KOREA UNIVERSITY

DSBA CS224n 2021 Study

[Lecture 18]
Future of NLP + Deep Learning

DSBA
Data Science & Business Analytics

고려대학교 산업경영공학과

Data Science & Business Analytics Lab

발표자 : 오수지

00

발표 목차

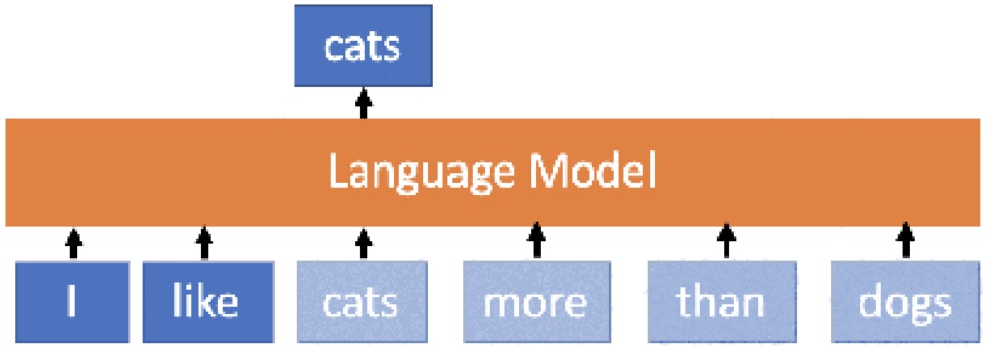
Contents

00 발표 목차
Contents

01 Large Language Models and GPT-3

GPT-1 : Improving Language Understanding by Generative Pre-Training

- ✓ 트랜스포머 디코더만 사용
- ✓ 단방향 모델 $\mathrm{P}(w_i) = P(w_i \mid w_0, \ldots, w_{i-1})$
- ✓ 일반적인 LM을 통해 pretrain
- ✓ Teacher Forcing 이용
- ✓ LM은 레이블이 필요없음
  - 대량의 데이터 확보 가능
  - Train data size : BooksCorpus(800M Words)



$$L = \sum_i \log P(w_i \mid w_0, \ldots, w_{i-1}; \theta)$$

$$h_0 = UW_e + W_p$$

$$h_l = Decoder\ block_i(h_{l-1})$$

$$P(w_i) = softmax(h_n W_e^T)$$

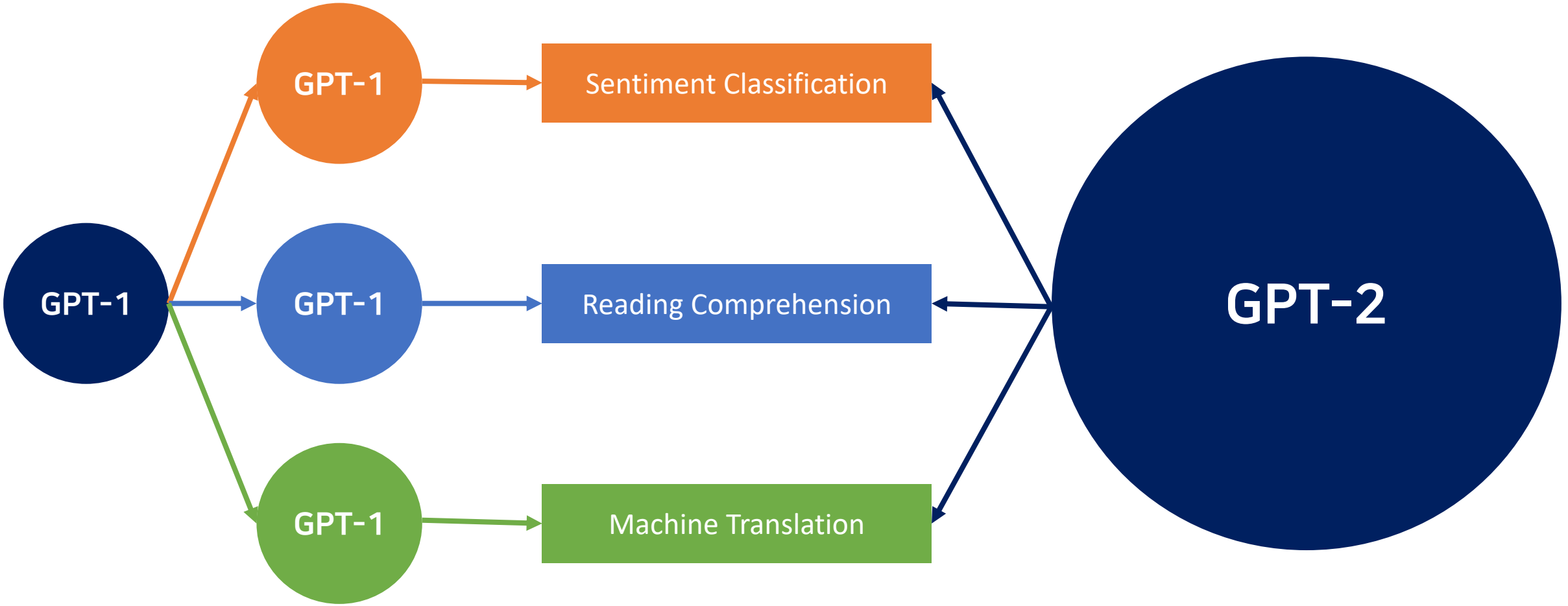
김재희 석박통합과정 CS224N 스터디 발표자료

No Supervised Learning Anymore!

