특집 01

언어자원 구축과 단어 임베딩 기법을 이용한 딥러닝 기반의 자연어 분석 및 생성

강승식 (국민대학교)

목 차

1. 서 론
2. 자연어처리 분야의 특성과 언어처리 기술
3. 딥러닝 기법을 이용한 한국어 언어처리 기술
4. 자연어 분석과 생성 기술의 연구 방향
5. 결 론

1. 서 론

4차 산업혁명 시대에 핵심적인 기술로 사용되고 있는 신경망 모델(neural network model)은 1980년대에 신기술로 각광을 받았으나 컴퓨팅 환경이 뒷받침되지 못함으로 인하여 실용적인 기술로 발전하지 못하는 한계가 있었다. 이 모델은 2000년대에 이르러 컴퓨터 하드웨어의 발전으로 처리 속도 및 기억장치 용량의 획기적인 증가, GPU를 이용한 병렬처리와 클라우드 컴퓨팅 환경의 변화에 따라 심층 신경망(deep neural network)이라는 딥러닝 기술로 발전하였다[1]. 초기의 딥러닝 기술은 전통적인 학습 기법으로 성능 향상이 어려웠던 영상처리 분야에서 이미지 검색과 이미지 분류 분야에 매우 적합하였고, 특히 전처리 과정에서 수작업으로 처리해야 했던 속성 추출 및 선택(feature extraction and selection) 문제를 해결함으로써 최적화된 모델을 생성하는데 매우 큰 기여를 하였다[2,3].

딥러닝은 대규모 학습 데이터가 구축되는 모든 분야에서 최적화된 모델을 생성하기 때문에 기계학습 분야에서 획기적인 방법론으로 등장하였는데, 자연어 처리 분야의 연구에서도 딥러닝 기법이 보편적으로 사용되고 있다[4,5]. 딥러닝을 자연어처리 연구에 적용하는 연구는 워드 임베딩(word embedding)을 기반으로 하고 있으며, 자연어처리의 핵심 연구 주제인 문장의 분석 및 생성 문제를 해결하는데 매우 중요한 방법론으로 사용되고 있다. word2vec 알고리즘으로 시작된 워드 임베딩 기법은 문맥정보에 따라 단어의 의미가 달라지는 문맥 기반의 임베딩으로 발전하여 최근에는 언어모델(Language Model)을 워드 임베딩에 적용한 ELMo(Embedding from Language Model)와 BERT(Bidirectional Encoder Representations from Transformers) 임베딩 기술이 개발되고 있다[6,7].

자연어처리 연구는 1980년대부터 기계번역을 중심으로 발전해 왔으며, 초기의 기계번역 연구는 컴퓨터 매뉴얼 등 기술적인 문서를 번역하기 위한 목적으로 시작되었다. 기계번역 방법론으로는 어휘사전과 변환사전, 구문구조 변환규칙(transformation rule), 생성사전과 생성규칙을 이용한 규칙기반(rule-based) 기법으로 시작하여 각 규칙들에 대한 사용빈도와 확률기법을 이용한 통계적 방식(stochastic approach), 예제기반(example-based) 방식으로 발전하였다[8]. 기존의 방식들은 기계번역의 품질을 향상시키는 한계를 극복하기 어려운 본질적인 문제가 있었는데, 학습데이터가 축적되면서 딥러닝 방식을 이용하게 되었고 구글 번역기를 비롯하여 기계번역에 딥러닝 기법을 적용함으로써 기존의 방식에 비해 현저하게 성능이 향상되어 실용적인 시스템으로 발전하였다[9].

딥러닝 방식의 기계번역은 피봇(pivot) 또는 중간언어 방식(interlingua approach)과 유사하게 104개 언어의 기계번역을 각각 양방향으로 개발하는 것이 아니라 하나의 통합 소프트웨어로 구현하여 n개 언어에 대해 n×(n-1)개의 번역기를 개발해야 하는 본질적인 문제를 해결하였다. 딥러닝은 기계번역 뿐만 아니라 품사 태깅(part-of-speech tagging), 구문 분석(parsing), 개체명 인식(named entity recognition), 자연어 생성(natural language generation) 등 자연어처리와 관련된 전반적인 문제를 해결하는 방법론으로 확장되어 실용적인 수준의 분석-생성 시스템을 개발하는 방법론으로 사용되고 있다.

국내에서 자연어처리 관련 연구는 전자통신연구원의 지식검색엔진 개발과 자동 통역, 엑소브레인 사업을 중심으로 수행되어 왔으며, 2017년 9월부터 2020년 12월까지 '차세대정보컴퓨팅 기술개발' 사업의 일환으로 '한국어 정보처리 원천기술'을 개발하는 연구가 수행되고 있다. 이 연구에서는 한국어 언어자원의 구축과 언어분석 모듈, 그리고 핵심 응용 기술을 개발하는 연구를 수행하고 있으며, 딥러닝 학습에 필수적인 대규모 원시 말뭉치와 더불어 품사 태깅, 구문 태깅 말뭉치를 구축하고, 언어 분석 모듈로 품사 태거, 구문 분석기, 개체명 태거 등을 개발하는 연구를 수행하고 있다. 핵심 응용 기술로는 자연어 생성 및 대화 시스템을 개발하고 있으며 이와 더불어 단어 임베딩 기술과 문맥정보를 이용한 임베딩 기법을 연구하고 있다.

