

인공지능 언어모델의 상식적 판단

송상헌
(고려대학교 언어학과)

세종어제훈민정음에

나랏말싸미랏어는

문자와로써스스디아니

호씨이런저글이런피

성이니르거저훈배이셔

무춤내제브들시러피디

훈글미하나라이네우하

야어엇비너가새글스미

느랏저네름이.거거스네

하영수.방그.거.거.네.네.네

아.아.아.아.아.아.아.아

상식 = 보통지식

사람들이 보통 알고 있거나 알아야 하는 지식. 일반적 견문과 함께 이해력, 판단력, 사리 분별 따위가 포함된다.

상식 ≠ 전문지식

적어도 성인이라면 누구나 알고 “있어야” 마땅할 정도로 우리가 사는 세계에 대한 보편적 지식을 말한다.

상식 → 사고력

주어진 상황에서 어떠한 일반적 결론을 도출할 수 있어야 ‘상식적 판단’을 갖추었다고 평할 수 있다.

상식의 보통성

- ① 그러므로, 우리의 역사는 분석의 역사가 아니라 종합의 역사요, **전문적 역사가 아니라 상식적 역사**다. 《역사와 민족》
- ② 정보통신 인프라라면 사무용 건물에나 적용될 것이라는 것이 **일반인의 상식**이다. 《월간중앙 5월호》
- ③ 저자는 매우 복잡한 현상이나 **전문용어**로 기술될 실험장치들까지도 **상식적인 수준에서** 이해시키고자 한다. 《천리안의 자연과학 서적 서평》

상식과 교육의 관계

- ① 이일보다도 한 수가 위라는 신립이 또 있으니 두 장수가 문경 새재쯤 지킬 것은 **당연한 상식 이하의 상식**이었다.《박종화, 임진왜란》
- ② 내 질문에 현구는 **너무나 상식적인 물음**이란 듯 멀뚱한 얼굴로 나를 올려다 보았다.《김원일, 노을》

설명을 위한 도구

- ① 그 양반이 태남이에게 보여 주던 그 깊은 관심과 자애는 **상식으로는 설명할 수 없는 거**여서 그 당시엔 노망으로 돌렸고...《박완서, 미망》
- ② 세 살 먹은 **어린애**가 본다 해도 이 사건은 **상식적으로 말도 안 돼요**. 《이호철, 문》

사회적 기대치

- ① 그것은 병원에서 진단할 것이지만 **워낙 사건이 상식 밖의 일이라서** 저의 질문에 대한 답변을 신용하기 위해서입니다.《김용성, 리빠똥 장군》
- ② 거기다가 아무리 확신범의 특질을 감안한다 할지라도 범행 동기는 너무도 **상식과 먼 거리에 있었다**.《이문열, 사람의 아들》

상식과 윤리

- ① 그는 결점을 늘어놓아 상대방 기분을 해치는 것만큼 **몰상식**은 없다고 생각했다.
- ② 그는 그녀가 공연장에서 떠들고 웃는 **몰상식한 행동**을 보이자 그만 나와 버렸다.
- ③ 교양이란 거죽에만 살짝 덮여졌을 뿐이고, 그 속에는 **온통 상식 이하의 것들**이 도사리고 있다. 《인간적인 진실로 인간적인》
- ④ 소위 해방군이라고 자처하는 저들이 **수준이 상식 이하**라는 사실에 놀라지 않을 수 없었습니다. 《끝이 없는 이 길을》
- ⑤ 상식적으로 용납할 수 없는 일을 범하거나 **통상적인 윤리로 받아들이기 어려운 일**을 묵과하는 태도는 명철한 생활인의 자세일 수가 없다. 《인간과 사회 -전통윤리와 현대풍조의 갈림길에서》

세계 지식의 패턴화

- ① 그 상품을 입수(入手)하자면 그만한 대가를 치뤄야 하는 것이 당연한 일이요, **영화를 구경하자면 관람료를 지불해야 하는 것은** 상식 이하의 상식이 아니고 무엇이던가. 《비석과 금강산의 대화》
- ② 안 해도 될 일을 해야 할 때는 **무슨 속셈이 있다고 보는 게** 상식 아니겠소? 《하오의 戀歌》
- ③ 평소 안경을 끼던 사람은 벗고 안 끼던 사람은 끼는 것이 **도피의 상식**이다. 《광야의 끝에서》
- ④ 남성이 만들어 낸 '**남성=사회역할, 여성=가정역할**'이라는 효율적인 역할 구분은 마땅히 여자는 여자답고 남자는 남자다워야 한다는 상식 세계의 터전을 다져 놓았다. 《여성의 일곱가지 콤플렉스》

