동국대학교 동서사상연구소
『철학·사상·문화』 제41호 2023. 01. 294-312쪽
http://dx.doi.org/10.33639/ptc.2023..41.014

인공 지능은 상식을 배울 수 있을까? *

박명관**

< 목 차 >

【요약문】 최근 컴퓨터 공학 학제의 일부로서 인공 지능 기술의 발달은 인공 지능 컴퓨터 시스템이 전통적으로 사람이 해 온 언어 처리(이해 및 생성)를 사람처럼 수행할 수 있도록 하고 있다. 사람처럼 언어를 처리하는 시스템을 신경망 언어 모델이라고 하는 바, 신경망 언어 모델을 활용하여 인공 지능의 본래의 정의에 맞게, 인공 지능 컴퓨터 시스템이 인간의 학습 능력, 지각 능력, 추론 능력을 갖추게 하려고 노력하고 있다. 인간의 학습, 지각, 추론 능력을 일반적으로 상식(common sense)이라고 정의할 때, 본 연구의 목적은 먼저 인공 지능을 실질적으로 구현하기 위하여 신경망 언어 모델에서 신경망 상식 모델을 개발하는 최근의 시도들을 먼저 살펴보고, 다음으로 인간의 지능을 컴퓨터로 구현할 때 적용 방법론 및 컴퓨터 구현 가능성으로 인하여 인간 지능과 인공 지능이 서로 괴리가 예상되는 바, 상식의 측면에서 두 지능의 차이를 가져오는 원인 그리고 이 차이점을 극복하기 위한 시도들을 살펴보고 평가하는데 있다.

【주제어】 인공지능, 상식, 신경망 언어/상식 모델, 언어, 지식, 지능

* 이 논문은 2021년도 동국대학교 연구년 지원에 의하여 이루어졌음
** 동국대학교 영어영문학부 교수

1. 서론

인공 지능 기반 컴퓨터(이하, 인공 지능)는 최근 여러 분야에서 급진적인 발전을 보여 왔으며 한때는 불가능하게 보였던 바둑 및 포커와 같은 게임에서 인간 챔피언을 완전히 이겼다. 음성 인식, 기계 번역, 사진 태깅(tagging)과 같은 다양한 영역의 발전은 일상이 되었다. 그러나 Davis & Marcus(2015)에서 지적한 것처럼, 현재의 인공 지능은 근본적인 것이 여전히 빠져 있다. 그것은 일반적인 상식이다.

상식(common sense)은 일반적으로 우리가 가지고 있는 지식으로, “사람들은 돈을 잃어버리고 싶어하지 않는다”, “돈을 지갑에 보관한다”, “주머니에 지갑을 보관한다”, “칼은 물건을 자른다”, “물건은 이불을 덮어도 사라지지 않는다” 등과 같이 일반 사람들이 보통 알고 있는 일종의 기본 지식이다. 우리는 상식이 없이는 일상 세계를 이해하기 어렵다. 마찬가지로 인공 지능이 상식을 갖추지 못하면 소설, 뉴스 기사, 영화 등을 이해하기 어렵다.

인공 지능은 거의 모든 사람이 갖고 있는 상식이 결여되어 있으며 상식 기능을 갖고 있는 인공 지능을 개발하는 효과적인 방법이 아직 일반화되지 않고 있다. 이 문제를 해결하기 위하여 Davis & Marcus(2015)가 지적한 이후 인공 지능 기술 분야에서 인공 지능의 상식 습득에 관한 대표적인 시도로서 2015년 이후 빠르게 발전한 신경망(neural-network) 언어 모델을 사전 훈련하여 상식을 생성/추론하는 신경망 상식 모델을 개발해 왔다. 본 연구에서는 이 신경망 상식 모델의 특징을 살펴보면서 현재 이 모델의 성격과 성능, 그리고 사람의 상식과 인공 지능의 상식의 차이, 다음으로 이 차이의 원인과 이를 해소하기 위한 시도들을 살펴볼 예정이다.

2. 상식

2.1 상식의 정의

상식은 1973년판의 『The Shorter Oxford English Dictionary』에서 네 가지 의미로 정의되고 있다. 먼저, 전통적 의미는 “오감의 공통된 결합 또는 중심으로

간주되는 내적 감각", 두 번째 의미는 "보통의, 정상적인 또는 평균적인 이해", 그렇게 이해하지 않으면 사람이 "멍청하거나 미쳤다고 간주될 수 있는" 상태이다. 세 번째 의미는 "인류의 일반적인 감각 또는 공동체의 일반적인 감각(이것의 두 하위 의미는 좋은 건전한 실용적 감각과 일반적인 현명함)"이다. 네 번째는 철학적 의미로 "주요 진리(를 파악하는) 능력"이다.

또한 상식의 의미는 추가적으로 다양하게 정의되어 왔다. *Merriam-Webster Online Dictionary*는 "상황이나 사실에 대한 단순한 인식에 기초하여 내린 건전하고 신중한 판단"으로 정의한다. 『Cambridge Dictionary』는 "우리 모두가 합리적이고 안전한 방식으로 살 수 있도록 도와주는 기본적 수준의 실용적인 지식과 판단"으로 정의한다. van Holthoorn & Olson(1987, p. 9)은 "다소 보편적이고 반성이나 논쟁 없이 유지되는 지식, 판단, 심미안으로 구성된다"고 정의하고, Lewis(1967, p. 146)는 "보통 사람의 기본적인 정신적 복장"이라고 본다.

상식에 대한 다양한 이해와 정의에도 불구하고, 인공 지능에 구현하려고 하는 상식(또는 단순히 감각)은 영어 Wikipedia(https://en.wikipedia.org/wiki/)가 제시하는 일반적 정의에 가깝다. Wikipedia에서 상식은 "일상적인 문제에 관한 정상적이고 실제적인 판단 또는 거의 모든 사람이 공유하는(즉, 공통적인) 방식으로 인식, 이해 및 판단할 수 있는 기본적인 능력"으로 정의된다. 따라서 Davis & Marcus(2015)가 말하는 인공 지능이 상식이 결여되어 있다라고 하는 것은 현재의 인공 지능이 상식을 가진 일반 사람과 달리 일상적으로 직면하는 문제에 대해 정상적이고 실제적인 인식, 이해, 판단을 할 수 없음을 의미한다.

