KIPS Tr. Software and Data Eng.
Vol.5, No.9 pp.419~424 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.9.419

Named Entity Recognition for Patent Documents Based on Conditional Random Fields

Lee Tae Seok† · Shin Su Mi†† · Kang Seung Shik†††

ABSTRACT

Named entity recognition is required to improve the retrieval accuracy of patent documents or similar patents in the claims and patent descriptions. In this paper, we proposed an automatic named entity recognition for patents by using a conditional random field that is one of the best methods in machine learning research. Named entity recognition system has been constructed from the training set of tagged corpus with 660,000 words and 70,000 words are used as a test set for evaluation. The experiment shows that the accuracy is 93.6% and the Kappa coefficient is 0.67 between manual tagging and automatic tagging system. This figure is better than the Kappa coefficient 0.6 for manually tagged results and it shows that automatic named entity tagging system can be used as a practical tagging for patent documents in replacement of a manual tagging.

Keywords : Conditional Random Fields, Named Entity Recognition, Patent Corpus, Kappa Coefficient, 10-Fold Cross Validation

조건부 랜덤 필드를 이용한 특허 문서의 개체명 인식

이 태 석† · 신 수 미†† · 강 승 식†††

요 약

특허 정보검색에서는 검색 정확도를 높이거나 유사 특허들을 검색하기 위한 목적으로 청구항 등 특허 기술 문서의 내용을 대표하는 개체명 인식이 필요하다. 본 연구에서는 특허 개체명을 자동으로 인식하기 위하여 기계 학습 기법에서 태깅 문제 해결에 매우 우수한 성능을 보이는 조건부 랜덤 필드 기법을 이용하는 특허 개체명 인식 방법을 제안하였다. 개체명 태깅이 되어 있는 특허 문서 말뭉치에서 66만 어절을 학습용 데이터로 사용하여 특허 개체명 시스템을 구축하고, 7만 어절을 평가용 데이터로 사용하여 성능 평가를 하였다. 실험 결과에 의하면 개체명 인식 정확도는 93.6%이고, 개체명 인식 성능을 수작업 태깅 결과와 비교하여 일치도를 평가했을 때 카파 계수는 0.67로 나타났다. 이 카파 계수 값은 두 사람의 수작업 태깅 결과에 대한 카파 계수 0.6 보다 높은 것으로 특허 개체명 인식 시스템이 수작업 태깅을 대신하여 실용적으로 활용될 수 있음을 확인하였다.

키워드 : 조건부 랜덤 필드, 개체명 인식, 특허 말뭉치, 카파 계수, 10-등분 교차 검증

1. 서 론

자연어 처리 분야에서 품사 태깅(part-of-speech tagging)에 관한 연구는 단순히 품사를 부착하는 것으로부터 개체명 태그(named entity tag)를 부착하는 문제로 발전하였다[1]. 개체명 인식은 기본적으로 인명, 지명, 기관명을 대상으로 하고 있으며, 그 외에도 날짜, 시간, 가격과 같은 특정 수치 등으로 확대되어 왔다. 이때, 개체명 인식을 위한 학습 말뭉치는 품사 태깅 말뭉치(POS-tagged corpus)에 개체명 태그를 추가로 부착하는 방법으로 구축하게 된다.

품사 태깅을 비롯한 태깅 기법에 주로 사용되는 방법은 규칙 기반 접근법과 통계적 접근법을 기반으로 하는 기계 학습 기법으로 구분할 수 있다[2]. 규칙 기반 접근법은 언어 정보를 생성 규칙 형태로 표현하고 이를 적용하여 태깅을 수행하는데, 태깅 규칙은 언어 전문가가 수동으로 구축한다. 이에 비해, 기계 학습 기법에서는 태깅 작업에 적합한 기계 학습 모형을 정립하고 태깅 말뭉치로부터 추출된 자질 정보를 이용하여 태깅을 수행한다.