01 Large Language Models and GPT-3

GPT-2 : Language Models are Unsupervised Multitask Learners

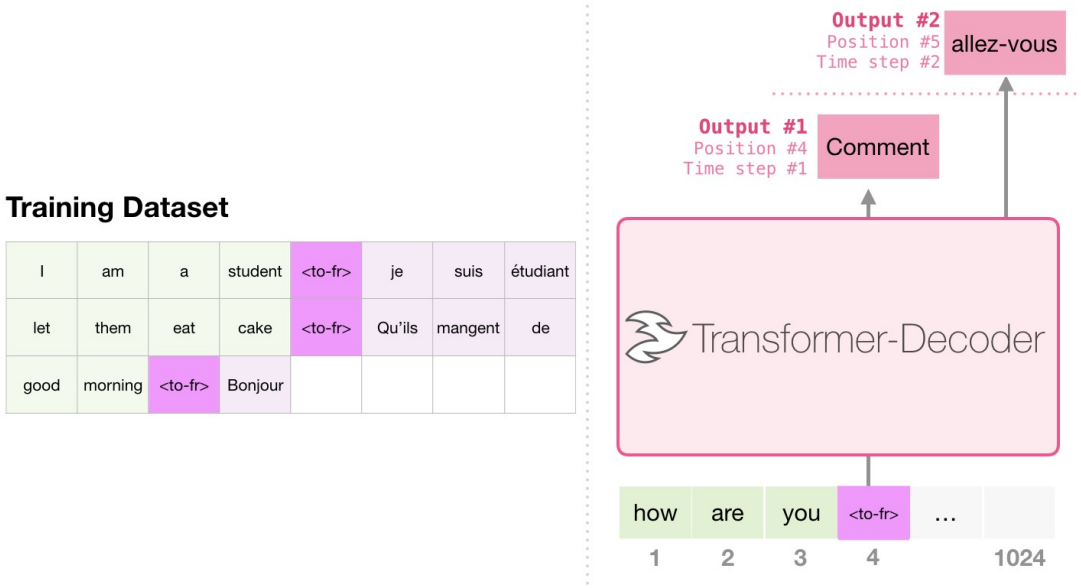
허민석님 유튜브

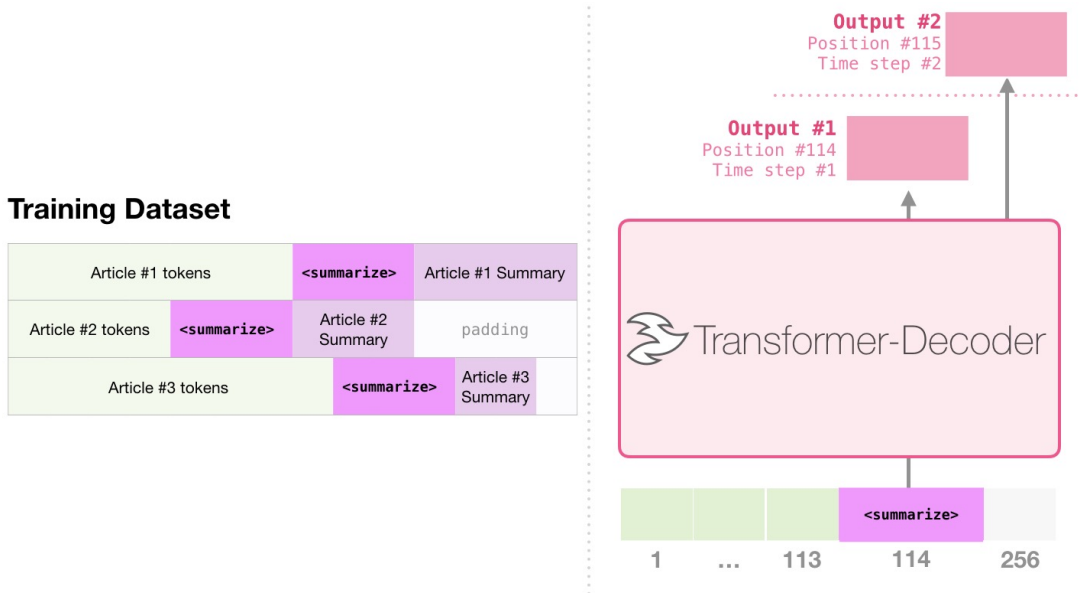


01 Large Language Models and GPT-3

GPT-2 : Language Models are Unsupervised Multitask Learners

jalammar.github.io

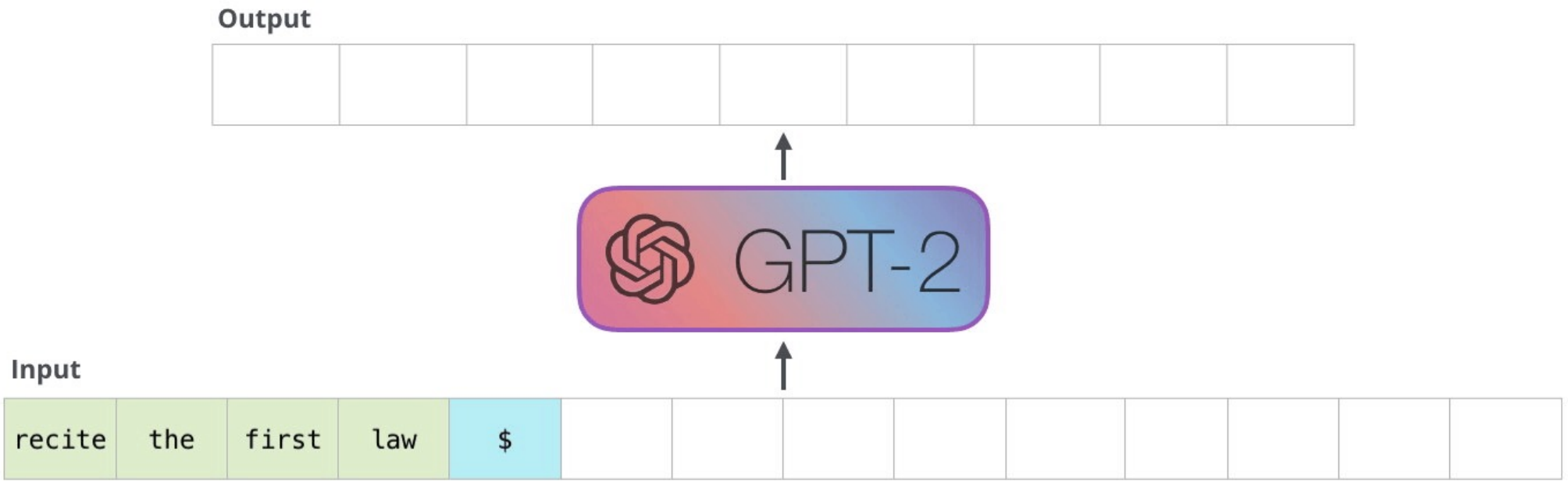




01 Large Language Models and GPT-3

GPT-2 : Language Models are Unsupervised Multitask Learners

jalammar.github.io



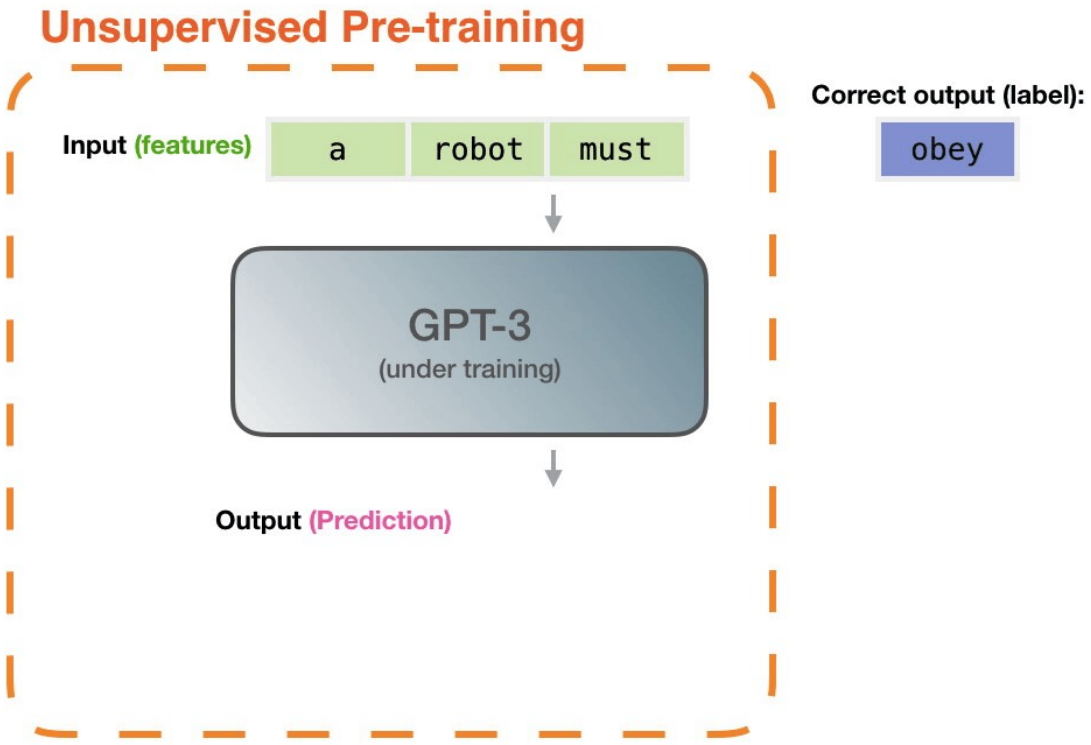
Few shot?

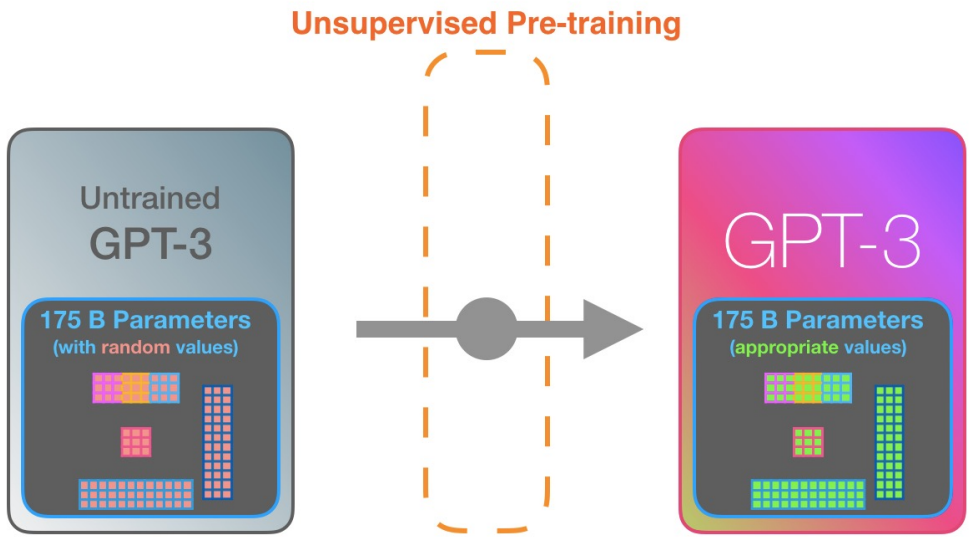
01 Large Language Models and GPT-3

GPT-3 : Language Models are Few Shot Learners

01 Large Language Models and GPT-3

GPT-3 : Language Models are Few Shot Learners





01 Large Language Models and GPT-3

What's new about GPT-3?

