CS11-711 Advanced NLP

Conditioned Generation

Daniel Fried and Robert Frederking
with slides from Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site

https://cmu-anlp.github.io/

Language Models

- Language models are generative models of text

$$s \sim P(x)$$

Title: Star’s Tux Promise Draws Megyn Kelly’s Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, `This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'"

Text Credit: Brown et al., 2020. GPT-3 paper

Conditioned Language Models

- Not just generate text, generate text according to some specification

Input *X*	Output *Y* (**Text**)	Task
English	Japanese	Translation
Document	Short Description	Summarization
Utterance	Response	Response Generation
Image	Text	Image Captioning
Speech	Transcript	Speech Recognition
Structured Data	NL Description	NL Generation

Formulation and Modeling

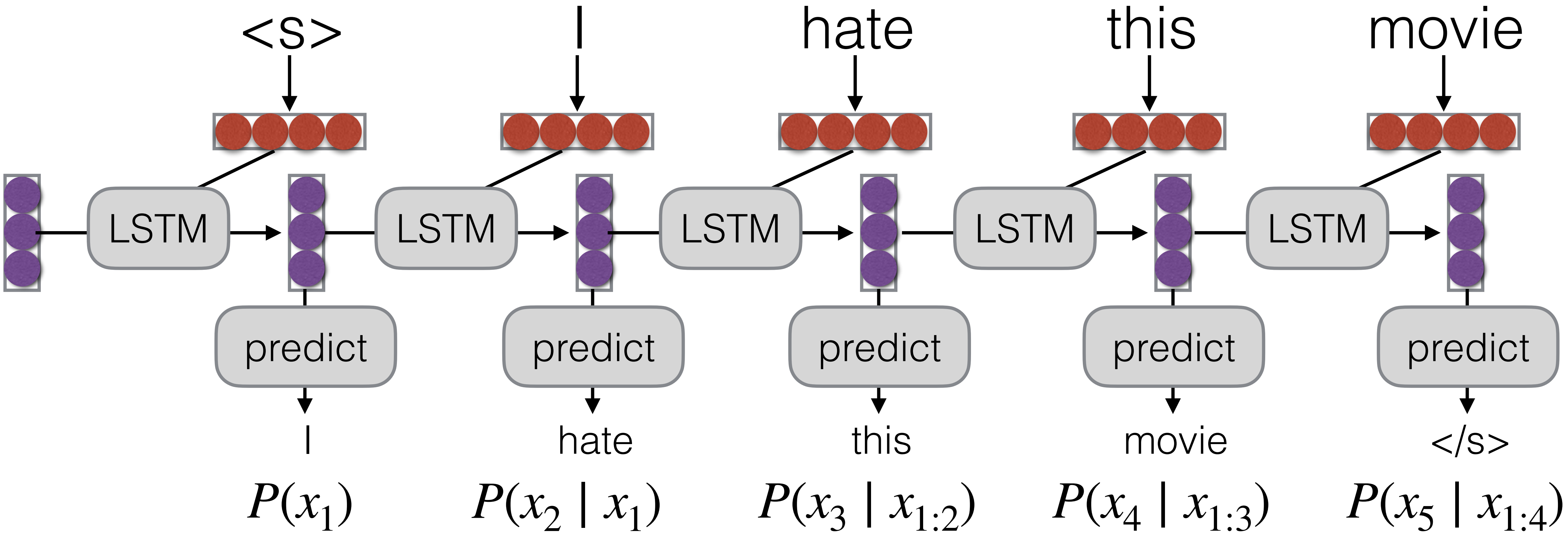
Calculating the Probability of a Sentence

$$P(X) = \prod_{i=1}^{I} P(\underbrace{x_i}_{\text{Next Word}} \mid \underbrace{x_1, \ldots, x_{i-1}}_{\text{Context}})$$

Conditional Language Models

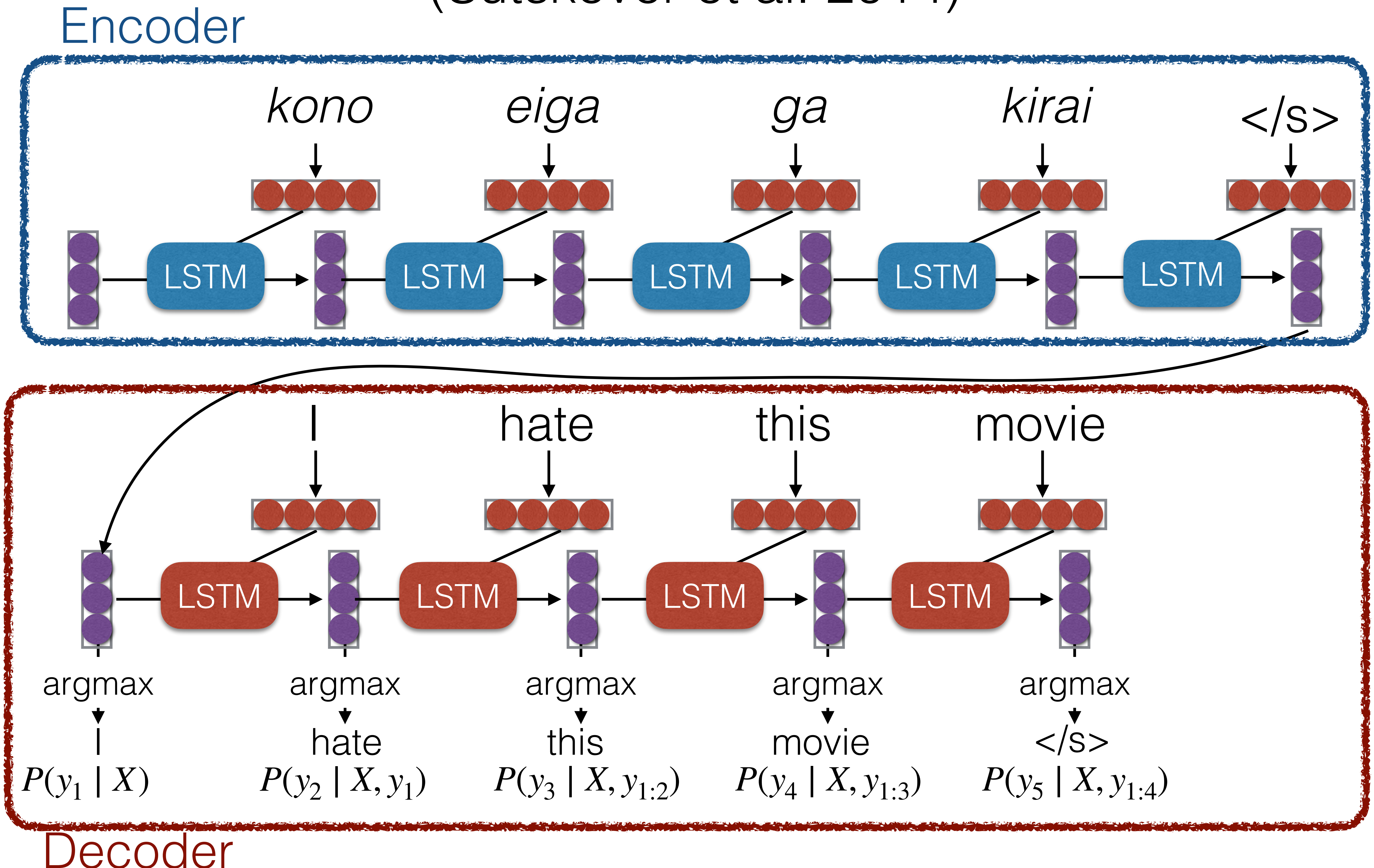
$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid \underset{\substack{\uparrow \\ \text{Added Context!}}}{\underline{X}}, y_1, \ldots, y_{j-1})$$

