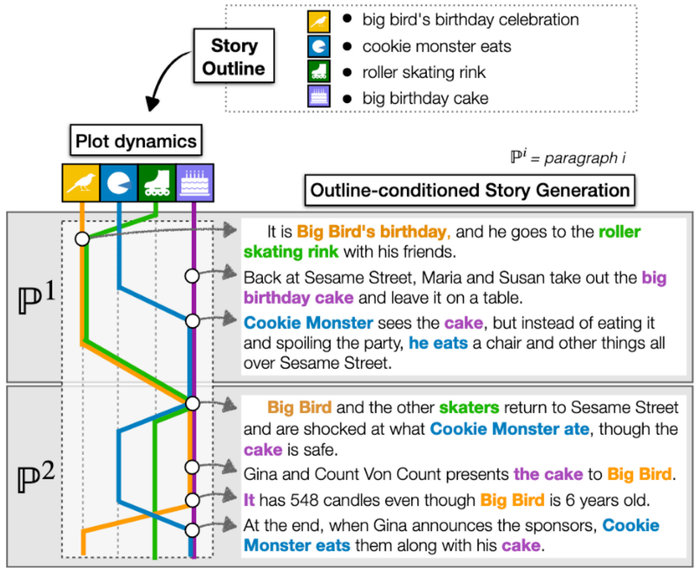
Natural Language Generation

COMP3361 — Week 10

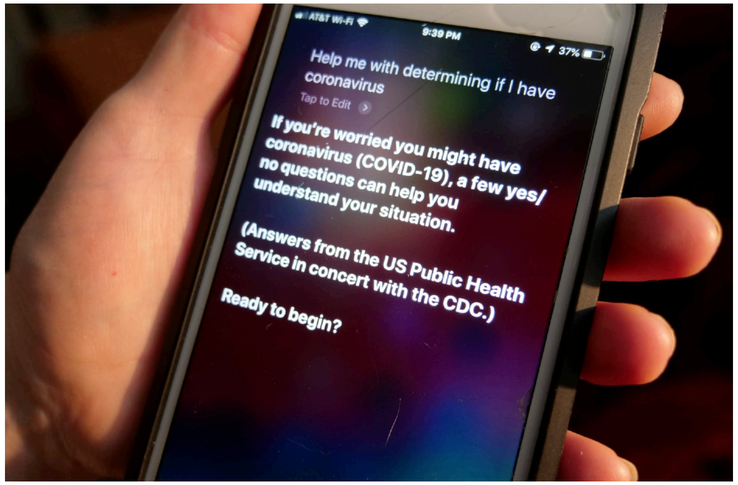
Lingpeng Kong

Department of Computer Science, The University of Hong Kong
Many materials from Stanford CS224n with special thanks!

Natural Language Generation



Story Generation

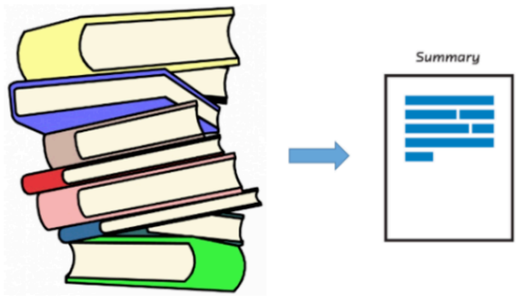


Dialogue Systems

Spring Couplets

Natural Language Generation

Machine Translation



Abstractive Summarization

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Image Captioning

Template-based NLG

Source Entity: Cotto

type[coffee shop], rating[3 out of 5],
food[English], area[city centre],
price[moderate], near[The Portland Arms]

System Generation:

Cotto is a coffee shop serving English food in the moderate price range. It is located near The Portland Arms. Its customer rating is 3 out of 5.

Description

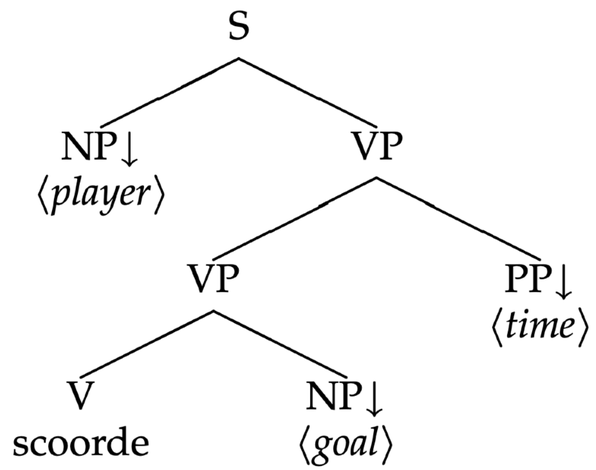
Game Commentary

Template-based NLG (Demo)

Thanks Tencent AI lab :-)

Template-based NLG

$S =$



$E =$ *player* ← ExpressObject (*currentgoal.player, P, nom*)
goal ← ExpressObject (*currentgoal, P, gen*)
time ← ExpressTime (*currentgoal.time*)

$C =$ Known (*currentmatch.result*) ∧ *currentgoal* = First (*notknown,goallist*) ∧
currentgoal.type ≠ *owngoal*

$T =$ goalscoring

(Kees van Deemter et al., 2003)

Template-based NLG

Why we like template-based NLG?

