한국빅데이터학회지
제6권 제1호, 2021, pp. 63-81 https://doi.org/10.36498/kbigdt.2021.6.1.63

딥러닝 중심의 자연어 처리 기술 현황 분석

Analysis of the Status of Natural Language Processing Technology Based on Deep Learning

박상언†

경기대학교

요 약

자연어 처리는 최근 기계학습 및 딥러닝 기술의 발전과 적용으로 성능이 빠르게 향상되고 있으며, 이로 인해 활용 분야도 넓어지고 있다. 특히 비정형 텍스트 데이터에 대한 분석 요구가 증가함에 따라 자연어 처리에 대한 관심도 더욱 높아지고 있다. 그러나 자연어 전처리 과정 및 기계학습과 딥러닝 이론의 복잡함과 어려움으로 인해 아직도 자연어 처리 활용의 장벽이 높은 편이다.

본 논문에서는 자연어 처리의 전반적인 이해를 위해 현재 활발히 연구되고 있는 자연어 처리의 주요 분야와 기계학습 및 딥러닝을 중심으로 한 주요 기술의 현황에 대해 살펴봄으로써, 보다 쉽게 자연어 처리에 대해 이해하고 활용할 수 있는 기반을 제공하고자 한다. 이를 위해 인공지능 기술 분류체계의 변화를 통해 자연어 처리의 비중 및 변화 과정을 살펴보았으며, 기계학습과 딥러닝을 기반으로 한 자연어 처리 주요 분야를 언어 모델, 문서 분류, 문서 생성, 문서 요약, 질의응답, 기계번역으로 나누어 정리하고 각 분야에서 가장 뛰어난 성능을 보이는 모형들을 살펴보았다. 그리고, 자연어 처리에서 활용되고 있는 주요 딥러닝 모형들에 대해 정리하고 자연어 처리 분야에서 사용되는 데이터셋과 성능평가를 위한 평가지표에 대해 정리하였다. 본 논문을 통해, 자연어 처리를 자신의 분야에서 다양한 목적으로 활용하고자 하는 연구자들이 자연어 처리의 전반적인 기술 현황에 대해 이해하고, 자연어 처리의 주요 기술 분야와 주로 사용되는 딥러닝 모형 및 데이터셋과 평가지표에 대해 보다 쉽게 파악할 수 있기를 기대한다.

■ 중심어 : 자연어 처리, 기계학습, 딥러닝, 데이터셋, 평가지표

Abstract

The performance of natural language processing is rapidly improving due to the recent development and application of machine learning and deep learning technologies, and as a result, the field of application is expanding. In particular, as the demand for analysis on unstructured text data increases, interest in NLP(Natural Language Processing) is also increasing. However, due to the complexity and difficulty of the natural language preprocessing process and machine learning and deep learning theories, there are still high barriers to the use of natural language processing.

In this paper, for an overall understanding of NLP, by examining the main fields of NLP that are currently being actively researched and the current state of major technologies centered on machine learning and deep learning, We want to provide a foundation to understand and utilize NLP more easily. Therefore, we investigated the change of NLP in AI(artificial intelligence) through the changes of the taxonomy of AI technology. The main areas of NLP which consists of language model, text classification, text generation, document summarization, question answering and machine translation were explained with state of the art deep learning models. In addition, major deep learning models utilized in NLP were explained, and data sets and evaluation measures for performance evaluation were summarized.

2021년 07월 14일 접수; 2021년 08월 10일 수정본 접수; 2021년 08월 19일 게재 확정.

† 교신저자 (supark@kyonggi.ac.kr)

We hope researchers who want to utilize NLP for various purposes in their field be able to understand the overall technical status and the main technologies of NLP through this paper.

■ Keyword : NLP, machine learning, deep learning, data set, evaluation measures

Ⅰ. 서론

딥러닝의 급격한 발전과 함께 자연어 처리에서의 딥러닝 기술 활용도 크게 증가하고 있으며, 이로 인해 자연어 처리의 성능도 다양한 분야에서 빠른 속도로 향상되고 있다. 자연어 처리는 학문 분야의 구분 없이 거의 대부분의 학문 영역에서 사용될 수 있다는 점에서 그 활용도가 매우 높다고 할 수 있다. 특히 뉴스와 보고서, SNS(Social Networking Service) 등 비정형 텍스트 데이터를 대상으로 하는 분석에 대한 요구가 늘어남에 따라 자연어 처리에 대한 관심은 더욱 증가하고 있다.

최근에는 챗봇이 공공분야의 대민 서비스를 비롯하여 다양한 업무 분야에서 활발하게 개발됨에 따라 자연어 처리의 활용 분야도 더욱 넓어지고 있다. 챗봇 분야에서의 최근 자연어 처리 연구 현황을 살펴보면, 특허상담 분야의 자동상담 서비스에서 기계독해의 성능을 향상시키기 위해 BERT(Bidirectional Encoder Representations from Transformers)를 이용하여 특허상담 질의에 대한 정답을 결정하는 연구가 있으며[1], 쇼핑몰의 지능형 챗봇을 구현하기 위해 패션쇼핑분야의 개체명 인식 사전을 구축한 연구[2] 등이 있다. 다른 분야에서의 활용 현황으로, 임상의사결정 지원시스템의 구축을 위해 자연어 처리에 기반하여 녹내장 사례에 대한 지식베이스를 구축함으로써 정확도를 향상시킨 연구[3]가 있으며, 소프트웨어 요구사항 분석 과정에서 자연어 처리와 기계학습를 이용함으로써 소프트웨어 요구사항 명세서의 요구사항에 대한 자동 분류를 수행한 연구가 있다[4]. 또한 한의학 분야에서는 자연어 처리와 기계학습을 이용한 한의변증진단 기술의 개발에 관한 연구[5]가 있는데, 동의보감의 증상과 변증에 대한 서술을 대상으로 자연어 처리를 적용하여 증상과 변증의 관계를 분석하고 증상으로부터 변증을 예측하는 시스템을 구현하였다. 이상과 같이 자연어 처리는 특허, 의학, 소프트웨어공학, 한의학 등 다양한 분야에서 활발하게 사용되고 있다.

그러나 자연어 처리는 기본적으로 언어에 대한 이해와 기술에 대한 이해를 동시에 필요로 하며, 특히 최근에는 기계학습과 딥러닝 기법에 대한 이해를 요구하기 때문에, 활용하고 싶어도 기술로 인한 장벽이 높은 편이다. 본 논문에서는 다양한 학문 분야에서 자연어 처리에 관심을 갖고 이를 활용하고자 하는 연구자들에게, 자연어 처리의 주요 분야에 대한 설명과 기술 및 연구 현황을 제공함으로써 보다 쉽게 자연어 처리 분야를 이해할 수 있도록 하고자 한다. 특히 자연어 처리에 최근 다양하게 활용되고 있는 딥러닝 기술을 중심으로 설명함으로써 기술적인 이해를 높이고자 한다.

이를 위해 먼저 2장에서 자연어 처리를 포함한 전체 인공지능의 분류체계와 특허 중심의 기술개발 현황을 살펴봄으로써 인공지능 분야에서 자연어 처리의 비중와 의의를 먼저 살펴본다. 그리고 3장에서 자연어 처리에서 가장 많이 활용되고 있는 머신러닝 분야의 현황을 살펴봄으로써, 기본적인 머신러닝 알고리즘과 성능 평가를 위한 평가지표에 대해 설명한다. 다음으로 4

장에서는 자연어 처리의 주요 분야 별로 각 분야의 내용과 활용 현황에 대해 살펴보고, 자연어 처리에서 주로 활용되고 있는 딥러닝 기법에 대해 정리하고 최신의 기술 동향을 알아본다. 그리고 자연어 처리의 성능 평가를 위해 주로 활용되고 있는 데이터셋과 평가 척도에 대해 정리한다. 마지막으로 5장에서는 본 연구의 결론에 대해 정리하고자 한다.

Ⅱ. 인공지능 분류체계 현황 조사

인공지능은 인간의 학습능력, 추론능력, 지각능력, 자연언어의 이해능력 등을 컴퓨터 프로그램을 실현한 기술로 정의된다. 두 분류로 나누어서 보면 먼저 사람처럼 보고 듣고 말하고 행동하는 기계를 만드는 연구와 사람의 사고방식을 흉내 내서 문제를 해결하고자 하는 연구가 있다. 자연어 처리는 인공지능의 정의에서도 큰 부분을 차지하고 있으며, 의사소통, 지각, 학습, 추론의 도구로 활용된다는 점에서 매우 중요한 분야임을 알 수 있다.