2. 자연어처리 분야의 특성과 언어자원의 구축

자연어처리 분야는 언어의 특성 때문에 국외 연구기관 및 산업체의 진입 장벽이 높은 특징이 있다. 21세기 세종계획은 1997년부터 10년간 국어정보화 사업으로 한국어 말뭉치를 구축, 배포하는 목적으로 수행되었다. 2011년에는 기구축 언어자원과 우수한 고급 인력을 활용하여 수준 높은 언어자원을 구축하는 세종계획 후속사업을 계획하였으나 실행되지 못하였다[10]. 기업에서 개발하는 언어처리 기술은 상업적인 SW 개발을 추구하기 때문에 핵심 요소 기술 개발을 수행하기 어려운 한계가 있으며, 전자통신연구원을 중심으로 국가사업으로 대규모 연구개발이 이루어지고 있으나 언어자원 구축 결과물과 언어처리 모듈이 여러 학문분야에서 활용되어야 하는 요구를 충족시키기 어렵다. 이에 비해, 구글을 비롯한 등 각국의 글로벌 업체들은 차세대 핵심 요소 기술 확보에 중점을 두어 언어처리 관련 분야에 대한 연구-개발 투자를 확대해 왔다. <표1>은 2011년 세종계획 후속사업의 기획에서 언어처리

〈표 1〉 언어처리 산업의 국내외 환경 요인 분석

강점 Strength	약점 Weakness
· 한국어 정보화 요소 기술에 대한 다년간의 Knowhow 축적 · 한글 처리의 복잡도에 따른 외국업체 진입장벽이 높음 · 기구축 결과물(세종 말뭉치 등) 활용 · 우수한 고급 인력 (국어학, 언어학 분야)	· 정부, 민간 투자 부족 (공공 성격이 강하지만 투자 효과가 비가시적) · 중장기 투자 미흡 (투자 대비 효용성) · 언어자원 결과물 공개 및 검증 미흡 · 사명감을 가진 연구자 (장인 정신 필요)
기회 Opportunities	**위협 Threats**
· 언어처리 관련 SW 시장규모가 급속히 확대 · 스마트폰 환경에서 언어처리 중요성 증가 및 자체 product 실용화 가능 · 응용 SW 성능 향상에 주요 역할 인지 · 빅데이터, 딥러닝 등 NLP 분야의 수요확대	· 국외업체의 관심 증가 및 자체 해결 움직임 (원시자료 수집 용이) · 우수 인력의 해외 유출 · 수요자 구입 꺼림 (무료 사용을 원함) · 자체 product 아닌 보조 역할

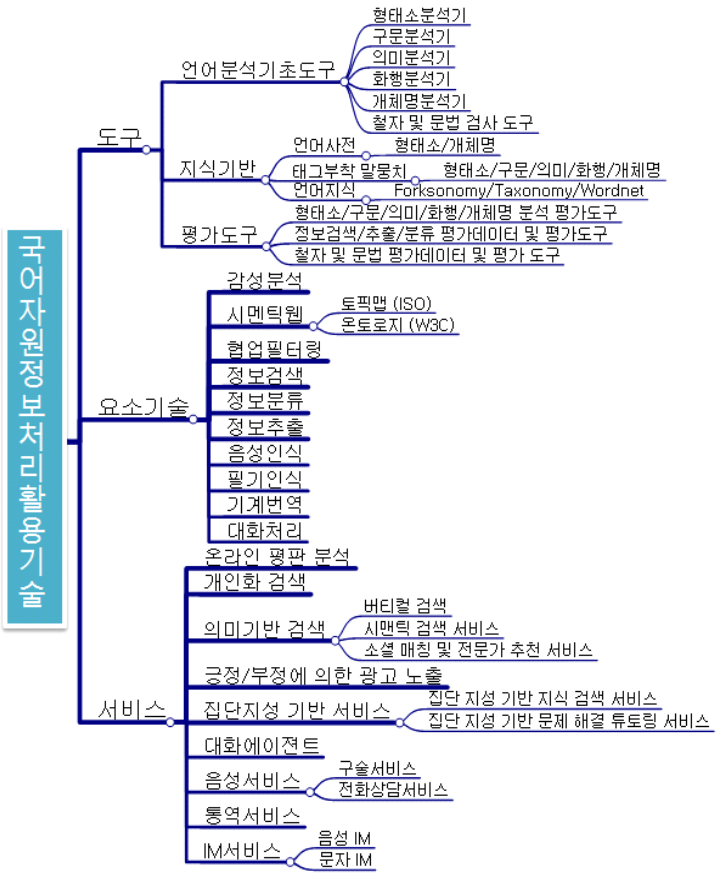
산업의 국내외 환경 요인을 SWOT 기법으로 분석한 것이다.

자연어처리 분야는 (그림 1)과 같이 어휘사전과 언어자원 구축, 그리고 언어 분석 모듈을 기반으로 응용 시스템을 개발하는 내용으로 구성된다. 특히, 한국어 언어자원은 언어처리 연구 분야의 특성을 반영하여 체계적으로 구축의 필요성이 증가하고 있다[10]. 그 이유를 살펴보면, 언어자원의 효율적인 관리, 보급, 확산 체계의 구축이 필요하다. 영어권에서는 다양한 언어자원들을

(그림 1) 자연어처리 분야의 연구 주제: 언어자원과 분석 모듈

구축해 왔으며, 미국 펜실베니아 대학의 LDC(Linguistic Data Consortium)에서 언어자원을 체계적으로 관리하여 배포하고 있다. 이와 더불어, 디지털 환경에서 대용량 한국어 언어자원 구축의 필요성이 높아지고 있는데, 한국어의 경우에는 정확도 및 신뢰도가 높은 한국어 언어자원이 절대적으로 부족하다. 이에 비해, 타 언어의 경우에는 각종 언어자원을 구축하고 홈페이지를 통해 각 언어에 관한 어휘, 용례, 빈도 정보 등 다양한 언어정보를 제공하고 있다.