1. Good for a specific domain.
2. Controlable.
3. High quality sentences.

Why we don’t like template-based NLG?

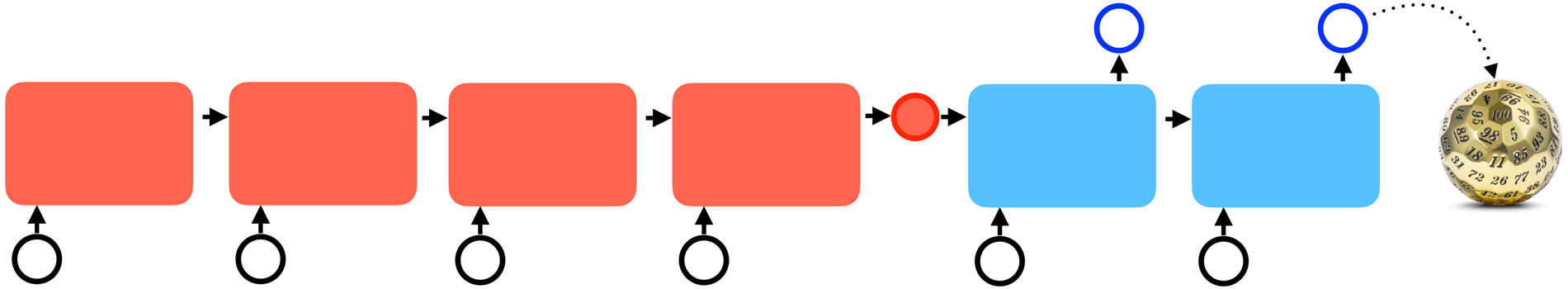
1. Domain transfer is difficult — you will have to build for every single application.
2. Not open-ended.
3. Require a lot human effort.

NLG as (Conditional) Language Modeling

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(y_1 \dots y_n \mid x_1 \dots x_m) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