개체명 인식은 그 목적에 따라 여러 가지 응용 분야에 활용되고 있다. Wang(2009)은 의료 분야에서 임상 기록에 대한 개체명 인식 연구의 결과로 기계 학습 기법을 이용하였

† 정 회 원 : 한국과학기술정보연구원 정보서비스실 책임연구원
†† 정 회 원 : 한국과학기술정보연구원 정보서비스실 선임연구원
††† 종신회원 : 국민대학교 컴퓨터공학부 교수
Manuscript Received : January 11, 2016
First Revision : April 4, 2016
Accepted : April 18, 2016
* Corresponding Author : Kang Seung Shik(sskang@kookmin.ac.kr)

으며, 의학 및 약학 분야의 특허 문서에서 명사구를 인식함으로써 검색의 정확도를 높이는 등 다양한 분야에서 개체명 인식에 관한 연구를 하고 있다[3, 4].

특허 관련 분야에서는 특허 검색의 성능을 향상시키거나 특허 문서를 자동으로 분류하기 위하여 자연어 처리 기법의 중요성이 강조되고 있다[5]. 그 이유는 특허 문서로부터 의미 있는 정보를 정확하고 신속하게 분류하고 특성을 파악하는 것이 바로 기업의 경쟁력과 이익에 직결되기 때문이다. 따라서 방대한 양의 특허들에 대한 선행 조사 및 분석을 통해 자사의 제품과 관련된 기술에 대한 적절한 로드맵을 만들거나, 미래의 경쟁력을 준비하는데 특허 분석과 선행 기술 조사에 많은 비용을 투자하고 있다.

개체명 태그가 부착되어 있는 특허 말뭉치로부터 조건부 랜덤 필드(CRF: conditional random fields)에 기반을 둔 학습 모델을 생성하는 방법은 입력 데이터 열에 대한 태그 부착 문제를 해결하는 응용 분야에서 은닉 마르코프 모델(hidden Marcov model)이나 최대 엔트로피 마르코프 모델(maximum-entropy Markov model) 등에 비하여 우수한 성능을 보이고 있다[6-8]. 본 논문에서는 조건부 랜덤 필드 기법을 이용하여 특허의 요약, 청구항 등 특허 문서에서 개체명을 인식하는 방법을 제안한다.

2. 개체명 인식 기법

개체명 인식에 대한 초기의 연구에서는 경험적인 규칙으로 개체명을 인식하는 규칙 기반 방법이 사용되었다. 이 기법에서는 인식하고자 하는 개체명의 앞뒤 문장 구조를 규칙으로 기술하여 개체명을 인식한다. 개체명 태깅된 대규모 말뭉치가 구축된 후에는 학습 말뭉치로부터 개체명 인식 규칙을 자동으로 추출하는 기계 학습 기법이 주로 사용된다.

개체명 인식을 위해서는 개체명이 문장 내에서 출현할 때 앞뒤 어휘들에서 발견되는 특징이나 자질들을 조사하여 활용할 필요가 있다. 어떤 자질들이 개체명 태깅에 유용하게 활용되는지에 관한 자질 선택(feature selection) 연구에 의해 구문 패턴이나 문형 자질, 단어 수준 자질, 사전 조사 자질들이 제시되었다.

구문 패턴이나 특정 문형으로부터 추출되는 자질은 “city such as”, “organization such as”와 같이 문장 뒤에 나오는 단어를 개체명으로 추출한다. “The president of Apple eats an apple.”라는 문장의 예에서 Fig. 1과 같이 <“대문자로 시작 여부”, “단어의 길이”, “단어”> 형태로 정보를 요약하여 “대문자로 시작되는 단어는 개체명 후보이다”, “3단어 보다 긴 것은 기관명 후보이다”와 같은 규칙을 사용할 수 있다.

단어 수준의 자질로는 대소문자, 구두점, 숫자 패턴, 형태소, 품사, 구 길이, 토큰 길이, n-gram, non-alpha 등의 정보들이 중요한 자질로 사용된다. 사전 조사 자질로는 용어 사전, 불용어, 약어, 조직명, 부처명, 항공사, 교육 기관 등에서 이미 목록화된 자질을 사용한다.

