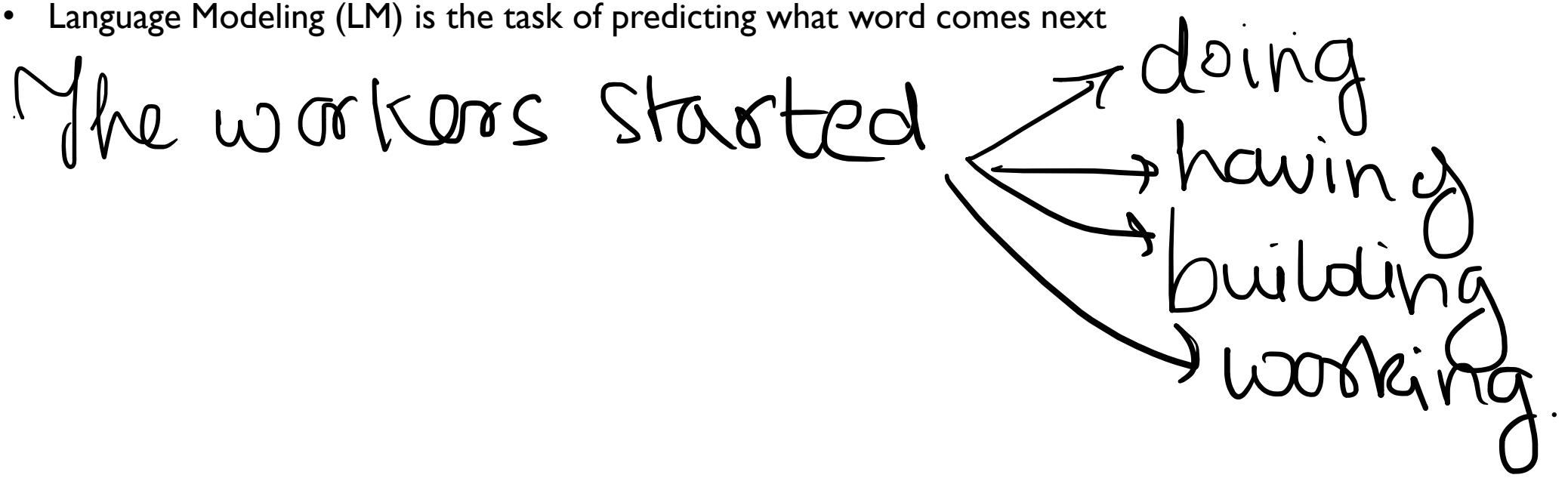
Presenter –
MUSKAN RATHORE
(DATA AND APPLIED SCIENTIST, MICROSOFT)

NLP – LANGUAGE MODELING

LANGUAGE MODELING

- Language Modeling (LM) is the task of predicting what word comes next



- Given a sequence of words $x^1, x^2, x^3, \ldots, x^t$, compute probability distribution of the next word x^{t+1}

