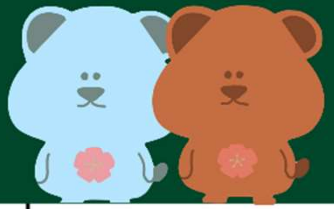


Week18_Future of NLP + Deep Learning

발표자: 황채원, 임세영

목차

01 Extremely large language models and GPT-3

02 Compositional representations and systematic generalization

03 Improving how we evaluate models in NLP

04 Grounding language to other modalities

Extremely large language models and GPT-3

Extremely large language models and GPT-3

GPT(Generative Pretrained Transformer)

- Generative : 생성모델. 데이터 전체의 분포를 모델링하는 기법
- Pretrained : 학습데이터량 많아 다른 모델에 비해 좋은 성능
- Transformer : Transformer Decoder를 사용한다. 예측하는 단어와 그 이후를 못보게 설계되었기 때문(masked self-attention)

Extremely large language models and GPT-3

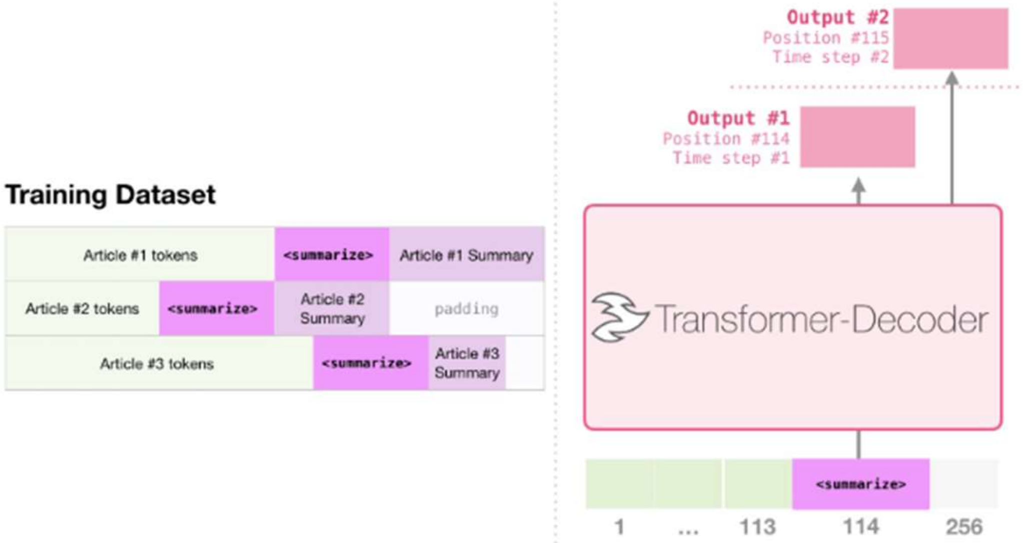
GPT-1

- Fine tuning 단계가 존재한다
  -> 시간, 돈이 많이 든다
  -> 다른 태스크에는 적용이 불가능하다.

Extremely large language models and GPT-3

GPT-2

- Fine tuning 단계없이 바로 태스크에 접근이 가능하다
- 단순히 앞의 단어로만 예측을 수행했던 GPT-1과 달리 태스크 정보를 추가로 입력받는다



Training Dataset

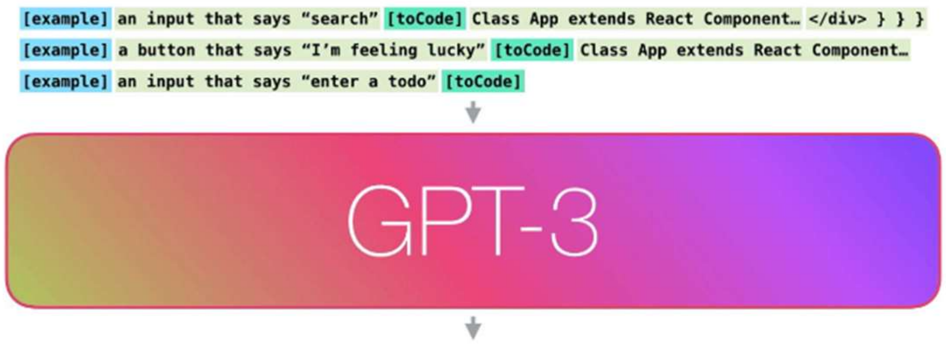
Article #1 tokens	<summarize>	Article #1 Summary	
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens	<summarize>	Article #3 Summary	

https://github.com/yookyungkho/DSBA_CS224N_2021/blob/main/slides/dsba_cs224n2021_lec18_suzieoh.pdf