<true, 3, “the”>, <false, 9, “president”>, <false, 2, “of”>, <true, 5, “apple”>, <false, 4, “eats”>, <false, 2, “an”>, <false, 5, “apple”>

Fig. 1. Example of rule patterns

3. 조건부 랜덤 필드

입력 데이터 열을 분할하고 태그를 부착하는 문제에 대한 해결 방법으로 은닉 마르코프 모델이 광범위하게 사용되어 왔다. 이 모델은 입력 데이터 열과 태그 열 사이의 결합 확률(joint probability)을 이용하는 생성 모델(generative model)이다. 이 모델은 학습 데이터로부터 최적의 모델을 생성하는데, 상호 작용하는 자질을 표현하거나 멀리 떨어진 입력 데이터 열 사이의 의존 관계를 표현하기 어려운 단점이 있다. 이러한 단점을 극복하기 위하여 은닉 마르코프 모델의 제약 조건을 완화하여 범용 모델로 확장하는 조건부 확률 모델이 제안되었다.

조건부 랜덤 필드 모델은 조건부 확률을 최대화하기 위해 훈련된 비방향성 그래프 모델(undirected graph model)이다. 일반적인 그래프 구조를 가진 조건부 랜덤 필드 모델의 하나인 선형 체인 구조의 조건부 랜덤 필드 모델은 입력 데이터 열에 태그 열을 결합하는 문제에 적합하며, 이 모델은 태깅 문제를 해결하는데 활용된다.

$x = x_1 \cdots x_n$ 를 입력 데이터 열에 대한 확률 변수라고 하고, $y = y_1 \cdots y_n$ 를 입력 데이터 열에 대응하는 태그 열의 확률 변수라고 하자.

매개 변수 $\Lambda = (\lambda_1, \lambda_2, \cdots, \mu_1, \mu_1, \cdots)$를 갖는 선형 체인 구조의 조건부 랜덤 필드 모델은 다음과 같이 조건부 확률로 정의된다.

$$P_\Lambda(y|x) = \frac{1}{Z(x)} exp(\sum_j \sum_{i=1}^{n} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \sum_{i=1}^{n} \mu_k s_k(y_i, x, i)) \quad (1)$$

Equation (1)에서 $Z(x)$는 입력 데이터 열에 대한 태그 열의 확률 값의 합이 1이 되도록 하는 정규화 상수이다. $t_j(y_{i-1}, y_i, x, i)$는 전이 자질 함수(transition feature function)이며, $s_k(y_i, x, i)$는 상태 자질 함수(state feature function)이다. λ_j와 μ_k는 각 자질 함수에 대한 가중치로서 태그가 부착된 학습용 데이터로부터 구할 수 있다. 매개 변수는 최대우도 추정법(maximum likelihood estimation)을 사용하

여 구하는데, 다른 알고리즘보다 수렴 속도가 빠른 BFGS (Broyden - Fletcher - Goldfarb - Shanno) 알고리즘이 주로 사용된다.

학습용 데이터로부터 매개 변수 Λ를 구하고 나면, 주어진 입력 데이터 열 x에 대하여 가장 가능성이 높은 태그 열 y는 Equation (2)와 같이 구할 수 있다.

$$y^* = \arg max_y P_\Lambda(x|y) \quad (2)$$

태그 열 y를 구하는 과정은 Viterbi 알고리즘을 사용한 동적 프로그래밍 방법을 이용하여 계산한다. 조건부 랜덤 필드 모델을 개체명 태깅에 적용할 때는 개체명 태깅 문제를 기본 명사구 인식(base NP chunking) 문제와 같이 시작-중간-끝 태그(begin-inside-outside tag)를 부여하는 태그 부착 문제로 간주하였다. 이 때, 기계 학습을 위한 자질로는 단어의 품사, 어휘, 단어의 길이, 문장에서의 단어 위치 등과 같은 다양한 언어 정보를 사용하였다.

4. 특허 개체명 인식

특허 문서에서 추출하고자 하는 개체명들은 기술명(technology name), 서비스명(service name), 제품명(product name)이다. 이러한 개체명들은 기술 흐름을 파악하고 특허 맵(patent map)을 작성하기 위해 활용된다. 그런데 특허 문서에 대해 개체명 태그가 부착되어 있는 특허 개체명 말뭉치 사례가 많지 않다.