$P(x^{t+1} | x^t, \ldots, x^1)$

Where x^{t+1} can be any word in vocabulary V={w1,w2,…,wV}

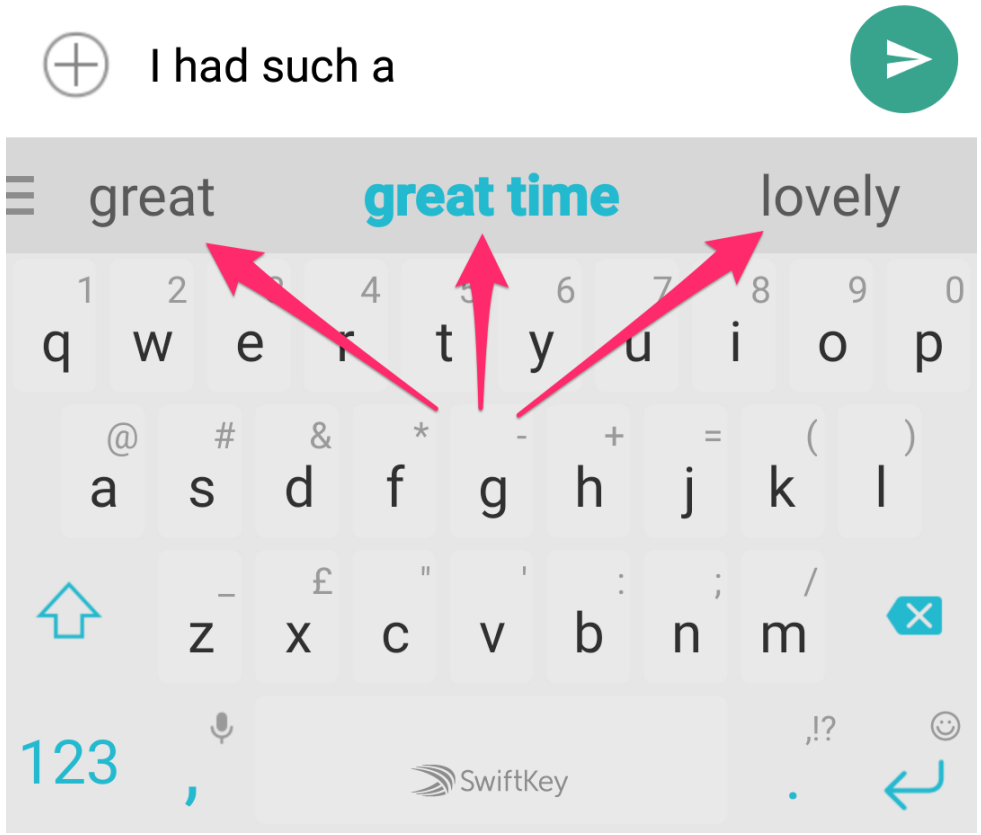
LANGUAGE MODELING

- You can also think of a Language Model as a system that assigns probability to a piece of text.
- For example, if we have some text $x^1, x^2, x^3, \ldots, x^T$, then the probability of this text (according to the Language Model) is:

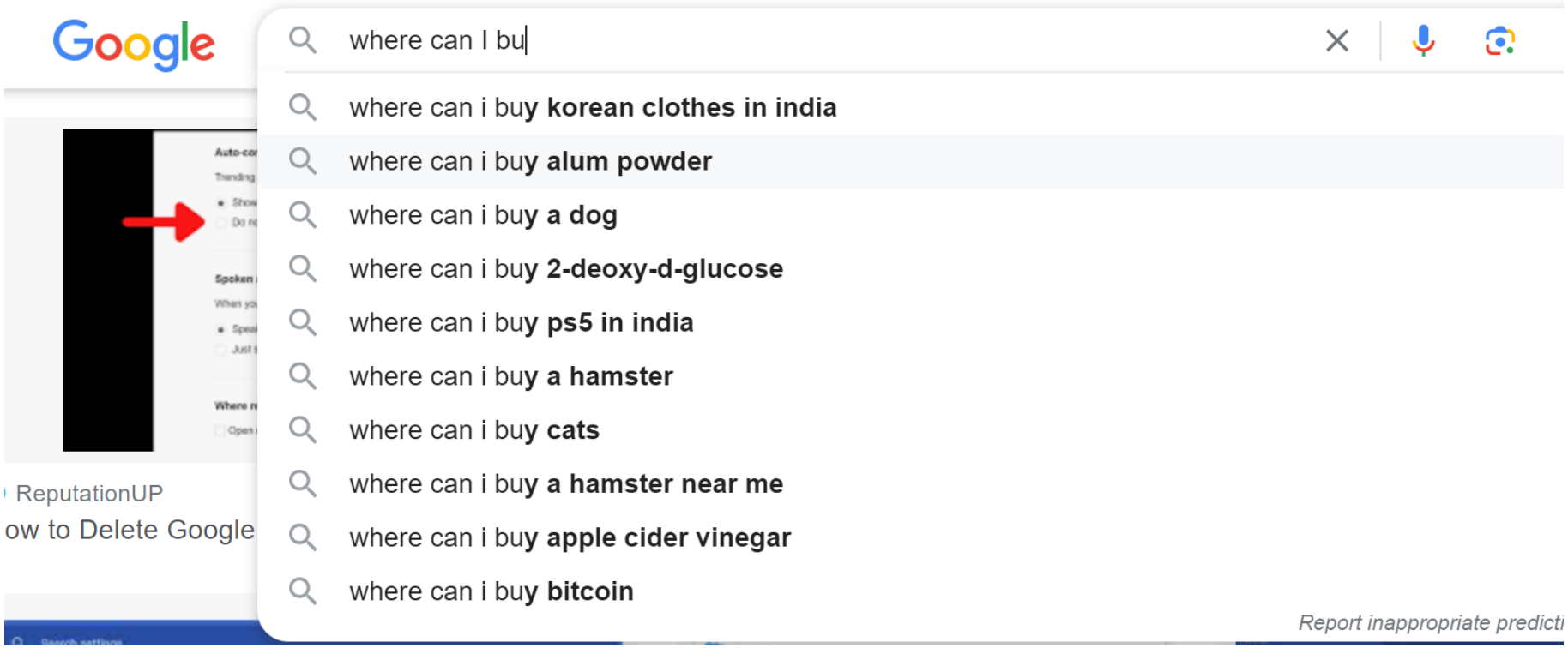
$$P(x^1, \ldots, x^T) = P(x^1) \cdot P(x^2 | x^1) \cdot P(x^3 | x^2, x^1) \cdots P(x^T | x^{T-1}, \ldots, x^1)$$

$$= \prod_{t=1}^{T} P(x^t | x^{t-1}, \ldots, x^1)$$

Language Model.



APPLICATIONS – KEYBOARD TEXT PREDICTION



APPLICATIONS – SEARCH BAR PREDICTION

- Question: How to learn a Language Model?
- Answer (pre- Deep Learning): learn a **n-gram Language Model**!

