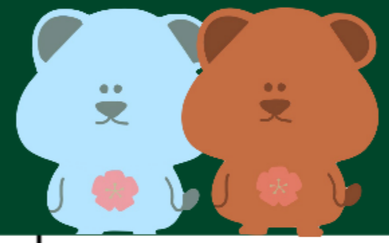


Future of NLP + Deep learning

김수한 문예지

EWHA EURON

#02 Compositional Representations and Systematic Generalization

#01 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

#01 용어

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

산 넘어 마을 = 넘다(산, 도착지)

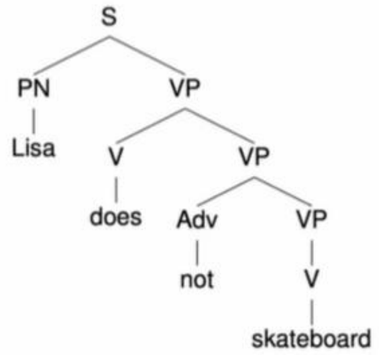
산 넘어 산 ≠ 넘다(산, 도착지)

#02 Compositional Representations and Systematic Generalization

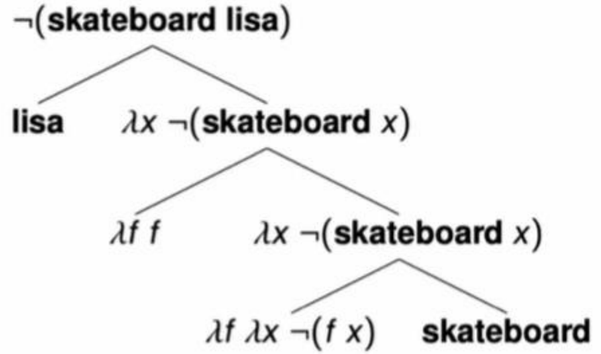
#02 Are neural representations compositional?

- **Compositionality of representations**

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



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



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

#02 Compositional Representations and Systematic Generalization

#02 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.

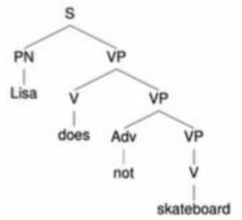
Given a dataset $\mathcal{X}$ of inputs x_i with derivations $d_i = D(x_i)$, compute:

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

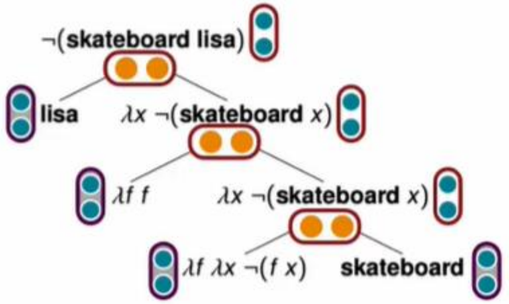
Then we can define datum- and dataset-level evaluation metrics:

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

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

#03 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

#03 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!**

Train set	**Test set**
Who directed Inception?	Did Greta Gerwig direct Goldfinger?
Did Greta Gerwig produce Goldfinger?	Who produced Inception?
...	...

#02 Compositional Representations and Systematic Generalization

#03 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.