Extremely large language models and GPT-3

GPT-3

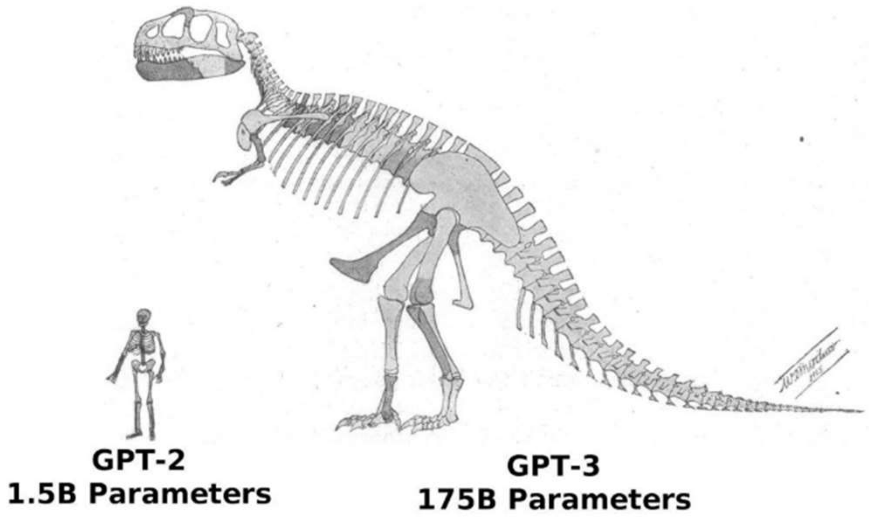
- Few shot : 추가로 몇 가지 예제를 입력해준다
- 175B 엄청나게 큰 크기의 모델을 엄청난 양의 데이터로 학습하였다



[example] an input that says "search" [toCode] Class App extends React Component… </div> } } }

[example] a button that says "I'm feeling lucky" [toCode] Class App extends React Component…

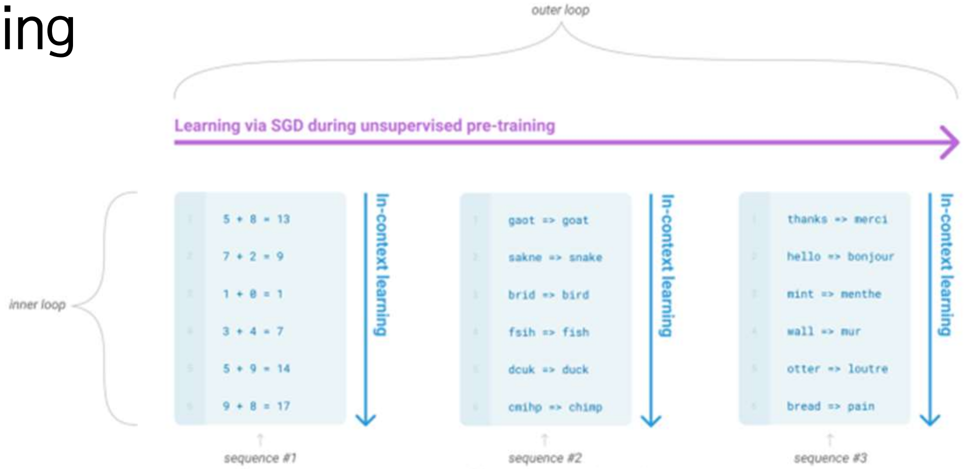
[example] an input that says "enter a todo" [toCode]



Extremely large language models and GPT-3

GPT-3

- GPT-2와 구조적으로 큰 차이는 없지만 sparse attention patterns
- Meta-learning : 모델이 훈련하는 동안 패턴 인식 능력을 스스로 개발하여 원하는 태스크에 빠르게 적응할 수 있도록 모델을 학습시키는 방법
- Innerloop에서의 학습 = in-context learning



Source: https://arxiv.org/pdf/2005.14165.pdf

Extremely large language models and GPT-3

GPT-3의 장점

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

GPT-3의 단점

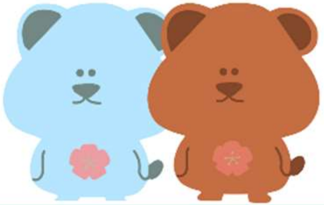
- Structured problems that require multiple steps of reasoning
    - RTE, Arithmetic, Word problems, Analogy making