오류가능성

- ① 어제의 상식만으로는 **오늘의 문제**를 풀 수 없다. 《정보교육》
- ② 우리들 사이에 **통용되어온 상식에 어떤 모순**이 있지 않은가? 《사회를 보는 논리》
- ③ 그리하여 어릴 때부터 가족, 학교, 매스컴 등을 통해 **여성이 열등하다는 것은 상식**으로 굳어 갔다. 《여성의 일곱가지 콤플렉스》
- ④ 잘못된 **의약 상식**과 보신에 대한 집착이 결합된 결과라고 할 수 있다. 《사회를 보는 논리》

상식의 일반적 특성

- A. 상식은 역사적 배경을 통해 문화인이라면 누구나 **암묵적으로 공유**하고 있는 보통 지식이다.
- B. 그 내용을 궁금해하거나 혹은 그에 반하는 행동을 하는 것은 **분별력** 있는 사람에게는 이상하게 받아들여지게 된다.
- C. 상식적 판단은 예상 가능한 범주 안에 들어야 하므로, 우리 사회의 여러 현상에 대한 **설명의 도구로서 기능**한다.
- D. 많은 경우 상식은 **윤리도덕, 예의범절, 규범질서** 등과 같은 선상에서 취급된다.
- E. 때에 따라 관찰을 통해 **유추적 패턴**을 만들어 상식적 판단을 생성할 수도 있다.
- F. 이 과정이 항상 완벽한 것은 아니기 때문에 상식은 **틀릴 수 있고 또 그러기에 수정될 수 있다**.

딥러닝 기반 언어 모형의 발전 동향

I heard a rumor ...

- GPT3의 서사구조
- 완벽하다... 사람보다 뛰어나다...

전기자동차

전기가 동력원이며 내연기관 대신
전동기로 구동력을 발생시킨다.