target source

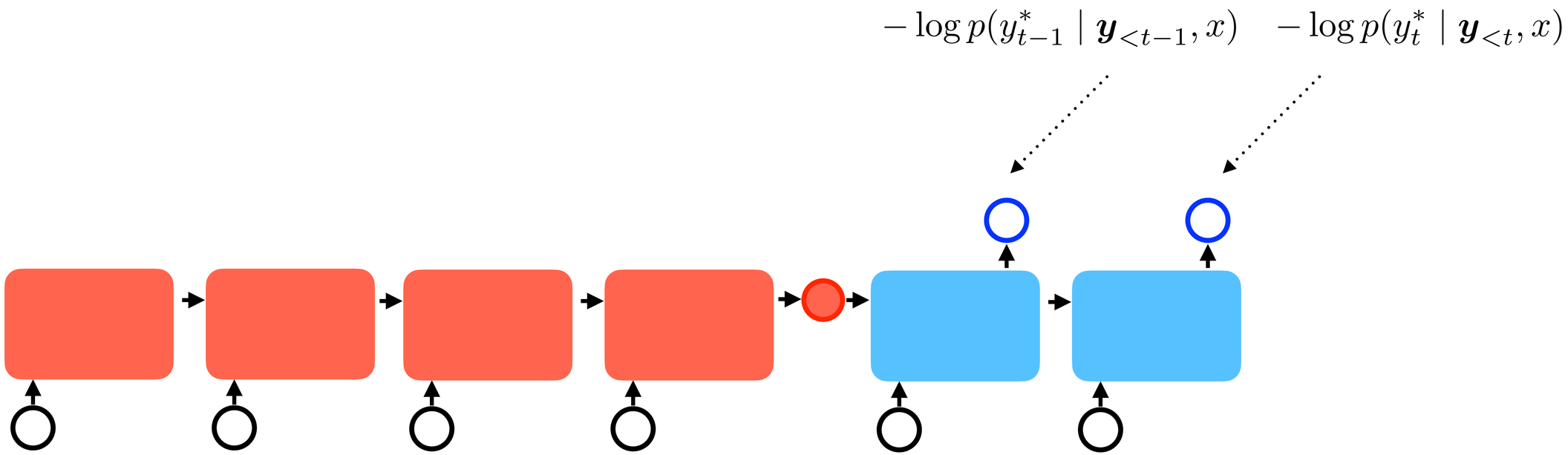
Conditional Language Model



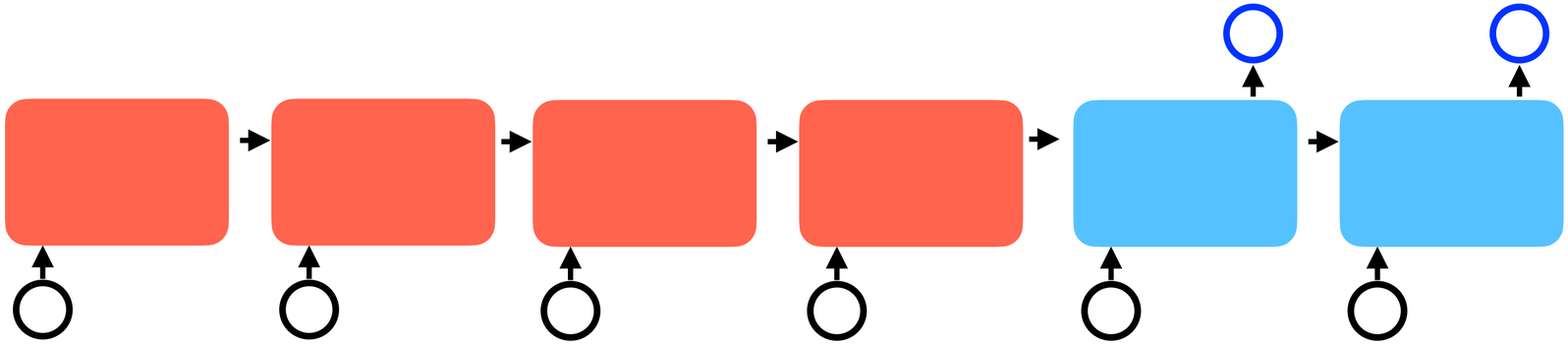
$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$

NLG as (Conditional) Language Modeling

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(y_1 \dots y_n \mid x_1 \dots x_m) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$