2.2 인공 지능이 구현하는 상식

최근 인공 지능이라는 용어가 회자되고 있으며, 그 개념이 매우 혼동스럽게 사용되고 있다. 나무위키(https://namu.wiki/w/)의 사전적 정의에 따르면, 인공 지능(人工知能) 또는 Artificial Intelligence(AI)는 일반적으로 인간의 학습 능력, 지각 능력, 추론 능력이 필요한 작업을 할 수 있도록 컴퓨터 시스템을 구현하려는 컴퓨터 공학의 세부 분야 중 하나이다. 인간을 포함한 동물이 갖고 있는 지능 즉, 자연 지능(natural intelligence)에 대응되는 개념이다. 다시 말해, 인공 지능은 인간 지능의 기능을 갖춘 컴퓨터 시스템이며, 인간 지능을 기계 등

에 인공적으로 구현한 것이다. 일반적으로 이는 범용 컴퓨터에 적용한다고 가정한다. 이 용어는 또한 그와 같은 지능을 컴퓨터 시스템에 구현할 수 있는 방법론이나 실현 가능성 등을 연구하는 과학 기술 분야를 지칭하기도 한다.

인공 지능의 정의에서 중요한 것은 인간의 학습 능력, 지각 능력, 추론 능력이 필요한 작업을 사람을 대신하여 수행할 수 있도록 인간의 지능을 구현한 컴퓨터 시스템이 인공 지능이다라는 점이다. 그런데 여기서 고려할 만한 것은 지식과 지능의 구분이다. 먼저 지식은 '배우거나 실천하여 알게 된 명확한 인식이나 이해'를 말하고, 지능은 '사물이나 현상을 이해하고 대응하는 지적 능력'을 말한다. 따라서 지능은 지식의 인식이나 이해를 넘어 사물이나 현상을 대응, 대처를 하는 행위(agency)의 개념이 추가되어 있다. 사람을 대신하여 사람이 일상적으로 처리하는 작업을 수행할 수 있도록 사람의 학습 능력, 지각 능력, 추론 능력을 갖춘 컴퓨터 시스템을 인공 지식이 아니라 인공 지능으로 부르게 된 이유이다.

바로 이 지점에서 인공 지능이 상식을 필요로 한다. 인공 지능을 도입하는 목적이 인간을 대신하여 인간이 하는 일을 수행하고자 할 때 인간이 가지고 있는 일반적인 학습, 지각, 추론 능력이 필요한 바, 이 세 가지 학습, 지각, 추론을 가능하게 하는 기본적인 능력으로서 인식, 이해, 판단 능력을 아우르는 상식이 없다면,[1] 인공 지능은 사람을 대신하여 사람이 해야 하는 일을 정상적으로 실제할 수 없음을 의미한다. 따라서 컴퓨터 공학의 한 분야로서 인공 지능 분야의 연구자들은 인간이라면 누구나 가진 상식을 컴퓨터 시스템에 학습시키는 효과적인 방법론 그리고 이에 따르는 상식의 습득 방법론의 실현 가능성을 고려해 왔다.

인공 지능에 상식을 학습시키는 것은 큰 도전적 작업이다. 앞서 언급한 것처럼, 상식은 사람이 일상적인 문제를 대처할 때의 일반적인 인식, 이해, 판단 능력으로 세분화되기도 하지만, 각각의 능력을 규정하기도 힘들고, 인공 지능에 각각의 능력을 어떤 방법으로 학습시켜야 할지를 추정하기도 힘들다. 이와 같은 인공 지능에 상식을 학습시키는 난제에 관련하여 최근 미국 U. of Washington의 Yejin Choi 교수 연구팀은 인공 지능의 상식 습득이라는 도전적 문제 해결을 위해 '생각의 전환'을 바탕으로 인간 경험의 총체적 자료가 글로 쓴 언어 자료에

1) 상식과 지능 둘 다 우리가 일상 생활에서 내리는 좋은 판단과 결정과 관련한 자질이다. 그럼에도, 상식과 지능은 또 구분된다. 상식은 실용적인 문제에 있어서 좋은 감각과 건전한 판단력인 반면, 지능은 지식과 기술을 습득하고 적용하는 우리의 능력이다.

있다고 보면서 인공 지능이 학습하여야 할 상식은 언어 텍스트 자료에서 도출되어야 한다고 제안한다(이와 관련하여, Yejin Choi 교수 연구팀의 발표/출판해온 Bosselut et al.(2018); Choi (2022); Forbes et al.(2019); Hwang et al.(2022); Rashkin et al.(2018); Sap et al.(2019); Shwartz et al.(2018); West et al.(2022) 등을 참고).

구체적으로, Yejin Choi 교수는 언어, 더 정확히 언어 텍스트는 명백하게 부정확성과 가변성에도 불구하고 세계가 작동하는 방식에 대한 방대한 양의 상식적인 사실과 규칙을 포괄할 수 있을 만큼 충분히 표현력이 풍부하고 견고하다고 본다. 결국 인간이 세상에 대한 지식을 습득하는 것은 논리적 형식이 아니라 언어라고 본다. 그리고 언어가 세상에 대한 지식을 제공한다는 것은 언어의 모호함과 책, 뉴스, 심지어 과학 문헌에 보고된 지식의 불일치에도 불구하고 사실이다. 따라서 인간 수준 지식 습득의 규모와 복잡성에 맞추기 위해 인공 지능은 언어를 직접적으로 통합하지 않고는 더 멀리 진보할 수 없다.

또한 인공 지능에 상식 기능을 추가하는 개발 방법은 방대한 코퍼스, 컴퓨팅(computing), 그리고 신경망 딥러닝(deep learning 혹은 심층 학습)을 사용하기 이전에도 시도되었다. 딥러닝은 언어학/심리학 전문가들이 구축해 놓은 인간의 심볼릭(symbolic) 상식 지식 그래프(commonsense knowledge graph)와 융합된 방대한 양의 원시 텍스트를 사용하여 신경망 상식 모델을 훈련할 수 있는 완전히 새로운 방법론이 될 수 있다.[2] 다시 말하지만, 언어를 활용하여 컴퓨터 시스템에 상식을 학습시키는 것은 신경망 언어 모델에서 신경망 지식 모델로의 강력한 전이 학습(transfer learning)을 허용하므로 심층 신경망의 경험적 혁신으로부터 큰 이점을 얻을 수 있다.