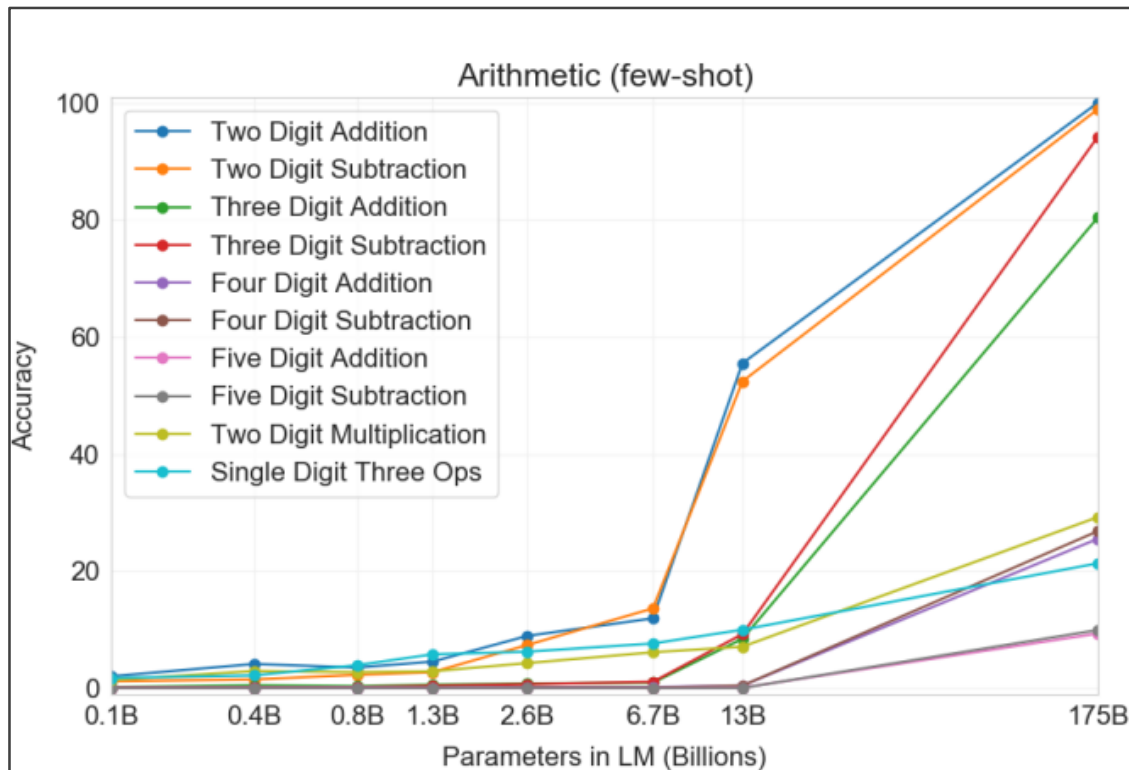


How to Cosplay
#3 Allison



beyond the **surface** languages

- Brown, Tom, et al. "Language models are **few-shot learners**." Advances in neural information processing systems 33 (2020): 1877–1901.
- Wei, Jason, et al. "**Finetuned language models** are **zero-shot learners**." arXiv preprint arXiv:2109.01652 (2021).
- Kojima, Takeshi, et al. "**Large Language Models** are **Zero-Shot Reasoners**." arXiv preprint arXiv:2205.11916 (2022).
- Srivastava, Aarohi, et al. "**Beyond the Imitation Game**: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).



GPT3의 사칙연산
(Brown et al. 2020, 22쪽)

beyond the **surface** languages

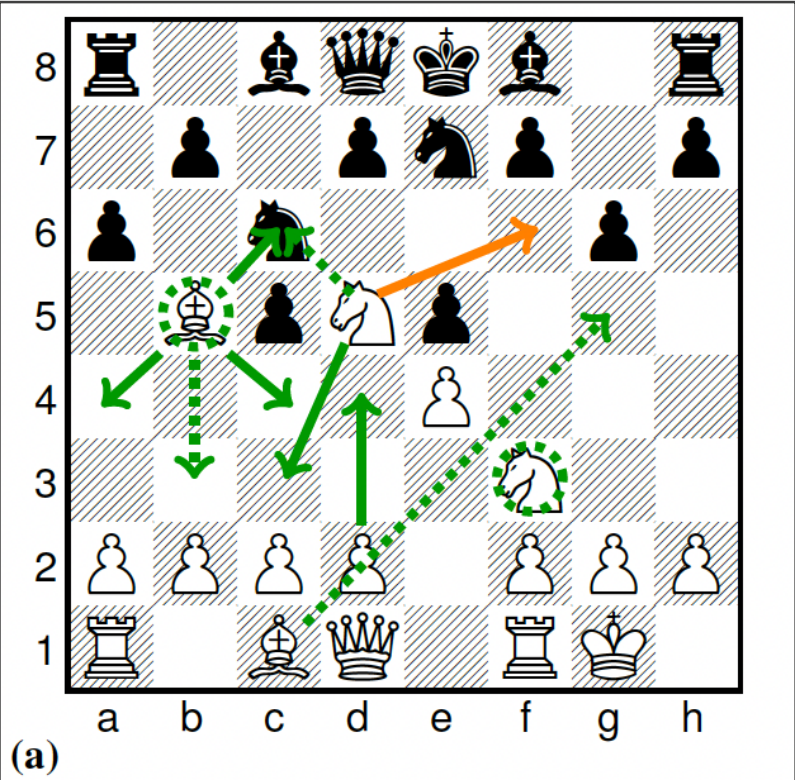
- Brown, Tom, et al. "Language models are **few-shot learners**." Advances in neural information processing systems 33 (2020): 1877–1901.
- Wei, Jason, et al. "**Finetuned language models** are **zero-shot learners**." arXiv preprint arXiv:2109.01652 (2021).
- Kojima, Takeshi, et al. "**Large Language Models** are **Zero-Shot Reasoners**." arXiv preprint arXiv:2205.11916 (2022).
- Srivastava, Aarohi, et al. "**Beyond the Imitation Game**: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).