최근 인공지능 중에서도 기계학습(머신러닝)과 딥러닝에 대한 관심이 높은데, 기계학습은 인공지능의 한 분야로 사람의 직접적인 지시 없이 컴퓨터가 학습을 통해서 문제를 해결할 수 있도록 하는 알고리즘이나 통계적 모형에 관한 연구를 말한다. 기계학습에는 회귀분석, 로지스틱 회귀분석, SVM(Support Vector Machine), 결정 트리, 랜덤 포레스트, 클러스터링, 유전 알고리즘, 강화학습, 인공신경망 등 다양한 분야가 있으며 이 중에서 인공신경망을 심화하여 층이 깊은 다층 네트워크를 사용하는 방법론을 딥러닝이라고 한다.

이 장에서는 기존의 인공지능 분류체계를 조사함으로써 시대에 따라 주요 관심사가 변하면서 어떻게 인공지능 분류체계가 변화되어 왔는지 살펴보고, 그 안에서 자연어 처리가 차지하는 비중에 대해 알아보고자 한다. 또한 특허 현황을 통해 자연어 처리의 기술 개발 현황을 살펴봄으로써 기술 현황에 대한 이해를 높이고자 한다.

2.1 2018년 이전 인공지능 분류체계

<표 1>은 2018년 이전의 주요 인공지능 분류체계를 요약한 것이다. 2006년의 정의를 보면 지금과는 많이 다른 관점에서 인공지능을 분류하고 있다. 반면 2015년부터는 체계가 현재와 비슷하나, 응용프로그램 인터페이스가 인공지능 기술로 분류되어 있거나, 인지 및 이해 쪽에 치우쳐 있어 관점의 차이가 있는 것을 볼 수 있다. 주목할 것은 2015년부터 자연어 처리가 언어 이해 등의 분류로 인공지능의 주요 영역을 차지하고 있는 것이다.

2.2 특허청의 4차 산업혁명 기술체계

<표 2>에서 보는 바와 같이, 2018년에 제시된 특허청 기술분류는 음성을 언어에 포함함으로

〈표 1〉 2018년 이전 인공지능 분류체계 요약

구분	인공지능 기술분류 체계	문제점
Waltz, D. (2006)[6]	전문가 AI, 자율로봇, 인지 보조, AI 이론/알고리즘, AI 튜링 검사	현대의 관점에서 볼 때 인공지능 기술에 대한 분류로 보기 어려움
Russel et al. (2016)[7]	자연어 처리, 지식 표현, 자동 추론, 기계학습, 컴퓨터 시각, 로봇공학	로봇공학은 기술 자체라기보다 응용분야에 가까움
한국지식재산연구원 (2016)[8]	학습 및 추론, 상황 이해, 언어 이해, 시각 이해, 인식 및 인지	분류가 인지 및 이해 쪽에 치우쳐 있음

써 Tractica의 분류와 다른 관점을 보이고 있다[9]. 이는 음성 인식이 소리를 언어로 변환한 후의 처리가 동일하기 때문으로 해석된다. 이 분류에서 학습과 추론 분야에 기계학습만 존재하기 때문에, 전체를 포괄하거나 분류가 상호 배타적이라고 보기는 어렵다. 또한 상황 인식과 응용 분야는 인공지능 기술 관점에 적합하지 않고 쓰인 그대로 응용에 가까운 분류로 보는 것이 타당하다.

〈표 2〉 특허청 인공지능 기술분류

구분	중분류	소분류
인공지능	학습과 추론	기계학습
	언어 이해	자연어 처리
		인간과 기계 간의 대화 모델링 또는 관리
		음성인식
	시각 인식	물체 인식, 예. 얼굴
		행동 인식
		장소 인식
	상황 인식	인간 감정 또는 기분의 인식
		사건 또는 사고의 인식
	응용 분야	의료; 건강관리
		상담; 개인 비서 서비스

2.3 과학기술정책연구원의 인공지능 기술분류

과학기술정책연구원에서는 2018년에 인공지능 기술분류를 인식과 분석 두 가지 관점으로 두고, 2차원의 표로 정리하였다[10]. 분석의 경우, 전체를 분류, 군집화, 생성 및 의사 결정으로 두고 그 밑을 다시 비기계학습과 기계학습으로 나눈 후, 기계학습은 다시 신경망과 비신경망으로 분류하였다. 인식은 이미지・영상, 신호, 텍스트・언어로 구분하고 있다.

인공지능의 분류를 체계화하고자 하였으나, 분석의 분류가 지나치게 복잡하고 회귀는 가장 많이 쓰이는 기계학습 알고리즘이면서 분류(classification)과는 성격이 다른 데에도 불구하고 분류에 속하며, 표 안의 셀 각각이 하나의 분류가 되기 때문에 지나치게 많은 분류가 존재한다는 문제점들이 있다. 따라서 보다 단순하고 직관적인 분류체계가 필요한 것으로 판단된다.

2.4 과학기술일자리진흥원의 인공지능 기술분류

과학기술일자리진흥원에서는 2019년에 <표

〈표 3〉 과학기술정책연구원(STEPI) 인공지능 기술분류

Level 1	2. 분석									
1. 인식	Level 2	분류 (Classification)			군집화 (Clustering)			생성 및 의사 결정 (generation/decision-making)		
		비 기계학습	기계학습		비 기계학습	기계학습		비 기계학습	기계학습	
			비 신경망	신경망 (딥러닝)		비 신경망	신경망 (딥러닝)		비 신경망	신경망 (딥러닝)
	이미지・영상	SIFT, SURF	SVM, HMM, ...	CNN, GAN, ...	Rule -based model	GMM, CRF, ...	CNN, GAN	Wiener filter	GMM, CRF	CNN, RNN, ...
	신호	DTW, VQ, ...	GPR, SVM, ...	RNN, DBN, ...	N-gram model, ...	K-means , GMM, ...	RNN VAE, ...	Time series model	ADP, ...	DQN, DRQN, ...
	텍스트・언어	Dictionary -based model	HMM, LDA, ...	CNN, RNN, ...	Rule -based model	LSA, PLSA	CNN, RNN, GAN	LP, GA	Q-learning	RNN, GAN, VAE

〈표 4〉 과학기술일자리진흥원(COMPA) 인공지능 기술분류

중분류	소분류	세분류	요소 기술
인공지능	학습지능	머신러닝	베이지안 학습, 인공신경망, 딥러닝, 강화학습, 앙상블 러닝, 판단 근거 설명
		추론/지식표현	추론, 지식표현 및 온톨리지, 지식처리
	단일지능	언어지능	언어분석, 의미이해, 대화 이해 및 생성, 자동 통역・번역, 질의응답(Q/A), 텍스트 요약・생성
		시각지능	영상 처리 및 패턴 인식, 객체 인식, 객체 탐지, 행동 이해, 장소/장면 이해, 비디오 분석 및 예측, 시공간 영상 이해,비디오 요약
		청각지능	음성분석, 음성인식, 화자 인식/적응, 음성합성, 오디오 색인 및 검색, 잡음처리 및 음원 분리, 음향인식
	복합지능	행동/소셜지능	공간 지능, 운동 지능, 소셜 지능, 협업 지능
		상황/감정이해	감정이해, 사용자 의도 이해, 뇌신호인지, 센서 데이터 이해, 오감인지, 다중 상황 판단
		지능형 에이전트	에이전트 플랫폼, 에이전트 기술, 게임 지능, 모방창작 지능
		범용 인공지능	상식 학습, 범용 문제해결, 평생 학습, 도덕-윤리-범 지능

4>와 같이 인공지능을 학습지능, 단일지능, 복합지능으로 분류하고, 텍스트, 이미지, 음성은 단일지능의 세분류로 정리하였으나, 단일지능의 정의와 하부 분류가 모호한 점이 있다[11]. 요소기술의 내용으로 볼 때 복합지능은 사실성 기술보다 복합적인 적용에 가깝다고 할 수 있다.

2.5 인공지능 기술분류 정리

이상과 같이 인공지능의 다양한 기술분류를 살펴보았으나, 대부분 문제점이 존재한다. 이는 기술에 대한 분류가 알고리즘과 데이터의 두 가지 관점에서 분류될 수 있고, 두 관점의 경계를 명확히 하기가 어렵기 때문이다. 과학기술정책연구원에서는 이 두 관점을 두 개의 차원으로 놓고 분류를 만들었으나, 그 결과 너무 복잡하고 중복되는 분류체계가 만들어졌다.

이상에서 조사된 최대한 기존 분류체계를 포괄하면서 각 분야가 분리될 수 있도록 분류체계를 정리해보면, 먼저 상위 분류로 기계학습, 추론, 딥러닝 그리고 기타로 분류할 수 있을 것이다. 기계학습은 원래 신경망과 딥러닝을 포함하고 있으나 편의상 딥러닝을 제외한 나머지 기계학습 기술들로 정의하는 것이 바람직하다. 그리고 이미지 처리, 자연어 처리, 음성 처리는 대부분 딥러닝에 기반하고 있으며 동시에 딥러닝 대부분을 차지하므로 딥러닝의 하부 단계로 정의할 수 있다.

