# MaxProb!

## A Tutorial at WING Group Meeting

## 19th Nov 2021

Yisong Miao

Tutorial Webpage: https://yisong.me/MLP/maxprob

# MaxProb Methods in NLP

## Agenda

Background and Tutorial Statement

MaxProb Methods in NLP

- Two representative works on (1) retrieval-based QA and (2) text generation.
- Recent results from our group (on discourse understanding).

Thank Liangming for giving me the pointer at very beginning.


# A blunt definition: Measuring how confident a model is in its prediction.

Credit: https://www.incimages.com/uploaded_files/image/1920x1080/getty_506910700_351729.jpg

$$\underset{i}{\operatorname{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

A more mathematical definition:
(A notion from ML community)

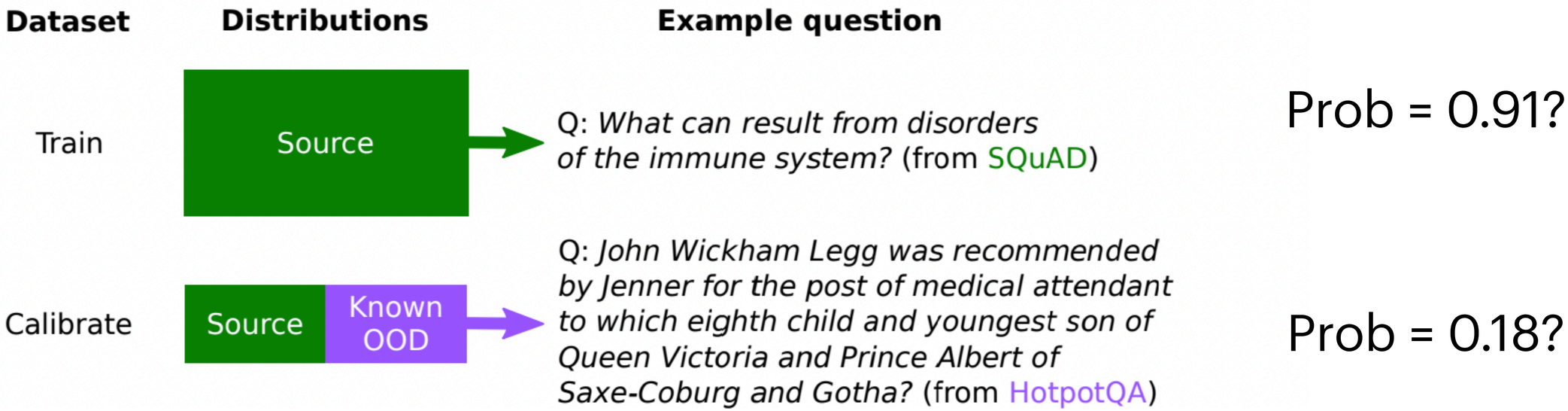
MaxProb is good enough for QA/Generation.
But it can be calibrated/adapted.

— Tutorial Statement

# Selective Question Answering under Domain Shift. ACL '20

Amita Kamath, Robin Jia, Percy Liang

# The Task — Selective QA




Prob = 0.91?

Prob = 0.18?

Selective QA: The model is able to abstain on Out-of-distribution data.