No.	Template	Accuracy
1	Let's think step by step.	78.7
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

**프롬프트에 따른 정확도 향상
(Kojima et al. 2022, 8쪽)**

beyond the **surface** languages

- Brown, Tom, et al. "Language models are **few-shot learners**." Advances in neural information processing systems 33 (2020): 1877–1901.
- Wei, Jason, et al. "**Finetuned language models** are **zero-shot learners**." arXiv preprint arXiv:2109.01652 (2021).
- Kojima, Takeshi, et al. "**Large Language Models** are **Zero-Shot Reasoners**." arXiv preprint arXiv:2205.11916 (2022).
- Srivastava, Aarohi, et al. "**Beyond the Imitation Game**: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).

[illegible]

언어 능력 → 학습 능력 → 추론 능력



- EBS 특집 다큐멘터리 '교육실험 프로젝트, 똑똑한 책읽기'
- 10주간의 '자기주도 독서실험' 10주 실험 후 **시험 성적 향상 눈에 띄어**



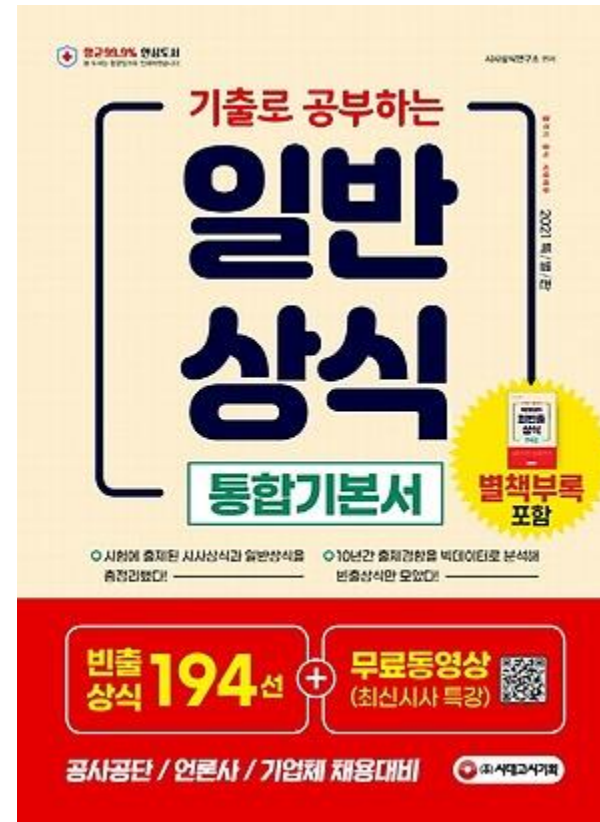
- **개념이해능력**: 본초자오선(本初子午線)
- 어설프게 배운 모국어는 아이의 **상상력과 추론 능력**에 도움이 되지 않는다

- 공부를 잘하려면 읽기 능력을 키워라
- 책 읽는 아이가 **학업역량, 자기효능감, 진로성숙도** 높다

상식: Common Sense

정의 | definition

- 사람들이 **보통** 알고 있거나 **알아야 하는 지식**. 일반적 견문과 함께 이해력, 판단력, 사리 분별 따위가 포함된다.
- 특정 사회의 구성원이 **특별한 교육을 거치지 않고도 암묵적으로 공유**하고 있는 지식체계
 - **세계 지식**:
 - 코트디부아르의 수도는?
 - 속초의 특산물은?
 - **도구**: 손톱깎이
 - **경험**: 비가 오면 관절이 아프다
 - **상대성**: 간짜장과 계란 프라이



상식은 언어가 아니다. 하지만 언어로 포장되어 있다.

Information Packaging

- Engdahl and Vallduví (1996)
 - 어떠한 정보를 담아서 전달하는 매개체로서 언어의 기능
 - 언어라는 형식에 포장되어 전달되는 정보가 세대를 이어 형성되면 → 상식

도구 논항

- Rissman et al. (2015)
 - Using instruments to understand argument structure: Evidence for gradient representation
 - 도구는 동사의 의미에 따라 표상되는 논항이라고 할 수 없음.