2.6 국내 인공지능 기술 개발 현황

인공지능 기술 개발 현황을 분석하기 위해 본 연구에서는 국내 인공지능 관련 주요기술의 특허 출원 동향을 살펴보았다. <표 5>는 2010년부터 2019년까지의 인공지능 분야별 특허 출원 수를 위에서 제시한 분류체계에 맞춰 보여주고 있다[12].

표를 보면 알파고가 등장한 2016년 이후 인공지능에 대한 관심이 급격히 증가하고, 이에 따라 인공지능 분야의 특허 출원량도 급격히 증가한 것을 보여준다. 2010년과 2019년을 주요 기술별로 비교해 볼 때, 학습 및 추론, 청각지능, 시각지능의 비중이 크게 증가하였다. 특히 시각지능의 경우, 언어지능과 청각지능을 합친 것보

〈표 5〉 국내 인공지능 주요기술 출원 동향

구 분	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	합계
기계학습/추론	12	37	25	29	58	54	184	351	598	996	2,344
자연어 처리	93	91	129	116	153	218	348	601	708	838	3,295
음성 처리	21	36	58	49	83	112	231	586	708	690	2,574
이미지 처리	105	120	163	228	272	373	697	1,242	1,652	1,764	6,616
기타	24	21	35	35	78	59	181	402	626	1126	2587
인공지능 전체	240	281	374	430	611	693	1,315	2,216	3,054	4,011	13,225

다 큰 비중을 차지하고 있는 것을 볼 수 있는데, 이는 언어지능이 사용 언어별로 차이를 보이는 것에 비해 시각지능은 모든 국가가 공통으로 활용할 수 있어 활용범위가 넓고 다양하기 때문인 것으로 해석 할 수 있다.

자연어 처리는 인공지능 초기부터 관심도가 높았으며, 다른 분야가 알파고 이후 급격히 증가한 것에 비하면 꾸준히 성장해 온 것으로 보인다. 최근 딥러닝으로 인해 성능이 많이 향상된 것을 감안할 때, 2020년 이후 더 많은 성장을 보일 것으로 기대되기도 한다.

Ⅲ. 기계학습 기술 현황

이 장에서는 인공지능 분야 중에서 최근 가장 많이 활용되고 있는 기계학습 분야의 기술 현황에 대해 정리하고자 한다. 기계학습은 자연어 처리에서 필수적으로 사용되는 기술이기도 하므로 본격적인 자연어 처리 현황을 살펴보기 전에 반드시 정리해야 하는 분야이기도 한다. 다만 이 논문에서는 자연어 처리에 대한 기술 현황이 주 내용이므로 기계학습에 대한 내용은 최소한으로 기술하고자 한다.

3.1 기계학습 주요 기술 분야

기계학습은 인공지능의 한 분야로 사람의 직접적인 지시 없이 컴퓨터가 학습을 통해서 문제를 해결할 수 있도록 하는 알고리즘이나 통계적 모형에 관한 연구를 말한다. <표 6>은 기계학습의 분류와 주요 기술을 보여준다. 먼저 가장 큰 분류로 학습할 때 정답을 사용하여 학습하는지의 여부에 따라 지도학습과 비지도학습으로 나뉜다. 다만 강화학습은 정답을 이용하는 것이 아니라 에이전트가 환경으로부터 받는 보상에 따라 학습되기 때문에 별도로 분류된다. 지도학습은 입력과 출력(혹은 정답)으로 구성된 학습 데이터를 이용해 학습한 후 주어진 입력에 대해 출력을 예측하고자 할 때 사용한다. 지도학습은 크게 회귀와 분류로 나뉘며, 회귀는 연속적인 숫자를 예측하는 반면 분류는 미리 정의된 클래스를 예측한다.

비지도학습은 출력(혹은 정답)이 없이 학습하는 모든 종류의 기계학습 알고리즘을 의미하며, 크게 군집화와 차원축소로 나뉜다. 군집화는 데이터를 비슷한 것끼리 그룹으로 묶는 것을 말하고, 차원축소는 데이터를 보다 작은 차원의 새로운 형태로 표현하여 원래 데이터보다 쉽게 해석할 수 있도록 만든다.

현재 파이썬 기반에서는 사이킷런이 머신러닝을 구현할 수 있는 도구로 가장 널리 쓰이고 있으며, 사이킷런에서는 분류, 회귀, 군집화, 차원축소의 넷으로 기계학습을 분류하고 있다[13]. 본 논문에서는 기계학습의 주요 기술 분야를 ‘분류’, ‘회귀’, ‘군집화’, ‘차원축소’, ‘강화학습’의 순으로 정리하고자 한다.

〈표 6〉 기계학습의 분류

대분류	소분류	주요 기술
지도학습	분류	k-최근접 이웃, 나이브 베이즈, 로지스틱 회귀, 결정 트리, 랜덤 포레스트, 그래디언트 부스팅, SVM, 신경망 등
	회귀	선형회귀분석, 릿지회귀, 라쏘회귀, 나이브 베이즈를 제외한 대부분의 분류 기술
비지도학습	군집화	k-평균, 병합군집, DBSCAN
	차원축소	PCA, NMF, tSNE 등
강화학습	(해당없음)	Q 러닝, SARSA, DQN(Deep Q-Network)

3.1.1 분류(Classification)

분류는 지도학습의 한 분야로, 주어진 입력 데이터에 대하여 미리 정의된 여러 클래스 중 하나를 예측하기 위해 사용된다. 머신러닝에서 가장 많이 사용되는 분야 중 하나로 다양한 알고리즘과 응용사례가 있다.

<표 7>은 분류에 사용되는 대표적 알고리즘에 대해 설명하고 있다.

3.1.2 회귀(Regression)

회귀는 지도학습의 한 분야로, 주어진 입력 데이터에 대하여 연속적인 숫자를 예측한다. 회귀에 사용되는 대표적 알고리즘으로 선형회귀분석이 있으며, 분류에 사용되는 알고리즘 중 나이브 베이즈를 제외한 대부분의 알고리즘이 회귀에도 사용이 가능하다.

3.1.3 군집화(Clustering)

군집화는 비지도 학습의 한 분야로, 비슷한 데이터들을 사전에 정의되지 않은 그룹으로 묶는 것을 말한다. 군집화에 사용되는 대표적 알고리즘으로 먼저 k-평균은 주어진 데이터를 k개의 군집으로 묶기 위해, 각 클러스터와 데이터들 간의 거리 차이의 분산을 최소화하는 방식으

〈표 7〉 기계학습 분류 알고리즘

알고리즘	내용
k-최근접 이웃	가장 단순한 머신러닝 알고리즘으로, 새로운 입력 데이터에 대해 가장 가까운 훈련 데이터 포인트를 찾아 그 출력(클래스)으로 예측
나이브 베이즈	특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기로, 텍스트 분류에서 특히 성능이 좋음.
로지스틱 회귀	선형회귀를 기반으로 하여 각 클래스에 대한 확률을 예측, 이진분류와 다중분류 모두 가능
결정 트리	트리 형태로 이루어진 입력에 대한 연속된 질문으로 클래스를 예측
랜덤 포레스트	결정 트리의 앙상블 모형으로, 여러 결정 트리의 묶음을 통해 예측성능을 향상
그래디언트 부스팅	결정 트리의 앙상블 모형으로, 랜덤 포레스트와 유사하나 다른 점은 이전 트리의 오차를 보완하는 방식으로 순차적으로 트리를 생성
SVM	클래스의 구분을 위해 가장 큰 폭을 가진 경계를 찾는 알고리즘으로 비선형 분류를 하기 위해서 주어진 데이터를 고차원 특징 공간으로 사상하여 처리
신경망	사람의 뇌에서 영감을 얻은 통계학적 학습알고리즘으로 일반적으로 학습을 위해 오차역전파법을 사용하는 다층 퍼셉트론을 가리킴. 층이 깊은 신경망이 현재 딥러닝으로 주목을 받고 있음

로 학습하는 것을 말한다. 둘째 병합 군집(agglomerative clustering)은 시작할 때 각 포인트를 하나의 클러스터로 지정하고, 종료 조건을 만족할 때까지 가장 비슷한 두 클러스터를 합쳐나감으로써 군집을 만드는 알고리즘이다. 마지막으로 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)은 입력 데이터 집합에서 가까이 있는 데이터가 많은 밀집 지역을 군집으로 만들어나가는 알고리즘으로 클러스터의 수를 정하지 않아도 된다는 장점이 있다. 이 외에도 다양한 군집화 알고리즘이 있다.