- 175 billion = 175,000,000,000 parameters
- Trained on 500 billion tokens

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

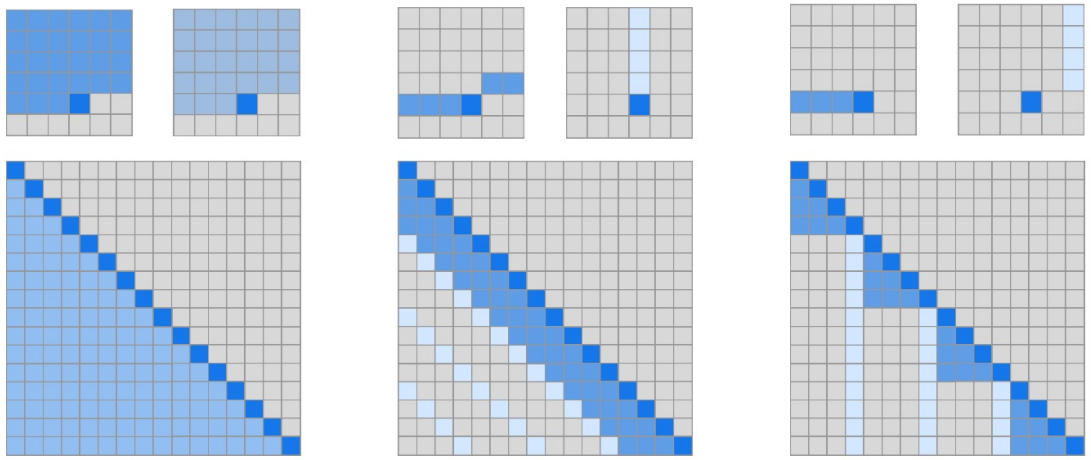
WIKIPEDIA
The Free Encyclopedia

Common Crawl

01 Large Language Models and GPT-3

What's new about GPT-3?

- Same architecture as GPT-2 with the exception of *locally banded sparse attention patterns*



(a) Transformer (b) Sparse Transformer (strided) (c) Sparse Transformer (fixed)

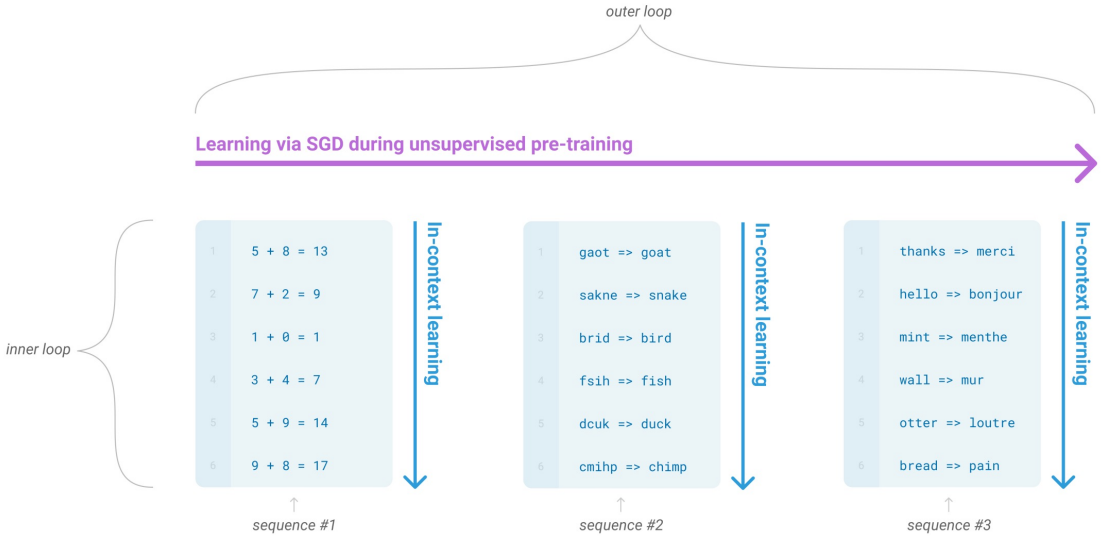
Figure 3. Two 2d factorized attention schemes we evaluated in comparison to the full attention of a standard Transformer (a). The top row indicates, for an example 6x6 image, which positions two attention heads receive as input when computing a given output. The bottom row shows the connectivity matrix (not to scale) between all such outputs (rows) and inputs (columns). Sparsity in the connectivity matrix can lead to significantly faster computation. In (b) and (c), full connectivity between elements is preserved when the two heads are computed sequentially. We tested whether such factorizations could match in performance the rich connectivity patterns of Figure 2.

Generating Long Sequences with Sparse Transformers (R Child et al, 2019)

01 Large Language Models and GPT-3

What's new about GPT-3?

- Meta-learning
  - 사람이 통제하던 기계학습 과정을 자동화함으로써 기계 스스로 학습 규칙(메타 지식)을 익힐 수 있게 하는 방법론
  - The model develops a broad set of skills and pattern recognition abilities at training time.



	data
1	5+8=13, 7+2=9, 1+0=1, 3+4=7, 5+9=14
2	Gaot=>goat, sakne=>snake, brid=>bird, fsih=>fish
3	Thanks=>merci, hello=>bonjour, mint=>menthe

01 Large Language Models and GPT-3

What's new about GPT-3?

- **몇 가지 궁금증들 (feat. 뇌피셜)**
  1. Pre-training의 In-context learning 과정에서 weight update가 발생하는가?
  2. Pre-training 시 input은 Teacher Forcing 방식인가? (아마 Yes)
  3. 하나의 sequence는 하나의 task로만 구성되는가?
     - E.g. 1+3=5, 7+2=9, 1+8=9, 10+1=11 vs 1+3=5, gaot→goat, thanks→merci

Figure 1.1: Language model meta-learning. During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

01 Large Language Models and GPT-3

What's new about GPT-3?

- Language Modeling
  - Penn Tree Bank
  - Story Completion (LAMBADA)
- Knowledge Intensive Tasks
  - E.g. Reading Comprehension (TriviaQA, CoQA)

(1) *Context:* "Yes, I thought I was going to lose the baby." "I was scared too," he stated, sincerity flooding his eyes. "You were ?" "Yes, of course. Why do you even ask?" "This baby wasn't exactly planned for."
Target sentence: "Do you honestly think that I would want you to have a _____ ?"
Target word: miscarriage

---

(2) *Context:* "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel. "He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said _____.
Target word: Gabriel

---

(3) *Context:* Preston had been the last person to wear those chains, and I knew what I'd see and feel if they were slipped onto my skin-the Reaper's unending hatred of me. I'd felt enough of that emotion already in the amphitheater. I didn't want to feel anymore. "Don't put those on me," I whispered. "Please."
Target sentence: Sergei looked at me, surprised by my low, raspy please, but he put down the _____.
Target word: chains

---

(4) *Context:* They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move.
Target sentence: Aside from writing, I 've always loved _____.
Target word: dancing