Surprisal



The chef stirred the savory stew with ...

a spoon
a whisk

7.993397
11.96577

4.100479
6.620835

언어 인공지능과 상식의 만남

예시 | example

- 식당 서빙 로봇
- 마트 관리 인공지능: 비가 오는 날
- 윤리의 문제
 - 당신은 **윤리적**인 사람입니까?
 - 당신은 **상식적**인 사람입니까?
 - 비윤리적 인간 / 몰상식한 인간
- Ask Delphi: 특정 시나리오에 따라 도덕적 여부를 판단하는 AI
 - AI는 사람이 준 데이터로 학습한다. 결국 AI의 윤리는 그걸 가르치는 사람의 문제다. 그걸 최대한 바로잡는 것, **상식에 부합하는 AI를 만드는 것이** 이 프로젝트의 목표다. (최예진 교수님 인터뷰, 2021.12.06 조선일보)



AI2 Allen Institute for AI

Delphi Ask Delphi

Delphi is a research prototype designed to model people's moral judgments on a variety of everyday situations. This demo shows the abilities and limitations of state-of-the-art models today.

To learn more: [paper](#) / [article](#) / [FAQ](#)

Model outputs should not be used for advice for humans, and could be potentially offensive, problematic, or harmful. The model's output does not necessarily reflect the views and opinions of the authors and their associated affiliations.

Try one of these examples:

- Ignoring a phone call if the phone call is urgent.
- Helping a friend in need if they break the law.
- Legitimizing racism for the greater good.
- Can I torture a cat if it makes me happy?

* Input a **situation** for Delphi to ponder:

Example: Robbing a bank.

Ponder

Stochastic Parrots



Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell. 2021. **On the Dangers of Stochastic Parrots**: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Pages 610–623.



메타의지

인공지능 언어모델은 자연 언어에 대한 실질적인 이해(understanding)를 하고 있는 것이 아니라 단순히 **확률에 기반하여 단어의 조합을 결과값으로 반환(return)**한다는 점에서 **확률적 앵무새**이다.

소통 가능성 (Communicability)

- 학습데이터의 문제
- 모라베크의 역설
- 아스퍼거 증후군: **행간의 의미**
- 맥락 지식: **항진 명제**



가추 추론: abductive reasoning

추론 (推論)

- 이미 **알고 있는 정보로부터** 논리적 결론을 도출하는 행위 또는 과정
- 어떠한 판단을 근거로 삼아 **다른 판단**을 이끌어내는 것
- 특정한 **명제**에서 다른 **명제**를 이끌어내는 과정

가설적 추론 (假推法, 歸推法)

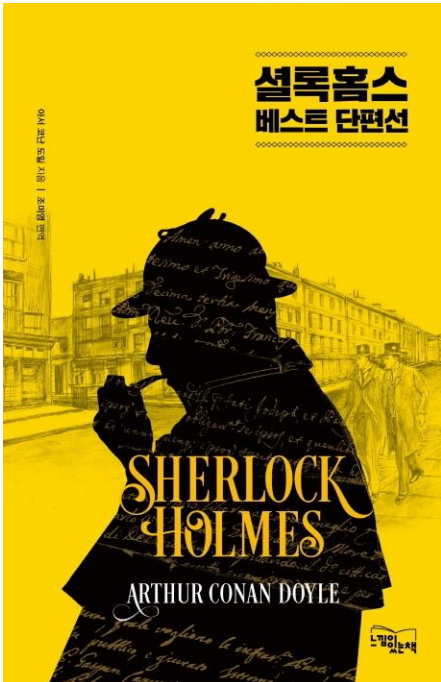
- (놀라운) 현상이나 사건을 **가장 잘 설명할 것 같은 가설**을 선택하는 방법
- 주어진 **사실로부터 시작해서** 가장 그럴 듯한 **최선의 설명**을 도출
- 의식하지 못하지만 일상 생활에서 가장 많이 사용하는 추론 방식

If you hear hooves,
think horse, not zebra.



Charles Sanders Peirce
(1839-1914)

설명적 가설: explanatory hypothesis



콩자루의 예 (퍼스의 전기 관점)

• 연역법

S1: 저 자루안의 콩은 모두 하얗다.
S2: 이 콩은 저 자루에서 나왔다.
S3: 그러므로 이 콩은 하얗다.

• 귀납법

S1: 이 콩들은 모두 저 자루에서 나왔다.
S2: 이 콩들은 모두 하얗다.
S3: 그러므로 저 자루안의 콩은 모두 하얗다.

• 가추법

S1: 저 자루안의 콩은 모두 하얗다.
S2: 이 콩은 하얗다.
S3: 그러므로 이 콩은 저 자루에서 나왔다.