산업체에서는 각자 개별적으로 해당 응용 목적만을 위해 구축하여 사용하고 있는데 국어정보 산업의 진흥을 위해서는 관련 산업체에서 공통적으로 필요로 하는 국어 기초자원의 구축이 필요하다. 또한, 기구축된 국어자원들에 대해 연구자들이 편리하게 활용할 수 있도록 다양한 언어처리 도구 개발이 필요하다. 국어자원을 연구자들이 편리하게 활용할 수 있으려면 연구자들이 활용하기 편리한 형태로 가공하여 제공하고, 각종 국어자원을 활용하는데 필요한 도구를 제공하며, 연구자들이 활용 중에 불편한 점을 반영하여 도구의 기능을 발전시키고 그 성능을 보완하여 사용자 편의성을 추구하는 작업이 지속되어야 한다. 마지막으로, 언어자원의 호환을 위한



(그림 2) 언어자원 및 언어처리 모듈과 그 활용 분야

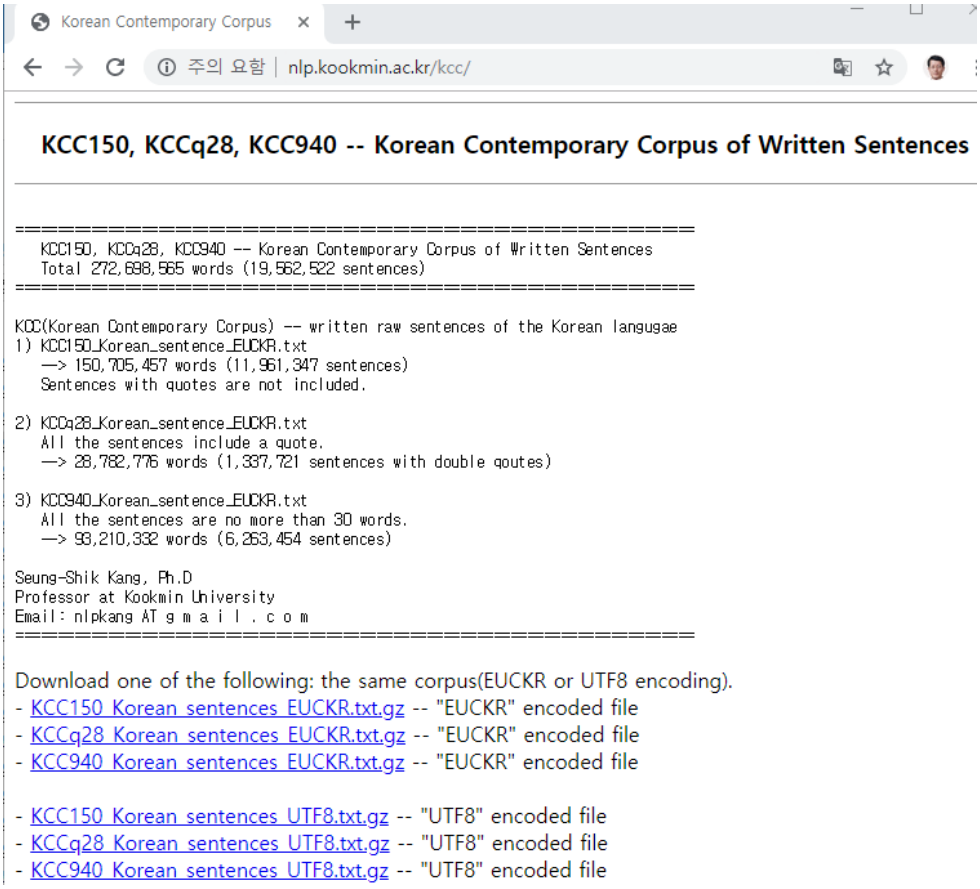
〈표 2〉 한국어 언어자원의 구축 및 공개

공개 언어자원	공개 장소	공개 내용
KoNLTK platform, github	http://konltk.org/ https://github.com/konltk/	한국어 정보처리 원천 기술 공개 플랫폼 (언어자원, 언어처리 모듈 등)
대규모 원시 말뭉치 (KCC150, KCCq28, KCC940)	http://nlp.kookmin.ac.kr/kcc/	한국어 원시 말뭉치 구축 및 공개 2억7천만 어절(2,000만 문장)
언어자원 활용 웹서버 구축	http://cqpweb.kr/	CQPweb in Korea
자동띄어쓰기 학습및 평가용 데이터	https://sites.google.com/site/koreanlp2018/task-1	학습 데이터 15,000 문장, 평가 데이터 1,000 문장
한국어 복합명사 분해 학습 및 평가용 데이터	https://sites.google.com/site/koreanlp2018/task-2	복합명사 분해 학습 및 평가용 데이터 구축
2018 차세정 언어 처리 경진대회	https://sites.google.com/site/koreanlp2018/home	자동 띄어쓰기, 복합명사 분해 경진대회, 학습데이터 및 평가데이터, 제출 시스템, 평가결과 공개 등
한국어 자동 띄어쓰기 모듈	https://github.com/ask4git/BUFS_KoSpacing	한국어 자동 띄어쓰기 모듈의 소스와 리소스 공개 (부산외대)
한국어 복합명사 분해 모듈	https://github.com/hyunyoung2/Hyunyoung2_Korean_Compound_Noun_Decomposition	복합명사 분해 모듈소스 공개
한국어 품사 태거	http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/	한국어 품사태깅 말뭉치를 이용한 품사태거 구현

효율적인 관리 체계로 기구축된 언어자원과 신규 구축되는 자원, 그리고 자원을 편리하게 활용하기 위한 언어처리 도구를 관리하고 배포하는 관리 체계가 필요하다.

언어자원은 단순히 대량의 원시문장들을 모아 놓은 원시말뭉치와 그 말뭉치에 품사 태그, 구문 태그, 의미 태그, 개체명 태그 등을 부착하는 태깅 말뭉치가 있다. 이러한 언어자원들은 해당 언어의 특성을 분석하거나 의미가 있는 정보를 추출하여 가공하는데 필수적이다. 4차 산업혁명에서 매우 중요시되는 지식처리 엔진을 개발하는데 있어서 언어자원와 언어처리 기술은 핵심적인 요소이며, 언어처리 기술을 이용하여 개발되는 응용 시스템의 개발은 언어자원의 구축과 언어처리 기술을 기반으로 한다. (그림 2)는 언어자원과 관련된 언어처리 기술을 언어처리 도구와 요소기술, 서비스로 나누어 도식화한 것이다 [10].