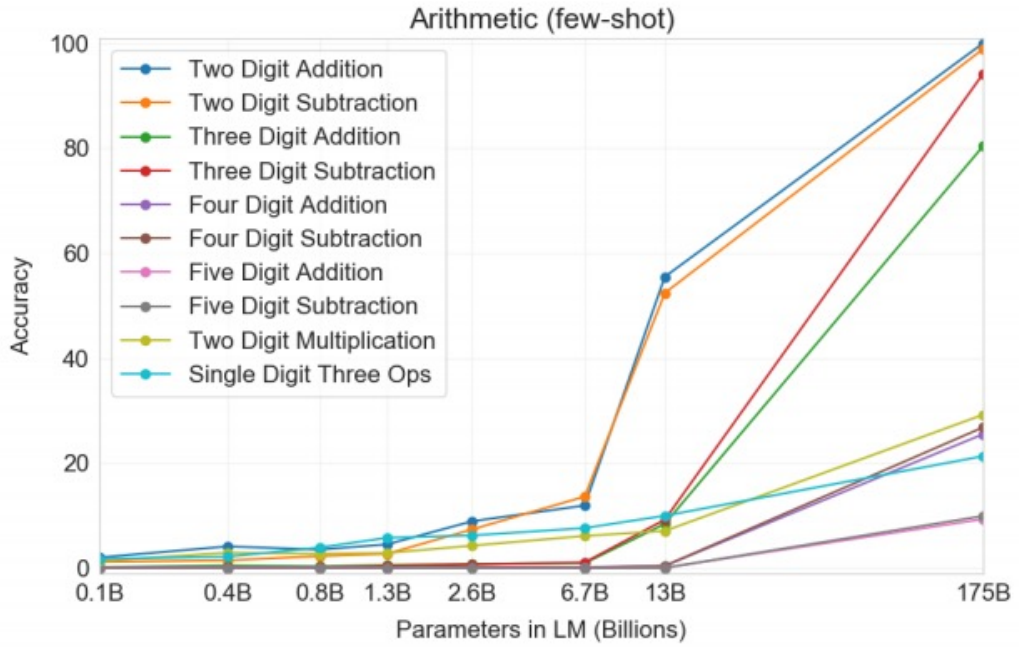
✔ Alice was friends with Bob. Alice went to visit her friend Bob
✖ Alice was friends with Bob. Alice went to visit her friend and took a...

The man went to the park. He was happy to leave his ___ . → house
Alice was friends with Bob. Alice went to visit her friend ___ . →

Large Language Models and GPT-3

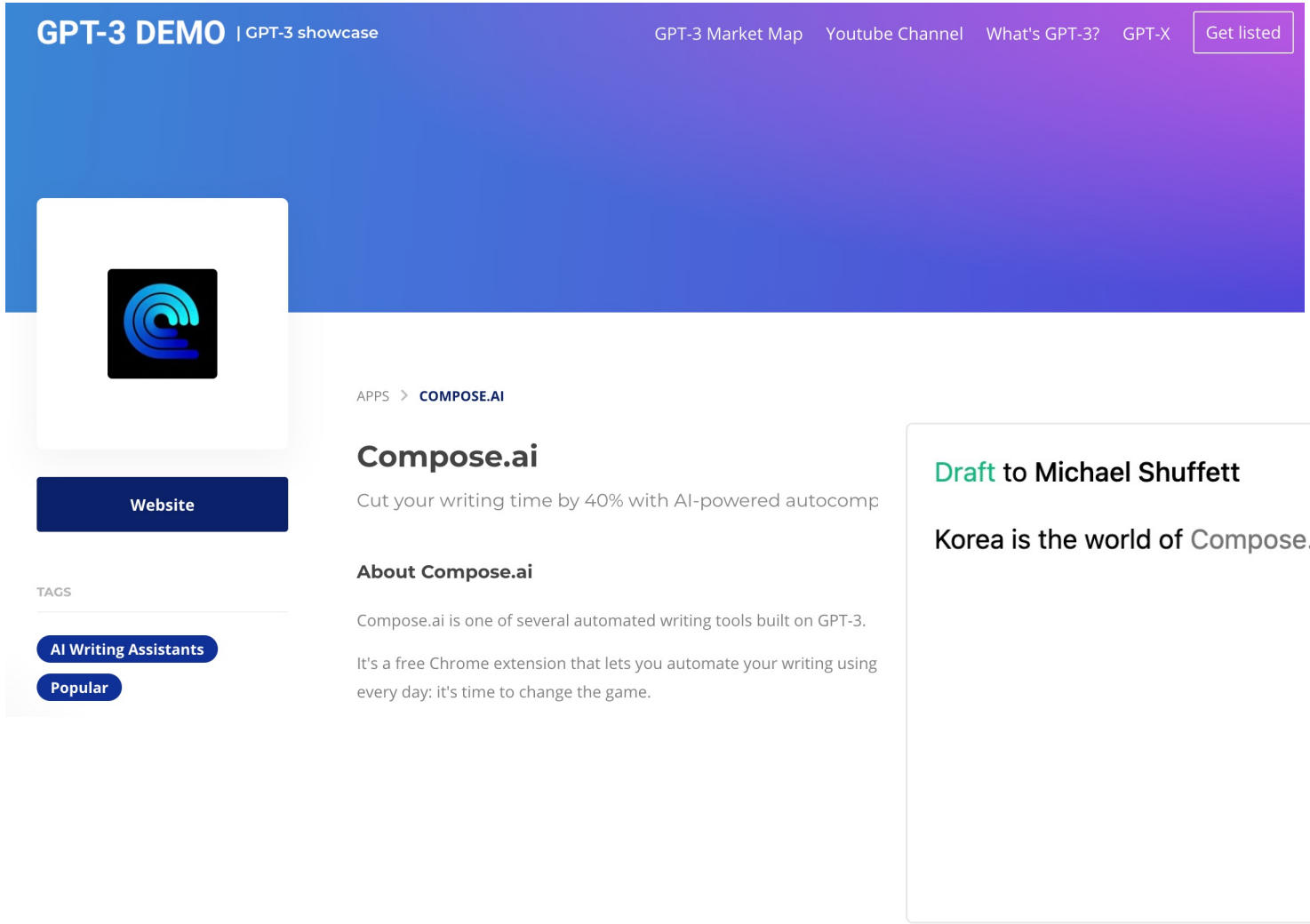
What's new about GPT-3?

- Structured problems that require multiple steps of reasoning
  - RTE, Arithmetic, Word problems, Analogy making



01 Large Language Models and GPT-3

Playing with GPT-3!



https://gpt3demo.com/

01 Large Language Models and GPT-3

Limitations and Open Questions

1. Seems to do poorly on more structured problems that involve decomposing into atomic / primitive skills.
2. Performing permanent knowledge updates interactively is not well studied.
3. Doesn't seem to exhibit human like generalization (systematicity).
4. Language is situated and GPT-3 is merely learning from text without being exposed to other modalities.

00

발표 목차

Contents

02 Compositional Representations and Systematic Generalization

용어

- Systematicity (체계성)
  - 사람이 이해하는 문장들 간엔 **확실하고 예측 가능한 패턴**이 있다.
  - E.g. 철수는 영희를 좋아한다. → 영희는 철수를 좋아한다.

Stefan Frank

Imagine you meet someone who only knows two sentences of English:

Could you please tell me where the toilet is?
I can't find my hotel.

So (s)he does not know:

Could you please tell me where ***my hotel*** *is?*
I can't find ***the toilet****.*

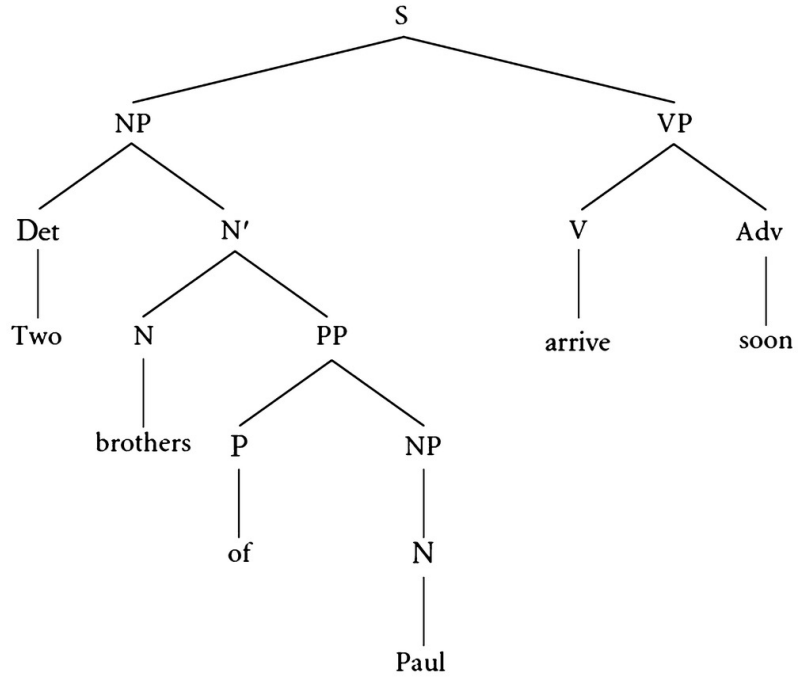
This person has no knowledge of English but simply memorized some lines from a phrase book.