KRL '20

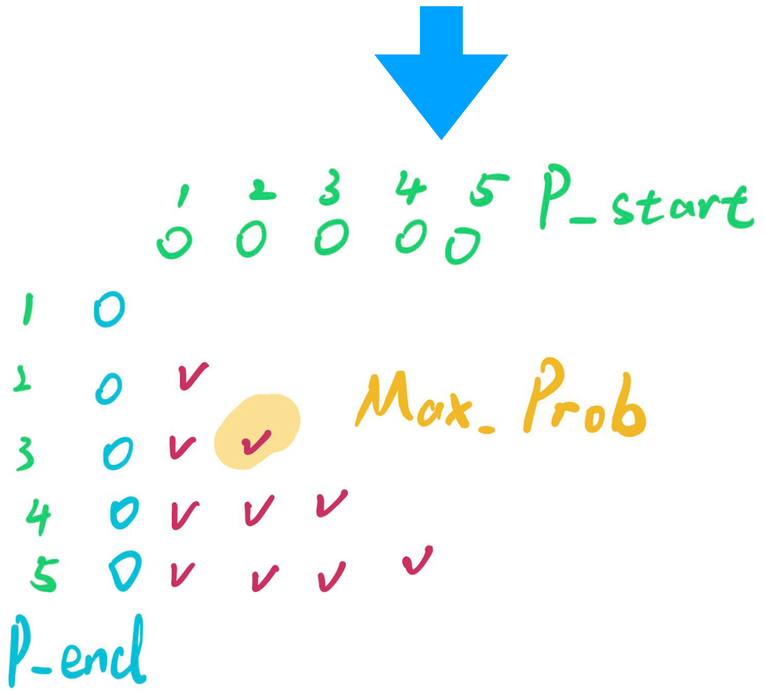
# Methods

## MaxProb and Calibrator

- MaxProb

Purely unsupervised.

$$c_{\text{MaxProb}} = f(\hat{y} \mid x) = \max_{y' \in Y(x)} f(y' \mid x)$$




"Calibrator"

— Supervised.

— Known OOD.

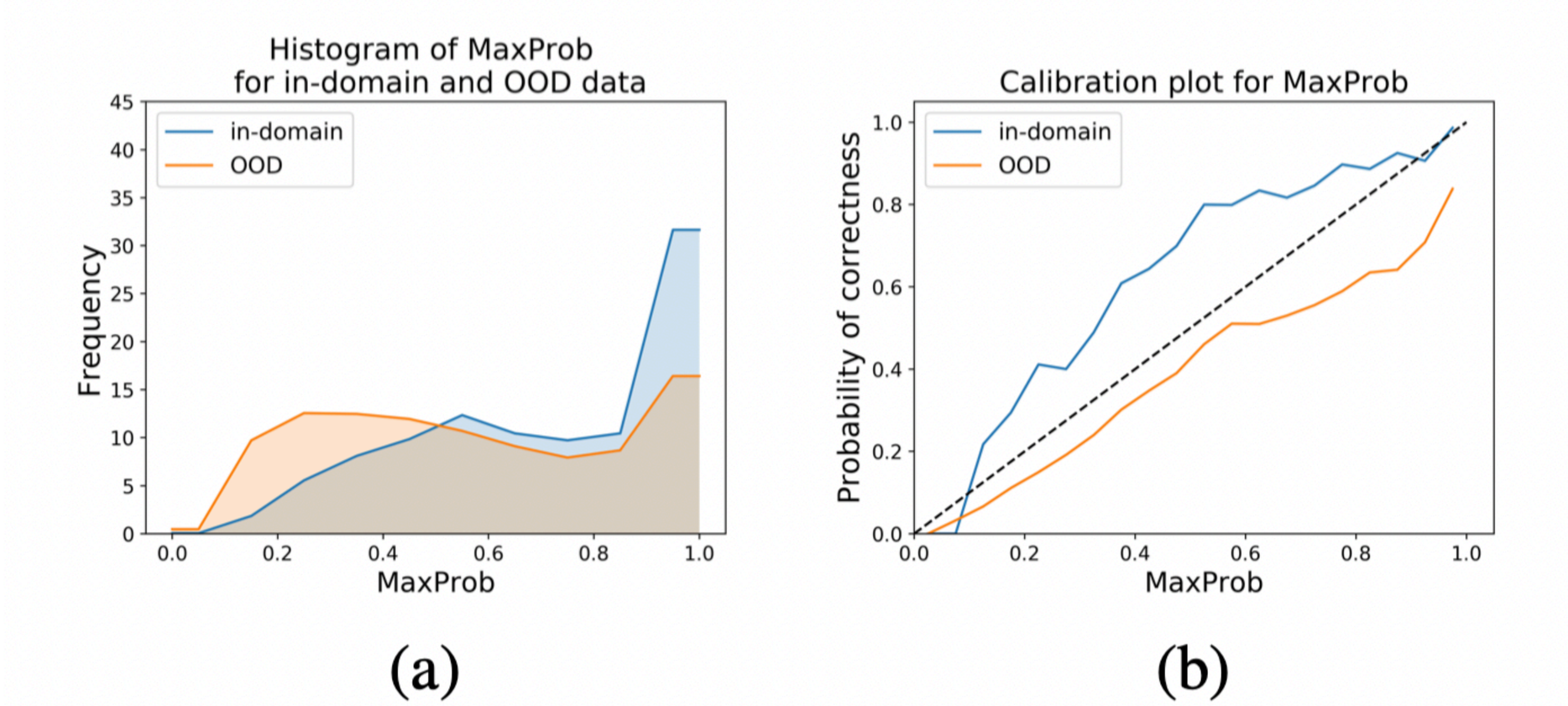
— Any textual feature.