언어자원 구축과 언어처리에 관한 연구는 기계번역으로 시작되어 정보검색 분야에서 색인, 문서 분류, 클러스터링, 문서요약 등에 관한 주제로 확대되었다. 자연어 처리 연구가 기계번역, 정보검색 등 응용 시스템 개발을 목적으로 진행됨으로 인하여 언어자원 구축과 언어분석, 생성 등 본질적인 문제를 해결하는데 체계적인 연구가 수행되지 못하는 제약이 있었다. 빅데이터 분석과 딥러닝 기술이 중요한 기술로 등장하면서 한글 텍스트 데이터 분석 기술이 매우 중요한 핵심 요소 기술로 인식되었고, 이에 따라, 과학기술정통부가 지원하는 '차세대 정보컴퓨팅 사업'에서 2017년 9월부터 자연어처리 분야에 대한 국책 연구과제가 수행되는 계기가 되었다.



(그림 3) 한국어 원시 말뭉치 2억7천만 어절(1950만 문장)

정보통신 분야의 '원천기술 연구개발'을 목적으로 하는 '차세대 정보컴퓨팅 사업'의 '한국어 정보처리 원천기술' 연구과제는 연구자들이 개별적으로 연구-개발해 왔던 국내 자연어처리 분야의 연구자들이 체계적으로 수정, 보완, 확장 및 완성도를 높여서 언어자원과 언어처리 기술을 공개하고, 딥러닝 등 최신기술을 한국어 언어처리 연구에 적용하는 연구를 목적으로 진행되고 있다. <표 2>는 '한국어 정보처리 원천기술' 개발사업의 1단계 연구의 성과물로 구축된 언어자원들을 공개한 내용이다[11].

이 과제의 1단계 연구결과물로 구축된 2억7천만 어절의 원시 말뭉치는 2019년 1월에 공개되었으며, 딥러닝을 위한 워드 임베딩 등 대규모 말뭉치가 필요한 연구에 자유롭게 사용할 수 있도록 하였다[1]. 말뭉치를 구성하는 문장들은 주로 뉴스기사에 관한 것이다. KCC150은 1억5천만 어절(1200만 문장) 규모로 일반적인 뉴스기사 문장들로 구성되었고, KCCq28은 뉴스기사에서 인용이 포함된 문장들을 별도로 모아서 2천8백만 어절(130만 문장) 규모로 구축한 것이다. KCC940은 인용이 포함된 문장을 별도로 구분하

1) http://nlp.kookmin.ac.kr/kcc/

(그림 4) CQPweb의 한국어 언어자원 활용 사이트

지 않고 9천3백만 어절(620만 문장) 규모로 추가 구축한 원시 말뭉치이다. 이 말뭉치들은 모두 2억7천만 어절(1950만 문장) 규모의 원시 말뭉치로 현대 한국어 문어체 문장들로 구성되었다. 이 말뭉치는 뉴스기사들을 대상으로 하였기 때문에 한국어의 다양한 문장과 어휘들을 포괄하지 못하는 제한이 있으나 한국어의 다양한 언어 현상을 연구하는데 매우 유용하게 사용될 것으로 기대된다.

딥러닝 언어처리 기술에서는 단어 임베딩 기법으로 단어 벡터를 구하기 위해서는 대규모 원시 말뭉치가 필요하며, 영어의 경우에 언어 분석 모듈과 언어자원의 공유가 매우 활성화되어 있다. 독일어의 경우에도 국책 사업으로 D-SPIN이라는 "A German Infrastructure for Language Resource and Tools"를 추진하고 있으며, 유럽 각국의 언어처리 프로젝트로 CLARIN이라는 과제를 추진하여 언어자원 및 언어처리 모듈을 개발하고 있다.[2] 딥러닝 기술이 자연어처리 연구에서 매우 활발하게 진행되면서 각국의 언어자원을 체계적으로 구축하여 제공하고 있으며, 영어에 대해서는 LDC를 중심으로 매년 지속적으로 다양한 분야의 각종 말뭉치을 구축하고 있으며, 샘플 말뭉치는 무료로 제공하고 대용량 말뭉치는 유료 라이센스로 제공되고 있다.[3] nltk와 kaggle 사이트[4]에는 기계학습을 위해 100여개의

2) https://weblicht.sfs.uni-tuebingen.de/
3) https://www.ldc.upenn.edu/
4) http://www.nltk.org/, http://www.kaggle.com/

학습 데이터를 제공하고 있으며, kaggle.com은 1만6천여개의 기계학습 데이터셋을 공개하고 있는데 이중에서 텍스트 데이터는 600여개가 구축되어 있다. 딥러닝 언어처리 연구개발이 활성화되면서 언어처리 산업이 상업화되고 있고, 각국의 어휘사전과 빈도 정보, ngram 어휘데이터 및 각종 언어자원을 구축하여 판매하는 사례도 발생하고 있다.[5)]

이에 비해, 한국어에 대한 학습 데이터셋은 매우 부족하다. 세종계획 말뭉치가 연구자들에게 배포되었지만 사용자들이 국립국어원의 허락을 받는 과정이 필요하고, 1990년대에 구축된 말뭉치의 규모가 딥러닝을 위한 대규모 말뭉치로서 활용하기에 부족하다. 한국어 원시 말뭉치로 구축된 2억7천만 어절은 일반적인 한국어의 언어현상을 연구하는 목적과 더불어 딥러닝을 위한 언어처리 기술에서 매우 유용하게 활용될 것이다. (그림 4)는 '한국어 정보처리 원천기술' 연구결과물로 구축된 CQPweb[6)] 언어자원 웹서비스 사이트이다. CQPweb은 언어자원으로 구축된 원시 말뭉치와 태깅 말뭉치 등 각종 언어자원을 웹에서 web query를 통해 필요한 정보를 추출, 검색할 수 있는 웹서비스이다. 이 웹서비스는 다양한 말뭉치들로부터 어휘정보, 출현빈도, 연관관계 등의 언어정보를 정규식 기반의 질의문(CQP: Corpus Query Processor)을 통해 추출, 검색하는 기능을 제공한다.