설명 가능성

• 3단계 구성

- ① 어떠한 **사실**로부터
- ② 어떠한 **상황/사건**에 대하여
- ③ 가장 **그럴 듯하게 설명**할 수 있는 가설 생성

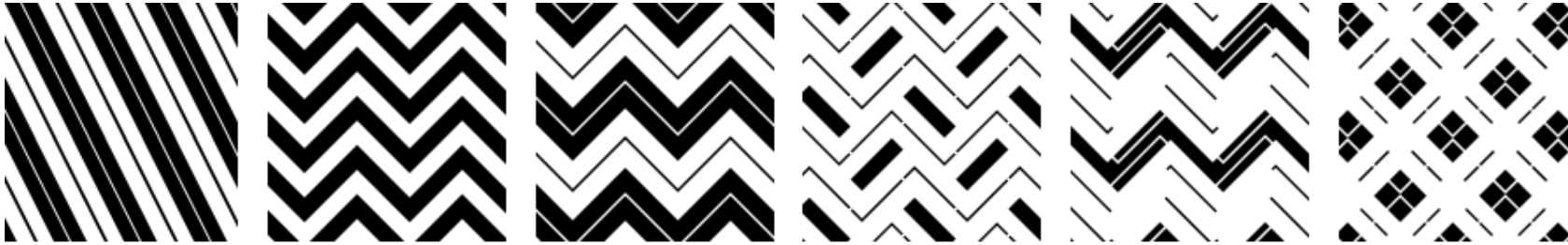
• 특징

- ① 정도성 degree: 확률모델
- ② 십중팔구 likelihood: 틀릴 수도 있다
- ③ 상대성 relativity: 사람, 문화, 맥락마다

• 가추추론이 재소환되는 이유

지식/상식을 확장하는데 기여하는 유일한 추론으로서, 과학발전에 기여함.

세계에 대한 지식과 경험을 패턴(Pattern)화



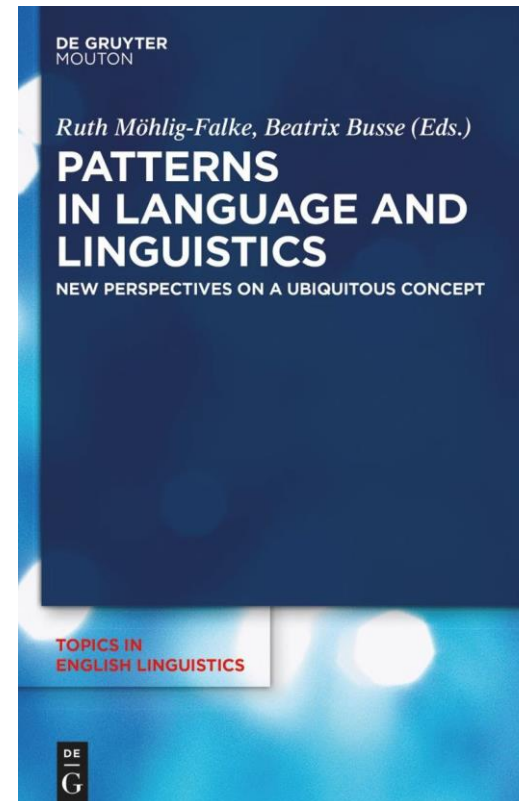
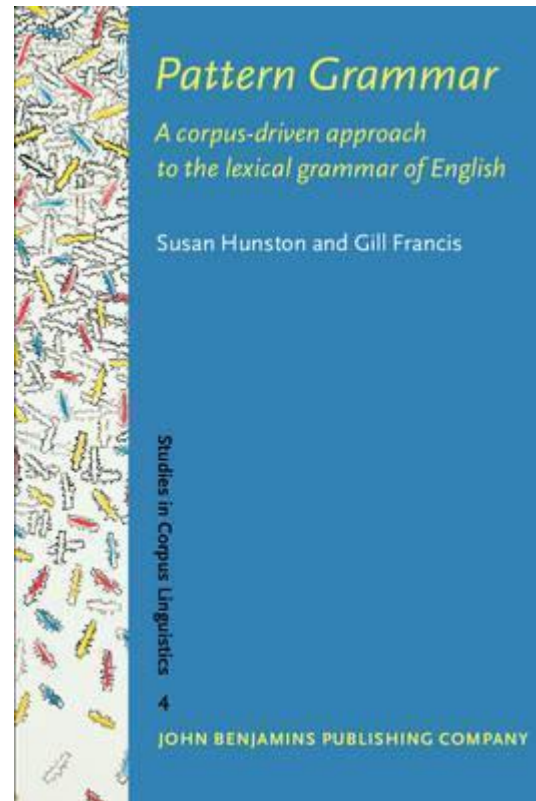
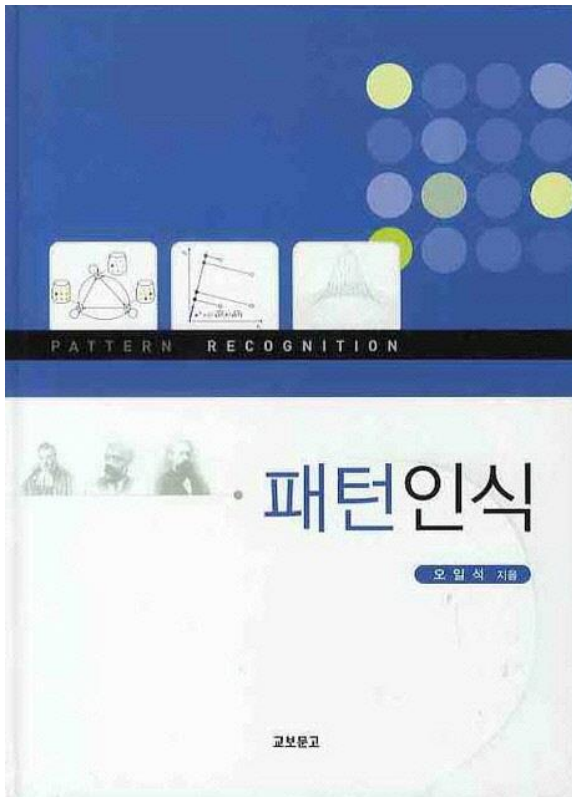
사람 vs. 컴퓨터

