심층학습을 이용한 음절태깅 기반의 외래어 인식 시스템

박호민†○, 김창현‡, 천민아†, 노경목†, 김재훈†
†한국해양대학교, 컴퓨터정보공학과
‡한국전자통신연구원
homin2006@hanmail.net, chkim@etri.re.kr, kmq7542@gmail.com, minah0218@kmou.ac.kr, jhoon@kmou.ac.kr

Loanword Recognition Using Deep Learning

Ho-Min Park†○, Chang-Hyun Kim‡, Min-Ah Cheon†, Kyung-Mok Noh†, Jae-Hoon Kim†
†Department of Computer Engineering, Korea Maritime and Ocean University
‡Electronics and Telecommunications Research Institute

요 약

외래어란 외국어로부터 들어와 한국어에 동화되고 한국어로서 사용되는 언어이다. 나날이 우리의 언어사용 문화에서 외래어의 사용 비율은 높아져가는 추세로, 전문분야에서는 특히 두드러진다. 그러므로 더 효율적이고 효과적인 자연언어처리를 위해서 문서 내 외래어 인식은 중요한 전처리 과정이다. 따라서 본 논문에서는 bidirectional LSTM(이하 bi-LSTM)-CRF 모형의 심층학습을 이용한 음절태깅 기반의 외래어 인식 시스템을 제안한다. 제안하는 시스템의 외래어 인식 학습 과정은 다음과 같다. 첫째, 학습용 말뭉치 자료의 한글 음절들과 공백, 마침표(.)를 토대로 word2vec을 통해 학습용 피쳐(feature) 자료를 생성한다. 둘째, 학습용 말뭉치 자료와 학습용 피쳐 자료를 결합하여 bi-LSTM 모형 학습 자료를 구축한다. 셋째, bi-LSTM 모형을 거쳐 학습된 결과물을 CRF 모형에서 로그 가능도(log likelyhood)와 비터비(Viterbi) 알고리즘을 통해 학습 결과물을 내놓는다. 넷째, 학습용 말뭉치 자료의 정답과 비교한 뒤 모형 내부의 수치들을 조정한다. 다섯째, 학습을 마칠 때까지 반복한다. 본 논문에서 제안하는 시스템을 이용하여 자체적인 뉴스 수집 자료에 대해서 높은 정확도와 재현율을 기록하였다.

주제어: 외래어, 음절태깅, bi-LSTM-CRF, word2vec

1. 서론

한글 및 한국어 자연언어처리에 있어서 해당 문서의 주제어를 찾아내는 건 그 문서를 파악하는데 중요한 요소이다. 따라서 문서 분류(classification)나 문서 조직화(organization), 또는 주제어 추출(keyword recognition) 등의 작업에 있어서 반드시 전처리(preprocessing) 과정에 포함하는 경우가 많다. 그렇기에 효율적인 주제어 찾기는 자연언어처리에 있어서 중요한 주제이며 관련 연구가 활발하게 현재까지도 진행되고 있다[1-3].

일반적으로 한 문서에서 중요한 뜻을 가지는 단어의 품사는 명사이며, 그래서 주제어 추출은 해당 문서의 명사들을 추려내 그 중 중요도가 높은 명사를 찾는 일로 간주된다[4].

여러 다양한 분야에서 인터넷을 통한 외국과의 활발한 학문적 교류로 인해서 사회 전반적으로 외래어를 사용하게 되는 경향이 두드러지고 있다. 그러나 외래어는 사용 분야와 적용 범위, 새롭게 만들어지는 주기가 짧고 다양할 수밖에 없다. 그렇기에 사전에 등재될 때까지 외래어는 미등록어가 된다. 이러한 현상은 미등록어 문제를 일으키고 그것은 한국어 자연언어처리에 있어서 큰 걸림돌이다[5]. 따라서 외래어 인식은 중요하고 반드시 필요한 전처리 과정이라고 할 수 있다.