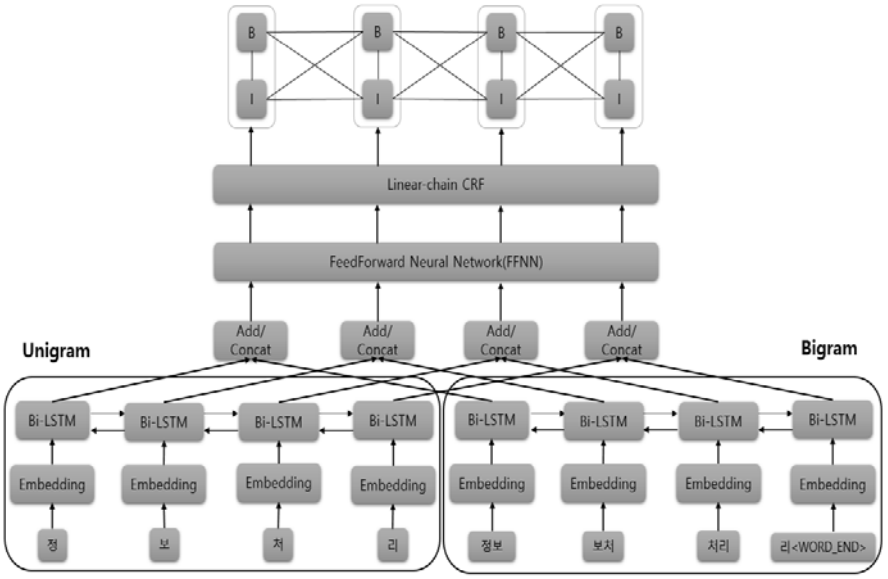
3. 딥러닝 기법을 이용한 한국어 언어처리 기술

딥러닝을 위한 언어처리에서는 단어와 문장, 또는 문서의 내용을 속성 벡터(feature vector)로 표현하는 작업이 필요하다. 기존의 자연어 처리 연구에서는 문서의 내용을 벡터로 표현하기 위하여 어휘사전(lexicon)을 구성하고 문서에 출현한 어휘들의 출현빈도(term frequency)와 문서빈도(document frequency)를 이용하여 문서 벡터(document vector)를 구성한다. 각 어휘들이 그 문서의 내용을 특징짓는 속성(feature)이 되고, 문서 벡터를 구성하는 속성 선택(feature selection)과 tf-idf 가중치 계산 기법에 의해 문서 벡터를 구성하는 것이다. 딥러닝에서 문서 벡터의 구성 방법은 단어 임베딩(word embedding) 기법을 이용하여 각 어휘들에 대한 단어 벡터(word vector)를 구하고, 단어 벡터를 기반으로 문장 벡터 또는 문서 벡터를 구하는 것이다. 딥러닝 기법을 이용하여 한국어 분석 및 생성 기술로 자동 띄어쓰기와 복합명사 분해, 의존 구문분석, 그리고 한국어 문장을 생성하는 연구를 수행하였다.

3.1 양방향 LSTM을 이용한 자동 띄어쓰기와 복합명사 분해

한국어 문장에서 어절들의 띄어쓰기 문제는 한글 맞춤법 규정에서 제시하는 띄어쓰기 규칙을 따르는데, 한국어의 경우는 영어와 달리 띄어쓰기 단위가 명확하지 않고 모호한 경우들이 발생한다. 따라서 다양한 유형의 실사례들에서 사용자들은 띄어쓰기를 무시하거나 경우에 따라서는 띄어쓰기와 붙여쓰기가 모호한 경우들이 다수 발생한다. 특히, 복합명사는 띄어쓰기와 붙여쓰기가 모두 허용될 수 있으므로 띄어쓰기가 일관성이 결여되는 경우가 매우 많다. 띄어쓰기를 무시하거나 또는 띄어쓰기의 일관성이 결여되는 문제를 해결하기 위해 자동 띄어쓰기 방법론이

5) https://www.lexicalcomputing.com/
6) http://cqpweb.kr/



(그림 5) 양방향 LSTM과 선형체인 CRF를 이용한 복합명사 분해

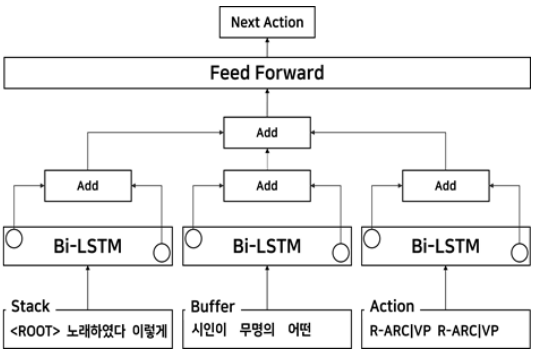
개발되었다.

초기에는 형태소 분석기와 어절 인식에 의한 띄어쓰기 방법론이 제시되었고, 원시 말뭉치가 구축되면서 ngram 방식의 통계적인 띄어쓰기 확률 계산 방법이 제안되었다. 원시 말뭉치를 이용한 기계학습 기법으로 CRF(Conditional Random Field)를 이용한 띄어쓰기 기법이 사용되었고, 최근에는 대규모 말뭉치를 이용한 딥러닝 기법을 적용함으로써 띄어쓰기 정확도를 향상시키는 연구가 진행되고 있다. (그림 5)는 음절단위 ngram 방식의 워드 임베딩 기법을 이용하여 입력 문장을 벡터화하여 딥러닝 기법으로 양방향 LSTM과 선형체인 CRF를 이용하여 자동 띄어쓰기 문제와 복합명사 분해 문제에 적용하는 방법이다 [12,13].