(One Type of) Language Model
(Mikolov et al. 2011)



(One Type of) Conditional Language Model

(Sutskever et al. 2014)



Encoder

kono *eiga* *ga* *kirai* </s>

LSTM LSTM LSTM LSTM LSTM

Decoder

I hate this movie

LSTM LSTM LSTM LSTM

argmax → I $P(y_1 \mid X)$

argmax → hate $P(y_2 \mid X, y_1)$

argmax → this $P(y_3 \mid X, y_{1:2})$

argmax → movie $P(y_4 \mid X, y_{1:3})$

argmax → </s> $P(y_5 \mid X, y_{1:4})$

Methods of Generation

The Generation Problem

- We have a model of $P(Y|X)$, how do we use it to generate a sentence?
- Two methods:
  - **Sampling:** Try to generate a *random* sentence according to the probability distribution.
  - **Argmax:** Try to generate the sentence with the *highest* score.

Ancestral Sampling

- **Randomly generate** words one-by-one.

```
while $y_{j-1}$ != “</s>”:
  $y_j \sim P(y_j \mid X, y_1, \ldots, y_{j-1})$
```

- An **exact method** for sampling from the model for P(X), no further work needed.

- Maximum likelihood training assumes samples are sampled from the underlying distribution =>
  ancestral samples are what your model thinks the training data looks like.

Greedy Search

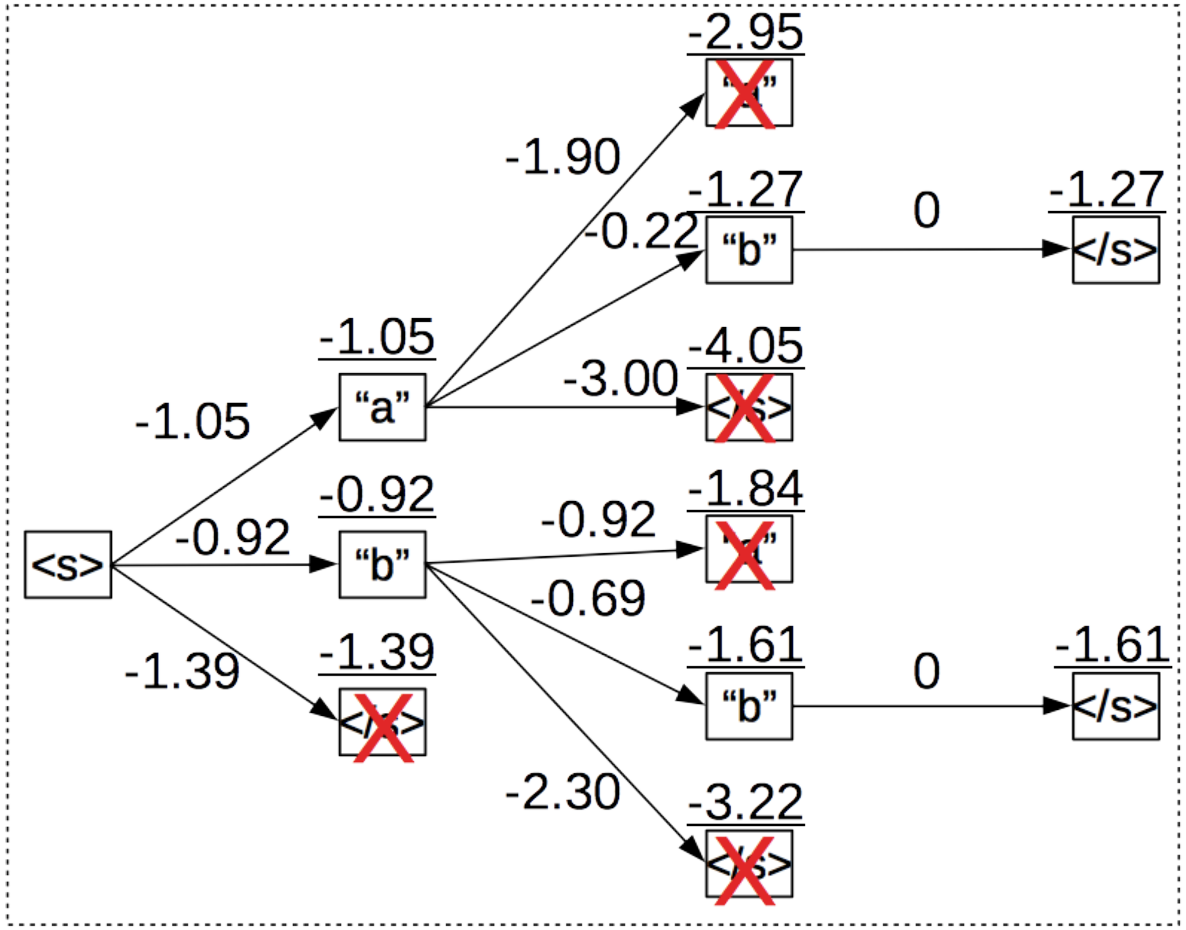
- One by one, pick the single highest-probability word

```
while $y_{j-1}$ != “</s>”:
    $y_j = \text{argmax}\, P(y_j \mid X, y_1, \ldots, y_{j-1})$
```

- Deterministic
- **Not exact, real problems:**
  - Will often generate the “easy” words first
  - Will prefer multiple common words to one rare word

Beam Search

- Instead of picking one high-probability word, maintain several paths



- Adjust the "beam size" *b* to spend more time on searching, find a better scoring hypothesis

Case Studies in Conditional Language Modeling

Translation

- Most widely used example of generation?
- e.g. Translation from **WMT shared tasks** (http://statmt.org/wmt21/)
  - News translation
  - Similar language translation
  - Biomedical translation
  - Low-resource multilingual translation
  - Translation efficiency
  - Terminology-aware translation
  - Lifelong learning

Summarization

- Take a longer text, and output a shorter text
- Can use *extractive* models that copy input, or *abstractive* models that generate from scratch
- Multiple granularities:
    - Sentence compression (sentence -> shorter sentence)
    - Single document summarization (document -> shorter text)
    - Multi-document summarization (multiple documents -> text)