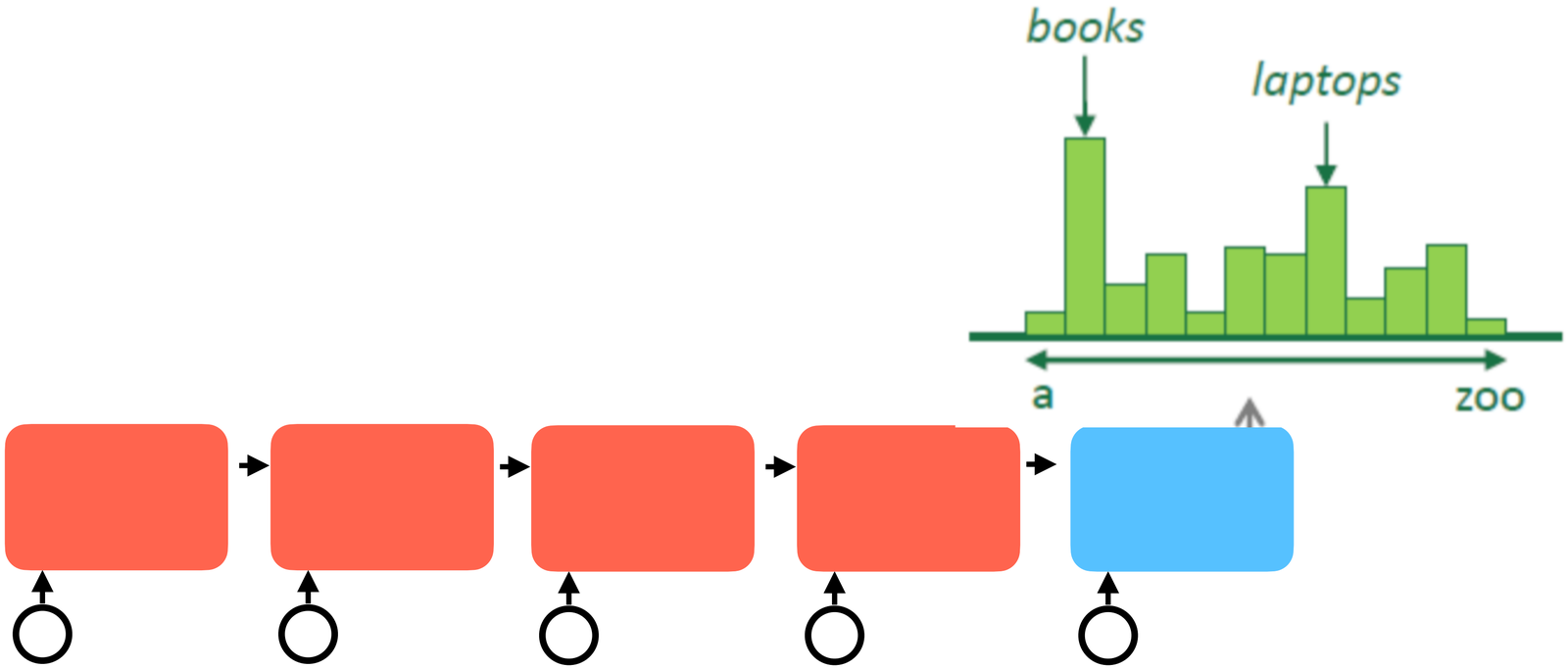
NLG as (Conditional) Language Modeling



Training from scratch or starting with

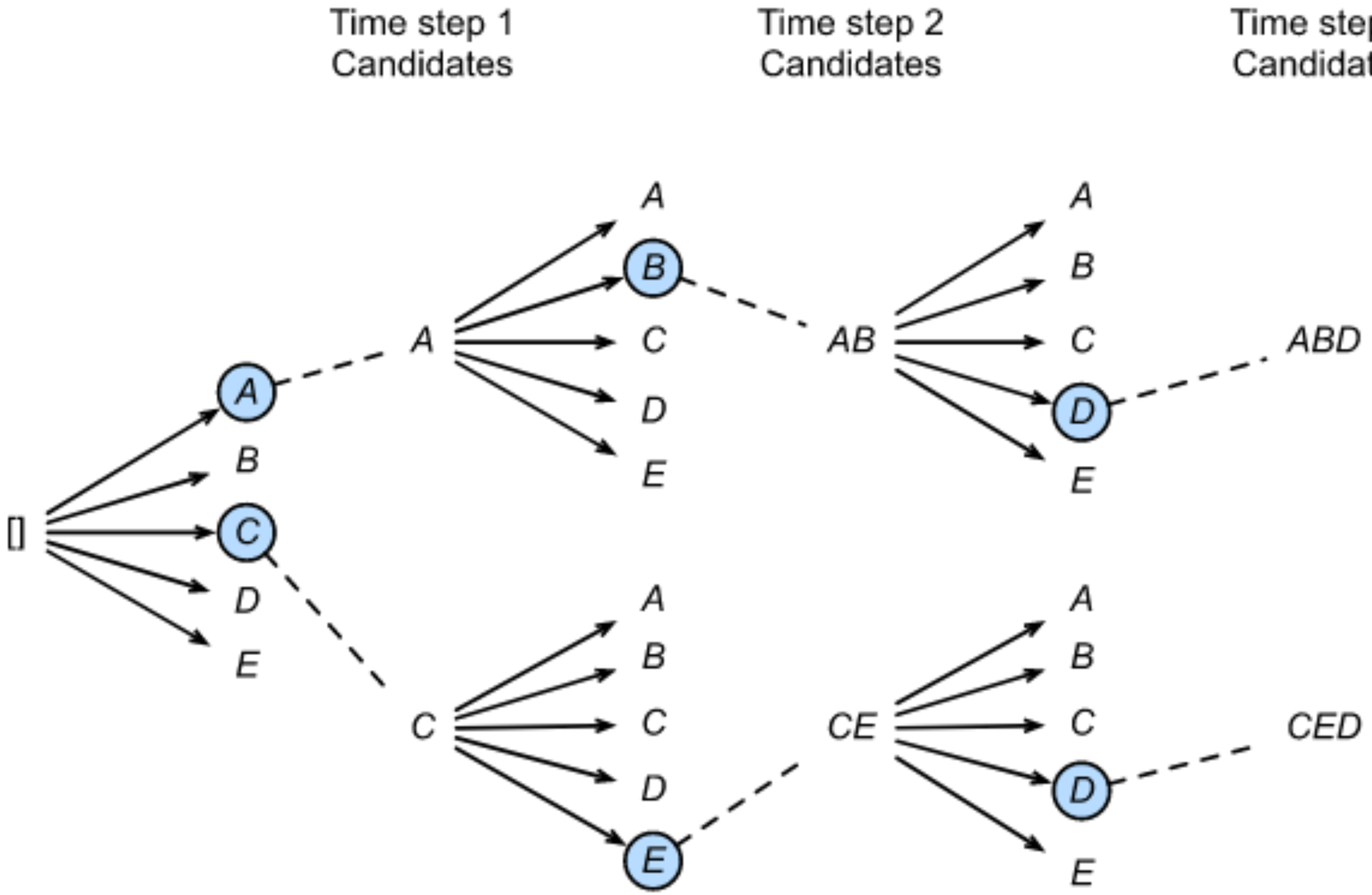
? The objective is the same.

