 경 민(GyeongMin Kim) [정회원]

· 2017년 8월 : 백석대학교 정보통신학부 정보보호학과(공학사)
· 2018년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석사과정
· 관심분야 : 딥 러닝, 자연어 처리, 지식표현
· E-Mail : totoro4007@korea.ac.kr

한 승 규(Seunggyu Han) [정회원]



· 2019년 2월 : 대구경북과학기술원(DGIST) 융복합대학 기초학부 (공학사)
· 2019년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석사과정
· 관심분야 : 딥 러닝, 자연어처리
· E-Mail: seunggyu-han@gmail.com

오 동 석(Dongsuk Oh) [정회원]

· 2014년 2월 : 충북대학교 정보통신공학부 (공학사)
· 2016년 2월 : 서강대학교 컴퓨터공학과 (공학석사)
· 2018년 8월 : 다이퀘스트 근무 전임연구원
· 2018년 3월 : NHN엔터네인먼트 근무 전임연구원
· 2019년 3월 ~ 현재 : Human-inspired 복합지능 연구센터 연구원
· 관심분야 : 인공지능, 지식표현, 자연어처리

임 희 석(HeuiSeok Lim) [종신회원]

· 1992년 3월 : 고려대학교 컴퓨터학과 (이학학사)
· 1994년 3월 : 고려대학교 컴퓨터학과 (이학석사)
· 1997년 3월 : 고려대학교 컴퓨터학과 (이학박사)
· 2008년 2월 ~ 현재 : 고려대학교 정보대학 컴퓨터학과 교수
· 관심분야 : 자연어처리, 뇌신경 언어 정보 처리
· E-Mail : limhseok@korea.ac.kr