Dialog Response Generation

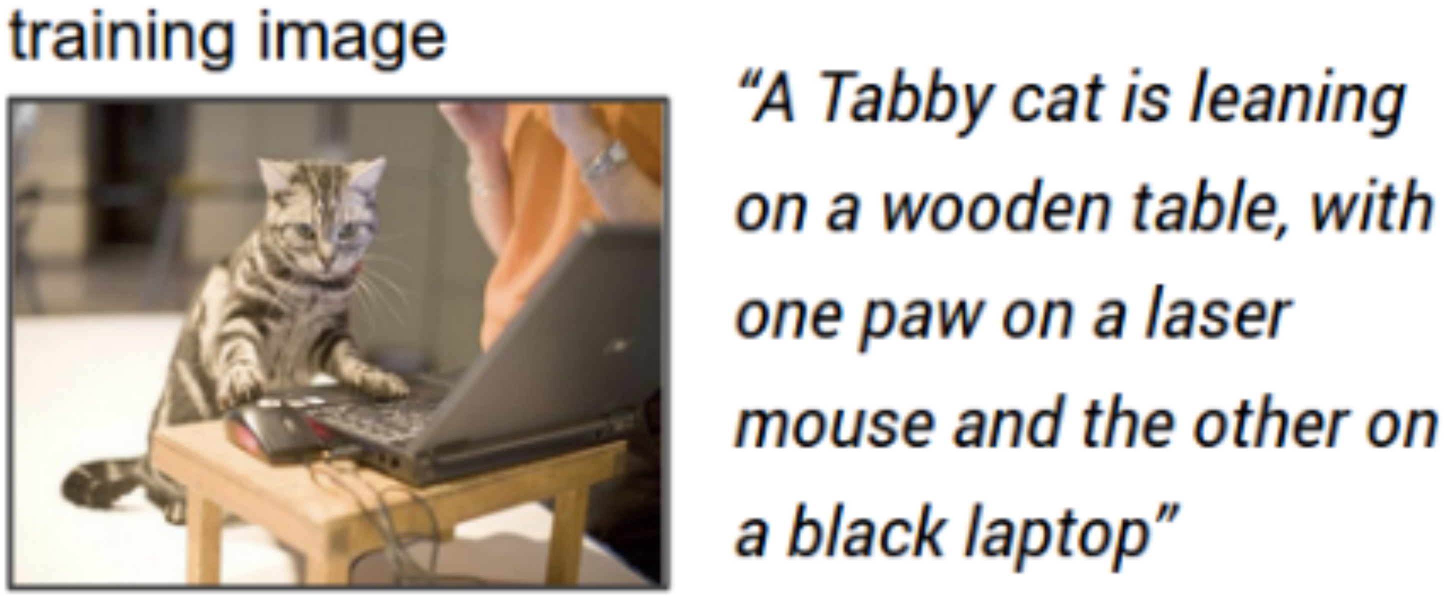
- Given an input utterance, generate a response
- e.g. Meena Chatbot (Adiwardana et al. 2020)

Human: Hi!
Meena: Hi there! How's it going?
Human: Pretty well.
Meena: I'm glad to hear it. :D
Human: What is your favorite book?
Meena: Oh, that's a toughie. Right now I'd have to say it's A Dance With Dragons, but Harry Potter is a close second.
Human: Oh I love Harry Potter
Meena: Me too! I've read the series at least five times over.
Human: Haha, same probably. Which one do you like the most
Meena: Probably Chamber of Secrets.
Human: Ah. I like Prisoner of azkaban best
Meena: Yeah, that's a good one too.
Human: Do you like the movies more or the books
Meena: I like the movies more. The books are good, don't get me wrong, but I like the pacing better in the movies.
Human: That is an unpopular opinion
Meena: I know. I'm the only one I know who prefers the movies.

Image Captioning

(e.g. Karpathy et al. 2015)

- Input is image features, output is text



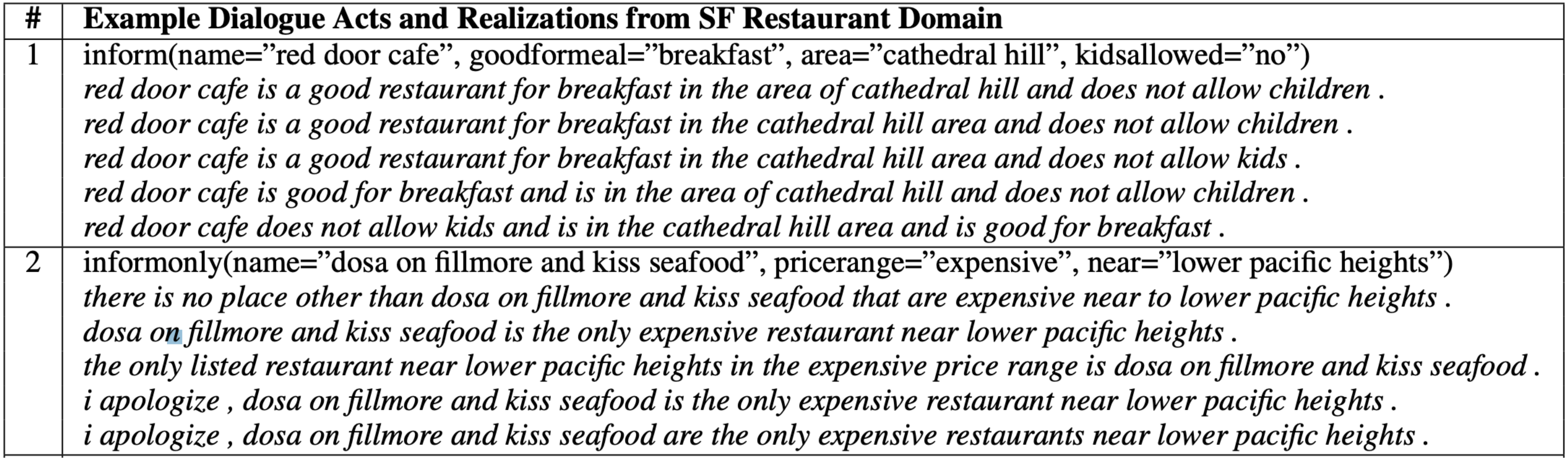
"A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop"

- Use standard image encoders (e.g. CNN, Transformers)
- Often pre-trained on large databases such as ImageNet

From Structured Data

(e.g. Wen et al 2015)