Extremely large language models and GPT-3

GPT-3 : Limitations and Open questions

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

Compositional representations and systematic generalization



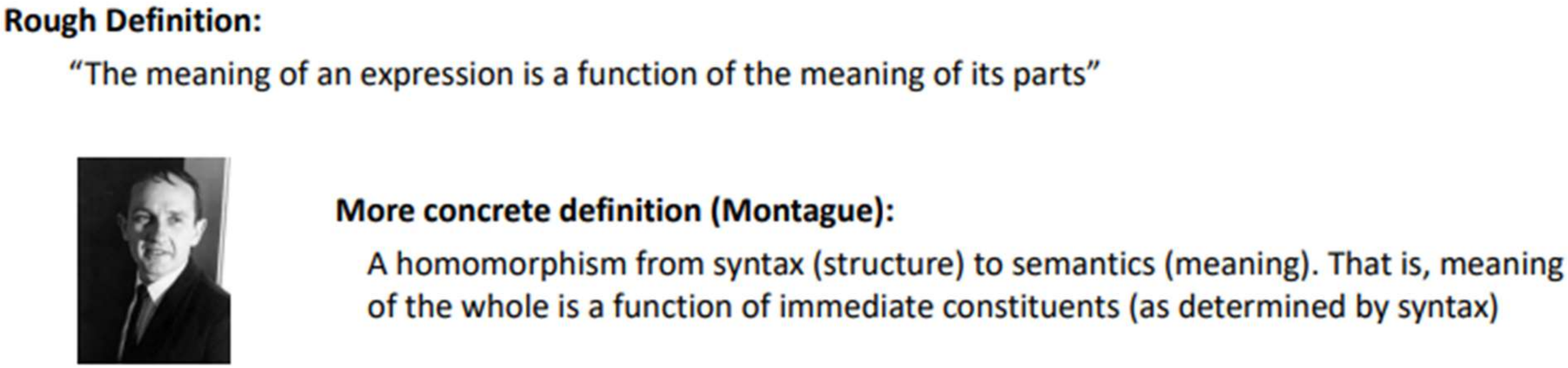
Compositional representations and systematic generalization

Systematicity = 체계성

- 사람이 이해하는 문장들 간엔 확실하고 예측 가능한 패턴이 있다
- ex. any speaker that understands the sentence “John loves Mary” should be able to understand “Mary loves John”.

Compositionality = 구성성

- 한 표현의 의미는 그것을 구성하는 부분들의 기능으로 이루어진다



Rough Definition:

“The meaning of an expression is a function of the meaning of its parts”

More concrete definition (Montague):

A homomorphism from syntax (structure) to semantics (meaning). That is, meaning of the whole is a function of immediate constituents (as determined by syntax)

Compositional representations and systematic generalization

Are human languages compositional?

Brown Cow = *Brown objects* ∩ *Cows*

Red Rabbit = *Red objects* ∩ *Rabbits*

Kicked the Ball = *Kicked(Ball, Agent)*

Red Herring ≠ *Red things* ∩ *Herring*

Kicked the bucket ≠ *Kicked(Bucket, Agent)*

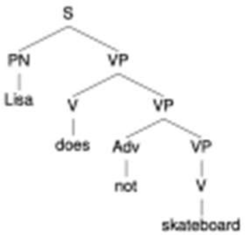
그럼에도 불구하고 표현의 구성성은 체계성의 유용한 선행 조건이다.

Compositional representations and systematic generalization

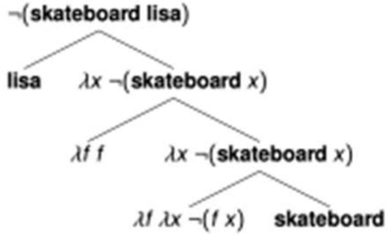
Are neural representations compositional?

TRE [Andreas 2019]: Compositionality of representations is about how well the representation approximates an explicitly homomorphic function in a learnt representation space