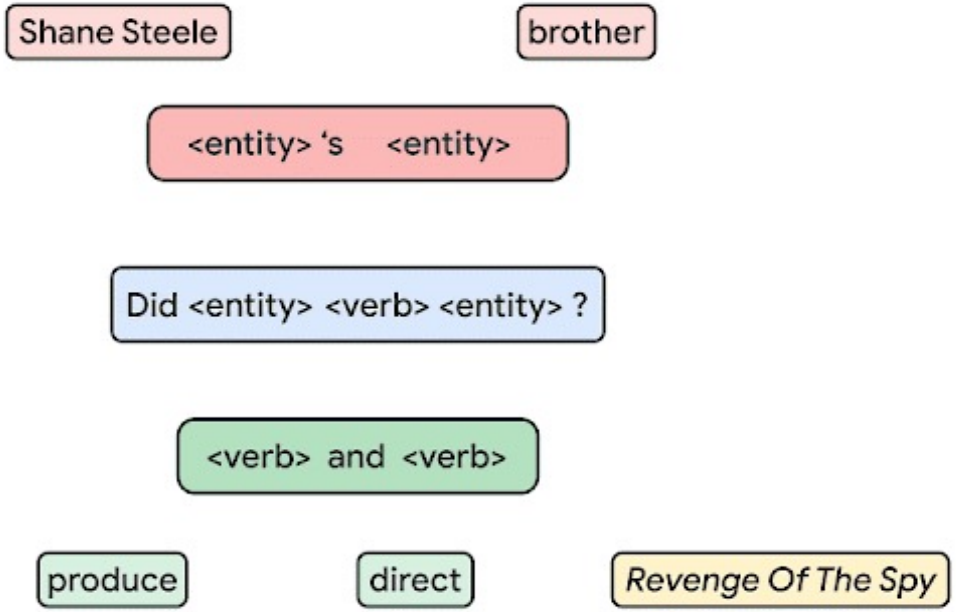
- Human language behavior is (more or less) **systematic**: if you know some sentences, you know many.
- Sentences are not atomic but made up of words.
- Likewise, words can be made up of morphemes. (e.g., *un* + *clear* = *unclear*, *un* + *stable* = *unstable*, …)
- It **seems like** language results from applying a set of rules (grammar, morphology) to symbols (words, morphemes).

02 Compositional Representations and Systematic Generalization

용어

- Compositionality (구성성)
  - 한 표현의 의미는 그 표현을 구성하는 구성 요소들의 의미와 구조로 구성된다.





02 Compositional Representations and Systematic Generalization

Are human languages compositional?

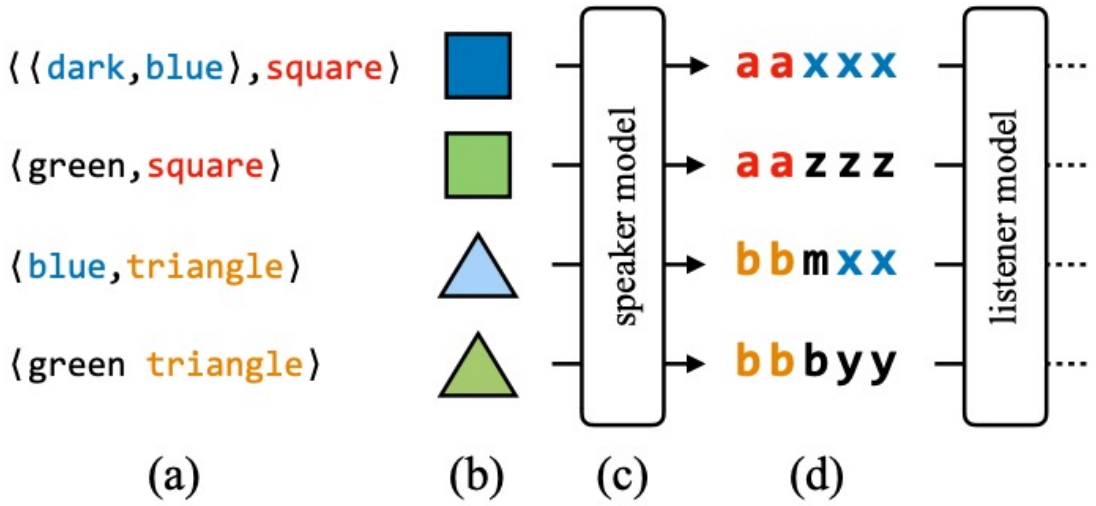
산 넘어 마을 = 넘다(산, 도착지)

산 넘어 산 ≠ 넘다(산, 도착지)

02 Compositional Representations and Systematic Generalization

Are neural representations compositional?

- Communication game
  - Is this encoding scheme compositional?
  - To what extent can we analyze the agents' messages as being built from smaller pieces (e.g. pieces xx meaning blue and bb meaning triangle)?



Measuring Compositionality in Representation Learning (Jacob Andreas, ICLR 2019)

02 Compositional Representations and Systematic Generalization

Are neural representations compositional?

Tree Reconstruction Error (TRE)

First choose :

- a distance function $\delta : \Theta \times \Theta \to [0, \infty)$ satisfying $\delta(\theta, \theta') = 0 \Leftrightarrow \theta = \theta'$
- a composition function $* : \Theta \times \Theta \to \Theta$

Define $\hat{f}_\eta(d)$, a *compositional approximation to* f with parameters η, as:

$$\hat{f}_\eta(d_i) = \eta_i \qquad \text{for } d_i \in \mathcal{D}_0$$
$$\hat{f}_\eta(\langle d, d' \rangle) = \hat{f}_\eta(d) * \hat{f}_\eta(d') \qquad \text{for all other } d$$

$\hat{f}_\eta$ has one parameter vector η_i for every d_i in $\mathcal{D}_0$; these vectors are members of the representation space Θ.

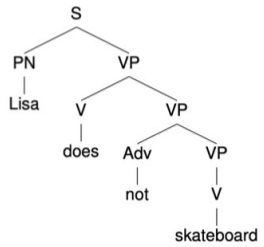
Given a dataset $\mathcal{X}$ of inputs x_i with derivations $d_i = D(x_i)$, compute:

$$\eta^* = \arg\min_\eta \sum_i \delta\big(f(x_i), \hat{f}_\eta(d_i)\big) \tag{2}$$

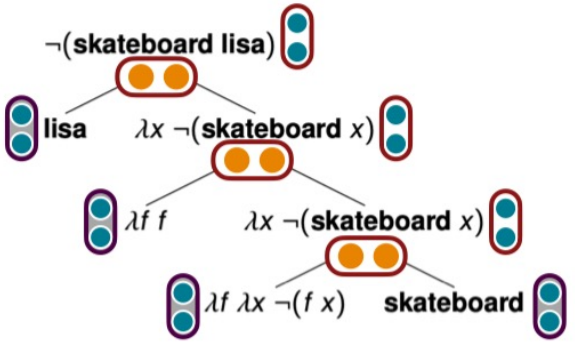
Then we can define datum- and dataset-level evaluation metrics:

$$\mathrm{TRE}(x) = \delta\big(f(x), \hat{f}_{\eta^*}(d)\big) \tag{3}$$
$$\mathrm{TRE}(\mathcal{X}) = \frac{1}{n} \sum_i \mathrm{TRE}(x_i) \tag{4}$$

Lisa does not skateboard =
⟨Lisa, ⟨does, ⟨not, skateboard⟩⟩⟩



NN(Lisa does not skateboard) ≈
f(v(Lisa), f(v(does), f(v(not), v(skateboard))))