— Any classifier can do the trick.

KRL '20

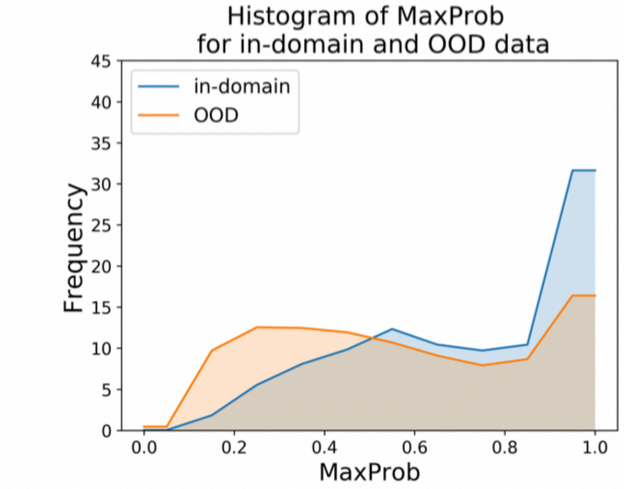
# MaxProb can be overconfident!

## — Results

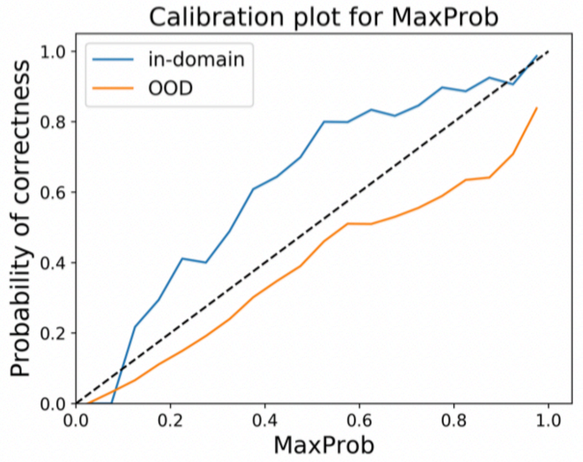



KRL ‘20

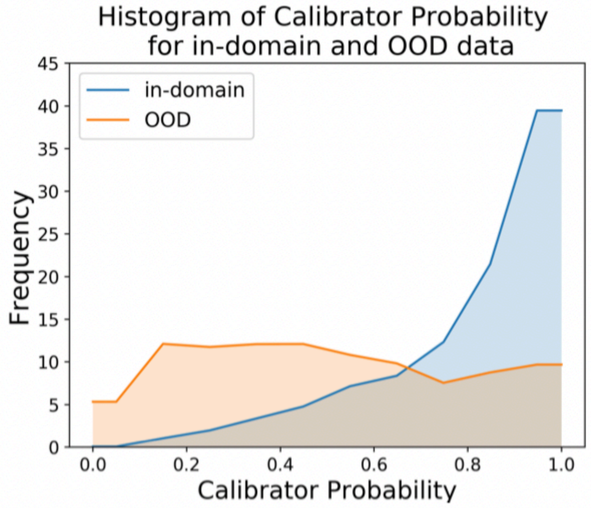
# MaxProb can be overconfident!



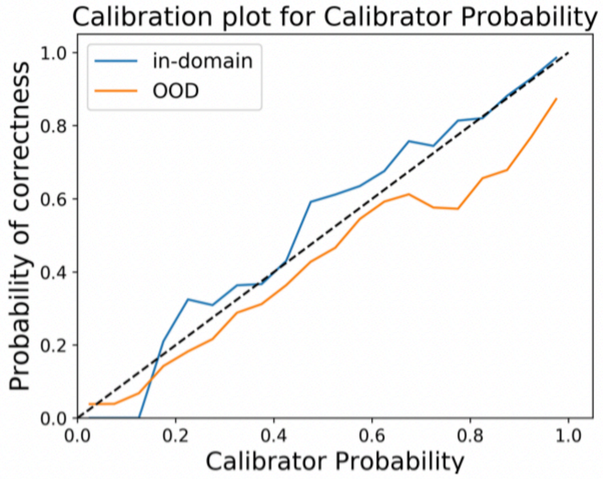

(a)




(b)




(c)




(d)

KRL '20

# Summary

MaxProb measures how confident a model is.