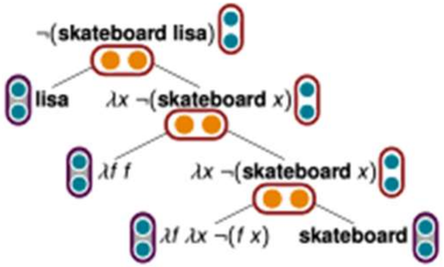
Lisa does not skateboard = ⟨Lisa, ⟨does, ⟨not, skateboard⟩⟩⟩



m(Lisa does not skateboard) = ⟨m(Lisa), ⟨m(does), ⟨m(not), m(skateboard)⟩⟩⟩



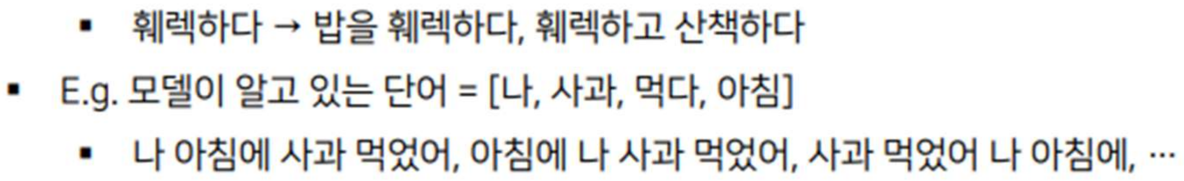
NN(Lisa does not skateboard) ≈ f(v(Lisa), f(v(does), f(v(not), v(skateboard))))



Compositional representations and systematic generalization

Compositional Generalization = 구성일반화 능력

이미 알고 있는 구성물로부터 무한한 수의 조합을 생산하고 이해할 수 있는 능력



  - 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, …

모델이 좋은 구성일반화 성능을 갖고 있는지 검증하기 위한 trainset?

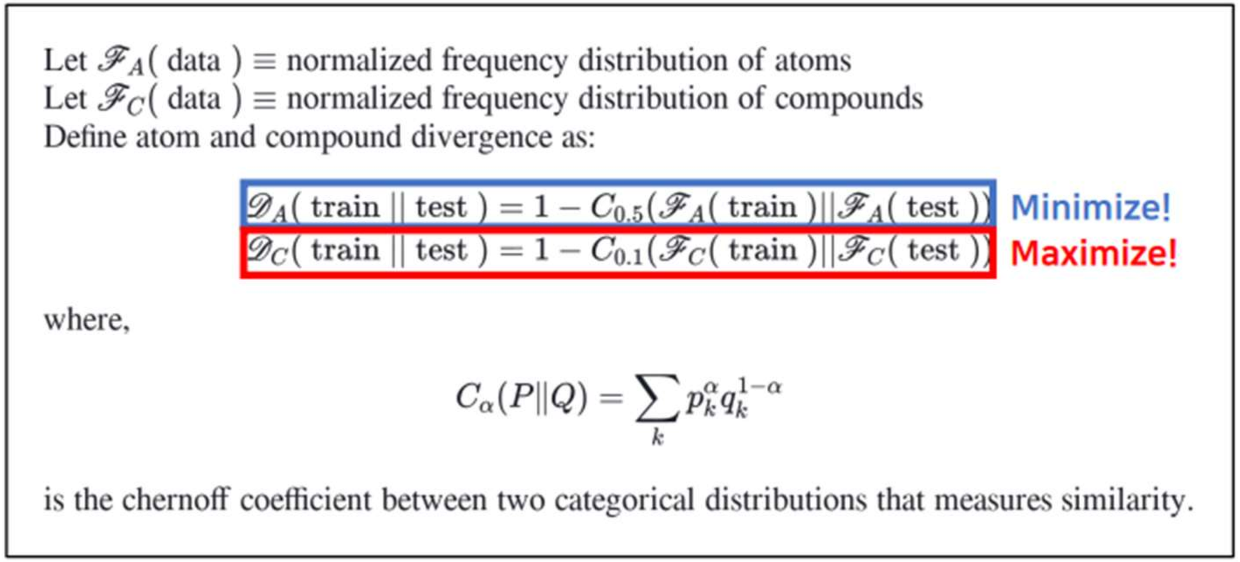
1. Similar atom(=가장 작은 구성 단위) distribution
2. Different compound(=구성결과) distribution

-> 발생하는 성능차이가 진짜 compound distribution에 의한 것인지 알 수 있어야한다.

Compositional representations and systematic generalization

모델이 좋은 구성일반화 성능을 갖고 있는지 검증하기 위한 trainset?