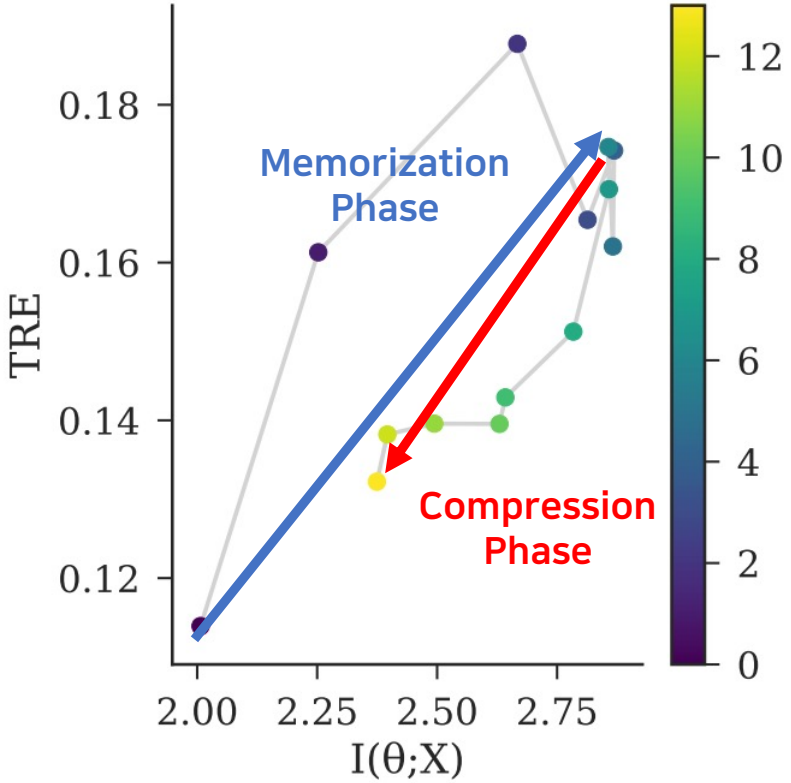


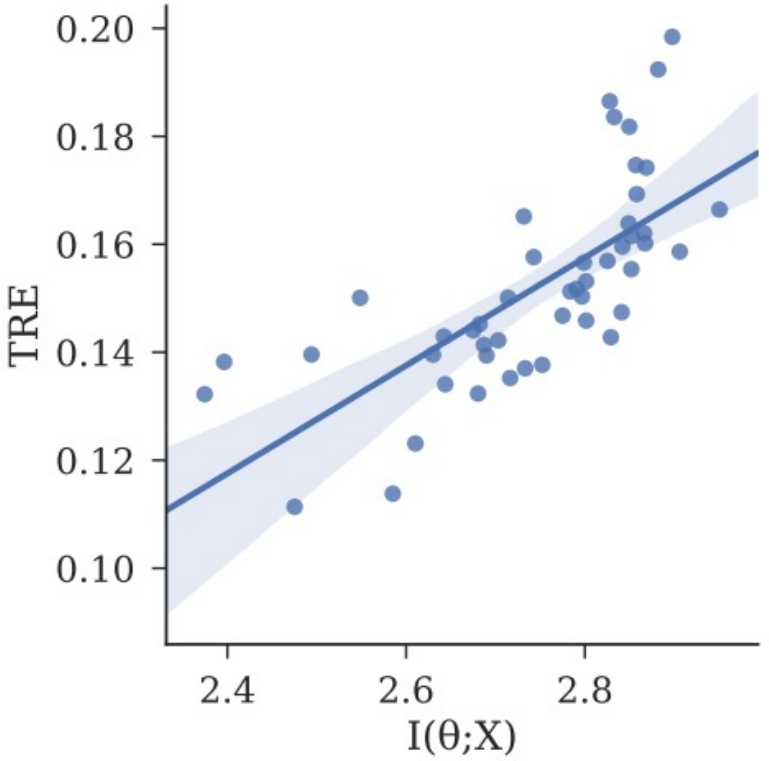
Measuring Compositionality in Representation Learning (Jacob Andreas, ICLR 2019)

02

Compositional Representations and Systematic Generalization

Are neural representations compositional?





Measuring Compositionality in Representation Learning (Jacob Andreas, ICLR 2019)

02 Compositional Representations and Systematic Generalization

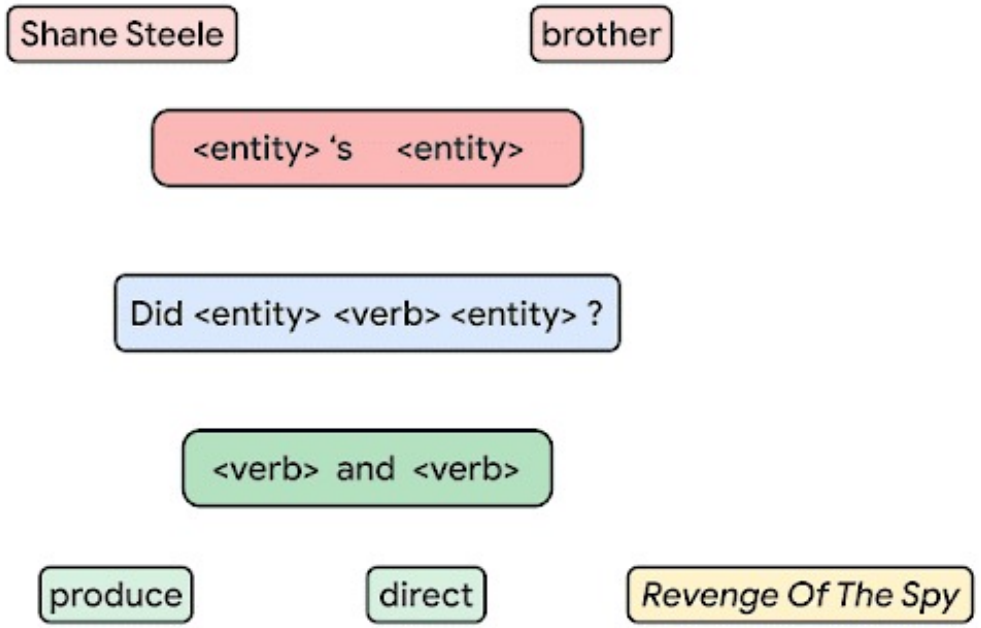
Do neural networks generalize systematically?

- **Compositional Generalization**
  - The capacity to understand and produce a potentially infinite number of novel combinations of known components.
    - 훼렉하다 → 밥을 훼렉하다, 훼렉하고 산책하다
  - E.g. 모델이 알고 있는 단어 = [나, 사과, 먹다, 아침]
    - 나 아침에 사과 먹었어, 아침에 나 사과 먹었어, 사과 먹었어 나 아침에, …
- **Questions**
  1. Do neural networks (including large transformers) generalize systematically on challenging benchmarks involving realistic language?
  2. Can we create a dataset split that explicitly tests for this kind of generalization?

02 Compositional Representations and Systematic Generalization

Do neural networks generalize systematically?

ai.googleblog.com



Train:
- Did Christopher Nolan produce Goldfinger?
- Who directed inception?

Test:
- Did Christopher Nolan direct Goldfinger?
- Who produced Goldfinger?

Atoms:
- produce
- direct
- inception
- goldfinger
- Christopher Nolan
- Who [predicate] [y]?
- Did [x] [predicate] [y]?