Unsupervised.

Tend to be overconfident.

KRL ‘20

# Surface Form Competition: Why the Highest Probability Answer Isn't Always Right

**Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, Luke Zettlemoyer**

Also covered by **Vered** in our NLP Seminar 2021:
https://wing-nus.github.io/nlp-seminar/speaker-vered
https://www.youtube.com/watch?v=L5CaFBRiRw4

# Surface Form Competition

## The problem to address in this paper.

**A human wants to submerge himself in water, what should he use?**

Humans *select* options

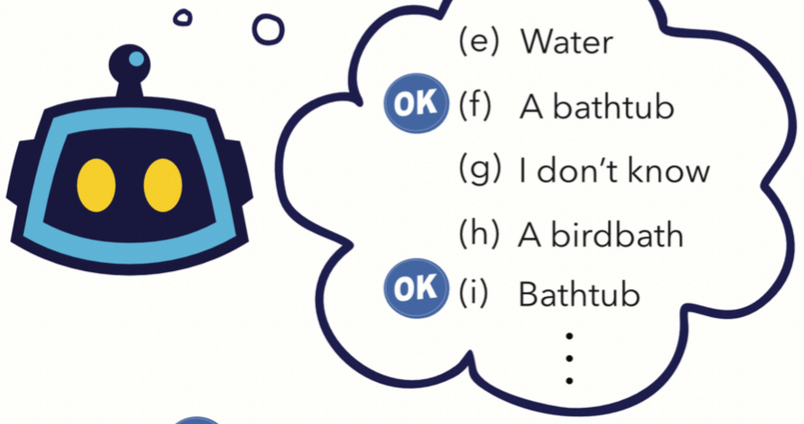
✗ (a) Coffee cup

✓ (b) Whirlpool bath

✗ (c) Cup

✗ (d) Puddle

Language Models assign probability to *every possible string*




OK = right concept, wrong surface form

# Zero-shot Learning

## The task and notations

Without the loss of generality, let's take Natural Language Inference (NLI) for example.

**X** (Premise): The bar closed because

$\mathbf{y_1}$ (hypothesis): it was crowded.

$\mathbf{y_2}$ (hypothesis): it was 3am.

Which is the correct hypothesis, $\mathbf{y_1}$ or $\mathbf{y_2}$ ?

# The Math!

## Baselines

- Language Model.

$$\underset{i}{\operatorname{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

$$P(\mathbf{y}_i|\mathbf{x}) = \prod_{j=1}^{\ell_i} P(y_i^j|\mathbf{x}, y_i^1, \cdots, y_i^{j-1})$$

- Length Normalised strategy.

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} \log P(y_i^j|\mathbf{x}, \mathbf{y}^{1\cdots j-1})}{\ell_i}.$$

**Core claim: Direct probability is not an adequate zero-shot scoring function due to surface form competition.**

HWSCZ '21

**x** (Premise): The bar closed because

$\mathbf{y_1}$ (hypothesis): it was crowded.

$\mathbf{y_2}$ (hypothesis): it was 3am.

Which is the correct hypothesis, $\mathbf{y_1}$ or $\mathbf{y_2}$ ?

# The Math!

## PMI and DC-PMI

- PMI

$$\mathrm{PMI}(\mathbf{x}, \mathbf{y}) = \log \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} = \log \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{x})}.$$

**How likely the hypothesis becomes given the hypothesis.**

- Domain-conditional PMI

$$\mathrm{PMI}_{\mathrm{DC}}(\mathbf{x}, \mathbf{y}, \mathrm{domain}) = \frac{P(\mathbf{y}|\mathbf{x}, \mathrm{domain})}{P(\mathbf{y}|\mathrm{domain})}$$

$$= \frac{P(\mathbf{y}|\mathbf{x}, \mathrm{domain})}{P(\mathbf{y}|\mathbf{x}_{\mathrm{domain}})}$$