---

2) 뉴로-심볼릭 인공 지능(Neuro-symbolic AI)는 신경망(neural) 및 기호(symbolic) 인공 지능 아키텍처를 통합하여 각각의 장단점을 보완하여 학습(learning), 인지(cognitive), 추론(reasoning) 모델링이 가능한 견고한 인공 지능을 제공한다고 본다. Henry Kautz, Francesca Rossi, 그리고 Bart Selman도 신경망 및 기호 인공 지능 아키텍처의 합성을 주장했다. 이들의 주장은 Daniel Kaneman의 책 『Thinking Fast and Slow』에서 논의된 두 가지 종류의 사고를 통합/합성할 필요성에 기초하고 있다. Kaneman은 인간의 사고를 시스템 1과 시스템 2의 두 가지 요소로 설명한다. 시스템 1은 빠르고 자동적이며 직관적이며 무의식적이다. 시스템 2는 더 느리고 단계적이며 명시적이다. 시스템 1은 패턴 인식(pattern recognition)에 사용된다. 시스템 2는 계획, 연역, 숙의적 사고를 다룬다. 이 관점에서, 신경망 딥러닝은 첫 번째 종류의 인지를 가장 잘 처리하는 반면, 상징적 추론은 두 번째 종류의 인지를 가장 잘 처리한다. 둘 다 학습하고, 추론하고, 인간과 상호 작용하여 피드백을 받아들이고, 질문에 답할 수 있는 강력하고 신뢰할 수 있는 인공 지능을 구현하는 데에 필요하다.

물론, 인간이 일상적인 문제에 대해 일반적인 인식, 이해, 판단은 직관적이며, 이와 같은 직관적인 추론은 노력 없이 이루어진다. 인간은 일상에서 경험하는 거의 모든 사물, 사람, 사건에 대해 무의식적으로 항상 직관적 추론(intuitive inference)을 한다.[3] 우리가 부분적으로만 관찰하는 장면의 큰 그림 맥락에 대해 즉각적인 판단을 내리는 것이 직관적인 추론이다. 또한 우리는 타인의 동기 및 의도, 그리고 그들의 정신적, 감정적 상태가 무엇인지를 추론한다. 직관적인 추론이 매우 자연스럽고 노력을 들이지 않기 때문에 인공 지능에게도 쉬울 것이라고 우리는 가정한다.

Choi(2022)에 따르면, 직관적 추론은 (가능한 모든 대안 사건에 대해 전적으로 분별적인 추론(discriminative inference)과 달리) 즉각적이고 생성적이다. 직관적 추론의 공간은 무한하므로 (고정된 미리 정의된 레이블 세트와 달리) 이를 설명하기 위해 자연어 전체를 필요로 한다. 하지만 직관적 추론은 특히 함의의 경우에는 본질적으로 예측적이므로 추가적인 맥락에서 거의 항상 기각될 수(defeasible) 있다. 또한 직관적인 추론은 물리적 그리고 사회적 세계가 어떻게 작동하는지에 대한 풍부한 배경 지식에서 도출된다.

우리가 직관적인 추론을 언어로 전달할 때, 대안적 추론(의 일부)을 명시적으로 인정하지 않고, 단어 하나하나 즉시적으로 가장 가능성이 높은 직관적 추론을 언어로 표현하여 생성하는 것과 거의 같다. 이것은 우리가 사용하는 "소리내어 생각하는(speak aloud)" 방법과 유사다. 마음 속에서 먼저 생각의 다음 부분을 끝내거나 다가올 문장의 정확한 표현을 계획하지 않고도 생각의 다음 단어를 말해 나아간다.

직관적인 추론에 대한 이러한 관찰은 언어를 통해 즉각적인 생성 추론을 처리할 수 있는 전산 추론 생성 모델의 필요성을 제기한다. 해결해야 할 주요 문제는 직관적인 추론을 생성할 수 있는 규모(scale)이다. 인간이 자신의 생각을 전달하는 방식과 마찬가지로 직관적인 추론을 즉각적으로 생성하는 방법을 학습할 수 있는 새로운 기계 학습(machine learning) 모델과 알고리즘(algorithm)을 필요로 한다. 실제로, 이러한 단어별 생성은 오늘날 신경망 언어 모델에서 텍스트 생

3) 우리의 추론 대부분은 직관적인 추론이다. 이 경우 추론의 이유에 대해 주의하거나 의식 없이 추론의 결과를 사실이라고 믿는다. 이와 달리, P_1, P_2, …P_n에서 C를 추론한다면, 이것은 반사적인 추론(reflective inference)이다. 이 경우 추론 C를 믿는 이유에 대해 주의하고 의식한다.

성이 작동하는 방식이다. 예를 들어, OpenAI의 GPT-3(Generative Pre-trained Transformer 3)은 음성과 유사한 텍스트를 생성하기 위해 딥러닝으로 학습한 언어 모델로 모든 대안 문장을 명시적으로 열거하지 않고 한 번에 한 단어만 샘플링하여 놀랍도록 일관된 단락을 생성한다. 신경망 언어 모델의 발전은 언어 기반 즉석 생성 추론 시스템을 구축하기 위한 강력한 기술 기반을 제공한다.

3. 신경망 언어 모델에서 신경망 상식 모델로

최근 Allen Institute for AI 및 University of Washington에서 Yejin Choi 교수가 주축이 되어 상식 모델링으로 발전한 COMET(COMmonsEnse Transformers)은 언어, 즉 방대한 텍스트 문장을 미리 학습한 신경망 언어 모델을 이용하여 전이 학습(transfer learning)을 통해 즉각적으로 추론을 생성한다. COMET은 “인간의 공통적 일상 경험의 사회적 측면과 물리적 측면을 글로 표현한 텍스트의 방대한 상식 저장소”를 통해 훈련되었다. COMET에 어떤 문장을 입력하면, 이 문장의 주성분(즉, 주어, 목적어 등)과 과거, 미래 및 현재 사건, 인물 및 조건과의 관계를 예측한다.