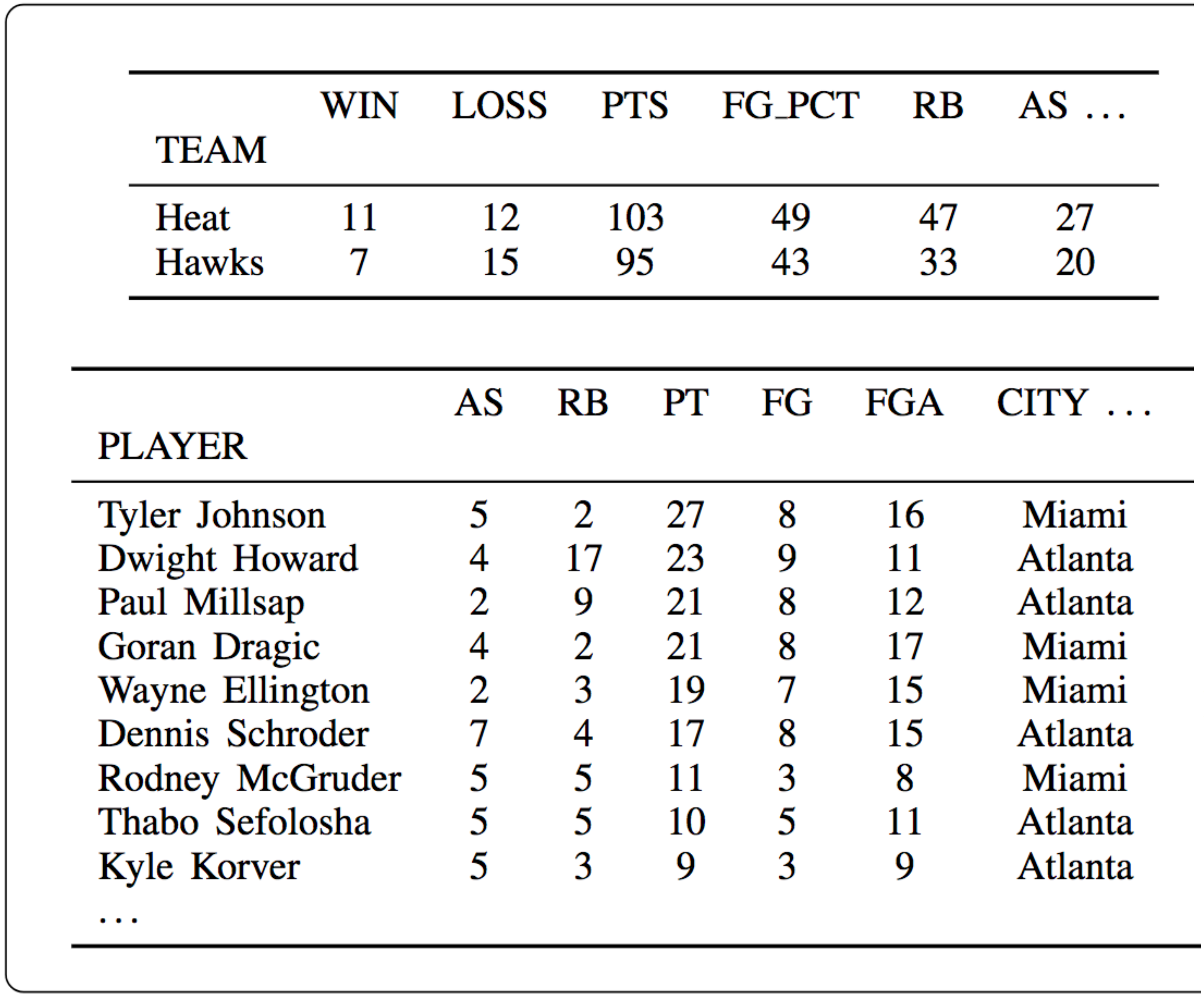
- When you say “Natural Language Generation” to an old-school NLPer, it means this



#	**Example Dialogue Acts and Realizations from SF Restaurant Domain**
1	inform(name=”red door cafe”, goodformeal=”breakfast”, area=”cathedral hill”, kidsallowed=”no”) *red door cafe is a good restaurant for breakfast in the area of cathedral hill and does not allow children .* *red door cafe is a good restaurant for breakfast in the cathedral hill area and does not allow children .* *red door cafe is a good restaurant for breakfast in the cathedral hill area and does not allow kids .* *red door cafe is good for breakfast and is in the area of cathedral hill and does not allow children .* *red door cafe does not allow kids and is in the cathedral hill area and is good for breakfast .*
2	informonly(name=”dosa on fillmore and kiss seafood”, pricerange=”expensive”, near=”lower pacific heights”) *there is no place other than dosa on fillmore and kiss seafood that are expensive near to lower pacific heights .* *dosa on fillmore and kiss seafood is the only expensive restaurant near lower pacific heights .* *the only listed restaurant near lower pacific heights in the expensive price range is dosa on fillmore and kiss seafood .* *i apologize , dosa on fillmore and kiss seafood is the only expensive restaurant near lower pacific heights .* *i apologize , dosa on fillmore and kiss seafood are the only expensive restaurants near lower pacific heights .*

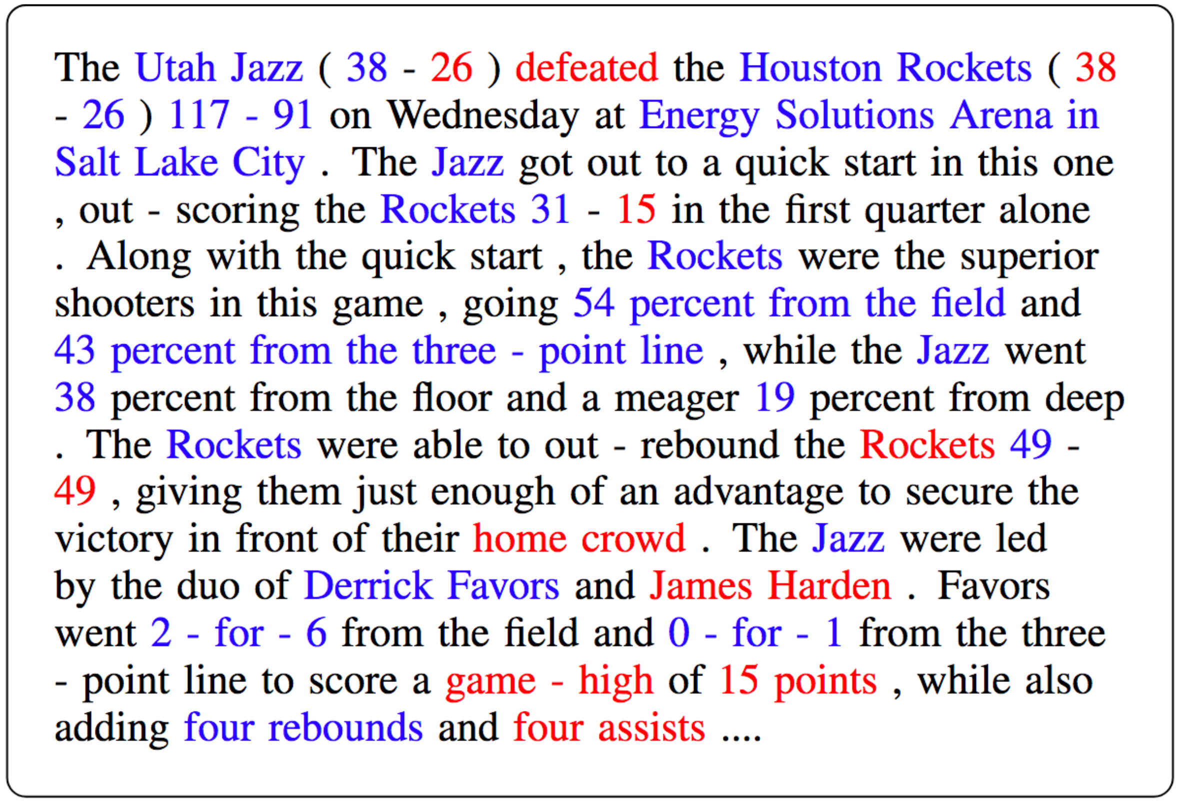
Still a Difficult Problem!

- e.g. "Challenges in data-to-document generation" (Wiseman et al. 2017)



TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS . . .
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

PLAYER	AS	RB	PT	FG	FGA	CITY . . .
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
. . .						