본 논문에서는 심층학습을 이용한 음절태깅 기반의 외래어 인식 시스템을 제안한다. 제안하는 시스템은 심층학습 모형인 bi-directional LSTM과 CRF 모형을 이용하여 외래어를 인식할 문서의 음절마다 태그를 부착하여 외래어를 인식한다.

적용되는 외래어의 범위로는 영어만을 규정했으며 그 이유는 외래어 중 가장 많이 사용되고[6] 현재까지의 관련 연구들 역시 영어 방면에 집중되어 있기 때문이다[5, 7-8].

본 논문의 구성은 다음과 같다. 2장에서 제안하는 심층학습을 이용한 음절태깅 기반의 외래어 인식 시스템에 대하여 소개하고, 3장에서는 결론 및 향후 연구에 대해 기술한다.

2. 심층학습을 이용한 음절태깅 기반 외래어 인식

심층학습은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습 알고리즘의 집합으로 정의된다. 풀어서 설명하면 큰 틀에서 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야라고 이야기할 수 있다[9].

텐서플로(Tensorflow)는 기계학습과 심층학습 모형을 프로그래밍 적으로 구현하기 위해 구글(Google) 사에서 제작 및 배포한 오픈소스 라이브러리이다. 프로그래밍 언어 중에서도 파이썬(Python) 위주로 제작되었으나 Java, Go, C언어 버전도 제공하며 운영체제별 Ubuntu, Mac OS X, Windows 버전 세 가지를 제공한다[10].

젠심(gensim)은 파이썬 프로그래밍 언어에서 문서의 분석 및 분류에 특화되어있는 오픈소스 라이브러리 모듈이다[11].

word2vec은 2013년 구글 사에서 Tomas Mikolov 외 (2013)[12]에서 제안된 단어 임베딩을 위한 기계학습 모형이다. 그를 위한 네트워크 모형 두 가지의 이름은 각각 CBOW(Continuous Bag-of-Words)와 skip-gram 모형이다.

그림 1에서의 CBOW 모형은 크게 입력층(input layer), 전개층(projection layer), 출력층(output layer)으로 이루어져 있으며 문서 내의 각각의 단어마다 앞·뒤의 단어들을 원-핫(one-hot) 변환을 실시하여 목표 단어를 맞추기 위한 학습망을 구성한다.

그림 2에서의 skip-gram 모형은 CBOW 모형과 마찬가지로 입력층, 전개층, 출력층으로 구성되지만 그와는 반대로 각각의 단어를 근거로 앞·뒤에 어떤 단어들이 등장할지 예측하여 맞추기 위한 학습망을 만들어서 멀리 떨어져있는 단어일수록 낮은 확률로 택하는 방법을 사용한다.

따라서 CBOW 모형과 Skip-gram 모형은 서로 반대의 방법을 취하고 있다고 볼 수 있으며 본 논문에서 제안하는 시스템에서는 Skip-gram 모형을 차용하여 음절 임베딩을 실시하였다.

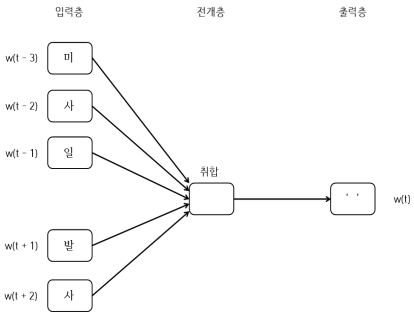

그림 1. CBOW 모형

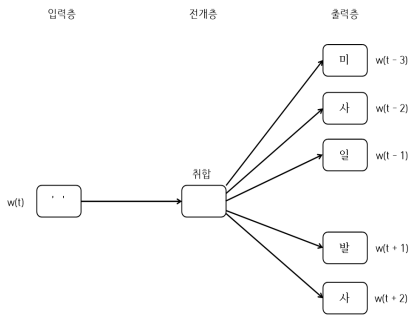

그림 2. Skip-gram 모형

그림3 에서의 LSTM(Long Short Term Memory)이란 인공신경망(artificial neural network)중 하나인 순환(recurrent)인공신경망(이하 순환신경망)에서 사용되는 뉴런 구조의 한 종류이다. 이전 단계의 출력이 다음 단계의 입력 자료가 되는, 기존 순환신경망의 한계인 장기의존성 문제를 해결하기 위해 고안되었다[13].