Ex: "The students opened their _____"

- Definition: A n-gram is a chunk of n consecutive words.
  - unigrams: "the", "students", "opened", "their"
  - bigrams: "the students", "students opened", "opened their"
  - trigrams: "the students opened", "students opened their"
  - 4-grams: "the students opened their"

- Idea: Collect statistics about how frequent different n-grams are, and use these to predict next word.

LANGUAGE MODELING

N-GRAM LANGUAGE MODELS

- We make a simplifying assumption: x^{t+1} depends not on all previous words, but only on the preceding n-1 words.

$$P(x^{t+1} | x^t, \ldots x^1) = P(x^{t+1} | \underbrace{x^t, \ldots, x^{t-n+2}}_{\text{Previous n-1 words}})$$

According to Bayes

$$P(A|B) = \frac{P(A.B)}{P(B)}$$

By Bayes' theorem,

$$P(x^{t+1} \mid x^t, \ldots x^{t-n+2}) = \frac{P(x^{t+1}, x^t, \ldots, x^1)}{P(x^t, \ldots, x^1)}.$$

Prob. of n-gram

Prob of n-1 gram

$$= \frac{\text{count}(x^{t+1}, x^t, \ldots, x^1)}{\text{count}(x^t, \ldots x^1)}.$$

LANGUAGE MODELING

Ex: "as the examiner started the clock, the students opened their _____"

Lets take example of 4-gram model, then to gte the word in space we will only consider last 3 words and not all words.

Ex: "~~as the examiner started the clock,~~ the students opened their _____"

$$P(w \mid \text{"students opened their"}) = \frac{\text{Count("students opened their } w\text{")}}{\text{Count("students opened their")}}$$

LANGUAGE MODELING

For example, suppose that in the corpus:

- "students opened their" occurred 1000 times
- "students opened their books" occurred 400 times
  → P(books | students opened their) = 0.4
- "students opened their exams" occurred 100 times
  → P(exams | students opened their) = 0.1

$P(w=\text{"exams"}) < P(w=\text{"books"})$

So we can predict $w = \text{"books"}$

Looking at the sentence "as the proctor started the clock, the students opened their _____"

Do you think our prediction is correct??

LANGUAGE MODELING

For example, suppose that in the corpus:

- "students opened their" occurred 1000 times
- "students opened their books" occurred 400 times
  → P(books | students opened their) = 0.4
- "students opened their exams" occurred 100 times
  → P(exams | students opened their) = 0.1

$P(w = \text{"exams"}) < P(w = \text{"books"})$
So we can predict
$w = \text{"books"}$

Looking at the sentence "as the proctor started the clock, the students opened their _____"
No – Why? – Because we left the context of "examiner", had we looked at all prev words and not just n-1 words we would know that w should be "exams" with higher probability and not "books"

LANGUAGE MODELING

SPARSITY PROBLEMS IN N-GRAM MODELS

Sparsity Problem 1

Problem: What if *"students opened their* $\boldsymbol{w}$*"* never occurred in data? Then $\boldsymbol{w}$ has probability 0!

(Partial) Solution: Add small δ to the count for every $\boldsymbol{w} \in V$. This is called *smoothing*.

$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

Sparsity Problem 2

Problem: What if *"students opened their"* never occurred in data? Then we can't calculate probability for *any* $\boldsymbol{w}$!

(Partial) Solution: Just condition on *"opened their"* instead. This is called *backoff*.

Note: Increasing *n* makes sparsity problems *worse.* Typically we can't have *n* bigger than 5.

Storage: Need to store count for all *n*-grams you saw in the corpus.

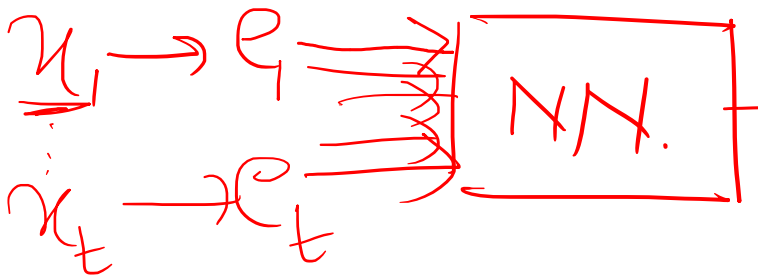
$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

Increasing *n* or increasing corpus increases model size!

STORAGE PROBLEMS IN N-GRAM MODELS

NEURAL LANGUAGE MODELING

$x_1, x_2 \ldots x_t \longrightarrow x_{t+1}$ ↪ Predicting next word using NN model.



$h = f(W.e + b_1)$

$\hat{y} = \text{softmax}(U.h + b_2)$

Adv
↪ No sparsity problem, No storage problem

Disadv ↪ ① Fixed window is too small.
② Sequentiality of words is not considered.

QUESTIONS?

THANK YOU !