**Remark 1** $P(\mathbf{y}|\mathbf{x}, \mathrm{domain}) = P(\mathbf{y}|\mathbf{x})$

**Remark 2** $P(\mathbf{y}|\mathrm{domain}) = P(\mathbf{y}|\mathbf{x}_{\mathrm{domain}})$

# The Math!
## Summary

- Domain-specific PMI

$$\mathrm{PMI}_{\mathrm{DC}}(\mathbf{x}, \mathbf{y}, \mathrm{domain}) = \frac{P(\mathbf{y}|\mathbf{x}, \mathrm{domain})}{P(\mathbf{y}|\mathrm{domain})} = \frac{P(\mathbf{y}|\mathbf{x}, \mathrm{domain})}{P(\mathbf{y}|\mathbf{x}_{\mathrm{domain}})}$$

**Remark 1** $P(\mathbf{y}|\mathbf{x}, \mathrm{domain}) = P(\mathbf{y}|\mathbf{x})$

**Remark 2** $P(\mathbf{y}|\mathrm{domain}) = P(\mathbf{y}|\mathbf{x}_{\mathrm{domain}})$

**x** (Premise): The bar closed because

$\mathbf{y_1}$ (hypothesis): it was crowded.

$\mathbf{y_2}$ (hypothesis): it was 3am.

Which is the correct hypothesis, $\mathbf{y_1}$ or $\mathbf{y_2}$ ?

$x_{domain}$ = “because”

$P(y|x_{domain})$ = P(y|”because”)

Back to our example!

$$\frac{P(\text{“it was 3am”}|\text{”The bar closed because”})}{P(\text{””it was 3am””}|\text{because})}$$

# Insights

## On designing $x_{domain}$

- SST: “[Illuminating if overly talky documentary] $_{P}$” [[The quote] has a tone that is] $_{DP}$ [positive.]$_{UH}$
- ARC: [What carries oxygen throughout the body?] $_{P}$[the answer is:] $_{DP}$[red blood cells.]$_{UH}$

For Seminar attendance yesterday: Is it possible to design *“better”* $x_{domain}$

- Continuous?
- Better “calibrated”?

# Comparing the two methods.

## In Selective QA (ACL '20) and Surface Form Competition (EMNLP '21)

### Calibrator

— Supervised.

— Known OOD.

— Any textual feature.

— Any classifier can do the trick.

KRL '20

### DC-PMI

$x_{domain}$ = "because"

$P(y|x_{domain}) = P(y|\text{"because"})$

HWSCZ '21

- Major difference: where the "calibration comes from?"
  - Calibrator: need known OOD to train a binary classifier.
  - DC-PMI: unsupervised to obtain $P(y|x_{domain})$.

# Conclusion

- MaxProb is Simple and effective measure of model confidence (I cannot think of something simpler!). But we need to calibrate/adapt it for usage.
- A classical ML method, being reused by fellas in the age of BERT. Similar renaissance happens in:
  - Information bottleneck method [1]
  - EM method [2]

[1] Specializing Word Embeddings (for Parsing) by Information Bottleneck. EMNLP ‘19

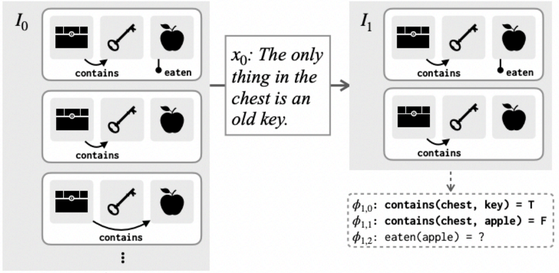
[2] Towards Interpretable Natural LanguageUnderstanding with Explanations as Latent Variables. NeurIPS ‘20

“I am hungry”.
Prob = ?

“Thank You”.
Prob = ?

“I want to play”.
Prob = ?

World State?



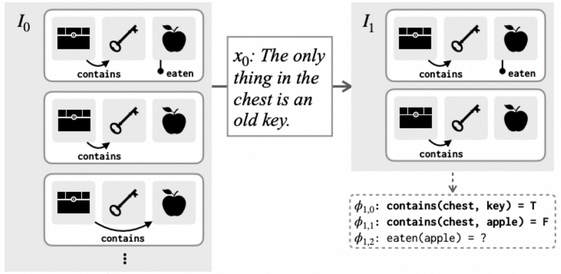

Inspiration drawn from LNA '21 to design this page.

"I am hungry".
Prob = 0.04

"Thank You".
Prob = 0.91

"I want to play".
Prob = 0.05

World State:{
Speaker: **IS** finishing a tutorial
Audiences: **IS** feeling happy
Time: **IS** Friday afternoon
Venue: **IS** Zoom
}




Inspiration drawn from LNA '21 to design this page.