본 논문에서는 사람이 직접 개입하여 개체명 태그를 부착하고 오류 검증 과정을 거쳐 구축한 한국과학기술정보연구원(KISTI)[1]의 특허 문서 말뭉치를 사용하였다. 이 특허 문서 말뭉치는 미국 등록 특허에서 분류별로 임의로 선별한 2,400건에 대해 7명의 작업자가 태깅을 한 자료이다. 이 말뭉치는 Table 1과 같이 361,211 문장에 대한 태깅 정보를 사용하였다. 이 말뭉치를 구축할 때 작업자의 의견 일치도를 코헨의 카파 계수(Cohen's kappa coefficient)로 계산하면 0.6으로 나타났다.

Table 1. Tagged sentences in USA patent by IPC

IPC class	Sentences
A	53,932
B	55,642
C	37,371
D	26,319
E	39,757
F	42,041
G	62,995
H	43,154
합계	361,211

실험에 사용할 특허 문서는 전기통신기술 분야의 특허 분류 H04* 분야의 17,531개 문장 중 품사 태그가 인식된 17,142개 문장이다. 각 문장은 Table 2와 같이 발명의 명칭(title), 요약서(abstract), 청구항(claim), 명세서(description)에서 추출하였으며, 85.36%가 명세서에 있는 문장이다.

Table 2. Sentences in patent fields

Patent fields	Sentences
description	14,964
claim	1,784
abstract	700
title	83
Total	17,531

특허 태깅 말뭉치를 단어 단위의 토큰으로 분리하면 Table 3과 같이 총 742,510개의 토큰들로 구성되어 있다. 이 중에서 개체명으로 태깅이 된 토큰은 91,440개(12.31%)이고, 태그가 부착되지 않은 토큰이 651,070개(87.69%)이다. Table 3에서 B-Product, B-Service, B-Technology 태그는 각각 제품명, 서비스명, 기술명이 시작되는 단어에 부착하는 태그이고, I-Product, I-Service, I-Technology 태그는 각 개체명의 중간 또는 끝 단어에 부착하는 태그이다. 각 개체명에 속하지 않는 단어들에는 O 태그를 부착한다.[2]

Table 3. Tagged tokens as named entities

Named entity tags	Tokens	Percent(%)
B-Product	24,541	3.31
B-Service	19	0.00
B-Similar_Product	13,547	1.82
B-Technology	12,793	1.72
B-unknown	2	0.00
I-Product	28,857	3.89
I-Service	1	0.00
I-Similar_Product	5,490	0.74
I-Technology	6,188	0.83
I-unknown	2	0.00
O	651,070	87.69
합계	742,510	100

1) http://www.kisti.re.kr/.

2) B-unknown과 I-unknown 태그는 수작업 태깅 과정에서 unknown 태그로 부착한 개체명이다. 서비스명으로 태깅된 것은 19개이다.

개체명 인식 성능이 어절 단위로 어느 정도 영향이 있는지 알아보기 위해 학습 자질을 어절 단위로 나누어 상태 자질과 전이 자질로 구분하여 CRF++를 사용하여 실험하였다.[3] Table 4는 각 자질들의 기여도를 나타내는 결과를 보여주는데 한 어절 단위, 두 어절 단위, 세 어절 단위 모두 x_i 위치 전후의 상태 자질과 전이 자질이 성능에 가장 많은 영향을 미치는 것을 알 수 있다. 또한, 세 어절 묶음 자질은 데이터 부족 현상으로 인식률에 큰 영향을 주지 못했다. Table 5는 각 자질들의 인식 성능 기여도가 높은 자질들을 선택한 것으로 최적 자질 선택은 F1 점수가 10 이상인 자질들만 모두 선택하였다.