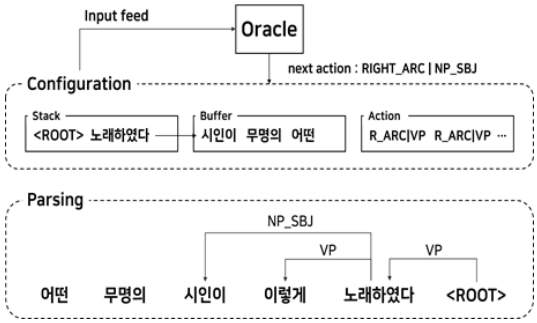
자동 띄어쓰기 실험 결과로, 음절 unigram 벡터만 사용한 정확도보다 음절 unigram과 음절 bigram을 함께 사용한 경우가 더 높은 성능을 보인다. 또한 어절 정확도, 어절 재현율, 공백 재현율, F1 Score에서도 음절 unigram만 사용한 경우보다 음절 unigram과 음절 bigram을 함께 사용한 경우에 어절 정확도와 어절 재현율, 공백 재현율 및 F1 Score에서 더 높은 성능을 보였다. 복합명사 분해 실험에서는 음절 벡터를 사용하였고, 복합명사 태그 정확도와 어절 재현율, 공백 재현율에서 음절 unigram 벡터보다 bigram 벡터를 함께 사용한 경우에 더 높은 성능을 보였다.

3.2 양방향 LSTM을 이용한 전이 기반의 의존 구문 분석

의존 문법에 의한 구문 분석은 문장을 구성하는 어절들의 지배-의존 관계를 파악하는 것으로 정의된다[14,15]. 전이 기반의 의존 구문 분석은 원거리 의존 관계 분석과 초기의 행동 오류가 전파 문제가 있다. 이러한 단점을 극복하기 위하여 다층 양방향 LSTM을 이용하여 환경 자질을 학습하는 방법을 사용하였다. Arc-eager 알고리즘



(그림 6) 행동 선택 모델의 구조



(그림 7) Arc-eager 전이 기반의 구문 분석 과정

에서 행동 선택 수는 의존 관계 수의 2배와 이동(shift), 축약(reduce)의 합이 된다. 따라서 행동 선택의 분류 성능을 높이기 위해 최대 행동 가지 수를 줄일 필요가 있고, 한국어의 지배소 후위 규칙을 적용하여 문장의 역방향으로 전이 기반 방식을 적용하였다.

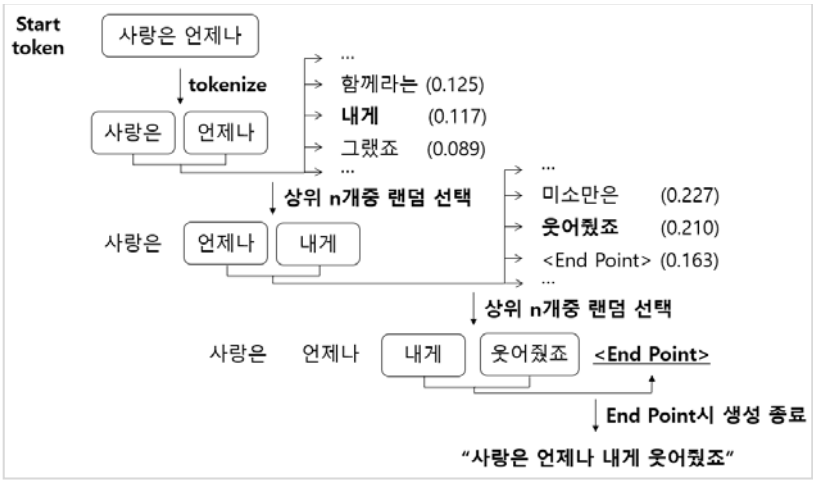
Arc-eager 전이 기반 방식에 따라 행동 선택 모델을 학습하기 위해서 (그림 6)과 같이 환경 자질을 스택과 버퍼, 그리고 이전 행동들로 선택하였다. 또한, 지배소 후위 규칙에 따라 버퍼를 역방향으로 취함으로써 학습 모델도 역방향으로 학습할 수 있도록 적용하였다. 먼 거리의 의존 관계를 학습하고 초기 행동의 오류 전파를 줄이기 위해서 문장의 순방향과 역방향을 학습하는 다층 양방향 LSTM을 사용한다. 행동 선택 모델의 결과는 한국어의 지배소 후위 규칙과 arc-eager 알고리즘에 따라 right-arc와 축약으로 한정한다. 실제 문장을 분석하기 위해 품사 정보를 부착하는 Tree Tagger[7]를 행동 선택 모델 신경망과 통합하는 방식으로 구현하였다. 행동 선택 모델은 스택, 버퍼, 이전 행동들을 입력으로 받아 다음 행동을 예측하는데 환경 정보를 추출할 수 있도록 우선 각 환경마다 양방향 LSTM을 적용한다.

(그림 7)은 arc-eager 전이 기반 알고리즘에 따라 실제 문장의 의존 관계를 분석하는 과정을 표현한 것이다. (그림 7)의 상단은 문장의 의존 관계를 VP(<ROOT>, 노래하였다), VP(노래하였다, 이렇게)와 같이 분석한 상태에서 그 시점의 환경 상태를 나타낸다. 행동 선택 모델은 환경 자질에 따라 적절한 다음 행동을 예측하고, 행동을 토대로 환경을 변화시킨다. 따라서 (그림 7)의 하단처럼 행동 선택 모델의 행동에 따라 문장의 의존 관계를 분석한다.