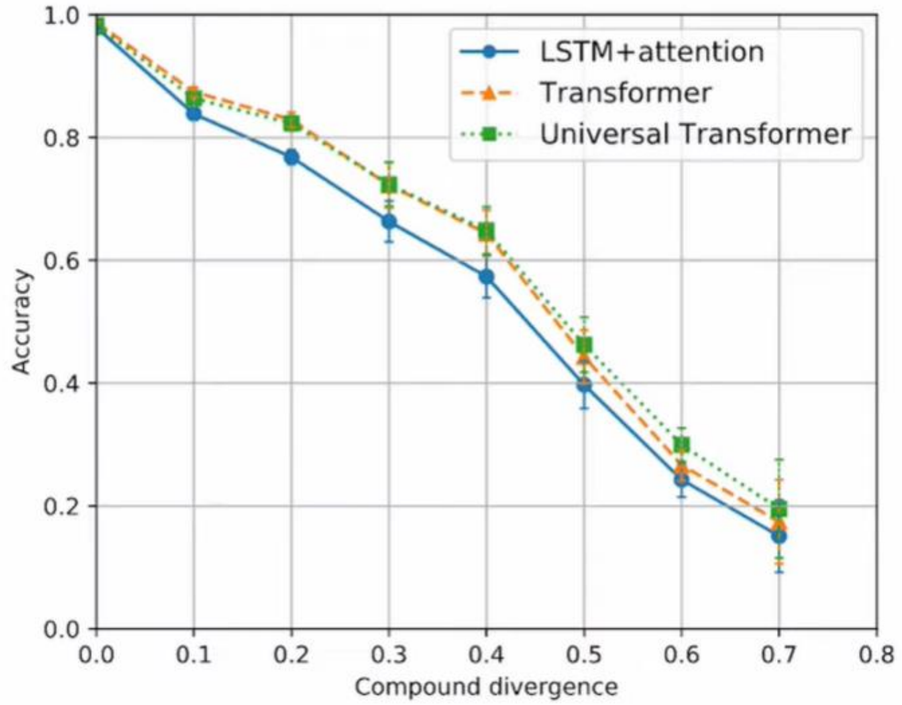
- The compound distributions of the train and test sets are very similar, then their compound divergence would be close to 0.
  → Not difficult tests for compositional generalization
- The compound divergence close to 1 means that the train-test sets have many different compounds.
  → Good test for compositional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

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

#02 Compositional Representations and Systematic Generalization

#03 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

#03 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

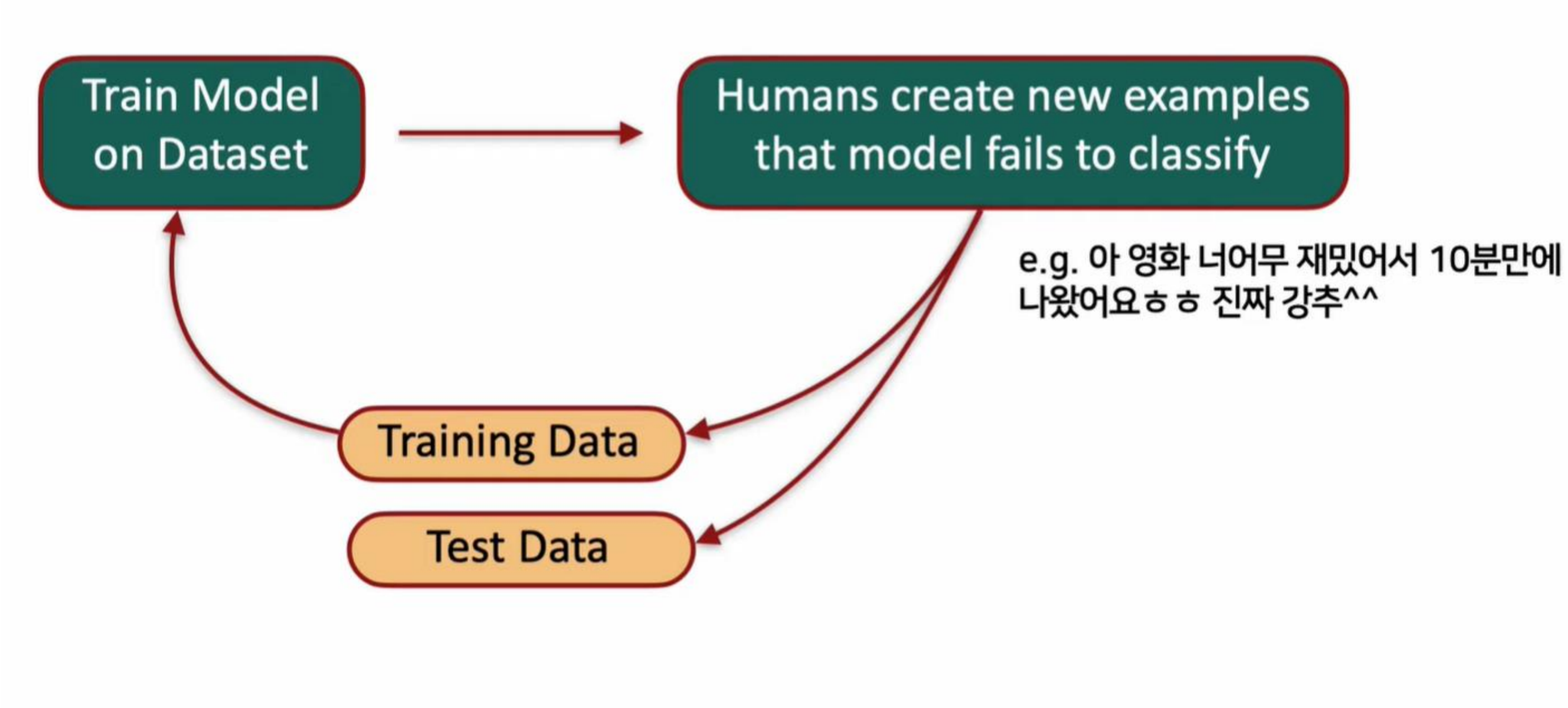
#03 Improving how we evaluate models in NLP