• 데이터 간섭 현상

- ① 사람은 한두번만으로도 대상에 대한 패턴 정보를 얻는 것이 가능하다.
- ② 컴퓨터는 충분한 양의 데이터를 통해서 학습을 해야 한다.

• 메타의지

- ① 사람은 내가 무엇에 대한 패턴을 만드는지를 의식적으로 인지한다.
- ② 컴퓨터는 자신이 무엇을 학습하는지를 모르고 수행한다.



사실과 추론 능력을 적재적소에 쓸 수 있는가?

가설을 도입하기 위한 예비 추정

- 연역법과 귀납법에 전제되는 가설들은 추론에 앞서 선형적으로 존재하는 것이 아니므로, 현상을 탐구하고 추론하는 과정에서 **가설의 형성 단계**를 간과할 수 없다. 가추추론은 **모든 탐구의 가장 첫 번째 단계에 해당**한다.

인간 뇌 작동의 고유 특성

- 몇 번의 시도 만에 **가장 그럴듯한 가설**을 제시할 수 있는 인간의 능력은 **진화의 과정에서 파생**되었다.
- 가추추론 능력은 **인간의 지성이 자연법칙의 이해와 설명에 특별히 적응된 결과물**이다.

입력과 출력

- 각각의 **질문과 답변이 1:1로 코드화되어 있는 것이 아니라**, **새로운 입력**이 제시되었을 때 이에 **적절하게 반응할 수 있는 새로운 출력**을 생성하여 도출할 수 있어야 한다.
- 가추추론의 또다른 의의는 **비록 그 가설이 참이 아닐지라도 개연성 있게** 설명했다는 데에 있다.

- Fann, Kuang Tih. 1970. Peirce's theory of abduction, Martinus Nijhoff, 1970.
- 이윤희. 2013. Semiotic Understanding of Mimetic Action from Peirce's Perspective: Towards Narrative Cognition and Symbolization. 기호학 연구 36: 197-217.
- 최재웅. 2020. 가추법과 일상대화. 2020 세미오시스 여름학교, 한국외대
<https://youtu.be/Kb3wwGpmJYQ> | https://youtu.be/bXbLfyL_No | <https://youtu.be/sWbGAjDsuBQ>

언어 추론

- Bhagavatula, C., ... Choi, Y. (2020). Abductive Commonsense Reasoning. In ICLR.

Given These Partial Observations ...



O₁ It was a very **hot** summer day.