How to perform decoding?



$$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Beam Search



https://d2l.ai/

Dive into Depp Learning

Beam Search or Pure Sampling?

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

degenerate repetition

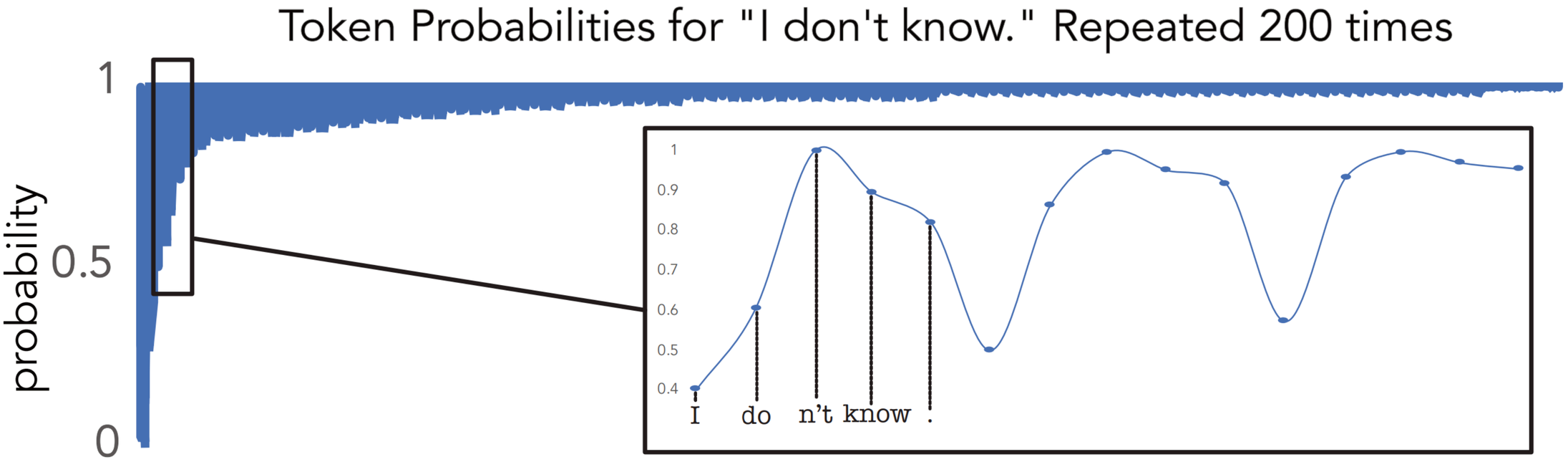
Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

incoherent gibberish

GPT-2 Large (774M parameters)

Holtzman et al., 2020

What happened?



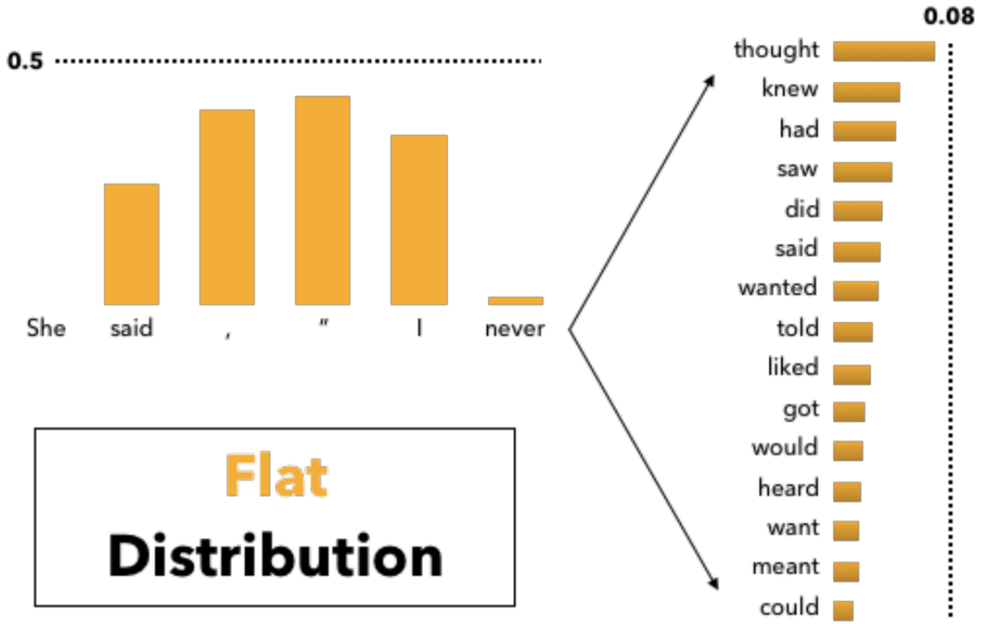
The probability of a repeated phrase increases with each repetition, creating a feedback loop.