3.1.4 차원축소(Dimensionality Reduction)

차원축소는 비지도학습의 한 분야로, 데이터를 새롭게 표현(일반적으로 원래 데이터의 차원보다 작은 차원으로 축소)하여 사람이나, 머신러닝 알고리즘이 보다 쉽게 해석할 수 있도록 하는 작업을 말한다. 고차원 데이터에 대해 특성의 수를 줄이면서 꼭 필요한 특징만으로 데이터를 표현하는 방법을 의미하며, 중요한 특성만을 골라내는 특성 선택(feature selection)과 기존 특성의 조합을 이용하여 새로운 특성으로 추출하는 특성 추출(feature extraction)의 두 가지 방법으로 구분한다. 특성을 2차원으로 축소하여 시각화하는 경우가 있으나, 일반적으로는 독립적인 작업으로 사용되기보다 분류나 회귀의 전처리 작업으로 사용되는 경향이 강하며, 주로 사용되는 알고리즘으로 PCA(Principal Component Analysis), NMF(Non-negative matrix factorization), t-SNE(t-Distributed Stochastic Neighbor Embedding) 등이 있다.

3.1.5 강화학습(Reinforcement Learning)

강화학습은 학습의 주체가 되는 에이전트가 주어진 환경에 대해 어떻게 행동해야 하는지를 학습하기 위한 방법론으로, 에이전트는 자신이 한 행동에 대해 보상을 받고 이에 대한 반응을 통해 자신의 선택 방법 혹은 전략에 대해 학습을 하게 된다. 입력과 출력의 쌍으로 이루어진 훈련 집합을 사용하지 않는다는 점에서 지도학습과는 분명한 차이가 있으며, 다양한 일련의 상황에서 자율적으로 행동하는 소프트웨어나 로봇을 만들기 위한 학습 방법으로 많이 활용되고 있다. 최근에는 딥러닝에 기반한 DQN(Deep Q-Network) 등을 통해 더욱 주목을 받고 있다.

3.2 기계학습 성능 평가를 위한 데이터셋 및 평가 지표

<표 8>은 기계학습 알고리즘을 테스트하거나 성능을 평가하기 위해 사용하는 대표적인 데이터셋을 보여주며, <표 9>는 기계학습 알고리즘의 성능 평가를 위한 대표적인 평가지표를 설명하고 있다.

Ⅳ. 자연어 처리 기술 현황

자연어 처리(Natural Language Processing, NLP)는 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 일을 말한다. 최근에는 머신러닝 혹은 딥러닝을 이용하여 자연어 문서를

〈표 8〉 기계학습 데이터셋

데이터셋	내용
Kaggle Datasets	50,000개 이상의 데이터셋과 분석에 사용된 400,000개의 코드 보유. 대표적인 예로 타이타닉 승객들의 생존 여부를 예측하기 위한 Titanic 데이터셋이 있음.
the UC Irvine Machine Learning Repository	머신러닝 커뮤니티 서비스로, 557개의 데이터셋 보유. 대표적인 예로 식물의 유형을 예측하기 위한 Iris 데이터셋이 있음.

〈표 9〉 기계학습 성능 평가지표

척도	내용
평균절대오차 (Mean Absolute Error: MAE)	회귀 성능을 측정하기 위한 지표로 실제 예측한 값과 모델이 예측한 값의 오차의 평균을 의미하는 지표. 0에 가까울수록 오차의 평균이 작다는 의미로서 학습시킨 모델이 더 정확하다는 것을 의미.
평균제곱근오차 (Root Mean Square Error: RMSE)	평균절대오차와 마찬가지로 회귀 성능을 측정하며, 예측값과 실제 값을 뺀 오차의 평균값에 제곱근을 지표로 사용. 해당 값 또한 MAE와 마찬가지로 0에 가까울수록 좋은 성능.
오차행렬 (Confusion Matrix)	오차행렬은 True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN)으로 구성된 행렬을 말하며 주로 분류 문제의 성능 평가를 위한 다양한 지표를 계산하기 위해 사용됨. 정확도(accuracy)는 (TP+TN)/(TP+TN+FP+FN)으로, 분류 성능을 측정하는 평가지표로 가장 많이 활용됨. 재현율(recall)은 TP/(TP+FN)으로 실제 positive 중 정확히 positive라고 식별된 사례의 비율을 말하고, 정밀도(precision)는 TP/(TP+FP)로 positive로 식별된 사례 중 실제 positive 사례의 비율을 말함. 클래스 간의 불균형이 심한 경우에는 정확도 대신 정밀도와 재현율을 사용함. F1 척도는 정밀도와 재현율의 조화평균으로 정밀도와 재현율을 조합하여 하나의 값으로 성능을 표현하고자 할 때 사용함.

처리하는 기술, 즉 전통적으로 텍스트 마이닝에 속하는 분야를 모두 포괄하고 있어 그 범위가 더욱 넓어졌다고 볼 수 있다. 이 장에서는 주로 딥러닝에 기반한 자연어 처리의 주요 분야와 각 분야에서의 대표적인 기술과 성능에 대해 정리한 후, 딥러닝 기반의 자연어 처리 기술에 대해 알아보고자 한다.

4.1 딥러닝 기반 자연어 처리의 주요 분야

4.1.1 언어 모델 (Language Model)

언어 모델은 문서에서 주어진 앞부분의 단어들의 시퀀스 즉 순서를 이용해 다음 단어의 예측을 수행하는 모형을 생성하는 프로세스를 말한다. 문서에 나타난 문맥을 학습하기 위한 가장 중요한 학습모형으로, 트랜스포머에 기반한 BERT, GPT(Generative Pre-trained Transformer)가 언어 모델에 기반하여 사전학습을 수행함으로써 언어의 구조와 문맥을 학습한 후에, 파인튜닝을 통해 다양한 자연어 처리 애플리케이션을 수행하는 방식으로 활용되고 있다.

현재 GPT-3가 이 분야에서 가장 뛰어난 성능을 보이는 알고리즘 즉 SOTA(state of the art)로 알려져 있다[14]. 아래 표는 GPT-3의 성능을 나타내며, LAMBADA(LAnguage Modeling Broadened to Account for Discourse Aspects) 데이터셋에서 기존에 비해 매우 향상된 결과를 보이고 있다.

〈표 10〉 언어 모델 GPT-3 성능 (정확도)

모형	LAMBADA	StoryCloze	HellaSwag
기존 SOTA	68.0	91.8	85.6
GPT-3 Zero-Shot	76.2	83.2	78.9
GPT-3 One-Shot	72.5	84.7	78.1
GPT-3 Few-Shot	86.4	87.7	79.3

4.1.2 문서 분류 (Text Classification)

문서분류는 주어진 문서에 대해 미리 정의된 클래스로 분류하는 작업으로 감성 분석, 스팸

메일 분류, 뉴스 기사 분류 등 다양한 응용 분야가 있다. CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), DBN(Deep Belief Network) 등 다양한 딥러닝 기법들이 개발되고 사용되었다. 최근 SOTA는 BERT에 기반한 DocBERT로 <표 11>과 같이 네 개의 분류 데이터 셋에서 SOTA를 기록하였다[15]. <표 11>은 머신러닝의 대표적 알고리즘 중 하나인 SVM과 딥러닝 알고리즘인 CNN, LSTM(Long Short-Term Memory) 그리고 DocBERT의 성능을 F1과 정확도로 보여주고 있다.

〈표 11〉 문서 분류 BERT 모델 성능

모형	Reuters/ F1	AAPD/ F1	IMDB/ 정확도	Yelp'14/ 정확도
SVM	86.1	69.1	42.4	59.6
CNN	80.8	51.4	42.7	66.1
LSTM	87.0	70.5	52.8	68.7
DocBERT	90.7	75.2	55.6	72.5

4.1.3 문서 생성 (Text Generation)

문서 생성은 사람이 쓴 것과 유사한 문장을 만들어내는 작업으로, 다른 대부분의 자연어 처리가 주어진 문장에 대해 새로운 문장을 만들어내는 시퀀스-투-시퀀스 형태의 작업인 것에 비해 문서 생성은 변환할 입력 데이터가 없다는 점이 가장 큰 차이이다. 시(문학) 생성, 농담 생성, 이야기 생성 등의 분야가 있으나, 도전적이고 사람들의 관심을 끌기에 좋은 소재인 것에 비해 아직 실질적인 응용 분야는 많지 않은 편이다. GPT, LSTM, GAN(Generative Adversarial Networks), VAE(Variational Auto-Encoder)등 다양한 딥러닝 방법이 사용되었으며, GPT-3의 등장으로 현재는 GPT-3가 가장 뛰어난 성능을 보이는 것으로 알려져 있으나 명확하게 성능 평가를 하기 어렵다는 점이 있다.