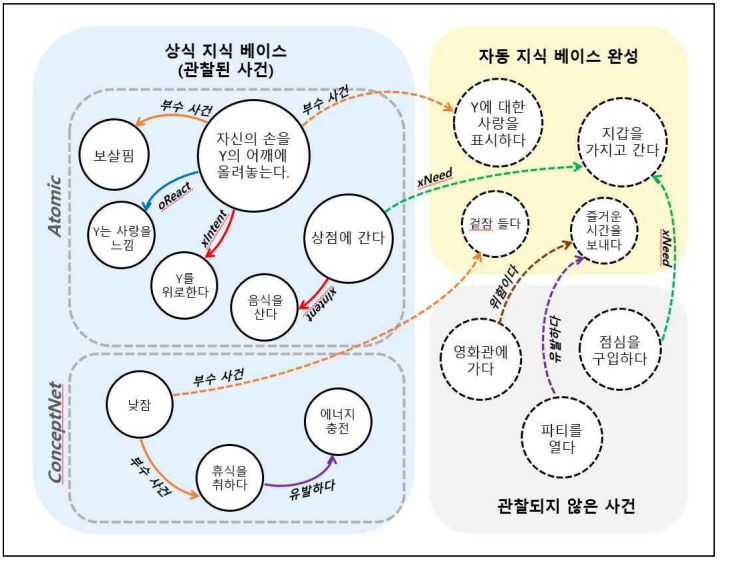
입력 문장 “Gary는 불쏘시게와 통나무를 쌓아놓고 성냥 몇 개를 떨어뜨린다”가 주어졌을 때 COMET은 즉각적으로 상식 추론을 생성한다. 이 신경망 상식 모델은 문장의 주어 Gary(즉, PersonX)가 "불을 피우고 싶어"할 수도 있고, 그렇게 하기 전에 Gary가 "라이터를 가져오기"를 원할 수도 있다고 예측한다. 이 특정한 예는 신경망 언어 모델의 상식 추론 기능의 한계에 대하여, 앞서 언급한 인지 과학자 Gary Marcus의 비판에 대한 응답이 될 수 있다. 실제로 사전 학습된(pre-trained) 신경망 언어 모델은 상식 지능이 견고하지 못하기 때문에, Yejin Choi 교수 등은 COMET와 같은 신경망 상식 모델의 개발을 추진하게 되었다.

지난 수십 년간의 대부분의 상식 추론 컴퓨터 시스템과 대조적으로, COMET의 기본 개념 프레임워크는 (논리 기반 형식주의와 반대로) 상식 지식의 언어 기반 형식주의와 (고정 범주 집합에 대한 분별적 예측과 반대로) 무한한 공간의 직관적인 추론을즉각적으로 생성하는 방식을 통합하고 있다. COMET은 세계가 어떻게 작동하는지에 대해 신경망 언어 모델이 상식 지식을 학습하기 위해 맞춤화

된 교과서로 볼 수 있는 사람의 심볼릭 지식 그래프인 ATOMIC(an ATlas Of MachIne Commonsense, Sap et al. (2019)) 위에 구축되었다.

특정 주제에 대해 선언적 지식(declarative knowledge)을 제공하는, 인간을 위해 작성된 교과서와 유사한 ATOMIC은 일상적인 사물과 사건에 대한 상식적 규칙과 사실에 초점을 맞춘 선언적 지식 저장소이다. ATOMIC에 인코딩된 지식의 예는 [그림 1]의 왼쪽 상단에 나와 있다.

<그림 1>



구축 당시 ATOMIC은 130만 개 이상의 상식적인 규칙과 사실을 기반으로 한다. 이 자료는 규모가 큰 것처럼 보이지만 실제로 130만 개의 규칙과 사실은 우리 인간이 세상에 대해 가지고 있는 모든 일반적 상식을 포괄하기에는 여전히 너무 제한적이다.

그러나 ATOMIC에서 훈련된 COMET은 ATOMIC에서 기술하는 제한된 범위의 심볼릭 지식을 훨씬 넘어서 상식 추론을 일반화할 수 있다. 따라서 [그림 2]의 (a)와 (b)와 같이 이전에 관찰할 수 없었던 상황에 대해 놀랍도록 정확한 상식 추론을 할 수 있다.[4)]

<그림 2>-(a) ATOMIC-2020에서의 두 사건의 상식 추론

사 건	관 계	추 론
사건①: 누군가가(X) 독감에 걸렸다	결과적으로 X는 [ ] 원한다	• 건강에 좋은 것을 먹고 물을 계속 마신다. • 주치의를 부르다 • 휴식을 취한다 • 물을 계속 마신다 • 닭죽을 먹는다
사건②: X 감기에 걸렸다	앞서서, X는 감기에 걸리려면 [ ] 필요하다	• 감기에 걸린 사람을 접촉했다 • 추운 외부에 있었다 • 특별한 상황이 없었다
	결과적으로 X는 [ ] 원한다.	• 약을 먹는다 • 약을 구입한다 • 잠을 잔다 • 낮잠을 잔다 • 휴식을 취한다

<그림 2>-(b) COMET & ATOMIC-2020에 의해 생성된 네 사건의 상식추론[5)]

사 건	관 계	추 론
사건①: X가 감기에 걸리다	앞서서, X가 [ ] 필요하다	• 특별한 상황이 없었다 • 오한이 있었다 • 열이 있었다 • 병이 있었다 • 약을 먹었다
사건②: X	유발하다	• X가 발진이 있었다

4) COMET에서 주어진 문장에 대해 추론을 유도하는 관계(relation)는 다음 9가지이다: xNeed(x가 필요한다), xWant(원한다), xIntend(의도한다), xReact(반응한다), xEffect(결과적으로 초래한다), xAttr(속성을 갖는다), oEffect(결과적으로 초래한다), oWant(원한다), oReact(반응한다). 여기서 x는 문장의 주성분(주로 주어)이고 o는 대상(일반적으로 목적어)가 된다.