그림 4에서의 bi-LSTM이란 LSTM으로 구성된 순환신경망의 학습에 있어서 입력 자료에 대한 정보량을 증가시키는 목적으로 소개된 알고리즘으로, 기존의 순환신경망에서는 미래에 입력될 정보가 현재 상태 정보에게 영향을 줄 수 없었지만 bi-LSTM은 입력열의 정방향과 역방향으로 순환하는 두 개의 학습망을 통해서 심층학습을 진행한다[14].

본 논문에서는 학습 말뭉치 자료에 대한 학습 모형으로 bi-LSTM을 사용하여 입력되는 문장의 구문 정보를 양방향으로 제공하여 좀 더 효율적인 음절태깅을 수행하였다.

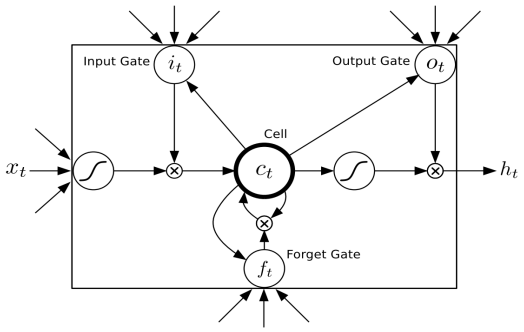

그림 3. LSTM 구조

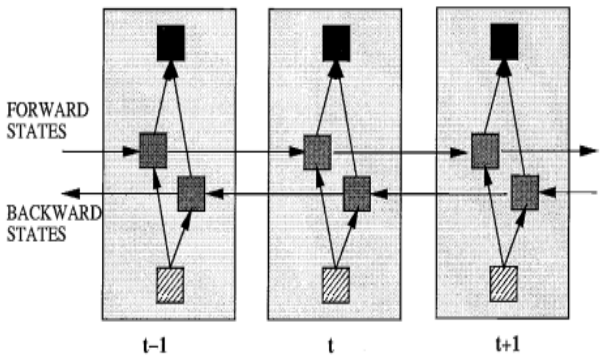

그림 4. bi-directional LSTM 망 구조 예시

CRF(Conditional Random Field)는 통계적으로 모형을 제작하는 방식으로 기계학습에서 모형의 구조에 따라 예측을 판단하는데 이용된다. 이론의 세부 내용으론 방향성이 없는 그래프를 제작하는데 그 그래프의 정점(vertex)은 입력되는 자료열들이 구성하며 각 정점마다 다른 정점으로 넘어갈 상태 전환 확률이 정점들을 연결하는 간선(edge)이 되어 그래프를 구성한다[8].

본 논문에서 제안하는 시스템에서는 bi-LSTM을 통해 학습되어 나온 결과값들을 이용하여 정점과 간선을 로그가능도(log likelyhood)로 구한다. 구하고난 뒤, 선형체인(linear chain) 형태 그래프를 제작하여 음절태깅을 진행한다.

비터비 알고리즘은 결과값이 나오지 않은 은닉 상태열로 구성된 그래프에서 가장 가능성이 높은(most likely) 예측 결과열을 생성하는 동적 프로그래밍 알고리즘이다. 은닉 상태열의 시작부터 끝까지 각 단계의 확률과 상태 전이 확률을 사용해서 해당 단계의 가장 가능성이 높은 결과를 선택하는데 그 선택된 결과들이 모여 예측 결과열(Viterbi path)이 된다[9].

본 논문에서 제안하는 시스템에서는 bi-LSTM을 통해 학습되어 나온 결과를 CRF를 통해 선형 체인 그래프로 만든 뒤 해당 알고리즘을 사용하여 각 학습 단계의 최종 결과물인 예측 결과열을 생성한다.