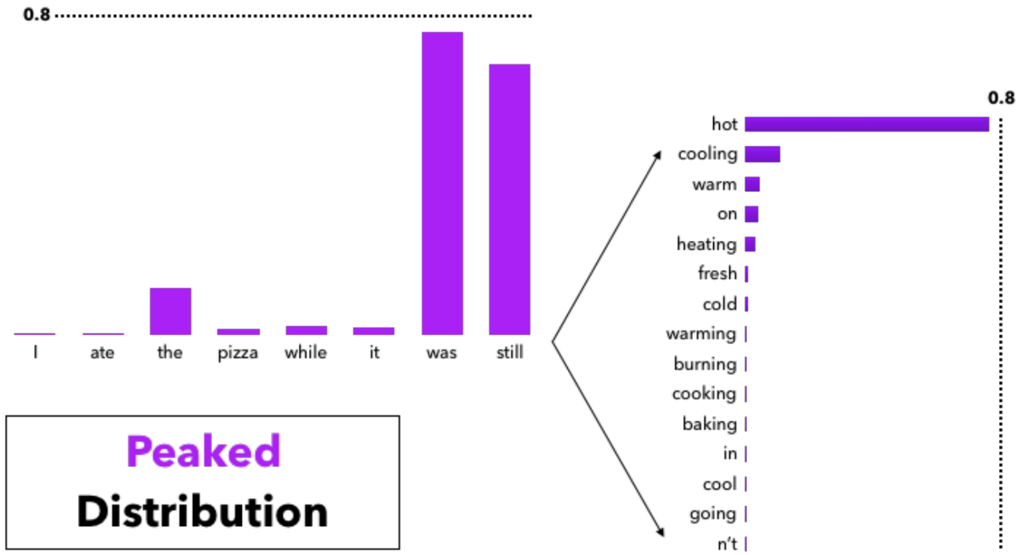
Holtzman et al., 2020

Top-K Sampling

Fan et al., 2018; Holtzman et al., 2020

Top-K Sampling

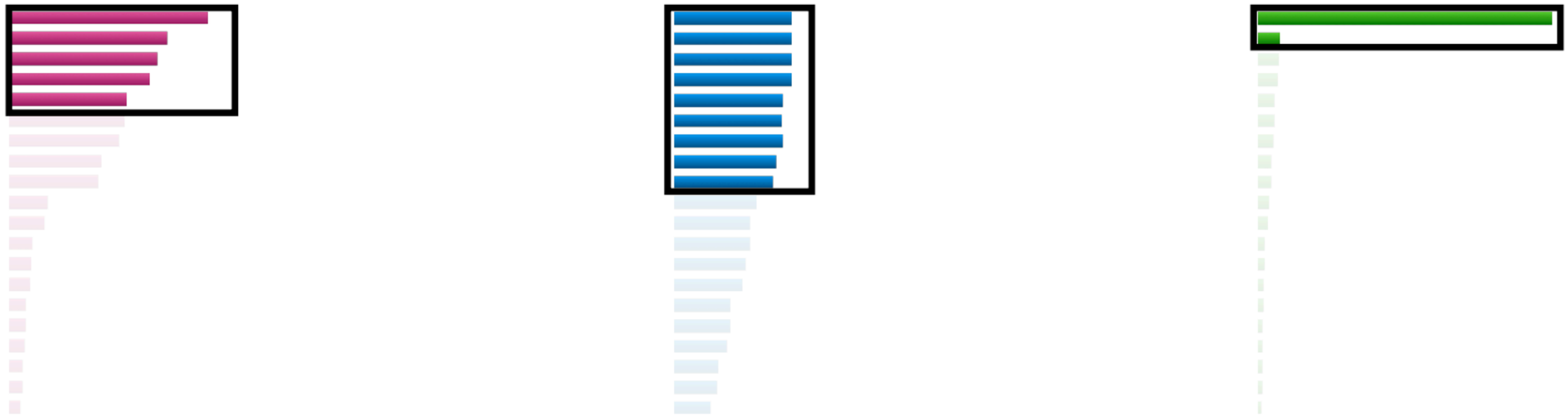




Holtzman et al., 2020

Top-p (nucleus) Sampling

To cut off by the cumulative probability mass, rather than the first K terms.