1. Similar atom(=가장 작은 구성 단위) distribution
2. Different compound(=구성결과) distribution



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

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

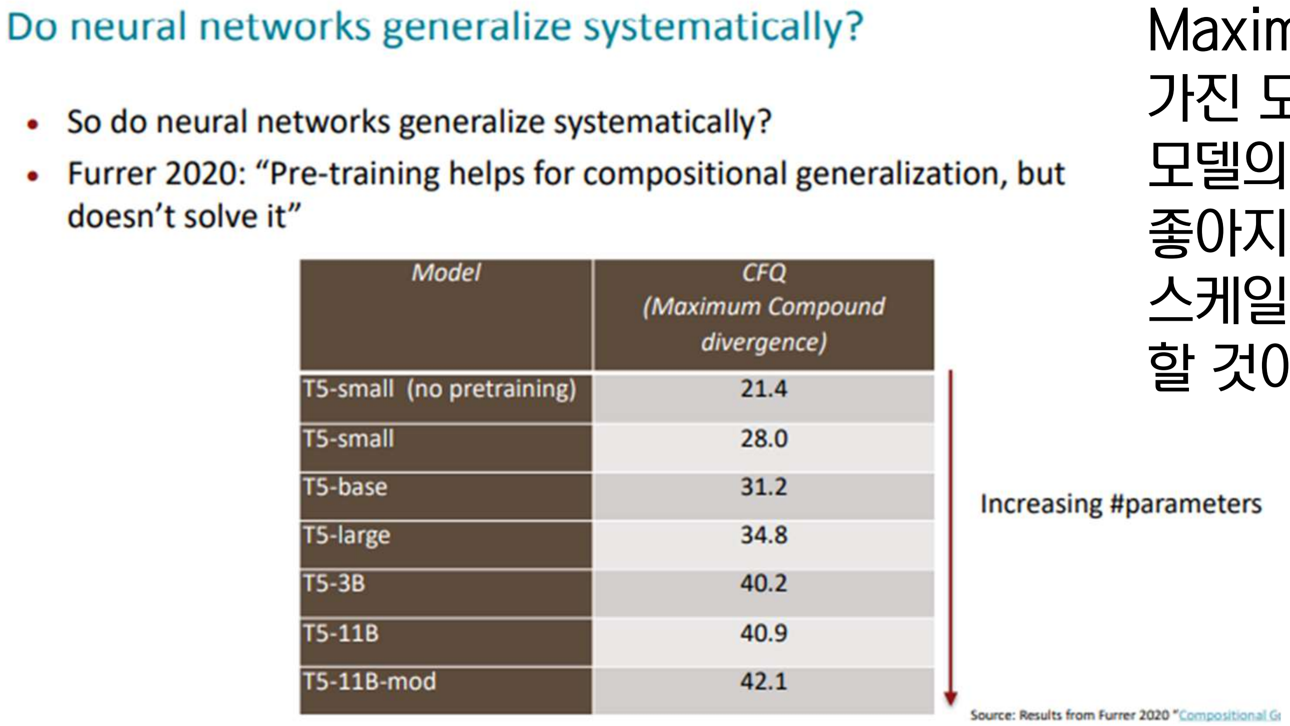
$$\boxed{\mathscr{D}_C(\text{ train } \| \text{ test }) = 1 - C_{0.1}(\mathscr{F}_C(\text{ train }) \| \mathscr{F}_C(\text{ test }))} \text{ Maximize!}$$

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.

Compositional representations and systematic generalization



Do neural networks generalize systematically?

- So do neural networks generalize systematically?
- Furrer 2020: “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

Increasing #parameters

Source: Results from Furrer 2020 “Compositional G

Maximum compound divergence를 가진 모델일 때
모델의 스케일을 늘릴수록 성능이 좋아지지만
스케일을 늘릴수록 계속 성능이 좋아지기만 할 것이라고 단정지을 수는 없다.