Table 4. Performance contribution by features

Features	State F1	Transition F1
x_{i-2}	1.09	**30.58**
x_{i-1}	3.46	**55.68**
x_i	**25.37**	**52.13**
x_{i+1}	2.74	**21.80**
x_{i+2}	1.03	8.45
x_{i-3}/x_{i-2}	2.48	**19.01**
x_{i-2}/x_{i-1}	4.09	**32.96**
x_{i-1}/x_i	**21.66**	**35.87**
x_i/x_{i+1}	**16.59**	**25.57**
x_{i+1}/x_{i+2}	5.41	**14.77**
$x_{i-3}/x_{i-2}/x_{i-1}$	5.57	**13.69**
$x_{i-2}/x_{i-1}/x_i$	**11.23**	**16.32**
$x_{i-1}/x_i/x_{i+1}$	**10.23**	**13.64**
$x_i/x_{i+1}/x_{i+2}$	8.87	**11.00**
$x_{i+1}/x_{i+2}/x_{i+3}$	4.79	7.67

Table 5. Finally selected feature set

State feature	Transition feature
x_i x_{i-1}/x_i, x_i/x_{i+1} $x_{i-2}/x_{i-1}/x_i$, $x_{i-1}/x_i/x_{i+1}$	x_{i-2}, x_{i-1}, x_i, x_{i+1} x_{i-3}/x_{i-2}, x_{i-2}/x_{i-1}, x_{i-1}/x_i, x_i/x_{i+1}, x_{i+1}/x_{i+2} $x_{i-3}/x_{i-2}/x_{i-1}$, $x_{i-2}/x_{i-1}/x_i$, $x_{i-1}/x_i/x_{i+1}$, $x_i/x_{i+1}/x_{i+2}$

3) https://taku910.github.io/crfpp/, CRF++: Yet Another CRF toolkit

5. 실험 및 성능 평가

특허 개체명 인식 시스템의 효용성을 검증하기 위하여 비교 대상 시스템으로 개체명 태그 사전을 이용하는 경우와 비교-평가를 수행하였으며, 자질 선택 기법의 효용성을 검증하기 위하여 기본 명사구 인식 방법의 자질 집합으로 구현한 경우와 성능을 비교하였다. 개체명 인식 실험은 기계학습 기법에서 말뭉치가 충분하지 않을 때 사용하는 10-등분 교차 검증(10-fold cross validation) 방식을 사용하였다. 전체 자료를 10등분하여 평가용 데이터로는 특허 문서 말뭉치의 10%에 해당하는 7만 어절을 발췌하고, 학습용 데이터는 나머지 90%에 해당하는 약 66만 어절(약 1.5만 문장)을 발췌하여 만들었다. 실험은 10등분된 자료를 평가용과 학습용으로 교대하여 10회 평균을 하였다.

자질 선택 효용성을 비교하기 위해 기본 명사구 인식 기법의 자질값과 특허 개체명 인식을 위한 최적화된 자질 집합으로 실험하였으며, 그 결과는 Table 6과 같다. 최적화된 CRF 개체명 인식 F1 점수가 기본 명사구 인식 자질 집합보다 3.29 높은 65.40로 나타났다. 또한, 학습 자료에서 수집한 개체명 사전으로 매칭한 경우보다 최적화된 CRF 개체명 인식이 1.6 더 높은 F1 점수를 보였다. 사전을 이용한 개체명 인식은 재현율 측면에서 좋은 점수를 보이지만, 정확률은 낮은 편이다. 기본 명사구 인식 기법의 성능이 사전 사용 단순 매칭보다 낮은 이유는 선택된 자질이 재현율을 저하시켰기 때문이다.

Table 6. Performance evaluation

	Acc.	Prec.	Recall	F1
CRF: feature selection	93.69%	65.65%	65.15%	65.40
CRF: base NP chunking	92.57%	68.36%	56.91%	62.11
Dictionary matching	92.14%	53.08%	79.94%	63.80

작업자 상호 의견 일치도를 평가하는 카파 계수를 이용하여 인식 시스템의 효용성을 검증하였다. N 은 토큰 라인의 수, n은 비교 평가자 수, k는 개체명 태깅 종류의 수, $\overline{P}$는 평가자들의 평가가 일치할 확률, $\overline{P_e}$는 평가자 평가가 우연하게 일치할 확률, $i = 1, \cdots N$, $j = 1, \cdots k$ 일 때, 카파 계수는 Equation (3)과 같이 계산된다.

$$Kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$$

$$\overline{P} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)$$

$$\overline{P_e} = \sum_{j=1}^{k}\left(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij}\right)^2, \quad 1 = \frac{1}{N}\sum_{j=1}^{k} n_{ij} \qquad (3)$$

평가 데이터는 N=72,342, n=2, k=11, $\overline{P}$=0.9359, $\overline{P_e}$ =0.8046 으로 카파 계수는 0.67로 나타났다. 개체명 태깅 시스템은 두 사람이 작업한 결과에 대한 의견 일치도인 0.6보다 높은 카파 계수를 보여주었다. 제안한 시스템은 사람이 작업한 수준과 유사한 결과를 얻을 수 있기 때문에 개체명 자동 인식을 통해 특허 개체명 말뭉치를 구축하는데 활용할 수 있다.