5) 두 상식 모델 ATOMIC-2020와 COMET-ATOMIC-2020의 트리플(사건-관계-추론 세 요소로 구성된) 사례는 두 모델 모두 적용 범위가 불충분하고 일반화가 부족하다. (a)에서 "누군가(X)가 독감에 걸렸다"에 대한 결과로 가능한 것은 대부분의 질병에 적용되지만 "누군가(X)가 감기에 걸렸다"의 결과에는 포함되지 않는다. 또한 감기의 근본적인 원인인 "아픈 사람 곁에 있었다는 것"이 포함되어 있다. 그러나 (b) COMET는 감기와 수두의 원인을 예측하지 못한다. 감기와 수두 둘 다 동일한 근본 원인을 가진 전염성 질병의 사례임에도 불구함에도 그러하다. "전염성 질환에 걸린다"에 대한 추상적이지만 직접적인 질문도 그렇다. 대신 "누군가(X)는 닭죽을 먹는다"와 같은 피상적인 텍스트 추론이 포함된다. 또한, COMET은 피부 질환의 공통적인 결과(즉, "발진이 있었다")를 생성하지만, 피부 질환 중 하나인 수두로 일반화하지는 못한다.

가 피부병에 걸리다		• X가 알레르기 체질이다 • 특별한 상황이 없었다 • 병이 있다
사건③: X가 전염병에 걸리다	앞서서, X가 [ ] 필요하다	• 특별한 상황이 없다 • 주치의를 보러 가다 • 병에 걸리다 • 병으로 아파왔다
사건④: X가 수두에 걸리다	앞서서, X가 [ ] 필요하다	• 주치의를 만났다 • 특별한 상황이 없다 • 수두에 걸렸다 • 의사를 만났다 • 병원을 갔다
	결과적으로 X는 [ ] 원한다	• 주치의를 만난다 • 의사를 만난다 • 병원을 간다 • 응급조치를 취한다
	결과적으로 X는 [ ] 판단한다	• X는 주치의를 만나다 • X는 독감에 걸린다 • X는 감기에 걸린다 • X는 닭고기 스프를 먹는다
	유발하다	• X는 수두에 알레르기 체질이다. • X는 주치의를 만났다 • X는 의사를 만났다 • X는 감기에 걸렸다 • 특별한 상황이 없었다

COMET의 상식 추론 일반화 능력은 언어의 신경망 언어 모델에서의 표현과 사람의 상식 지식의 심볼릭 표현 사이의 전산적 융합을 통해 달성된다. 실제로 COMET의 성공적인 상식 추론은 지식의 신경망 언어 모델에서의 표현과 사람의 상식의 심볼릭 표현의 혼합, 그리고 심볼릭 지식의 표현 매체로서의 언어의 활용에 기인할 수 있다. 지식과 추론 사이의 연속체를 인식하는 것도 중요하다. 지식과 추론은 일반적으로 별개의 지적 현상으로 간주되기 때문에 이것은 직관에 반하는 것처럼 보일 수 있다.

그러나 언어, 지식 및 직관적 추론에 대한 전산적 탐구는 광범위한 실제 사례를 접했을 때 지식과 추론 사이의 경계가 명확하지 않다는 것을 보여준다. 보다 구체적으로, “Gary의 불쏘시게 그리고 장작더미 쌓기”의 의도에 대해 추론할 때, 우리의 추론은 사람들이 일반적으로 불쏘시게와 장작으로 무엇을 하는지에 대한 기억된 상식 지식에 의존하고 있다.

이와 반대로, 사람들의 의도와 정신 상태, 사건의 원인과 결과, 사건의 전제 조건과 사후 조건에 대한 상식 추론의 일반적인 패턴은 모두 세상이 어떻게 작동하는지에 대한 우리가 기억하는 지식의 필수적인 부분이 된다. 요컨대, COMET은 상식 생성을 하는 인공 지능을 개발해 나아가기 위한 새로운 방법론으로서 언어, 지식 및 추론 사이의 뉴로 심볼릭 혼합을 보여준다. 이 조합이 없었다면, 이전에 관찰할 수 없었던 상황에 대해 유연하게 추론하는 COMET의 놀라운 일반화 능력을 얻을 수 없다.

4. 인공 신경망이 생성하는 상식에 대한 평가

Yejin Choi 교수가 신경망 언어 모델에서 신경망 상식 모델을 개발하는 시도는 대단한 성과를 이루었다. 신경망 언어 모델은 글로 쓰여진 인류의 지식 자원인 텍스트 코퍼스를 학습했다는 점에서 인류가 오래 동안 획득한 지식 총합을 습득했다고 말할 수 있다. 이를 바탕으로 신경망 상식 모델을 개발하는 시도는 지식에서 상식을 도출할 수 있다는 가정을 바탕으로 이루어진 것이다.

언어 지식에서 상식을 도출하는 획기적인 시도에도 불구하고, 먼저 인공 신경망 모델이 생성하는 상식과 인간이 구사하는 상식은 어떻게 다른가의 질문이 제기된다. 앞서 언급한 것처럼, 인공 지능의 개발은 지능 등 인간의 능력을 컴퓨터 시스템에 구현할 수 있는 방법론이나 실현 가능성 등의 제약을 받아 인간의 능력을 있는 그대로 구현하기 보다는 구현 시점에서의 기술 개발의 수준에 종속되게 된다. 따라서 신경망 모델의 상식과 인간의 상식이 서로 다르다라는 점은 어느 면에 있어서 자연스럽게 예상된다.

인간의 상식 능력과 관련하여, 인간은 유전적 특질(genetics)을 통해 사람이 갖는 상식이 어디서 비롯되는지에 대한 기초가 되는 지능을 물려받는다. 다시 말해, 인간은 어떤 면에서 유전적 지능을 통해 상식을 가지고 태어난다. 따라서 사람들은 어느 정도의 상식을 가지고 있고 이것은 사람들이 가지고 타고난 것이고 자연스러운(natural) 것이다. 사람들은 무엇을 하고 있다는 것을 의식하지 않고 자신의 상식을 사용한다. 이것이 자연스럽게 이루어지기 때문이다. 그러나 사람들은 성장하면서 끊임없이 자신의 지식과 기술을 확장하게 되는 데, 우리가 직면하는 많은 사물과 상황을 경험하면서 상식을 쌓게 된다. 사람들은 나이가 들수

록, 앞선 학습 경험에서 배웠을 것처럼, 여러 사물과 상황을 통해 상식을 갱신하게 된다. 결국 사람들은 유전적 지능을 통해 상식을 가지고 태어나지만, 상식 능력의 많은 부분이 또한 학습에서 나온다.