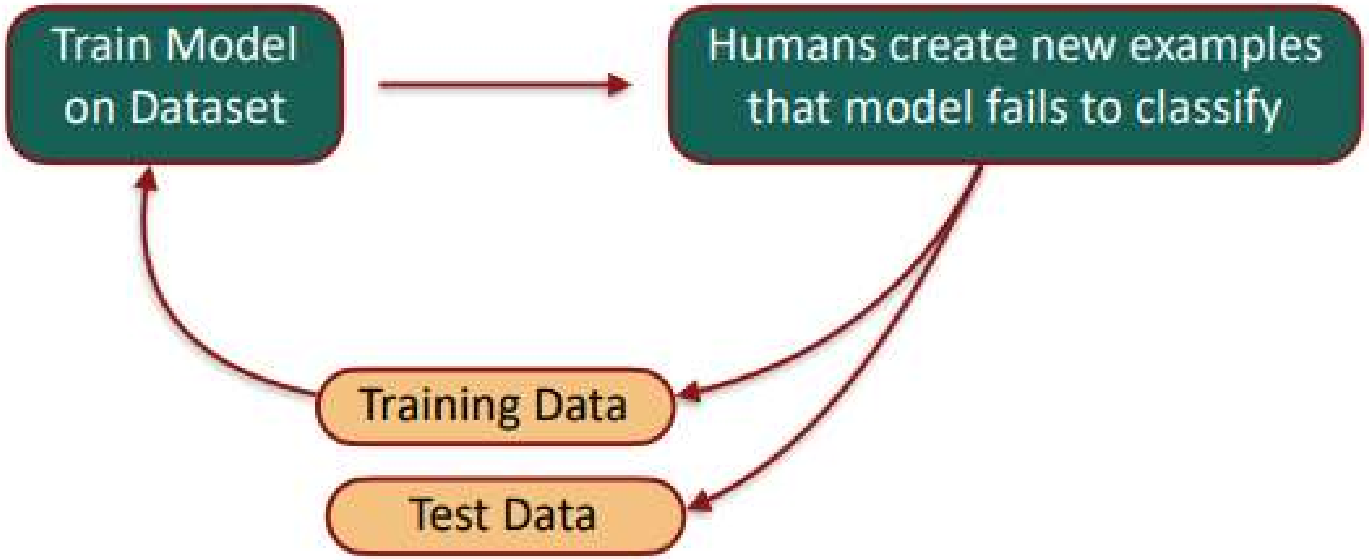
Improving how we evaluate models in NLP

Improving how we evaluate models in NLP

- 실제 세계에서의 모델 성능이 벤치마크 데이터셋에서의 모델 성능만큼 증가했을까?
- Task에 대한 모델의 이해도를 정확하게 측정하려면 어떻게 해야 할까?