- 벤치마크 데이터셋에서의 모델 성능은 날로 증가하는데 정말 실제 세계에서의 모델 성능도 그만큼 증가했을까?

- Task에 대한 모델의 이해도를 어떻게 하면 정확하게 측정할 수 있는가?

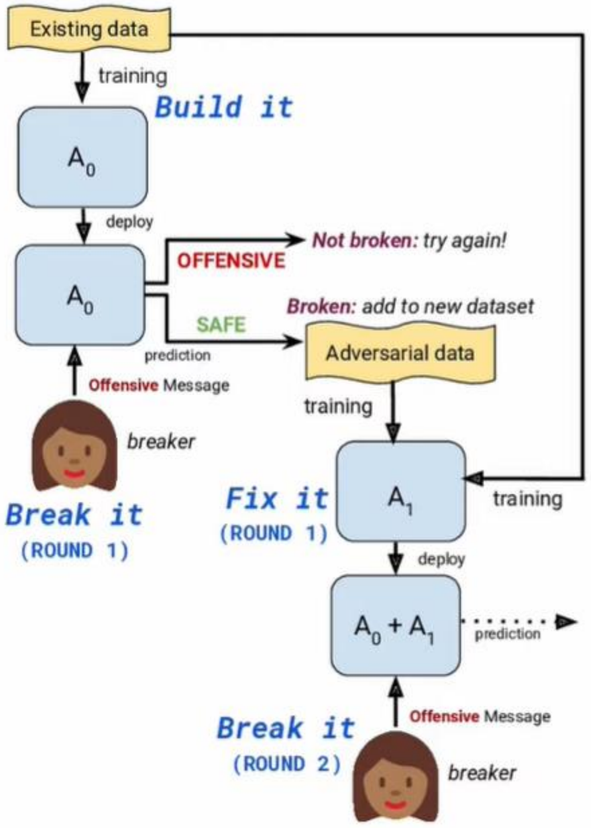
#03 Improving how we evaluate models in NLP

#01 Dynamic Benchmarks



#03 Improving how we evaluate models in NLP

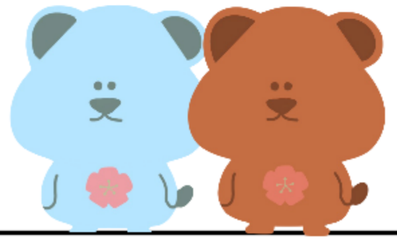
#01 Dynamic Benchmarks



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)

THANK YOU



EWHA EURON