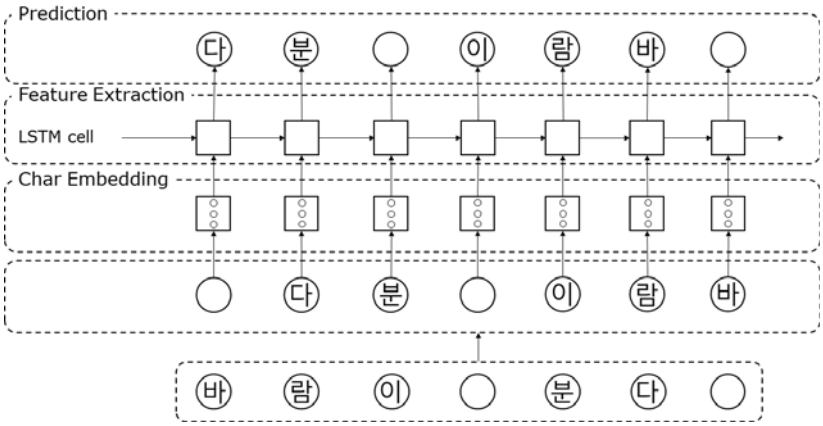
3.3 딥러닝 기법을 이용한 한국어 문장 생성

자연어 문장 생성은 문장을 생성하고자 하는 도메인과 목적에 따라 적합한 문장 생성 기법을 취하게 된다. 로봇저널리즘과 같이 스포츠 뉴스 도메인에서 경기결과에 따라 주요 내용을 뉴스 기사로 생성하거나, 지진 발생 뉴스기사, 일기예보, 기업의 분기별 보고서 등을 생성할 때는 문장의 주요 내용이 수치 데이터 형태로 주어지고 기존의 문장 데이터베이스에서 문장 템플릿을 생성하여 생성하고자 하는 주제에 적합한 문장

7) https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/



(그림 8) n-gram 기법을 이용한 문장 생성의 예



(그림 9) 한글 노래가사 생성을 위한 딥러닝 학습 구조도

을 생성하게 된다. 중국어에서 한시를 생성하거나 시, 소설, 에세이 등을 자유롭게 생성하는 방법으로는 학습 문장들을 딥러닝 기법으로 학습하고 주어진 주제 또는 키워드에 적합한 문장을 생성하는 기법이 사용된다[16]. 세익스피어 소설이나 해리포터를 학습 데이터로 하여 소설을 생성하는 간단한 방법으로 ngram 기법을 사용할 수 있다. ngram 기법을 이용한 문장 생성은 문장 s의 생성 확률 P(s)을 각 단어의 좌측 문자열이 주어졌을 때 문장 s에 대한 생성 확률을 계산하는 방법을 기반으로 한다.

$$P(s) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1w_2) \times \ldots \times P(w_{n-1}|w_1w_2w_3\ldots w_{n-2}) \times P(w_n|w_1w_2w_3\ldots w_{n-1})$$

이처럼 ngram 기법을 이용한 문장 생성은 이미 발화된 (n-1)개의 단어로부터 n번째 단어를 생성하는 방법으로 $P(w_n|w_1w_2w_3\ldots w_{n-1})$의 값이

가장 큰 w_n을 생성하게 된다. 학습 데이터로부터 ngram 확률을 구하고, 이를 기반으로 trigram 확률 기법에 의해 문장 생성 확률을 구하는 방법은 아래와 같다. (그림 8)은 ngram 확률 기법에 이용한 문장 생성 과정을 나타낸 예이다. ngram 기법을 이용한 문장 생성은 학습 데이터에 출현한 어휘들에 대해 출현 빈도가 높은, 그럴듯한 문장을 생성하는 것으로 학습 데이터셋에 출현한 ngram 데이터를 기반으로 문장을 생성한다.

딥러닝 이용한 문장 생성은 ngram 기법과 달리 학습 데이터셋으로부터 워드 임베딩을 통해 학습 모델을 구축하고, 이 학습 모델로부터 문장을 생성한다[17,18]. (그림 9)는 음절 단위의 임베딩 기법과 LSTM 언어모델을 이용한 한글 노래가사 생성 시스템의 구조와 학습 모델로부터 문자열을 생성하는 과정을 예시한 것이다.

4. 자연어 분석과 생성 기술의 연구 방향

딥러닝 기술은 CNN과 RNN이라는 기본적인 deep neural network 구조를 기반으로 양방향 LSTM, seq2seq 네트워크, GAN(Generative Adversarial Network)과 같이 문제 특성을 고려하여 다양한 응용 분야에 적합한 구조로 발전되고 있다. 언어처리 기법과 관련하여 word2vec이라는 단순히 단어 벡터를 구성하는 수준에서 발전하여, 최근에는 좌우 문맥정보를 고려하여 벡터를 구성하는 ELMo와 BERT 임베딩 기법이 개발되고 있으며, 최신 딥러닝 기술을 한국어 분석 및 생성 기술에 적용하는 연구가 진행되고 있다.

한국어 언어처리 기술은 최신 임베딩 기법과 더불어 대규모 언어자원을 기반으로 문맥정보를 이용한 상황인지 기술과 사용자 의도파악 기술을 개발하는 연구로 발전하고 있으며, 자연어처리의 핵심 응용 기술로는 자연어 추론(NLI: Natural Language Inference), NELL(Never Ending Language Learning) 등이 연구되고 있다. 언어처리 기술을 인공지능 분야에서 실제로 활용하기 위해 대화형 시스템을 개발하고, 인간과 기계의 대화에서 감정을 이해하여 인간처럼 소통하는 인공지능 기술을 개발한다. 이를 위해, 감정사전 등 언어자원을 확장 구축하고 이를 딥러닝 기법에 활용하는 연구를 수행한다. 또한, 문맥정보를 이용한 상황인지 기술과 사용자 의도파악 기술을 보이스피싱, 자살 예방, 심리 상담 등에 적용하여 현대사회에서 국가 사회적으로 매우 큰 비용을 지출하거나 심각한 문제를 야기하고 있는 사회문제 해결에 기여하는 연구를 수행하고 있다.