Holtzman et al., 2020

Evaluation of Text Generation

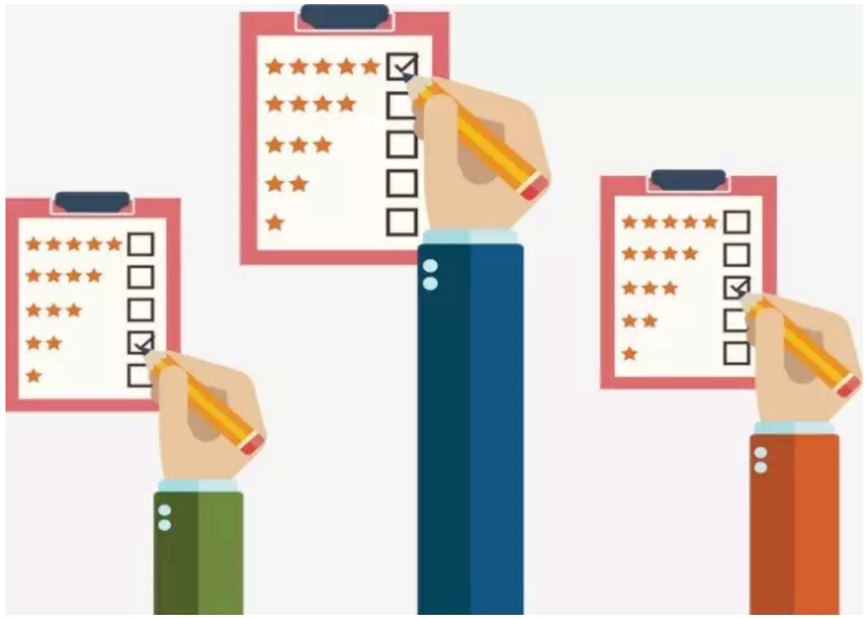
Human Centric Evaluation

Untrained Automatic Evaluation

Machine Learned Metrics

Celikyilmaz et al., 2020

Human-Centric Evaluation



Informativeness

Relevance

Fluency

Coherence

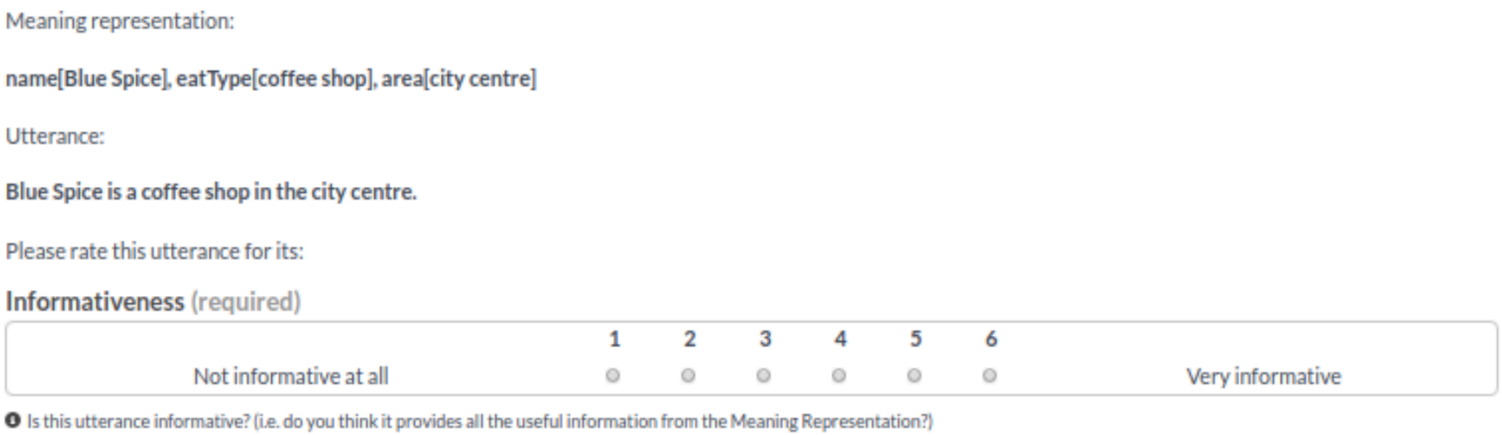
Factuality

Semantic Coverage

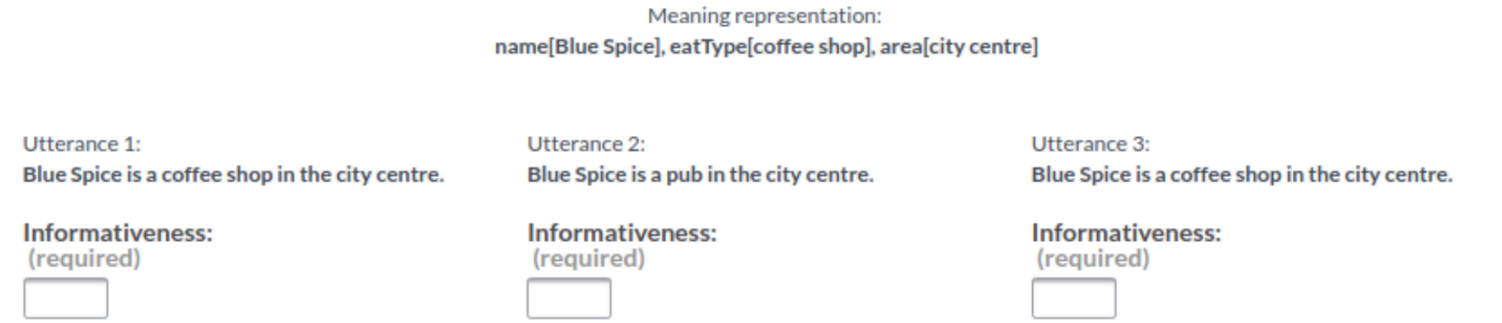
Adequacy

...

Human-Centric Evaluation



(a) Likert-scale question



(b) RankME-style question

Celikyilmaz et al., 2020

Untrained Automatic Metrics

Reference-based String matching

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

System output: The cat the cat on the mat.

Papineni et al., 2020

Untrained Automatic Metrics

Reference-based String matching

What you had for breakfast this morning?

Reference: I had one egg and some bread.