6. 결 론

특허 문서에서 기술명, 서비스명, 제품명에 대해 태깅한 데이터를 사용하여 태그 부착 문제에 뛰어난 성능을 보이는 것으로 알려진 조건부 랜덤 필드 모델을 이용하여 특허 개체명 인식 시스템을 구축하였다. 개체명 태깅된 특허 말뭉치로부터 학습용 데이터와 실험 데이터를 10-등분 교차 검증 방식으로 분할하여 학습 및 실험을 수행하였다. 7만 어절의 평가용 데이터로 개체명 인식 정확도를 평가한 결과 정확률이 93.69%, F1 점수가 65.40로 나타났다. 이는 단순 사전으로 매칭하는 방식이나, 기본 명사구 인식에 의한 자질 선택 방식보다 F1 점수가 각각 1.6, 3.29 만큼 향상된 것이다. 특히, 사람이 수작업으로 특허 개체명 말뭉치 구축했을 때 두 작업자간의 의견 일치도가 0.6인데 비해, 조건부 랜덤 필드에 의한 기계 학습 기법으로 태깅한 결과를 수작업 결과와 비교한 카파 계수 측정 결과는 0.67로 수작업 태깅한 결과와 유사한 수준임을 알 수 있다.

References

[1] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Lingvisticae Investigationes*, Vol.30, No.1, pp.3-26, 2007.

[2] S. Cucerzan and D. Yarowsky, "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence," *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pp.90-99, 1999.

[3] Y. Wang, "Annotating and Recognising Named Entities in Clinical Notes," *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pp.18-26, 2009.

[4] H. Gurulingappa, B. Muller, R. Klinger, H. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. Friedrich, "Patent Retrieval in Chemistry based on Semantically Tagged Named Entities," *Proceedings of the Eighteenth Text RETrieval Conference (TREC 2009)*, pp.1-9, 2009.

[5] D. Eisinger, G. Tsatsaronis, M. Bundschus, U. Wieneke, and M. Schroeder, "Automated Patent Categorization and Guided Patent Search using IPC as Inspired by MeSH and PubMed," *Journal of Biomed Semantics*, Vol.4, Suppl. 1, 2013.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proceedings of the 18th International Conference on Machine Learning*, pp.282-289, 2001.

[7] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *Machine Learning*, Vol.4, No.4, pp.267-373, 2011.

[8] H. Wallach, "Conditional Random Fields: An Introduction," CIS Technical Report MS-CIS-04-21, University of Pennsylvania, pp.1-9, 2004.

이 태 석

e-mail : tsyi@kisti.re.kr
1995년 경원대학교 전자계산학과(공학사)
2005년 고려대학교 컴퓨터학과(이학석사)
2016년 국민대학교 컴퓨터공학부(박사수료)
1997년~현 재 한국과학기술정보연구원 정보서비스실 책임연구원
관심분야 : 정보검색, 정보추출, 기계학습

신 수 미

e-mail : sumi@kisti.re.kr
1997년 홍익대학교 컴퓨터공학과(학사)
2014년 홍익대학교 컴퓨터공학과(박사수료)
1997년~현 재 한국과학기술정보연구원 정보서비스실 선임연구원
관심분야 : 데이터마이닝, 추천서비스

강 승 식

e-mail : sskang@kookmin.ac.kr

1986년 서울대학교 정보컴퓨터공학부 (학사)

1988년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년~2001년 한성대학교 정보전산학부 부교수

2001년~현 재 국민대학교 컴퓨터공학부 교수

관심분야 : 한국어정보처리, 자연어처리, 정보검색, 기계학습