문서 생성은 음성 대화 시스템에서, 사용자의 질문에 대한 답변을 생성하기 위해서 활용된다. 단순하게는 응답 문장 중 일부 값이 비어 있는 템플릿에 질문에 대한 답변 정보가 담긴 단어들을 결합함으로써 생성할 수 있으나, 최근에는 딥러닝에 기반한 생성에 대한 연구가 많이 진행되고 있다[16].

4.1.4 문서 요약 (Summarization)

문서 요약은 주어진 문서에서 중요하고 흥미있는 내용을 추출하여 요약문을 생성하는 작업으로 시퀀스-투-시퀀스의 전형적인 예이다. 추출적(Extractive) 요약과 추상적(Abstractive) 요약의 두 형태로 나뉘며, 추출적 요약은 원문에 있는 단어나 문장을 이용하는 반면, 추상적 요약은 원문에 없는 단어를 사용하거나 문장을 새로 생성하여 요약한다. 인코더-디코더 모형, BERT 등 다양한 딥러닝 방법이 사용되었다.

현재는 BERT에 기반한 모형이 CNN/Daily Mail과 New York Times 데이터셋에서 SOTA를 기록하고 있다[17].

<표 12>는 위 SOTA 모형과 기존의 추출적, 추상적 모형의 성능을 비교하여 보여 준다. SOTA라고는 하나 실질적으로 차이가 크지 않고 아직도 문서 요약은 쉽지 않음을 보여준다고 할 수 있다.

4.1.5 질의응답 (Question Answering)

질의응답은 주어진 문장(context)를 읽고, 주어진 문제(question)에 대해 올바른 답(answer)을 생성하는 작업으로 공학적으로는 문서 요약과 유사하다. CNN, LSTM, BERT 등 다양한 방법이 사용되었으며, 현재는 BERT를 이용한 모형

〈표 12〉 문서 요약 BERT 기반 모델 성능

모형	ROUGE-1	ROUGE-2	ROUGE-L
추출적 SOTA	41.59	19.01	32.86
추상적 SOTA	41.69	19.47	33.11
Zhang et. al	41.71	19.49	33.33

이 SQuAD(Stanford Question Answering Dataset) 1.1과 SQuAD 2.0 데이터셋에서 SOTA를 기록하고 있다[18].

<표 13>은 현재 SOTA를 기록 중인 위 BERTserini와 기존 모형의 성능을 비교해서 보여준다. 기존에 비해 비교적 향상된 재현율과 F1을 보인다.

〈표 13〉 질의응답 BERTserini 모델 성능

모형	전체일치도	F1	재현율
MINIMAL	34.7	42.5	64.0
BERTserini	38.6	46.1	85.8

질의응답과 관련하여 국내에서는 총 10만개 이상의 쌍으로 구성된 한국어 질의응답 데이터셋인 KorQuAD(Korean Question Answering Dataset) 2.0의 개발을 통해, 한국어에 대한 기계독해 연구에 대한 객관적인 기준을 제시하고자 한 연구가 있다[19]. 또한 특허상담 분야 자동상담 서비스에서 BERT를 이용하여 특허상담 질의에 대한 정답을 결정함으로써 기계독해의 성능을 향상시키기고자 한 연구가 있다[1].

4.1.6 기계번역 (Machine Translation)

기계번역은 한 언어로 작성된 문서를 다른 언어로 번역하는 작업으로, 두 언어를 모두 알고 있는 사람에게도 쉽지 않은 작업이다. RNN, CNN, BERT 등 다양한 기법이 사용되었으며, RNN의 경우 GRU(Gated Recurrent Unit)를 이용한 모형이 많이 사용되었다. 언어의 종류가 많기 때문에 어느 한 알고리즘이 SOTA라고 말하기 어려운 점이 있으나, 현재 IWSLT (International Workshop on Spoken Language Translation) 2014 데이터셋 중 영어에서 독일어와 불어로 번역하는 데이터셋에 대한 SOTA는 트랜스포머에 기반한 모형이다 [20].

<표 14>는 위 SOTA 모형의 성능을 이전의 트랜스포머 성능과 비교한 것으로 상당한 개선이 이루어진 것을 볼 수 있다.

〈표 14〉 기계번역 트랜스포머 기반 모델 성능

모형	IWSLT2014 (BLEU)		NMT (BLEU)
	EN-DE	EN-FR	
Transformer Large	47.57	56.15	60.95
Medina et. al	57.05	63.26	62.77

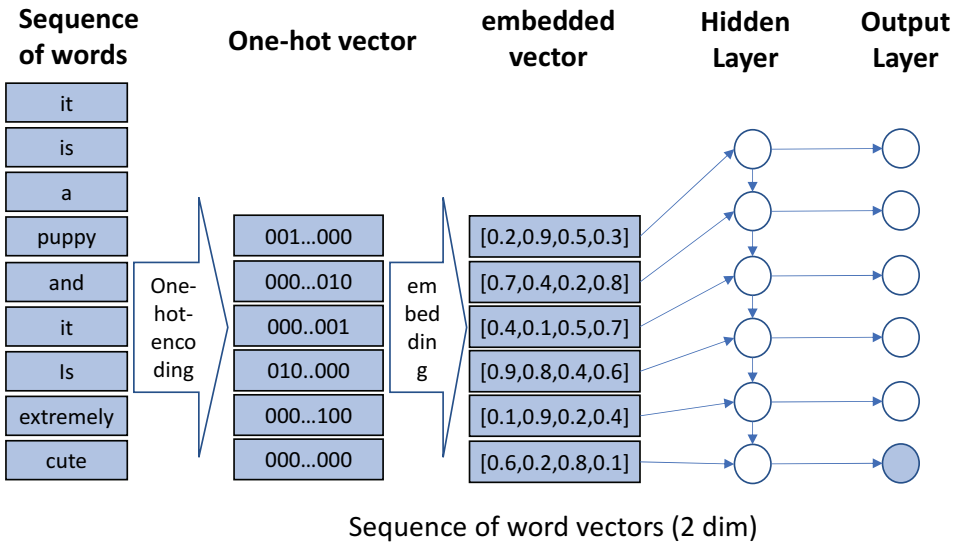
4.2 자연어 처리 주요 딥러닝 기법

이 절에서는 4.1절에서 설명된 자연어 처리 주요 분야에서 사용된 주요 딥러닝 기법에 대해 정리하고자 한다.

4.2.1 순환 신경망(Recurrent Neural Networks: RNN)

순환신경망은 주어진 입력값들의 순서가 의미가 있을 때, 즉 입력값이 앞의 입력값들에 의해 영향을 받는 경우에 사용하는 신경망으로, 시계열 값의 분석 등에서 활발하게 사용되고 있다. 자연어 처리에서는 문장을 이루는 단어들의 순서를 반영하기 위해 사용되는데, 근본적으로 문장은 단어들의 순서에 의해 문맥이 결정되기 때문에 문맥 파악을 위해서 순환 신경망이 사용되었다고 할 수 있다.

<그림 1>은 문서 분류에 RNN을 적용하는 예를 간단하게 보여주고 있다. 각 단어는 먼저 원핫벡터로 표현되었다가 짧은 길이의 밀집벡터로 다시 변환되어 RNN 모형에 입력으로 사용된다. 문서에 사용된 단어는 가변적이기 때문에 모형에 사용하기 위해서는 앞 혹은 뒤의 단어들을 잘라낼 수 밖에 없다. 예제에서는 앞의 세 단어를 자르고 6개의 단어를 사용하고 있으며, 만일 한 단어를 예제와 같이 크기 4의 벡터로 표현하면 문서는 [6, 4] 크기의 2차원 행렬로 표현되고 이 값이 뒤에 있는 순환신경망의 입력으로



〈그림 1〉 문서 분류 순환신경망 적용 예제

사용된다.

딥러닝의 기울기 소실 문제로 인해 RNN은 긴 문장에서는 앞 부분의 정보가 사라지는 효과가 결과가 나타나며, LSTM(Long Short-Term Memory)은 이를 보완하기 위한 모형이다. LSTM은 RNN에 비해 은닉 상태의 값에 셀 상태 값을 추가하여 장기 기억에 사용됨으로써 RNN이 장기 기억에 약한 것을 보완하였다. LSTM의 계산량이 많아 학습에 시간이 많이 걸리는 것을 보완하기 위한 변형으로 GRU(Gated Recurrent Unit)가 있다.