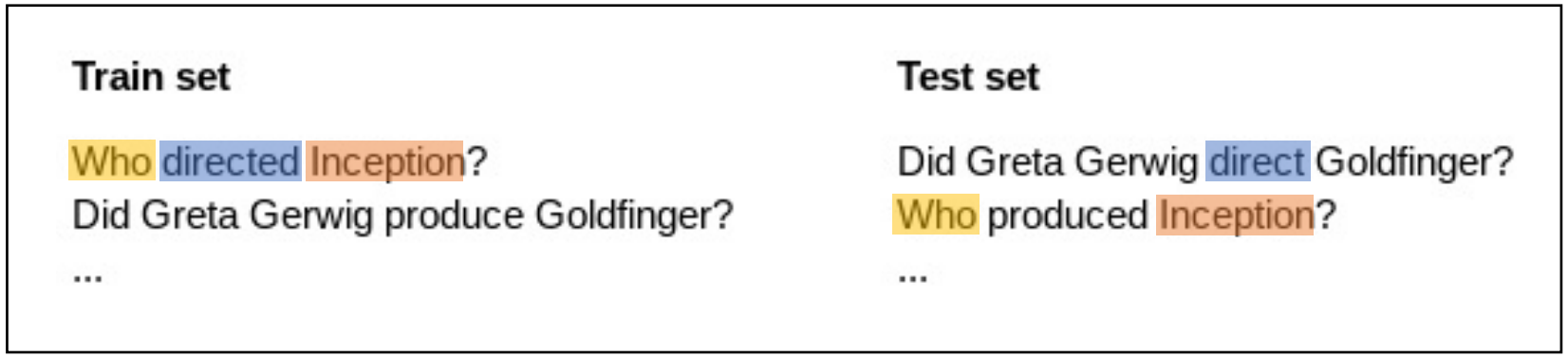
Compounds:
- Did Christopher Nolan [predicate] Goldfinger?
- Who directed [entity]?

Measuring Compositional Generalization: A Comprehensive Method on Realistic Data (Keysers et al, ICLR 2020)

02 Compositional Representations and Systematic Generalization

Do neural networks generalize systematically?

- Can we create a dataset split that explicitly tests for compositional generalization?
  - Ideal Compositionality Experiment
    1. Similar atom distribution: All atoms present in the test set are also present in the train set, and the distribution of atoms in the train set is as similar as possible to their distribution in the test set.
    2. Different compound distribution: The distribution of compounds in the train set is as different as possible from the distribution in the test set.
  - Split data into train / test such that **compound divergence is maximized and atom divergence is minimized!**



02 Compositional Representations and Systematic Generalization

Do neural networks generalize systematically?

Let $\mathcal{F}_A(\text{ data }) \equiv$ normalized frequency distribution of atoms
Let $\mathcal{F}_C(\text{ data }) \equiv$ normalized frequency distribution of compounds
Define atom and compound divergence as:

$$\mathcal{D}_A(\text{ train } || \text{ test }) = 1 - C_{0.5}(\mathcal{F}_A(\text{ train }) || \mathcal{F}_A(\text{ test }))$$ **Minimize!**

$$\mathcal{D}_C(\text{ train } || \text{ test }) = 1 - C_{0.1}(\mathcal{F}_C(\text{ train }) || \mathcal{F}_C(\text{ test }))$$ **Maximize!**

where,

$$C_\alpha(P \| Q) = \sum_k p_k^\alpha q_k^{1-\alpha}$$

is the chernoff coefficient between two categorical distributions that measures similarity.

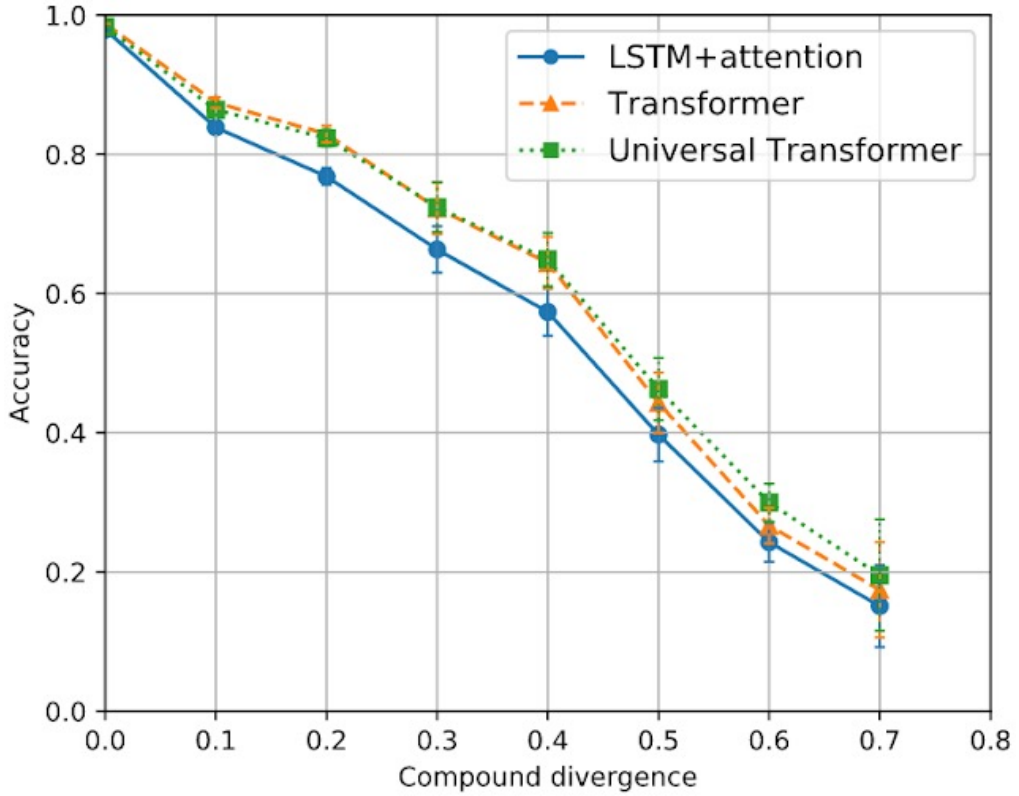
- The compound distributions of the train and test sets are very similar, then their compound divergence would be close to 0.
  → Not difficult tests for compositional generalization
- The compound divergence close to 1 means that the train-test sets have many different compounds.
  → Good test for compositional generalization

Measuring Compositional Generalization: A Comprehensive Method on Realistic Data (Keysers et al, ICLR 2020)

02 Compositional Representations and Systematic Generalization

Do neural networks generalize systematically?

- Do neural networks (including large transformers) generalize systematically on challenging benchmarks involving realistic language?



02 Compositional Representations and Systematic Generalization

Do neural networks generalize systematically?

- **Do neural networks (including large transformers) generalize systematically on challenging benchmarks involving realistic language?**
  - Pre-training helps for compositional generalization, but doesn't solve it.

Model	*CFQ (Maximum Compound divergence)*
T5-small (no pretraining)	21.4
T5-small	28.0
T5-base	31.2
T5-large	34.8
T5-3B	40.2
T5-11B	40.9
T5-11B-mod	42.1

00

발표 목차

Contents

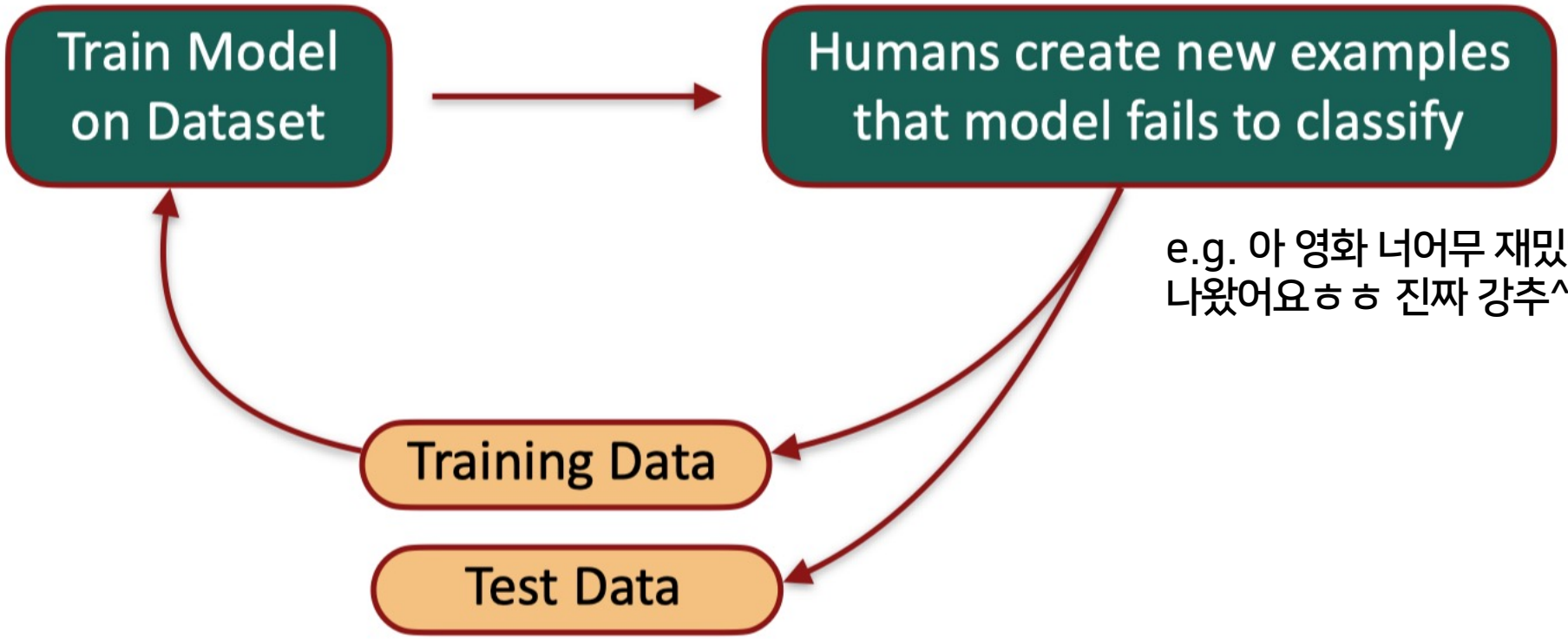
03 Improving how we evaluate models in NLP