3. 심층학습 음절태깅 기반 외래어 인식 시스템

문서 내에서 외래어를 인식하기 위해 일반적인 어절이나 단어적, 형태소적 접근이 아닌 음절에 따른 한국어('K' 태그)와 외래어('E' 태그) 분류로 접근하였다. 본 논문의 시스템에서는 젠심 모듈의 word2vec 모형으로 음절 임베딩을 시행하는 전처리 단계와 그러한 전처리 결과물로 실질적인 음절태깅을 위한 학습을 진행하는 bi-LSTM 모형, 학습 결과물을 다듬고 차원 축소(dimentionality reduction)를 진행하여 예측 결과를 도출해내 외래어를 인식하는 CRF 모형과 비터비 알고리즘을 이용한 후처리로 이루어져 있다.

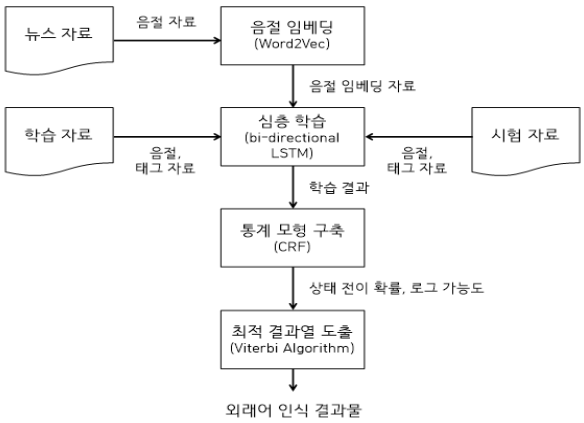

그림 5. 음절태깅을 이용한 외래어 인식 시스템

word2vec 모형에서 학습 자료로 사용된 뉴스 자료는 자체 수집한 뉴스 문서들을 이용하였다. 가능한 다양한 자료를 담기위해 문화, 경제, 연예, 국제, 과학, 지역, 정치, 사회, 스포츠의 9개 분야를 사용하였다. 분량은 약 2GB 정도이며 연합뉴스의 2017년도 분을 사용하였다.

음절 임베딩 단계에서는 skip-gram 방식을 이용한 word2vec 모형을 이용한다. 입력되는 각 음절에 대해 앞·뒤에 어떤 음절이 있을지를 예측한다. 가까이 있을수록 그 확률이 높아지고 멀리 있을수록 낮게 책정된다.

bi-LSTM 모형에서 학습 및 시험 자료로 사용된 자료는 자체 제작한 1만여 문장의 뉴스 보도 자료와 정답 자료를 사용했으며 80%를 학습에 사용했고 나머지 20%로 시험을 진행했다. 정답으로 쓰인 태그 종류는 총 네 가지로 표 1과 같이 설정하였다.

표 1. 태그의 종류

종류	설명
K	한국어 음절
E	외래어 음절
' '	띄어쓰기 된 부분
'.'	마침표

심층학습 단계에서는 학습 자료를 음절 단위로 분리하여 순차적으로 bi-LSTM 모형에 입력받는다. 그림 6과 같이 음절을 입력받아 학습 결과를 결합하여 다음 단계인 CRF 모형의 자질을 만들어낸다.

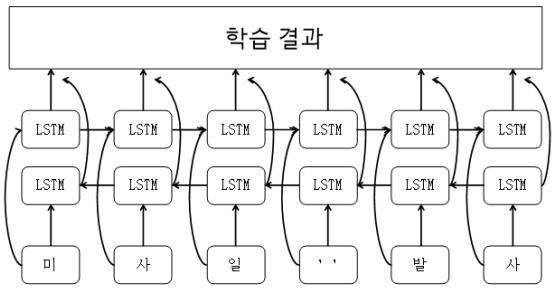

그림 6. bi-LSTM 학습 예시

통계 모형 구축 단계에서는 CRF 모형을 이용하여 선형 체인 형태의 그래프를 구성한다. 각 음절에 따른 정점과 학습 결과에 차원 축소를 진행한 값들을 간선으로 이용한다. 간선의 수치 계산 방법은 로그 가능도를 사용하며 그것을 최적 결과열 도출 단계에서 비터비 알고리즘을 적용할 때 각 정점에 대한 상태 전이 확률로 사용한다. 그렇게 비터비 알고리즘을 이용하여 동적 프로그래밍 방식으로 그림 7처럼 한 단계 한 단계 음절들에 대한 태그를 결정한 태그 예측열을 최종 결과물로 제출하게 되고 학습 과정에서는 정답과 결과물을 비교하여 학습율(learning rate)에 따라 내부 수치를 재조정하여 학습을 지속해 나간다.