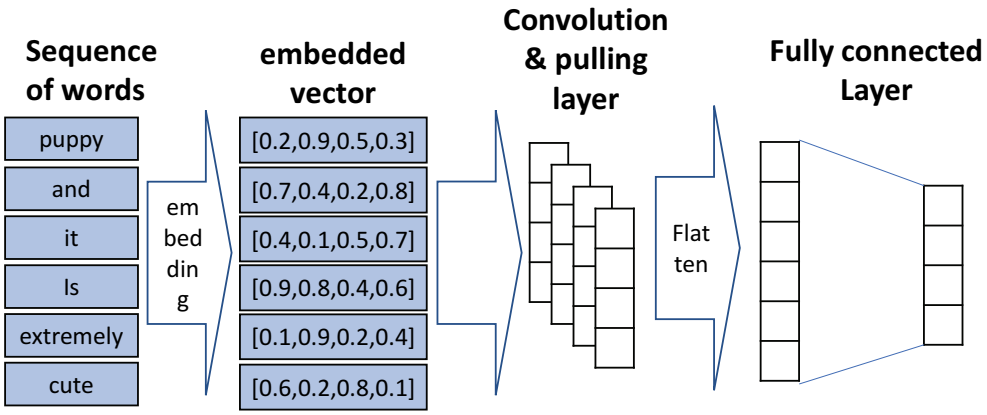
국내에서 이러한 순환신경망을 활용한 연구로, LSTM을 양방향으로 층을 쌓아 구현하는 양방향 LSTM을 이용하여 한국어 영화리뷰의 감성 분석을 향상시키고자 한 연구[21]가 있으며, 양방향 LSTM과 CRF(Conditional Random Field)를 이용하여 자동 띄어쓰기와 품사 태깅을 향상시키고자 한 연구[22]가 있다.

4.2.2 합성곱 신경망 (Convolutional Neural Networks, CNN)

합성곱 신경망은 원래 이미지 처리를 위해 개발된 신경망이지만, CNN이 이미지의 주변 정보를 학습한다는 점을 이용하여 텍스트의 문맥을 학습하는 연구를 선보였다[23]. 이 모형이 의외로 뛰어난 성능을 보이게 되면서 자연어 처리에서의 활용 분야가 넓어지게 되었다.

CNN은 합성곱층(convolution layer)와 풀링층(pooling)으로 구성되며, 합성곱층은 2차원 이미지에서 특정 영역의 특징을 추출하는 역할을 하는데, 이는 연속된 단어들의 특징을 추출하는 것과 유사한 특성이 있다.

<그림 2>는 CNN을 이용한 문서 분류의 예를 보여준다. 이미지와 달리 텍스트는 단어들의 1차원 시퀀스로 표현되므로 1D CNN 모형을 사용한다. 따라서 컨볼루션과 풀링 층의 모양이 일반적인 CNN과는 달리 1차원 형태를 갖는다. 컨볼루션과 풀링을 반복한 결과를 1차원 벡터로 평탄화하고 인공신경망을 이용한 분류기를 거치면 문서를 분류할 수 있다.



〈그림 2〉 합성곱신경망을 이용한 문서분류 예제

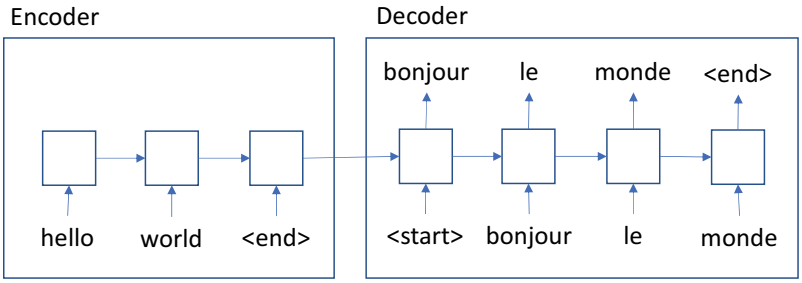
4.2.3 시퀀스-투-시퀀스(Sequence-to-Sequence) 모형

시퀀스-투-시퀀스 모형은 RNN 혹은 LSTM에 기반한 모형으로 번역에서 사용하는 대표적인 모형이다. 인코더와 디코더로 구성되어 있으며, 인코더는 시퀀스를 입력으로 받아 RNN 모형으로 문맥 정보를 축적하고, 디코더는 이 문맥 정보로부터 시퀀스를 출력한다[24]. RNN, CNN 기반 모형과의 확연한 차이는, 두 모형이 하나의 값만 예측하는 것에 비해 시퀀스-투-시퀀스 모형은 순서가 있는 여러 값을 예측한다는 것이다.

<그림 3>은 시퀀스-투-시퀀스 모형을 이용한 번역 예제를 보여준다. 인코더는 “hello world”를 입력받아 프랑스어로 RNN을 통해 문맥정보를 축적한다. 디코더는 이 문맥정보를 은닉층으로 넘겨 받아, <start> 신호와 함께 번역을 시작한다. 처음 생성된 단어인 bonjour는 다음 단어를 생성하기 위한 입력으로 다시 이용되고 문맥정보와 함께 다음 단어인 le를 생성한다. 이와 같은 단계를 반복하여 완전한 문장을 생성해 낸다.

4.2.4 어텐션 메커니즘 (Attention Mechanism)

시퀀스-투-시퀀스 모형은 긴 문장의 문맥 정보를 RNN의 은닉 층에 압축하는 과정에서 정보 손실이 발생하며, 문장을 생성할 때에도 기울기 소실 문제로 인해 장기 기억정보가 잘 전달되지 않는 문제가 있다. 이러한 문제는 LSTM을 사용해도 많이 개선되기 어렵다. 특히 번역에서는 앞 문장의 특정 단어가 생성하는 문장의 대응되는 단어에 직접적이고 가장 큰 영향을 미치는

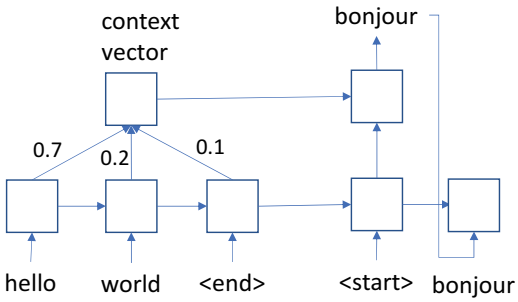


〈그림 3〉 시퀀스–투–시퀀스 모형을 이용한 번역 예제

경우가 많은데, RNN에서는 이러한 복합적인 정보가 하나의 셀에 압축되기 때문에 문제가 발생한다.

어텐션 메커니즘은 이러한 문제를 해결하기 위해 고안되었으며, 디코더에서 특정 단어를 생성 혹은 예측할 때 이 단어에 영향을 미치는 원래 문장의 단어에 보다 집중해서 직접적인 연결을 하기 위한 방안이다[25].

<그림 4>는 <그림 3>과 동일한 예제에 어텐션 메커니즘이 적용된 모습을 최대한 단순화하여 보여준다. 그림은 첫 단어인 bonjour를 생성하는 시점에서의 어텐션 메커니즘을 보여준다. 번역 대상인 문장의 단어들로부터 어텐션 정보를 담고 있는 문맥 벡터(context vector)를 생성하는데 bonjour는 당연히 hello의 영향을 가장 많이 받기 때문에 hello로부터의 가중치가 0.7로 가장 높다. 이 문맥 벡터와 <start>로부터 올라온 정보를 결합하여 bonjour가 생성된다. 그 뒷부분은 생략되어 있으나, <그림 3>과 동일하게 진행된다. bonjour는 다음 단어를 예측하기 위해 다시 입력으로 사용되고, 이로부터 계산된 은닉층의 값은 새로운 문맥 벡터와 결합되어 다음 단어 예측에 사용된다.



〈그림 4〉 어텐션 메커니즘을 이용한 번역 예제

4.2.5 트랜스포머 (Transformer)

기존의 어텐션 메커니즘은 RNN 구조 위에 어텐션을 추가한 반면, 트랜스포머는 시퀀스-투-시퀀스 모형을구조를 따르면서도 RNN을 사용하지 않는다는 점에서 차이가 있다. 그 외에 기존 어텐션 메커니즘과의 가장 중요한 차이점은, 기존에는 인코더로부터 디코더로 어텐션 메커니즘이 동작하는 반면, 트랜스포머에는 인코더나 디코더 내부에서 동작하는 셀프 어텐션(self-attention)이 구현되어 있다는 것이다[26]. 즉, 일반적인 어텐션은 번역 대상 문장으로부터 생성하는 문장으로의 어텐션이지만, 셀프 어텐션은 번역 대상 문장 안에서의 어텐션이기 때문에 앞에 셀프가 붙는다고 생각할 수 있다.