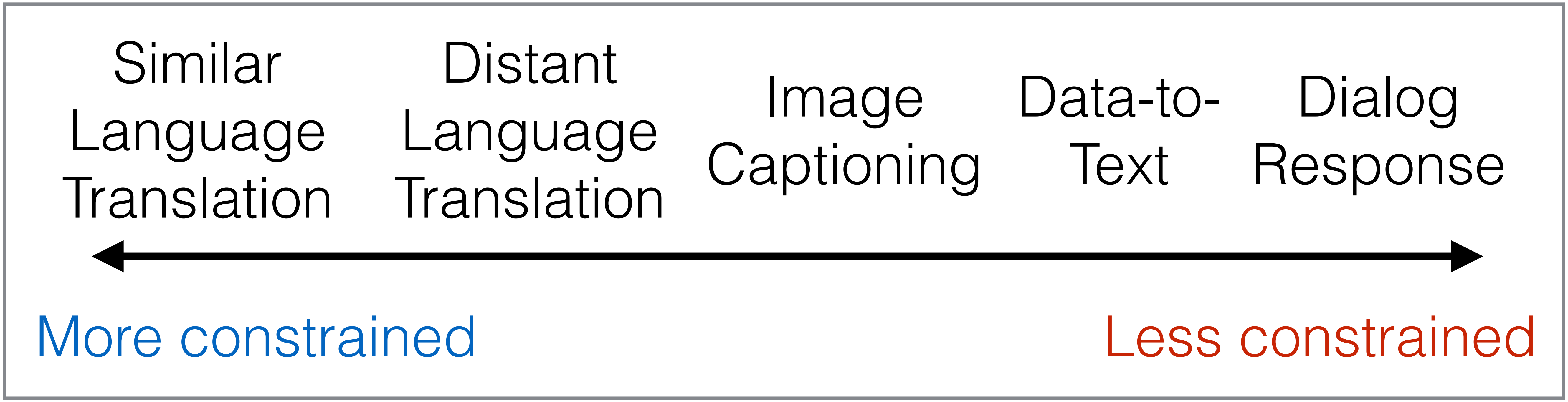
The Utah Jazz (38 - 26) defeated the Houston Rockets (38 - 26) 117 - 91 on Wednesday at Energy Solutions Arena in Salt Lake City . The Jazz got out to a quick start in this one , out - scoring the Rockets 31 - 15 in the first quarter alone . Along with the quick start , the Rockets were the superior shooters in this game , going 54 percent from the field and 43 percent from the three - point line , while the Jazz went 38 percent from the floor and a meager 19 percent from deep . The Rockets were able to out - rebound the Rockets 49 - 49 , giving them just enough of an advantage to secure the victory in front of their home crowd . The Jazz were led by the duo of Derrick Favors and James Harden . Favors went 2 - for - 6 from the field and 0 - for - 1 from the three - point line to score a game - high of 15 points , while also adding four rebounds and four assists ....

Figure 2: Example document generated by the Conditional Copy system with a beam of size 5. Text that accurately reflects a record in the associated box- or line-score is highlighted in blue, and erroneous text is highlighted in red.

- Focused evaluation using, e.g. information extraction

Level of Constraint on Output

- Given the conditioning, the outputs can be more or less constrained, very rough approximation below



Similar Language Translation

Distant Language Translation

Image Captioning

Data-to-Text

Dialog Response

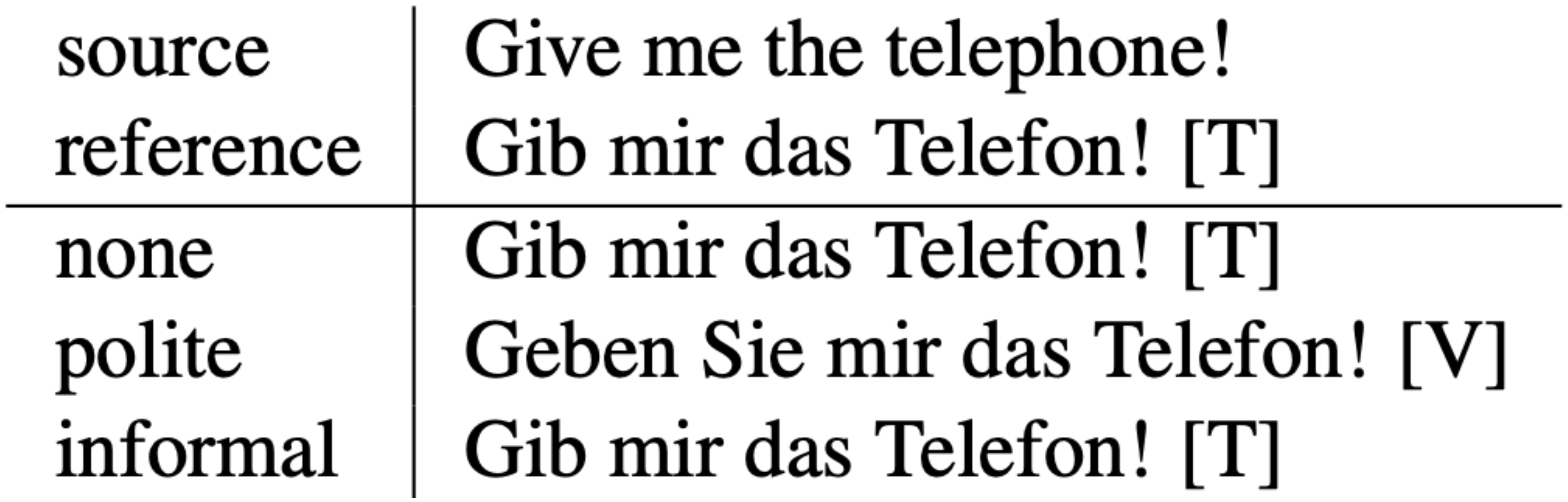
More constrained

Less constrained

- More freedom = more flexibility, but often more difficulty in modeling and evaluation

Controlled Generation

- Add a further constraint in addition to content-based ones
- **Politeness/Style Control:** Take an input *X* and a label indicating style, etc. (e.g. Sennrich et al. 2016)



source	Give me the telephone!
reference	Gib mir das Telefon! [T]
none	Gib mir das Telefon! [T]
polite	Geben Sie mir das Telefon! [V]
informal	Gib mir das Telefon! [T]

- **Personalization:** Take an input *X* and side information about the speaker (e.g. Hoang et al. 2016)
- etc. etc.