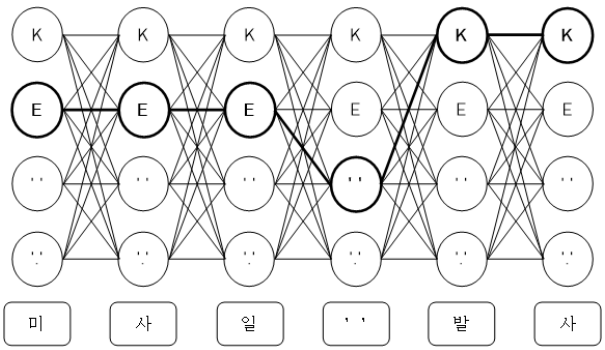

그림 7. CRF 그래프 구축 및 Viterbi Path 추적

4. 실험 결과

word2vec 모형 학습에 이용한 뉴스 자료는 연합뉴스의 2017년도 분의 문화, 경제, 연예, 국제, 과학, 지역, 정치, 사회, 스포츠의 9개 분야를 약 2GB 정도 수집하여 사용하였다. bi-LSTM-CRF 모형의 학습에 사용된 학습 자료와 시험 자료는 자체 제작한 1만여 문장의 KBS 뉴스 보도 자료와 그에따른 외래어, 한국어 태깅 결과 자료를 사용했다. 비율은 80 : 20으로 나누어 활용했다.

평가 방식은 단순 음절태그 예측 정확도(accuracy)-정확률(precision)-재현율(recall)-f1 measure 값, 한글 음절 임베딩 피쳐의 차원 수(50개, 100개), 태그 개수(2개('K', 'E'), 4개('K', 'E', ' ', '.'))에 따라 세 가지 방법으로 진행하였다. 음절 임베딩 자료 종류는 표 2의 내용과 같다.

표 2. 제작한 음절 임베딩 자료 종류

종류	설명
버전 1	임베딩 차원 = 50, 태그 수 = 2
버전 2	임베딩 차원 = 50, 태그 수 = 4
버전 3	임베딩 차원 = 100, 태그 수 = 2
버전 4	임베딩 차원 = 100, 태그 수 = 4

표 3. 방법에 따른 학습 결과

	정확도	정확률	재현율	f1 measure
버전 1	82.21	80.65	78.44	79.53
버전 2	85.62	84.37	83.85	84.11
버전 3	87.78	85.54	83.84	84.68
버전 4	90.53	88.39	86.19	87.28

표 4. 방법에 따른 실험 결과

	정확도	정확률	재현율	f1 measure
버전 1	76.43	75.07	73.59	74.32
버전 2	77.99	75.41	74.44	74.92
버전 3	80.21	79.54	78.93	79.28
버전 4	83.55	81.79	80.17	80.97

음절 임베딩의 차원이 50차원인 것보다 100차원일 경우에 평균적으로 높은 수치를 기록했으며, 태그를 2개 사용한 것 보다 띄어쓰기와 마침표를 넣어서 최소한의 문맥적 의미를 제공한 태그가 4개인 버전이 평균적으로 높은 수치를 기록했다. 이는 외래어 인식을 위한 올바른 음절태깅에 있어서 가능한 다양한 정보가 학습 모형으로 하여금 신뢰도 높은 예측을 하게 만든다는 것을 의미한다.