셀프 어텐션의 필요성을 보여주는 극단적인 예가 문장 내에서 사용되는 it과 같은 대명사의 해석이다. “The animal didn’t cross the street because it was too tired”와 같은 문장에서 it은 animal을 말하고 따라서 셀프 어텐션에서 animal로부터의 가중치가 가장 클 것이다. 반면, “The animal didn’t cross the street because it was too wide”라고 되어 있다면 여기서 it은 street를 의미하고 따라서 street로부터의 가중치가 가장 크게 나타난다. 트랜스포머는 이와 같이 문장 내에서 단어들 사이의 가중치가 문맥에 따라 달라질 수 있으며 이와 같은 정보가 문맥 정보에 포함되어야 함을 보여준다.

트랜스포머의 인코더와 디코더는 서로 다른 구조를 사용하는데, 그 이유는 인코더의 경우 문장이 완전하게 주어지기 때문에 어느 방향으로든 어텐션이 가능하지만, 디코더에서는 문장이 생성되고 있기 때문에 역방향의 어텐션은 있을 수 없기 때문이다.

4.2.6 GPT (Generative Pre-Trained Transformer)

GPT는 트랜스포머에 기반한 모형으로 트랜스포머에서 디코더만 사용하기 때문에 인코더-디코더 어텐션이 없고, 단방향 셀프 어텐션만 사용하여 학습한다는 특징이 있다[27]. 언어 모델(language model)에 기반하여 학습을 수행하

며, 문장에서 앞에 있는 단어들의 시퀀스로 다음 단어를 예측하는 방식으로 학습하여 모형에 언어의 구조와 문맥을 이해시킨다.

GPT-1은 1억 개 정도의 파라미터로 시작하였으나, GPT-2에서 14억 개, GPT-2의 변형에서 80억 개, 170억 개 등을 거쳐 GPT-3[28]에서는 1750억 개의 파라미터로 모형이 계속 커지고 있다. 언어 모델을 이용해 학습된 가중치를 포함한 모형을 기반으로 하여, 전이학습을 통해 문서 분류, 문서 생성, 문서 유사도 계산 등 다양한 분야에서 파인 튜닝을 함으로써 각 분야에 활용할 수 있도록 지원하고 있다.

GPT-3은 언어 모델의 특성을 잘 살려서 문서 생성에 특화된 성능을 보였으며, GPT-3이 생성한 가짜 뉴스 기사를 실제 뉴스 기사와 거의 분간하지 못할 정도로 뛰어난 성능을 보여 화제가 되었다.

4.2.7 BERT (Bidirectional Encoder Representations form Transformer)

BERT는 트랜스포머 구조에 기반한 모형으로, 동일하게 트랜스포머에 기반한 모형인 OpenAI GPT와의 차이는 양방향 트랜스포머 인코더를 사용한다는 점이다[29]. 이러한 구조는 GPT와 다른 방식의 학습 형태를 갖도록 하고 있는데, GPT는 앞 단어들 만으로 다음 단어를 예측하는 형태로 학습하기 위해 현재 위치의 오른쪽 단어들에 대해 마스킹을 이용하여 오른쪽 단어들로부터의 어텐션을 배제하는 반면, BERT는 문장에서 임의의 위치에 있는 단어들을 마스킹하고 이 단어들을 예측하기 위해 문장의 전체 단어로부터의 어텐션을 사용하는 방식으로 학습한다.

이 외에 BERT는 트랜스포머에 세그먼트 임베딩을 추가하여 문장을 구분할 수 있도록 하였으며, 각 단어 혹은 토큰의 위치를 기억하는 위치 임베딩도 추가하였다. GPT와 마찬가지로 다양한 자연어 처리 애플리케이션에 전이학습(transfer learning)을 이용할 수 있도록 지원한다. 따라서 네트워크 모형뿐만 아니라 사전학습된 가중치가 포함된 모형을 사용함으로써 학습에 소요는 자원과 시간의 부담을 줄이고 안정적인 활용이 가능하다.

BERT는 현재 수많은 자연어 처리 애플리케이션에서 가장 뛰어난 성능을 보이는 모형이기도 하다. 따라서 많은 자연어 처리 작업에서 활용되고 있는데, 이를 위한 다양한 BERT 라이브러리들은 Hugging Face(https://huggingface.co/) 인공지능 커뮤니티 사이트에서 설치하여 사용할 수 있다. BERT 라이브러리를 사용하는 간단한 사례부터 파인튜닝 적용방법까지 상세하게 설명되어 있기 때문에 쉽게 활용이 가능하다. 한국어에 대해 특화되어 설계된 BERT 모형 및 사전학습 모형으로는 SKT Brain의 KoBERT (https://github.com/SKTBrain/KoBERT)와 이를 경량화한 DistillKoBERT(https://github.com/monologg/DistilKoBERT) 그리고 ETRI의 KorBERT (https://aiopen.etri.re.kr/service_dataset.php) 등이 있다. 최근에는 기존 모델에 비해 더 적은 파라미터를 사용함으로써 계산량을 줄이면서도 비슷한 성능을 유지하는 KR_BERT가 소개되었다[30].

4.3 자연어 처리 성능 평가를 위한 데이터셋 및 평가지표

<표 15>는 분야 별로 자연어 처리 성능평가를 위한 데이터셋을 정리한 내용이다. 설명된 데이터셋은 주로 논문에서 새로 제시한 알고리즘의 평가를 위해 많이 사용된다.

<표 16>은 자연어 처리 분야 중에서 문서 요약과 기계번역에서 사용되는 측정 지표에 대한 설명이다. 그 외 문서 분류와 같은 분야는 <표 9>에서 설명한 머신러닝에서 사용하는 지표들

〈표 15〉 자연어 처리 데이터셋

분야	데이터셋명	내용
문서 분류	Reuters	1987년에 수집한 10,787개의 로이터 뉴스 기사에 대해 90개의 뉴스 카테고리로 분류한 데이터셋. 멀티라벨 즉 하나의 기사에 여러 카테고리가 할당되어 있음.
	AAPD(Arxiv Academic Paper Dataset)	55,840건의 학술저널 초록에 대해 54개의 분야로 분류한 데이터셋.
	IMDB	IMDB 사이트의 리뷰에 대해 1부터 10까지의 점수가 매겨져 있는 데이터셋.
	Yelp	Yelp Dataset Challenge를 통해 수집된 데이터로, 리뷰에 대해 1부터 5까지의 점수가 달려 있음.
문서 요약	CNN/DailyMail	CNN 신문기사와 그에 대한 요약(Highlight)의 쌍으로 이루어짐. 문서 요약의 성능 평가를 위해 사용되는 대표적인 데이터셋.
	The New York Times Annotated Corpus	CNN/DailyMail과 같이 문서 요약의 성능 평가로 사용되는 데이터셋. 1987년부터 2007년 사이에 쓰여진 뉴욕 타임즈 기사에 대해 도서관 학자가 쓴 요약문을 제공.
질의 응답	SQuAD (Stanford Question Answering Dataset)	주어진 문장(context)를 읽고, 주어진 문제(question)에 대해 올바른 답(answer)을 생성할 수 있는지 테스트하기 위한 데이터셋. 자연어 처리 중에서 질의응답(question & answer)의 성능을 측정하기 위한 데이터셋으로 사용되며, v1.1과 v2.0이 있음.
기계 번역	IWSLT Evaluation 2014～2020	The International Conference on Spoken Language Translation (IWSLT)은 매년 개최되는 학회로 기계번역에 대한 공개 평가대회가 함께 열림. IWSLT Evaluation 데이터는 이 평가대회에서 사용된 데이터셋으로 다양한 주제 및 언어에 대해 제공됨
자연어 추론	SNLI (The Stanford Natural Language Inference) Corpus	자연어 추론 평가를 위한 데이터셋로, 전제와 가설로 이루어진 한 쌍의 문장을 읽고 둘 간의 관계가 함의(entailment), 모순(contradiction), 중립(neutral) 중 어디에 해당하는지 분류하는 성능을 테스트. 이를 다양한 장르와 범위로 확장한 MultiNLI(MultiGenre NLI)가 있음. 한국어에 대해서는 이와 유사한 기능을 하는 KorNLI 데이터셋이 있음.
개체명 인식	OntoNotes	개체명 인식에 대한 성능을 측정하기 위한 데이터셋으로 개체나 특정 사건에 대해 상호참조 주석을 포함. (2) 뉴스 기사, 전화 대화, 웹 블로그, 방송 등 다양한 공개 말뭉치를 기반으로 만들어졌으며, 개체명 외에 상호참조, 참조 해소, 구문 분석 등의 정보를 담고 있음.