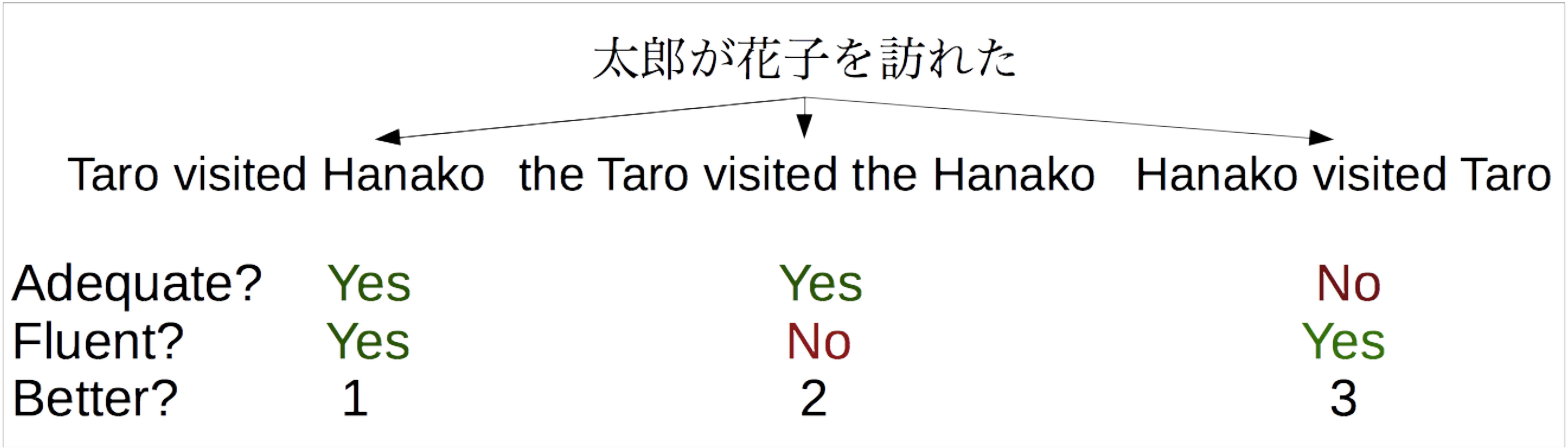
How do we Evaluate?

Basic Evaluation Paradigm

- Use parallel test set
  - *Unlike classification, may have multiple reference outputs per input*
- Use system to generate translations
- Compare target translations w/ reference
  - *Comparison typically harder than in classification*

Human Evaluation

- Ask a human to do evaluation



太郎が花子を訪れた

	Taro visited Hanako	the Taro visited the Hanako	Hanako visited Taro
Adequate?	Yes	Yes	No
Fluent?	Yes	No	Yes
Better?	1	2	3

- Final goal, but slow, expensive, and sometimes inconsistent

Human Evaluation Shared Tasks

- **Machine Translation**

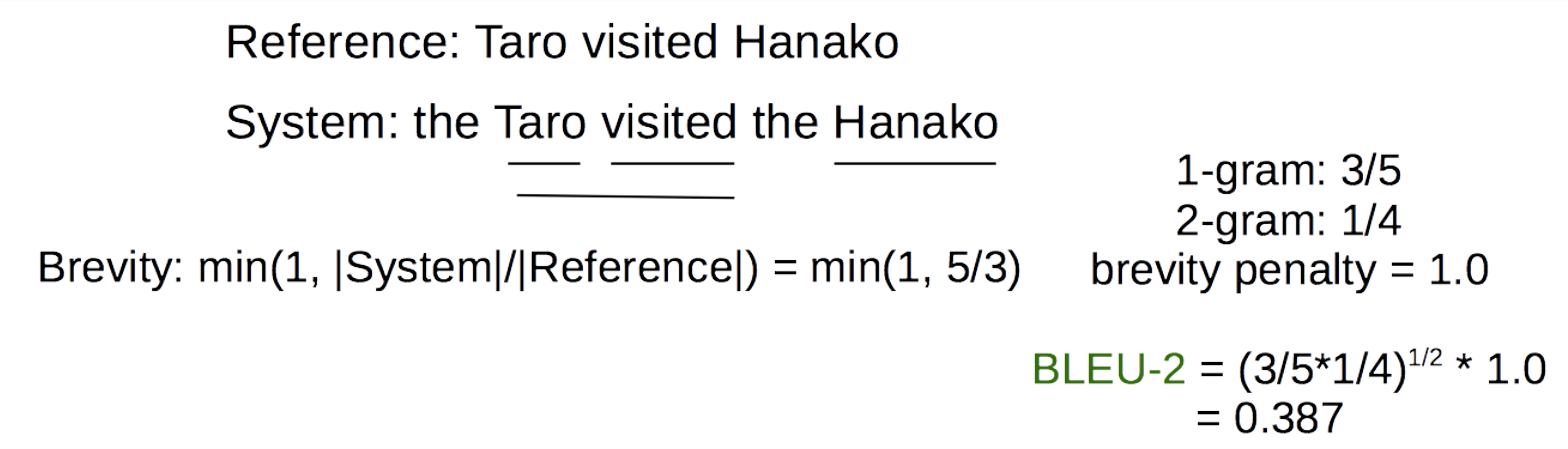
  - Conference on Machine Translation (WMT) shared tasks
    http://www.statmt.org/wmt20/

- **Composite Leaderboard**

  - GENIE leaderboard for QA, summarization, MT
    https://genie.apps.allenai.org/

BLEU

- Works by comparing n-gram overlap w/ reference



Reference: Taro visited Hanako

System: the Taro visited the Hanako

1-gram: 3/5
2-gram: 1/4

Brevity: min(1, |System|/|Reference|) = min(1, 5/3)

brevity penalty = 1.0

BLEU-2 = $(3/5*1/4)^{1/2} * 1.0$
$= 0.387$

- **Pros:** Easy to use, good for measuring system improvement
- **Cons:** Often doesn't match human eval, bad for comparing very different systems

Embedding-based Metrics

- Recently, many metrics based on neural models
  - **BertScore:** Find similarity between BERT embeddings (unsupervised) (Zhang et al. 2020)
  - **BLEURT:** Train BERT to predict human evaluation scores (Sellam et al. 2020)
  - **COMET:** Train model to predict human eval, also using source sentence (Rei et al. 2020)
  - **PRISM:** Model based on training paraphrasing model (Thompson and Post 2020)
  - **BARTScore:** Calculate the probability of source, reference, or system output (Yuan et al. 2021)

Perplexity

- Calculate the perplexity of the words in the held-out set *without* doing generation

- **Pros:** Naturally solves multiple-reference problem!

- **Cons:** Doesn’t consider decoding or actually generating output.

- May be reasonable for problems with lots of ambiguity.

Which One to Use?

- **Meta-evaluation** runs human evaluation and automatic evaluation on the same outputs, calculates correlation
- Examples:
  - **WMT Metrics Task** for MT (Mathur et al. 2021)
  - **RealSumm** for summarization (Bhandari et al. 2020)
- Evaluation is hard, especially with good systems! Most metrics had no correlation w/ human eval over best systems at some WMT 2019 tasks

Revisiting Inference

Limitations of Search

- If your underlying model is bad, finding a *better scoring hypothesis* can equal *worse generations*!
- *Search errors* can hide *model errors*

e.g. in machine translation, more search leads to short hypotheses (Stahlberg and Byrne 2019)