물론 현재 개발된 인공 신경망 언어/상식 모델은 경험적 학습을 바탕으로 언어 지식과 상식을 습득한다. 신경망 모델을 구동하는 내재된 신경망 알고리즘이 인간에게서 유전적 특질을 통해 어어받는 생득적 능력에 대응할 수 있다. 따라서 인간의 상식과 인공 신경망 모델의 상식을 역공학(reverse engineering)[6] 등을 통하여, 특히 후자가 구현하는 것처럼 특정한 문장이 주어졌을 때 추론하는 상식의 양상에 관해 면밀한 비교가 필요하다고 본다. 이에 관한 체계적 비교 연구로 Yejin Choi의 연구팀에서 AUTOMIC-2020 그리고 COMET & ATOMIC-2020은 주어진 문장에 대하여 사람의 상식 추론에 대비하여 약 43% 비슷하다고 보고하고 있다. 이것은 아직도 상식 추론에 있어 신경망 상식 모델이 인간 수준에 도달하기 위하여 개선할 점이 많다는 것을 보여주며, 특히 사람이 쉽고 자연스럽게 상식을 추론하는 생득적 능력으로서의 인간의 상식 추론 능력에 신경망 모델의 상식 추론 알고리즘이 필적하지 못하다고 볼 수 있다.

신경망 모델의 아키텍처(architecture) 상의 문제와 함께, 일반적으로 지적되는 것처럼, 신경망 모델의 성능은 신경망 모델이 학습하는 훈련 데이터세트의 성격과 연동된다. 특히, 신경망 모델의 상식 추론과 관련 이슈가 되는 것이 훈련 데이터세트의 소위 '보고 편향(reporting bias)'의 문제이다. 사람은 사물을 볼 때 실제 자주 일어나는 주요한 것보다, 본인이 봤을 때 흥미롭고 자세히 살펴보고 싶어하는 것을 보고(report)하는 경향이 있다. 이를 보고 편향이라고 말한다. 만약 인공 지능이 텍스트만 공부를 한다면 예를 들어 바나나는 대부분 노란색이란 사실을 모를 수 있다. 왜냐하면 사람들은 어떤 것의 '전형적인 것'을 바탕으로 세상을 이해하고 저장하기 때문이다. 사람들은 어떤 것이 이것의 전형성과 일치하면, 보통 언급하지 않는다.

---

6) 역공학(逆工學)은 이미 만들어진 장치, 프로세스, 시스템 또는 소프트웨어가 어떻게 작동하는지를 연역적 추론을 통해 이해하려는 과정이다. 다시 말해, 시스템 등이 어떻게 작동하는지 분석하고 이를 복제하거나 향상시키기 위해 시스템을 해부하는 과정이다. 역공학 절차는 역공학이 적용되는 객체에 특정하지만, 이 절차는 일반적으로 세 가지 기본 단계로 구성된다: 정보 추출, 모델링 및 검토. 정보 추출은 시스템의 작동을 설명하는 관련 정보를 수집하는 것을 말한다. 모델링(modeling)은 수집된 관련 정보를 추상적인 모델로 결합하는 것을 말한다. 검토는 정리된 정보의 유효성을 보장하기 위해 모델을 테스트하는 것을 말한다.

또 '선택 편향(selective bias)'이라는 개념도 있는데, 기계 학습을 구현하면서 훈련 데이터를 입력할 때, 실제 세상에서 일어나는 무작위 샘플의 훈련이 아니라 연구자가 흥미롭다고 생각하는 샘플을 쓰는 경향이 있다. 이에 따라 제한된 정보나 구체적이지 못한 정보를 바탕으로 결론을 내는 것을 '과잉 일반화'라고 한다. 또한 잘 회자되는 것처럼 역사적으로 오랫동안 성차별로 인하여, 예를 들어 인공 지능에게 물리학자의 사진을 학습시킨다면, 과거 대부분의 물리학자는 남자였기 때문에, 인공 지능은 잠재적으로 남자에게 편향된 알고리즘을 갖게 될 수 있다. 물론 기계 학습에서 딥러닝으로 넘어오면서 폭발적으로 많은 훈련 데이터 세트를 사용하게 되어 선택 편향의 문제는 다소 완화되었다. 또한 소량의 자료에 의한 과적화(over-fitting)-일반화의 문제도 최근 딥러닝 프레이밍워크 하에서 소량의 데이터에 의한 소위 '퓨샷 학습(few-shot learning)'을 채택하면서 완화되었다. 결국 편향을 충분히 예측하고 제거하려는 시도를 한다면, 딥러닝의 결과물은 사람의 행동 그리고 사람의 지능에 어느 정도 근접하며 우리가 기대하고 타당하고(reasonable) 합당한(fair) 인공 지능을 구현할 수 있을 것이다.

한편 현재 신경망 언어/상식 모델의 언어, 의미 학습은 분포 의미론(distributional semantics)을 기반으로 이루어진다. 분포 의미론은 대규모 표본의 언어 데이터에서 언어 표현 간의 분포 특성을 바탕으로 언어 표현 간의 의미적 유사성을 정량화하고 분류하기 위한 이론과 방법을 개발하고 연구하는 분야이다. 분포 의미론의 기본 개념은 소위 분포 가설(distributional hypothesis)로 요약될 수 있다: 분포가 유사한 언어 표현은 유사한 의미를 갖는다. 즉, 단어(의의미)는 같은 맥락에 출현하는 동료 단어에 의해 특징지어진다라는 1950년대 John R. Firth의 관찰에서 출발한다. 분포 가설은 최근 언어 습득/학습에서 유사성 기반 일반화(similarity-based generalization) 이론의 근거를 제공한다. 아이들은 비슷한 단어들의 분포로부터 단어들의 사용법에 대해 일반화함으로써 그들이 전에 거의 접하지 않았던 단어들을 사용하는 방법을 알아낼 수 있다. 분포 의미론은 선형 대수를 전산 도구 및 표상 프레임워크로 사용하는 것이 일반적이다. 기본 접근법은 고차원 벡터(vector)에서 분포 정보를 수집하고, 벡터 유사성 측면에서 분포/의미적 유사성을 정의하는 것이다. 분포 가설의 타당성 여부는 전산 모델링의 데이터 희소성(data sparsity) 문제, 그리고 상대적으로 빈곤한 입력이 주어졌을 때 (즉 잘 알려진, 자극의 빈곤(poverty of stimuli) 문제)) 어떻

게 아이들이 언어를 빠르게 배울 수 있는지에 대한 질문 둘 다 영향을 미친다.