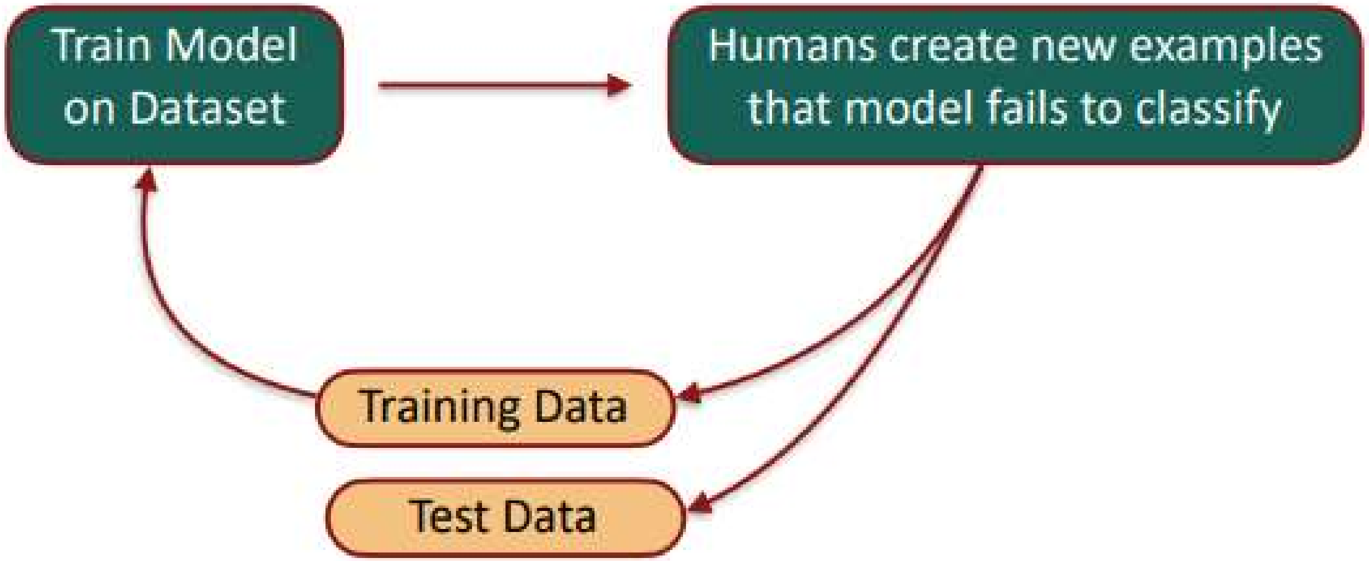
https://jiho-ml.com/weekly-nlp-29/

Improving how we evaluate models in NLP



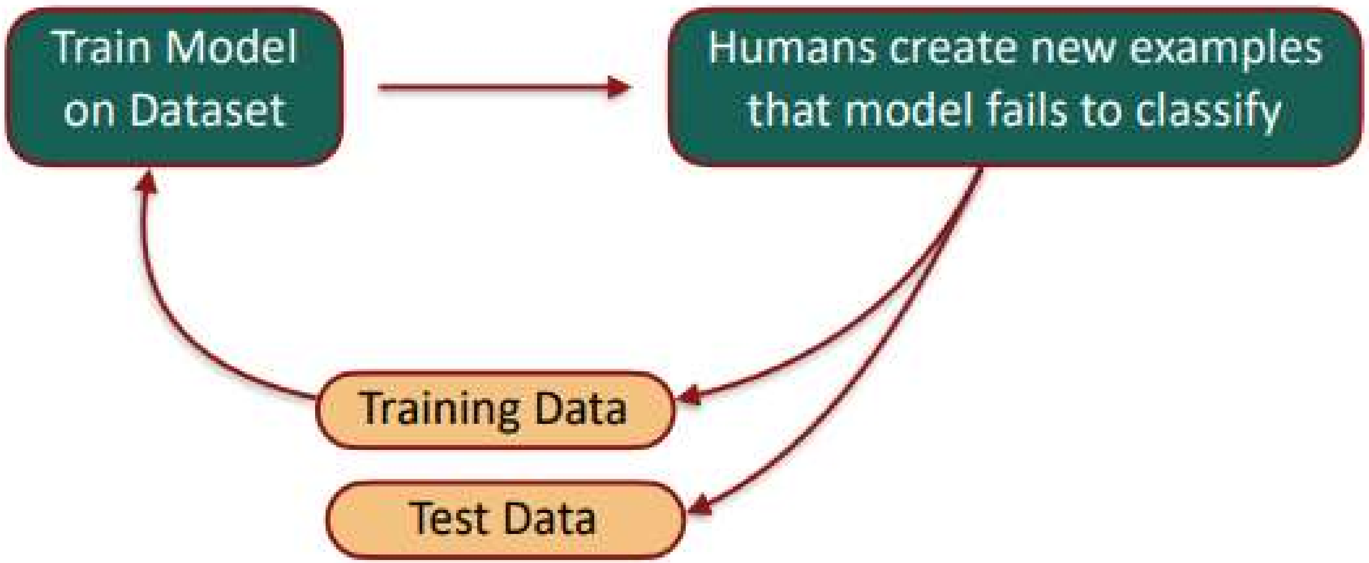
Overview of dynamic benchmarks

Improving how we evaluate models in NLP



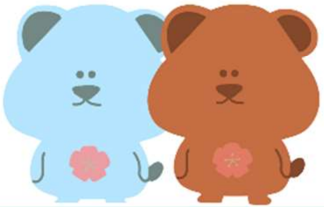
1. Start with a pre-trained model and fine-tune it on the original train / test datasets
2. Humans attempt to create new examples that fool the model but not other humans
3. These examples are then added into the train / test sets and the model is retrained on the augmented dataset

Improving how we evaluate models in NLP



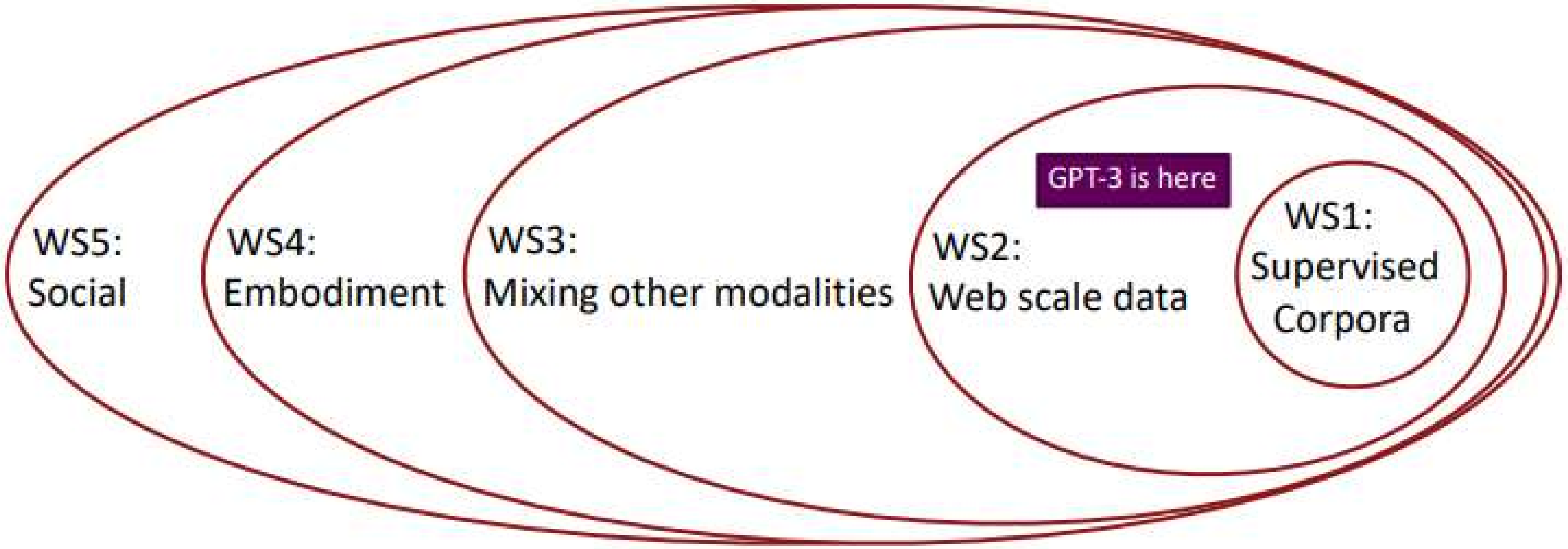
- Main Challenges: Ensuring that humans are able to come up with hard examples and we are not limited by creativity.
- Current approaches use examples from other datasets for the same task as prompts