5. 결론

본 논문에서는 심층학습을 이용한 음절태깅 기반의 외래어 인식 시스템을 제안하였다. 해당 시스템은 파이썬 프로그래밍 언어의 젠심 모듈을 이용해 word2vec 모형을 제작하여 한글 음절 임베딩의 피쳐를 제작하였고, 제작한 한글 음절에 대한 음절 임베딩 자료를 bi-LSTM과 CRF 모형을 이용하여 문서의 음절마다 'K'(한국어) 태그, 'E'(외래어) 태그를 부여해 외래어 인식을 수행한다.

제작한 시스템 내부의 word2vec 모형을 위한 학습용 자료로써 자체 수집한 뉴스 자료를 이용하였고, bi-LSTM-CRF 모형을 위한 학습용 자료로써 자체 제작한 음절태깅을 진행한 뉴스 말뭉치를 사용하였다.

하지만 가장 치명적인 약점은 학습 자료에 존재하지 않았던 외래어를 만났을 때 인식율이 낮았으며, 학습 단계에 있어서 어려움은 각 단계마다 과적합(overfitting)이 될 수 있다는 것과 말뭉치 내 외래어 음절보다 한국어 음절의 절대 개수의 높은 차이로 인해 음절 태깅의 결과가 한국어 태그로 편중(bias)될 수도 있다는게 있었다. 첫 번째로 제시한 약점과 편중 문제는 학습 말뭉치 자료의 추가적인 확보 및 정제로 어느정도 해결할 수 있을거라 생각하며 과적합 문제는 학습율 조정 및 학습율 감퇴 적용 등을 추가적으로 연구할 예정이다.

본 논문에서 제안하는 시스템의 개선을 위하여 향후 연구로 외래어 사전 추가, 학습 말뭉치 추가 확보 및 정제, 음절에 대해 추가적인 정보 제공 방법 연구 등을 진행하여 음절태깅을 이용한 외래어 인식 시스템의 성능을 향상시킬 계획이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

참고문헌

[1] 배경만, 김성현, 고영중, 김종훈, "자연어 기반 인터페이스에서 개체명 패턴을 이용한 효과적인 개체명과 주제어 인식 방법", 한국정보기술학회논문지, 제12권 제1호, 121-129, 2014.

[2] 주길홍, 이주일, 이원석, "효율적인 문서 검색을 위한 연관 키워드 추출 및 확산 클러스터링 방법", 한국정보기술학회논문지, 제9권 제6호, 155-166, 2011.

[3] 유은순, 최건희, 김승훈, "TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구", 한국컴퓨터정보학회논문지, 제20권 제2호, 121-129, 2015.

[4] 안희정, 김기원, 김승훈, "복합 명사구 합성 방법을 적용한 효과적인 도서 본문 주제어 추출", 한국컴퓨터정보학회논문지, 제22권 제3호, 107-113, 2017.

[5] 오종훈, 최기선, "은닉 마르코프 모델을 이용한 음차표기된 외래어의 자동인식 및 추출 기법", 인지과학, Vol.12 No.3, 19-28, 2001.

[6] 조남호, "한국어의 외래어 수용과 대응", 인문과학연구논총, 제35권 3호, 11-38, 2014.

[7] 박종혁, "유사 외래어 검출 알고리즘의 성능 향상", 충북대학교 석사학위논문, 2004.

[8] 고숙현, "문맥을 고려한 유사 외래어 검출 알고리즘", 충북대학교 석사학위논문, 2007.

[9] Y. Bengio, A. Courville, and P. Vincent., “Representation Learning: A Review and New Perspectives,” IEEE Trans. PAMI, special issue Learning Deep Architectures, 2013.
[10] [online]https://www.tensorflow.org, 2015.
[11] [online]https://radimrehurek.com/gensim, 2011.
[12] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”, 2013.
[13] Sepp Hochreiter and Jurgen Schmidhuber, “Long Short-Term Memory”, 1997.
[14] Mike Schuster and Kuldip K, “Bidirectional Recurrent Neural Networks”, 1997.
[15] John Lafferty, Andrew McCallum and Pernando C.N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, 2001.
[16] G. David Forney, Jr., “The Viterbi Algorithm: A Personal History”, 2005.