신경망 상식 모델의 상식 추론도 분포 의미론의 방법론을 따른다. 예를 들어, “철이가 배가 고프다”라는 문장이 주어졌을 때, 이 문장이 추론하는 문장/명제가 무엇인지를 예측하는 방식은 신경망 상식 모델에 질의-응답(Q&A)의 형식으로 질문(예를 들어, “철이가 배가 고프다”인 경우 “철이가 무엇을 원하니”)--을 주면, 신경망 상식 모델은 신경망 언어 모델이 학습한 문맥화된 표상을 바탕으로 “철이가 배가 고프다”와 분포적 확률의 측면에서 같이 출현하는 빈도가 높은 문장, 예를 들어 “철이가 음식을 먹고 싶다”를 통해, 신경망 상식 모델은 “철이가 원하는 것”이 “철이가 음식을 먹다”라는 응답으로 출력하게 된다.

앞서 살펴 본 것처럼, 최근 인공 지능은 대규모 신경망 언어 모델과 상식 지식 그래프(commonsense knowledge graph, CKG)을 통해 상식을 획득하고 모델링하는 데 진전을 이루었다. 그럼에도, 개념화, 즉 세상 존재물과 상황을 추상적 개념의 사례로 보고 이를 바탕으로 추론을 하는 것은 상식 추론을 하는 인간 지능의 필수적인 요소이다. 현재 개념화는 인공 지능 구현에 아직 완전히 도입되지 않았으며, 이러한 점에서 현재의 인공 지능 구현의 접근 방식은 현실 세계의 무수한 다양한 존재물과 상황에 대한 상식 지식을 다루는 데 비효율적이다.

이 문제를 해결하기 위해, He, Fang, Wang, & Song(2022)의 연구팀은 상식 추론에서 개념화의 가능한 역할을 연구하고, 추상 개념에 대한 추상적 지식을 습득하는 것으로부터 인간의 개념 유도를 재연하는 프레임워크를 형식화한다. 분류법 Probase을 활용하여 대규모 인간 주석으로서 상식 지식 그래프인 ATOMIC에 대한 상황별 개념화 도구를 개발한다. He et al. 은 사건 및 (사건, 관계, 추론으로 구성된) 트리플(triple) 레벨 둘 다에서 ATOMIC에 대한 개념화 유효성을 보장하는 데이터 세트에 주석을 달고, 언어적 자질을 기반으로 일련의 휴리스틱 규칙을 개발하고, 추상적 지식을 생성하고 검증하기 위해 일련의 신경망 모델을 훈련시킨다. 이러한 구성 요소를 기반으로 추상적 지식을 습득하는 파이프라인이 구축된다. 다음으로, ATOMIC에 대한 대규모 추상 CKG를 유도하여, 관찰하지 않았던 새로운 존재물 또는 상황에 대해 추론을 하게 된다. 또한 추상적인 트리플 레벨로 데이터를 실험하고 직접 보강함으로써 상식 모델링의 성능을 개선한다.

이와 같은 텍스트 중심의 신경망 언어/상식 모델의 기능을 개선하려는 시도와

함께 새로운 대안적 방식을 모색하기도 한다. 앞서 살펴본 것처럼, 신경망 상식 모델의 상식 추론 과정에서 우리가 확인할 수 있는 것은 사람은 언어를 '이해'하면서 이에 따라 어떤 문장/명제를 추론한다면 신경망 모델은 문장이 출현하는 맥락을 통해 가까이 출현하는 문장/명제를 추론 문장으로 제시한다는 점이다. 사람의 언어 사용과 달리 신경망 상식 모델에는 언어 이해가 결여되어 있다라는 점은 Bender & Koller(2020) 등이 오래 전에 지적하고 있다. 딥러닝의 신경망 모델을 바탕으로 한 연구 방법론은 신경망 모델의 학습의 결과를 평가하기 위한 정량적인 방법으로 여러 가지 metric을 제시하고, 이 값이 높을수록 "사람의 언어를 이해하는 것에 가까워진다"고 합의를 본 정도이지 복잡한 언어체계를 사람처럼 이해한다고는 말할 수 없다. 다시 말해, 딥러닝 기반 연구에서 신경망 모델이 언어를 "이해한다"는 표현의 정의가 과연 언어학에서의 "이해"와 같은 의미로 사용되었는지에 대해서는 의문을 갖게 한다.

Bender & Koller(2020)는 텍스트만으로 자연어의 진정한 이해를 달성하는 것은 불가능하며, 사람과 같은 의미 이해를 이루기 위해서는 어떤 종류의 의미론적 '지각 접지(perception grounding)'가 필요하다고 주장한다. 이를 채택하는 인지 및 로봇 공학 분야의 연구자들은 인공 지능 개발에서 접지 원리를 구체화하고 있다. 접지 원리의 근간은 인지 과학의 4E이다(이와 관련하여 장병탁(2018)을 참고). 인지 과학에서 확장형(**E**xtended), 체화형(**E**mbodied), 구현형(**E**mbedded), 행화형(**E**nactive) 인지, 즉 4E 인지의 제안자들은 인지 발달 이슈를 새로운 각도에서 접근한다. 인지 과학의 4E의 지지자들은 인간과 동물의 인지 발달 과정에서 없어서는 안될 자연 학습 환경의 속성에 중점을 두는데, 4E 인지 연구의 일반적 방법론적 토대는 상향식 이론화 및 모델링이다. 학습 로봇 에이전트가 새로운 행동을 배우고 실행할 수 있는 방법에 대하여 인간의 직관에 의존하기 보다는, 이 접근법은 자연주의적 조건에서 학습을 경험적으로 테스트과 시뮬레이션, 그리고 분석하는 설정만으로 '이미 제공된' 것과 학습 로봇 에이전트가 나머지를 달성하는 방법을 모색한다. 종종 이 접근 방식은 복잡한 지능형 행동이 학습 로봇 에이전트와 해당 환경의 직접적인 상호작용에서 발생할 수 있다는 관찰로 이어진다. 따라서 언어 지능을 구축하기 위한 4E 구성 요소에 대한 추가 분석은 학습자와 관련된 특정 아키텍처 또는 기술보다는 시뮬레이션 설계에 주로 초점이 맞춰진다.