O₂ He felt **much better!**

The More Plausible Hypothesis is ...

He decided to **run in the heat**.

H⁻ ❌ 🤖

He drank a glass of ice cold water.

H⁺ ✅

DEV

그림 추론

- Liang, C., Wang, W., Zhou, T., & Yang, Y. (2022). Visual Abductive Reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.



Event Descriptions

E_1 : A man falls off a running horse.

E_2 : The man lies on the ground and gets hurt.

E_3 : The man is taken to the hospital.

Legal Example

Premise (E_1 : , E_2 : ), Explanation (E_3 : )

Premise (E_1 : , E_3 : ), Explanation (E_2 : )

Illegal Example

Premise (E_2 : , E_3 : ), Explanation (E_1 : )

이야기 | STORY

언어 Language
맥락 Context
추론 Reasoning
소통 Communication

- 이야기를 이해하고 만들 수 있는가?
- 그림으로 설명할 수 있는가?

Q: 아래 담화에서 가운데 문장으로 더 적절한 것은 어느 것인가?

국립국어원, 2022년 이야기
완성 평가 말뭉치 연구 분석.

S1: 열이 38.5도가 나와서 병원을 찾았다.

(A) 병원에서는 **감기**라고 진단을 했다.

(B) 병원에서는 **폐암**이라고 진단을 했다.

S3: 집에 오면서 약국에 들러 **감기약**을 사서 왔다.

S1: 옷을 사러 친구랑 롯데백화점에 갔다.

(A) 친구는 **옷 하나**를 고르는 데 **한 시간이 넘게** 걸렸다.

(B) 친구는 **옷 전부**를 고르는 데 **십 분이 채 안** 걸렸다.

S3: 나는 기다리다가 지쳐서 그냥 나와 버렸다.

Q: 아래 담화에서 가운데 문장으로 더 적절한 것은 어느 것인가?

S1: 사냥꾼이 사슴을 향해 방아쇠를 당겼다.

(A) 사냥꾼이 쏜 총알이 사슴의 **다리에 맞았다.**

(B) 사냥꾼이 쏜 총알이 사슴을 **빚나갔다.**

S3: 총을 맞은 사슴은 절룩거리며 도망쳤다.

S1: 해영이는 집에 택배가 도착했다는 문자를 받고 서둘러 귀가했다.

(A) 집 앞에 도착했는데 택배 상자가 너무 컸다.

(B) 집 앞에 도착했는데 택배가 **없었다.**

S3: 택배회사에 전화하자 **분실** 택배로 처리되어 보상금을 받았다.

Q: 아래 담화에서 가운데 문장으로 더 적절한 것은 어느 것인가?

S1: 업무가 끝난 후 동료들과 회식을하기로 했다.

(A) 고기를 먹으러 근처의 **소고기 오마카세 집**으로 들어갔다.

(B) 고기를 먹으러 근처의 **돼지고기 식당**으로 들어갔다.

S3: **삼겹살**에 소주를 곁들이며 허심탄회한 시간을 보냈다.

S1: 친구가 과제를 제출하는 방법에 대해 아느냐고 물었다.

(A) 나는 메신저로 방법을 써서 보내 주었다.

(B) 나는 메신저로 그것도 모르냐며 면박을 주었다.

S3: 친구는 **고맙다고** 말하고는 온라인으로 과제를 제출했다.

인공주석물(annotation artifact)

- 표본 제작 과정에 투입된 여러 물질이나 기법 때문에 생성된 의도치 않은 부정적 부산물
- 찍기 실패 (이해능력이 아닌 패턴을 통해서 학습함)
- 과제(task), 주석자(annotator)

S1: 나는 축구공을 잠시 땅에 내려놓았다.

(A) 그러자 **축구공**이 밑으로 데굴데굴 굴러갔다.

(B) 그러자 **야구공**이 밑으로 데굴데굴 굴러갔다.

S3: 나는 축구공을 쫓아 얼른 뛰어갔다.

S1: 나는 **남자친구**의 우직한 모습을 보고 결혼을 결심했다.

(A) 남자친구에게 결혼을 하자고 프로포즈를 할 생각이었다.

(B) 남자친구가 결혼을 하자고 프로포즈를 해주기를 바랐다.

S3: 적당한 날 나는 남자친구에게 결혼하자고 프로포즈를 했다.

토론 & 질문