〈표 16〉 자연어 처리 성능 평가지표

분야	측정 지표	내용
문서 요약	ROUGE(Recall-Oriented Understudy for Gisting Evaluation)	전문가의 요약한 요약문과 자동으로 요약된 요약문을 비교하는 방법으로, 단어의 출현 순서와 일치하는 정도를 정확률(Precision)과 재현율(Recall), F-measure로 측정하는 방법. ROUGE-n은 비교할 때 n-gram 씩 순서를 고려하여 비교하여, 단어 출현 순서의 일치 여부까지 비교하여 측정함. 따라서 ROUGE-1은 두 문서 간에 겹치는 unigram의 수를 측정하는 지표이고, ROUGE-2는 겹치는 bigram의 수를 측정. ROUGE-L은 LCS 기법을 이용하여 최장 길이로 매칭되는 문자열을 측정함으로써 보다 유연한 성능 비교가 가능.
기계 번역	BLEU(Bilingual Evaluation Understudy)	기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여 번역에 대한 성능을 측정하는 방법. 기준이 되는 번역과 비교하여 단어의 수와 순서를 고려하여 점수를 계산.

을 동일하게 사용한다.

Ⅴ. 결론

자연어 처리는 최근 기계학습 및 딥러닝 기술의 발전과 적용으로 그 성능이 빠르게 향상되고 있으며, 이로 인해 의학, 한의학, 소프트웨어공학, 행정학 등 학문 분야를 막론하고 다양한 분야에서 활용되고 있다. 특히 SNS와 뉴스 및 각종 보고서 등으로 인해 비정형 텍스트 데이터가 폭발적으로 증가함에 따라 자연어 처리에 대한 관심도 더욱 높아지고 있는 추세이다. 그러나 자연어 전처리 과정 및 기계학습과 딥러닝 이론의 복잡함과 어려움으로 인해 아직도 자연어 처리 활용의 장벽이 높은 편이다.

본 논문에서는 자연어 처리의 전반적인 이해를 위해 자연어 처리와 관련한 인공지능의 분류와 기술개발 현황에 대해 살펴보고, 활발히 연구되고 있는 자연어 처리의 주요 분야와 기계학습 및 딥러닝을 중심으로 한 주요 기술의 현황에 대해 살펴봄으로써, 보다 쉽게 자연어 처리에 대해 이해하고 활용할 수 있는 기반을 제공하고자 하였다. 인공지능 기술의 분류체계와 관련하여 2006년부터의 변화과정을 봄으로써 자연어 처리와 관련한 기술이 어떻게 변화하였는지를 살펴보았으며, 특허 출원 동향의 변화를 통해 전체 인공지능 기술에서 자연어 처리가 차지하는 비중의 변화를 살펴보았다.

자연어 처리에 기계학습이 활발하게 적용됨에 따라, 3장에서는 기계학습의 주요 기술 분야와 성능 평가를 위한 데이터셋 및 평가지표를 살펴보았으며, 4장에서는 주로 딥러닝을 기반으로 한 자연어 처리 주요 분야를 언어 모델, 문서 분류, 문서 생성, 문서 요약, 질의응답, 기계번역으로 나누어 정리하고 각 분야에서 2020년을 기준으로 가장 뛰어난 성능을 보이는 모형들을 살펴보았다. 그리고, 자연어 처리에서 활용되고 있는 주요 딥러닝 모형들에 대해 정리하였으며, 마지막으로 자연어 처리 분야에서 사용되는 데이터셋과 성능평가를 위한 평가지표에 대해 정리하였다.

본 논문을 통해, 자연어 처리를 자신의 분야에서 다양한 목적으로 활용하고자 하는 연구자들이 자연어 처리의 전반적인 기술 현황에 대해 이해하고, 활용하고자 하는 목적과 관련 있는 주요 기술 분야를 파악하며, 어떤 딥러닝 기반의 모형을 활용할 수 있는지 그리고 어떤 데이터셋과 평가지표로 성능을 측정할 수 있는지 파악할 수 있기를 기대한다.

참 고 문 헌

[1] 민재옥, 박진우, 조유정, 이봉건, “BERT를 이용한 한국어 특허상담 기계독해”, 정보처리학회논문지, Vol. 9, No. 4, pp. 145-152, 2020.

[2] 김용기, 이창희, 이정민, “쇼핑몰 지능형 챗봇의 자연어 처리를 위한 패션쇼핑 개체명 인식 사전 구축”, 한국경영과학회 학술대회논문집, pp. 2744-2749, 2018.

[3] 장동진, Ubaid Ur Rehman, 정윤혜, Hafiz Syed Muhammad Bilal, 이승룡, “자연어처리 기반 진화형 임상의사결정지원시스템: 녹내장 진단 사례연구”, 한국통신학회지(정보와통신), Vol. 37, No. 9, pp. 34-39, 2020.

[4] 조병선, 이석원, “자연어처리와 기계학습을 이용한 요구사항 분석기술 비교 연구”, 한국컴퓨터정보학회논문지, Vol. 25, No. 7, pp. 27-37, 2020.

[5] 이승현, 장동표, 성강경, “자연어 처리 및 기계학습을 통한 동의보감 기반 한의변증진단 기술 개발”, 대한한의학회지, Vol. 41, No. 3, pp. 1-8, 2020.

[6] Waltz, David L., “Evolution, Sociobiology, and

the future of artificial intelligence", IEEE Intelligent Systems, Vol. 21, No. 3, pp. 66-69, 2006.

[7] Russell, Stuart, and Peter Norvig, "Artificial intelligence: a modern approach third edition", 2016.

[8] 곽현, 전성태, 박성혁, 석왕헌, "인공지능(AI) 기술 및 정책 동향", 2016 지식재산 이슈페이퍼, 한국지식재산연구원, pp. 35-76, 2016.

[9] 특허청, "4차 산업혁명 기술체계(Tech Tree)", 2018.

[10] 양희태, 최병삼, 이제영, 장훈, 백서인, 김단비, "인공지능 기술 전망과 혁신정책 방향", 과학기술정책연구원, 2018.

[11] 과학기술일자리진흥원, "인공지능(빅데이터) 시장 및 기술동향", 2019.

[12] 특허청, "4차 산업혁명 관련 기술 특허통계집", 2020.

[13] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[14] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al., "Language models are few-shot learners", arXiv preprint arXiv:2005.14165, 2020.

[15] Adhikari, Ashutosh, et al., "Docbert: Bert for document classification", arXiv preprint arXiv: 1904.08398, 2019.

[16] 박영민, 김학수, "음성 대화 시스템을 위한 심층 학습 기반 자연어 생성 기술", 정보과학회지, Vol. 39, No. 4, pp. 39-45, 2021.

[17] Zhang, Haoyu, Jianjun Xu, and Ji Wang, "Pretraining-based natural language generation for text summarization", arXiv preprint arXiv: 1902.09243, 2019.

[18] Yang, Wei, et al., "End-to-end open-domain question answering with BERTserini." arXiv preprint arXiv:1902.01718, 2019.

[19] 김영민, 임승영, 이현정, 박소윤, 김명지, "KorQuAD 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋", 정보과학회논문지, Vol. 47, No. 6, pp. 577-586, 2020.

[20] Medina, Julian Richard, and Jugal Kalita, "Parallel attention mechanisms in neural machine translation", 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018.

[21] 오영택, 김민태, 김우주. "Parallel Stacked Bidirectional LSTM 모델을 이용한 한국어 영화 리뷰 감성 분석", 정보과학회논문지, Vol. 46, No. 1, pp. 45-49, 2019.

[22] 김선우, 최성필, "Bidirectional LSTM-CRF 기반의 음절 단위 한국어 품사 태깅 및 띄어쓰기 통합 모델 연구", 정보과학회논문지, Vol. 45, No. 8, pp. 792-800, 2018.

[23] Yoon Kim, "Convolutional neural networks for sentence classification", arXiv preprint arXiv: 1408.5882, 2014.

[24] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le, "Sequence to sequence learning with neural networks", Advances in neural information processing systems, 2014.

[25] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473, 2014.

[26] Vaswani, Ashish, et al., "Attention is all you need." Advances in neural information processing systems, pp. 5998-6008, 2017.

[27] Radford, Alec, et al., "Improving language understanding by generative pre-training", 2018.

[28] Brown, Tom B., et al., "Language models are few-shot learners", arXiv preprint arXiv:2005. 14165, 2020.

[29] Devlin, 2018, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language

understanding", arXiv preprint arXiv:1810.04805, 2018.

[30] 이상아, 장한솔, 백연미, 박수지, 신효필, "소규모 데이터 기반 한국어 버트 모델", 정보과학회 논문지, Vol. 47, No. 7, pp. 682-692, 2020.

저 자 소 개

박 상 언(Park, Sang-Un)

·현재 경기대학교 소프트웨어경영대학 경영정보전공 교수로 재직 중이며, 한국과학기술원 전산학과에서 학사, 한국과학기술원 경영공학과에서 석사 및 공학박사학위를 취득하였다. 주요 관심분야는 텍스트마이닝, 딥러닝, 머신러닝, 프롭테크 등이다.