Grounding language to other modalities

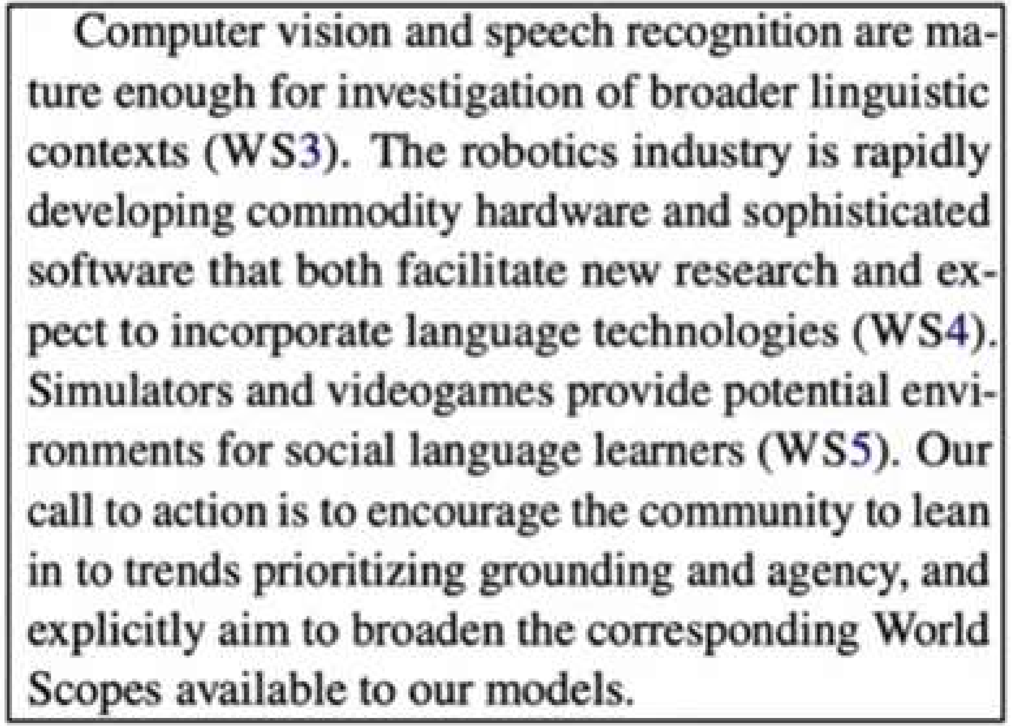


Grounding language to other modalities

- Many have articulated the need for using modalities other than text
- Bender and Koller [2020]: Impossible to acquire “meaning” (communicative
- intent of the speaker) from form (text / speech signal) alone
- Bisk et al [2020]: Training on only web-scale data limits the world scope of models.



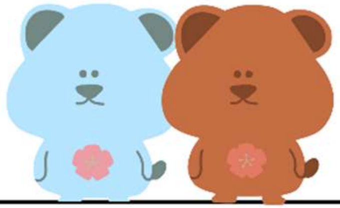
Grounding language to other modalities



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)

THANK YOU



EWHA
EURON