현재까지 구축된 언어자원과 학습데이터를 확장-보완하여 최근에 가장 중요한 이슈로 떠오르고 있는 최첨단 임베딩 기법으로 ELMo와 BERT 기술을 한국어 임베딩 기법으로 개발하고, 이를 기반으로 최신 언어처리 기술을 보이스피싱, 자살 예방 등에 적용하여 사회문제 해결에 기여하는 연구를 수행한다. 또한, 한국어 정보처리와 관련된 다양한 응용 시스템의 성능을 향상시키는 연구를 수행한다. 최신 딥러닝 기술과 관련된 한국어 정보처리의 핵심 요소 기술은 다음과 같다.

(1) 머신러닝 학습 및 평가를 위한 언어자원의 확장 구축
- 원시말뭉치, 품사-구문 태깅 말뭉치의 확장 구축
- 대규모 말뭉치를 이용한 딥러닝 기술 및 언어처리 기술 연구

(2) 최신 딥러닝 기술을 이용한 한국어 언어처리 기술 개발
- Language model과 좌우 문맥 정보를 이용하는 임베딩 기술
- ELMo와 BERT 임베딩 기법에 의한 한국어 문장의 임베딩 기술
- 선택적 임베딩 기법과 한국어 임베딩 기술을 언어이해 기술에 적용

(3) 언어처리 기술을 사회문제 해결에 적용하는 딥러닝 기술 연구
- 사회문제 해결을 위한 문맥정보와 상황인지 기술 연구
- 상황인지 기술을 이용한 사용자 의도파악 기술 연구
- 보이스피싱, 자살 예방 등 사회문제 해결에 적용

5. 결 론

1980년대에 기계번역으로 시작되었던 자연어처리에 관한 연구는 대규모 텍스트 데이터가 구축되고 딥러닝 기술이 보편화됨에 따라 딥러닝 기법을 이용하여 성능 향상 문제를 해결하는 방향으로 발전되어 왔다. 딥러닝 기법에서 가장 중요한 요소는 대규모 언어자원의 구축과 각 문제 해결에 적합한 학습 알고리즘의 구성이다. 차세대 정보컴퓨팅 사업의 일환으로 수행하고 있는 '한국어 정보처리 원천기술' 연구에서는 원시 말뭉치 등 대규모 언어자원을 구축하여 연구자들이 자유롭게 활용할 수 있도록 하고, 딥러닝 기반의 언어분석 모듈과 핵심 응용 기술을 개발하여 언어처리 산업의 활성화에 기여하고자 한다. 2017-2018까지의 1단계 연구성과물을 github와 웹사이트에 공개하였으며, 2019-202까지 2단계에서는 언어자원을 확장 구축하고 감성 데이터 등 최근에 중요한 이슈로 떠오르고 있는 분야의 언어자원을 구축하며, 최신 기술을 보이스피싱 등 사회문제 해결에 기여하는 연구가 진행되고 있다.

Acknowledgement

이 연구는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.NRF-2017M3C4A7068186).

참 고 문 헌

[1] I. Sutskever, J. Martens, and G. Hinton, "Generating Text with Recurrent Neural Networks", ICML-11, pp.1017-1024, 2011.
[2] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning", Nature, Vol.521, pp.436-444, 28th May 2015.
[3] J. Schmidhuber, "Deep Learning in Neural Networks: An overview", Neural networks, Vol.61, pp.85-117, 8th October 2014.
[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning based Natural Language Processing", IEEE, pp.55-75, 2018.
[5] T. Mikolov, I Sutskever, K. Chen, G. Corrado, and J Dean, "Distributed Representations of Words and Phrases and their Compositionality", arXiv:1310.4546v1, 2013.
[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, "Deep Contextualized Word Representations", NAACL, pp.2227-2237, 2018.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of NAACL-HLT 2019, pp.4171-4186, 2019.
[8] M. Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle", Artificial and Human Intelligence, 1984.
[9] K. Cho, B. Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", Proceedings of SSST-8, pp.103-111, 2014.
[10] 강승식, 한국어 정보화 세부 실행 계획 수립, 문화체육관광부 연구보고서, 2011년 7월.
[11] 강승식, 한국어 기계학습 및 평가용 Gold Standard 언어자원 뱅크 구축, 차세대정보컴퓨팅 기술개발사업 연구보고서, 2019년 5월.
[12] X. Zheng, H. Chen, and T. Xu, "Deep Learning for Chinese Word Segmentation and POS Tagging", EMNLP, pp.647-657, 2013.
[13] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging", arXiv preprint arXiv:1508.01991, 2015.
[14] D. Chen, and C. Manning, "A Fast and Accurate Dependency Parser using Neural Networks", EMNLP, pp.740-750, 2014.
[15] C. Dyer, M. Ballesteros, W. Ling, A. Matthews and N. A. Smith, "Transition-Based Dependency Parsing with Stack Long Short-Term Memory", ACL, pp.334-343, 2015.
[16] X. Zhang and M. Lapata, "Chinese Poetry Generation with Recurrent Neural networks", EMNLP, pp.670-680, 2014.
[17] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen and L. Carin, "Adversarial Feature Matching for Text Generation", In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp.4006-4015, 2017.
[18] T. H. Wen, M. Gasic, N. Mrksic, P. H. Su, D. Vandyke and S. Young, "Semantically Conditioned LSTM-Based Natural Language Generation for Spoken Dialogue Systems", arXiv preprint arXiv:1508.01745, 2015.

저 자 약 력

강 승 식

이메일 : sskang@kookmin.ac.kr

- 1986년 서울대학교 전자계산기공학과 (학사)
- 1988년 서울대학교 전자계산기공학과 (석사)
- 1993년 서울대학교 전자계산기공학과 (박사)
- 1994년~2001년 한성대학교 부교수
- 2001년-현재 국민대학교 소프트웨어학부 교수
- 관심분야 : 자연어처리, 한국어 정보처리, 정보검색, 텍스트마이닝 등