System1 output: I had one computer and some keyboards.

System2 output: I got up late so I did not even eat anything!

Papineni et al., 2020

Untrained Automatic Metrics

Reference-free Metrics (e.g. PPL)

The higher this quantity is, the better the language model is at modeling unseen sentences.

$$\prod_{i=1}^{m} p(x^{(i)})$$

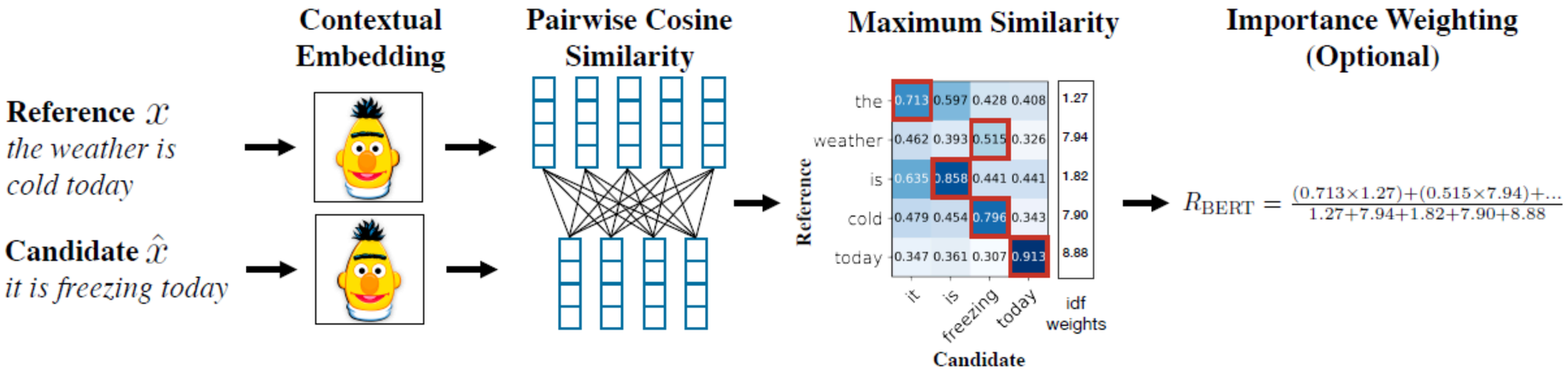
Perplexity on the test corpus is derived as a direction transformation of this.

$$\text{ppl} = 2^{-l}$$

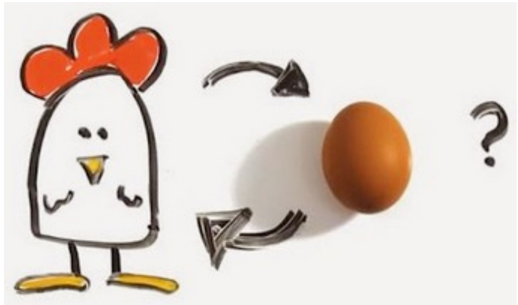
$$l = \frac{1}{M} \sum_{i=1}^{m} \log_2 p(x^{(i)})$$

M is the total length of the sentences in the test corpus.

Machine-Learned Metrics



Similarity between two machine-generated texts or between machine-generated and human-generated texts. These models can be viewed as digital judges that simulate human judges.



Sometimes a chicken egg problem

Celikyilmaz et al., 2020