5. 결론

지난 2013년에 본격적으로 도입된 인공 신경망 언어 모델은 글로 표현된 인류의 지적 자원인 대규모 텍스트 자료를 학습하고 문장에서 단어의 쓰임을 설명하고 예측하는 데 획기적인 성능을 발휘하고 있다. 인공 신경망 언어 모델은 10여 년간 학습 성능을 높일 수 있는 보다 발전된 알고리즘을 적용하면서 눈부시게 진화하였다. Yejin Choi 교수 연구팀은 지식 저장소로서의 인공 신경망 언어 모델을 바탕으로 특정한 문장이 추론하는 문장을 생성하는 인공 신경망 상식 모델을 최근 개발해 왔다. 본 연구에서는 인공 신경망 상식 모델이 생성하는 추론 상식의 성격을 살펴보고자 했다. 앞서 지적한 것처럼, 인공 지능 분야는 궁극적으로 인간 지능을 기계적으로 구현하려는 목표를 가지고 있지만 인공 지능 컴퓨터 시스템에 구현할 수 있는 방법론이나 실현 가능성의 제약을 받는다. 다시 말해, 컴퓨터 시스템에 인공 지능을 현시점에서 구현할 때 현재의 기술적 수준에 달려있다. 따라서 인공 지능을 구현하려는 시도로서 최근신경망 상식 모델에서 문장에서 추론하는 '상식'은 사람들이 문장을 읽으면서 산출하는 '상식'과 상이한 성격을 갖지만, 지적 자원의 총합으로서 인공 신경망 언어 모델의 지식 자원에서 상식 지식을 도출하려는 연구는 현재의 기술적 수준에서 최선의 시도라고 평가할 수 있다. 이와 같은 긍적적 평가와 함께, 신경망 상식 모델의 성능 개선을 위해 텍스트 데이터 상에서의 보고 편향 등의 문제를 추가적으로 해결할 필요가 있다, 또한 개념화, 즉 세상 존재물과 상황을 추상적 개념의 사례로 보고 이를 바탕으로 추론을 하는 것은 상식 추론을 하는 인간 지능의 필수적인 요소인 바, 신경망 상식 모델에서도 개념화 성능 개선을 모색하여야 하겠다. 다음으로, 현재의 신경망 상식 모델에서의 상식 추론이 확률 기반 분포 의미론 방법론으로 이루어지는 바 이를 극복하기 위하여 사람의 언어 의미 이해를 반영할 수 있는 체화형 인공 지능 구현의 필요성을 제기하였다.

참고문헌

장병탁, 「인간지능과 기계지능 – 인지주의 인공지능」, 『정보과학회지』 제36집 1호, 한국정보과학회, 2018.

Bender, E. M. & Koller, A., "Climbing towards NLU: on meaning, form, and understanding in the age of data", *Proceedings of the* 58th *Annual Meeting of the Association for Computational Linguistics*, in Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.463. https://www.aclweb. org/anthology/2020.acl-main.463

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A. & Choi, Y., "COMET: Commonsense transformers for automatic knowledge graph construction", in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, Florence, Italy, July 28-August 2, 2019, Association for Computational Linguistics, 2019.

Choi, Y., "The curious case of commonsense intelligence" in *Daedalus* Vol.151 No.2, 2022.

Davis, E. & Marcus, G., "Commonsense reasoning and commonsense knowledge in artificial Intelligence" in *Communications of the ACM* Vol.58 No.9, 2015.

Forbes, M., Holtzman, A. & Choi, Y., Do neural language representations learn physical commonsense? in *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation*, Montreal, Canada, July 24-27, 2019. cognitivesciencesociety.org.

He, M., Fang, T., Wang, W. & Song, Y., "Acquiring and modelling abstract commonsense knowledge via conceptualization", 2022. arXivpreprint arXiv:2206.01532.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A. & Choi, Y., 2020. (Comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs in *The*

Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI) 2021: 6384-6392. Virtual Event, February 2-9, AAAI Press, 2021.

Kahneman, D., *Thinking, fast and slow*. London: Penguin Books, 2011.

Lewis, C. S., *Studies in words*, Cambridge: Cambridge University Press, 1967.

Olson, D. R. & Holthoon, F. L. v., *Common sense: The foundations for social science*, Lanham: University Press of America, 1987.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A. & Choi, Y., "Event2mind: Commonsense inference on events, intents, and reactions" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (ACL) 2018: 463-473. Melbourne, Australia, July 15-20, 2018, Association for Computational Linguistics, 2018.

Sap, M., Bras, R. L., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A. & Choi, Y., "Atomic: An atlas of machine commonsense for 'if-then' reasoning" in *Proceedings of the AAAI Conference on Artificial Intelligence 33:* 3027-3035, 2019.

Shwartz, V. & Choi, Y., "Do neural language models overcome reporting bias? " in *Proceedings of the 28th International Conference on Computational Linguistics*(COLING), Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020.

West, P., Bhagavatula, C., Hessel, J., Hwang, J. D., Jiang, L., Bras, R. L., Lu, X., Welleck, S. & Choi, Y., "Symbolic knowledge distillation: From general language models to commonsense models" in *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2022.

Abstract

Can artificial intelligence learn common sense?

Park, Myung-Kwan*

The recent development of artificial intelligence technology as part of the computer science discipline has enabled artificial intelligence computer systems to perform the language processing (understanding and generation) that humans alone are able to do. This human-like system of language processing is called a neural-network language model, and the AI computer system leverages such a neural network language model to emulate human learning, perception, and reasoning skills, living up to the original definition of artificial intelligence. Encompassing learning, perception, and reasoning with commonsense, this work first looks at recent attempts to develop neural network commonsense models from neural network language models to practically implement more human-like artificial intelligence. Next, when implementing human intelligence mechanically, human intelligence and artificial intelligence are expected to differ from each other due to the mapping problems the methodology and implementation confront, so we examine and evaluate the causes of the differences between the two types of intelligence in terms of commonsense.

【Key words】 artificial intelligence, common sense, neural network language/common sense model, language, knowledge

* Professor, Dongguk University
** 논문접수일: 2022. 12. 20. 논문심사기간: 2022. 12. 26. ~2023. 01. 28. 게재확정일: 2022. 01. 31.