Introduction

- ✓ 벤치마크 데이터셋에서의 모델 성능은 날로 증가하는데 정말 실제 세계에서의 모델 성능도 그만큼 증가했을까?
- ✓ task에 대한 모델의 “찐” 이해도를 어떻게 하면 정확하게 측정할 수 있는가?

03 Improving how we evaluate models in NLP

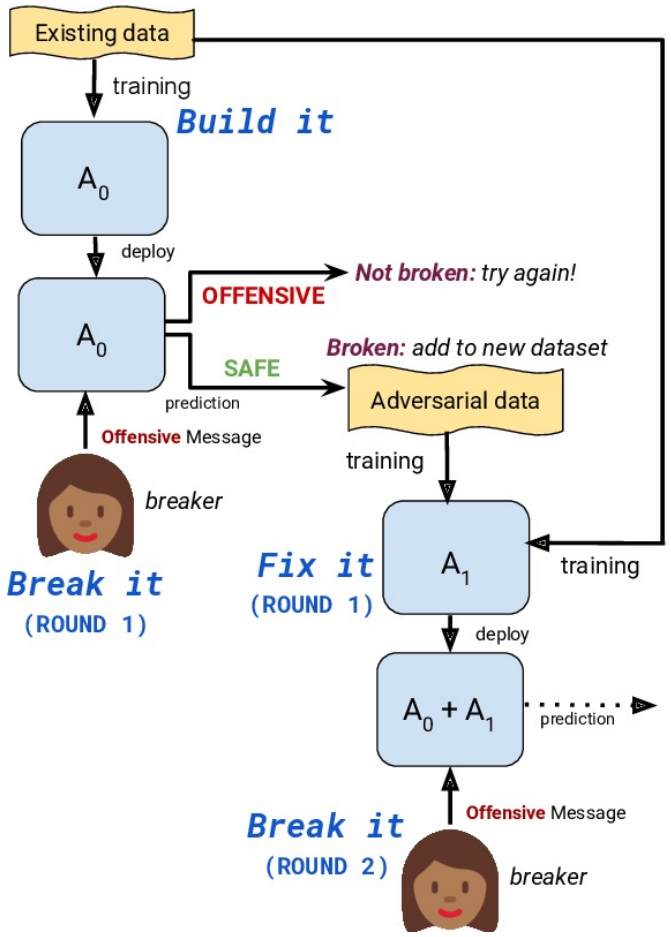
Dynamic Benchmarks



e.g. 아 영화 너어무 재밌어서 10분만에 나왔어요ㅎㅎ 진짜 강추^^

03 Improving how we evaluate models in NLP

Dynamic Benchmarks



1. **Build it** : 사용자의 공격적인 메세지를 감지할 수 있는 모델 개발
2. **Break it** : Crowdworker에게 모델은 "SAFE"하다고 생각하지만 Crowdworker는 "OFFENSIVE"하다고 생각하는 메세지를 만들어서 *"beat the system"* 해달라고 요청
3. **Fix it** : 2번 과정을 통해 모여진 예제들을 통해 모델을 재학습 → 적대적인 공격에 더 강건한 모델이 될 수 있도록!
4. **Repeat** : Break it – Fix it 을 계속계속 반복

Build-It Break-It Fix-It for Dialogue Safety (Dinan et al, EMNLP 2017)

00

발표 목차

Contents

04 Grounding Language to other modalities

Introduction

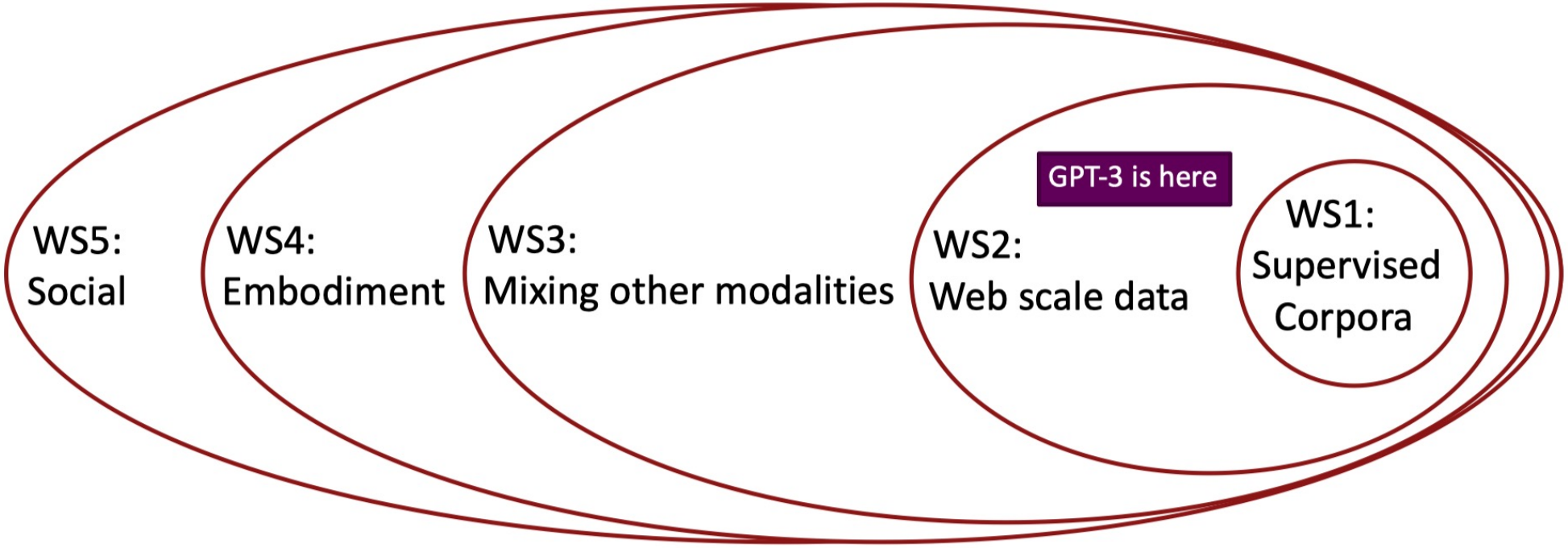
Grounding Language to other modalities

Introduction

You can't learn language from the radio

04 Grounding Language to other modalities

Introduction



Experience Grounds Language (Bisk et al, EMNLP 2020)

04 Grounding Language to other modalities

Introduction

> Computer vision and speech recognition are mature enough for investigation of broader linguistic contexts (WS3). The robotics industry is rapidly developing commodity hardware and sophisticated software that both facilitate new research and expect to incorporate language technologies (WS4). Simulators and videogames provide potential environments for social language learners (WS5). Our call to action is to encourage the community to lean in to trends prioritizing grounding and agency, and explicitly aim to broaden the corresponding World Scopes available to our models.

Experience Grounds Language (Bisk et al, EMNLP 2020)

감사합니다