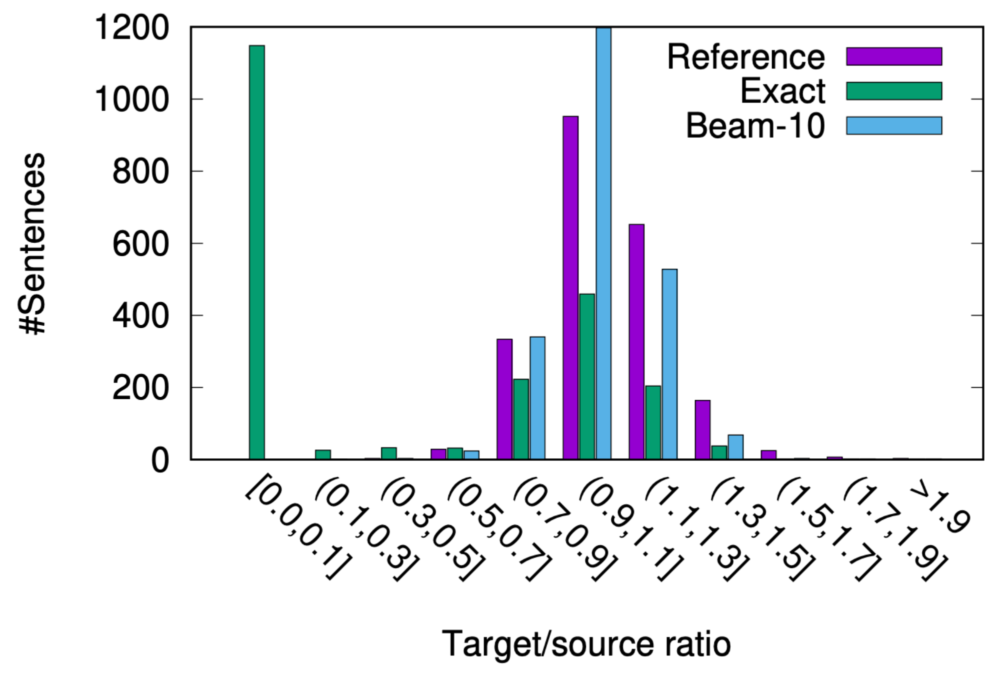


Figure 3: Histogram over target/source length ratios.

e.g. in open-ended generation, search leads to repetition (Holtzman et al. 2019)

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Limitations of Sampling

- Neural LMs that use a softmax assign non-zero probability to every word!

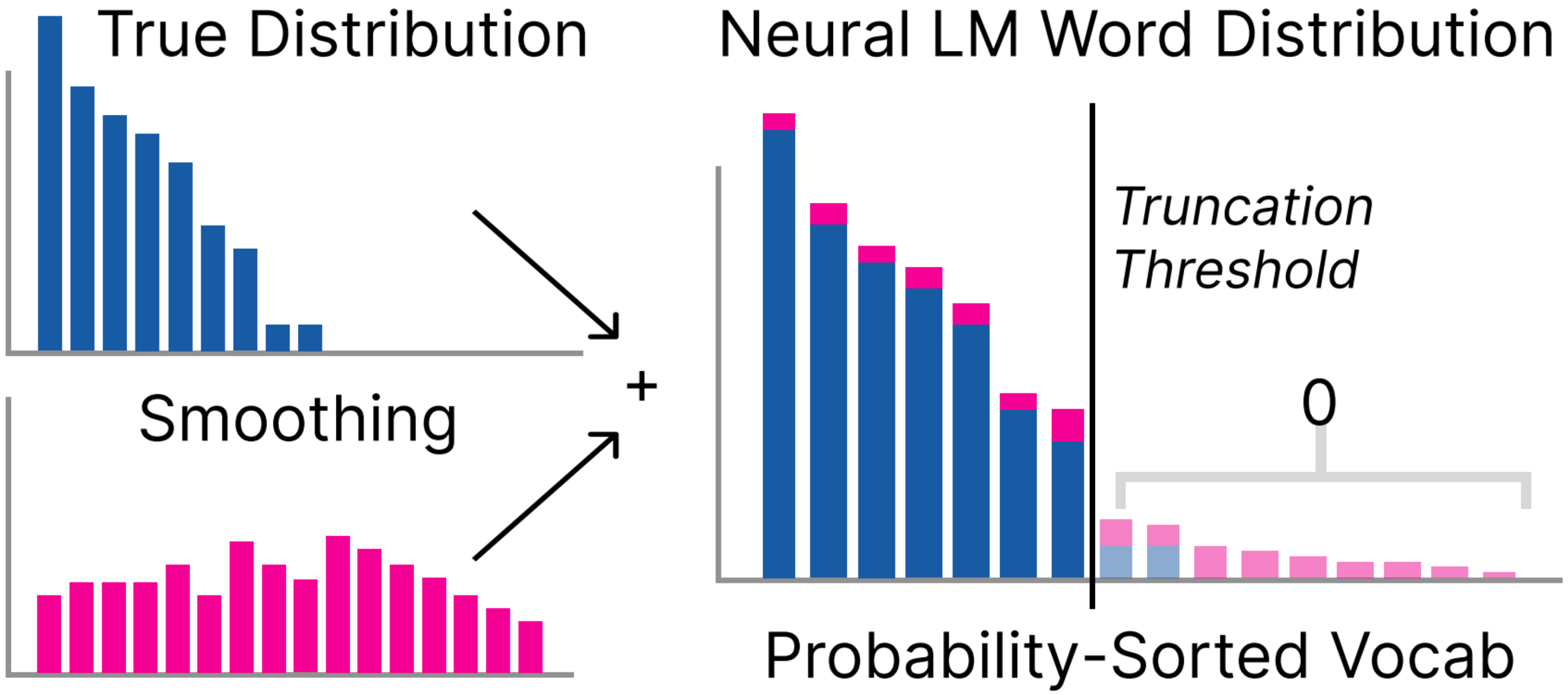


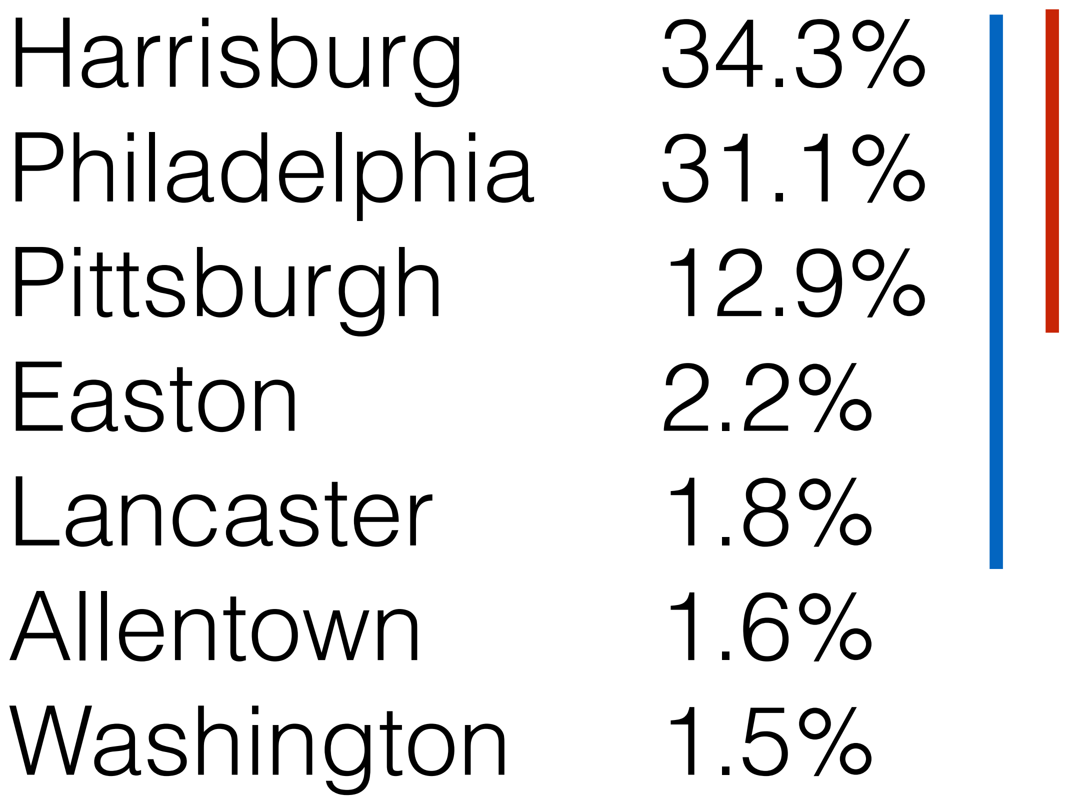
Figure 1: A neural LM as a mixture of the true distribution, and a uniform-like smoothing distribution. Truncation aims to approximate the true distribution support.

Hewitt et al. 2022.
Truncation Sampling as Language Model Desmoothing

Alternative 1:
Sample from a Truncated Distribution

- Remove the lowest-probability words at each time step.

$P(x_6 \mid \text{“The capital of Pennsylvania is”})$



Harrisburg	34.3%
Philadelphia	31.1%
Pittsburgh	12.9%
Easton	2.2%
Lancaster	1.8%
Allentown	1.6%
Washington	1.5%

Top-k Sampling
(e.g. k=5)
Fan et al. 2018

Nucleus (top-p) Sampling
(e.g. p=0.8)
Holtzmann et al. 2019

Alternative 2: Better Decision Rule

- minimum Bayes risk (e.g. Fernandes et al. 2022)

$$\mathrm{BayesRisk}(y|x) = \sum_{\tilde{y}} P(\tilde{y}|x)\mathrm{Error}(y, \tilde{y}) \qquad \hat{y} = \underset{y}{\mathrm{argmin}}\ \mathrm{BayesRisk}(y|x)$$

P(y | “What is your name”)	I don’t know	20.1%
	My name is Jane	10.4%
	My name is John	9.2%
	My name is Robert	8.3%

- Common method:
  - generate list of n candidates (using beam search or sampling)
  - rescore list of candidates

Toolkit: https://github.com/deep-spin/qaware-decode

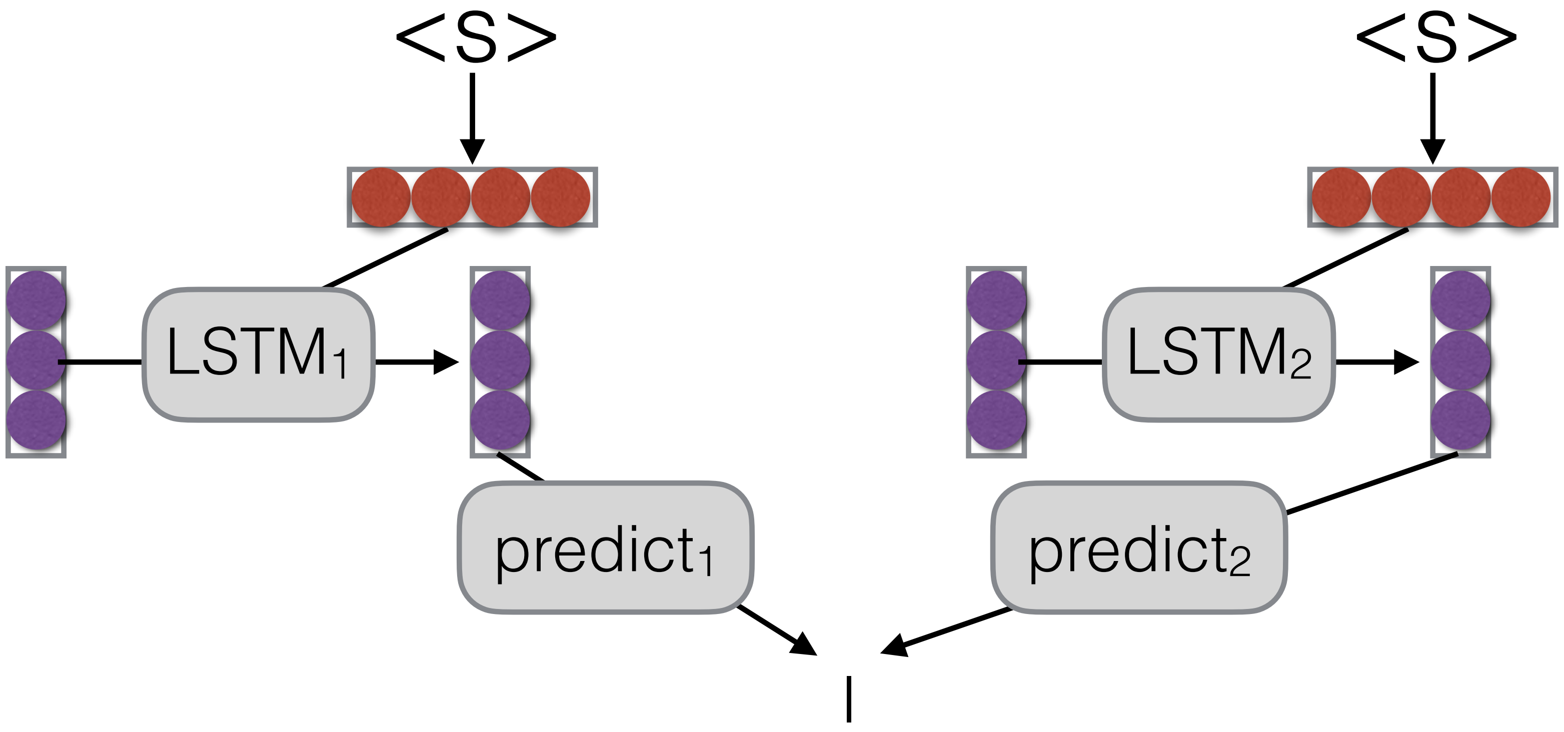
Alternative 3: Train Better Model!

- Your problems are because your model is scoring bad hypotheses highly, so fix it!
- Methods:
  - **Minimum risk training** (e.g. through reinforcement learning, enumeration)
  - **Margin-based training** (e.g. through ranking, “contrastive learning”)
- More in later classes

An Aside:
Model Ensembling

Ensembling

- Combine predictions from multiple models



- Why?
  - Multiple models make somewhat uncorrelated errors
  - Models tend to be more uncertain when they are about to make errors
  - Smooths over idiosyncrasies of the model

Linear Interpolation

- Take a weighted average of the M model probabilities

$$P(y_j \mid X, y_1, \ldots, y_{j-1}) =$$

$$\sum_{m=1}^{M} \underbrace{P_m(y_j \mid X, y_1, \ldots, y_{j-1})}_{\text{Probability according to model } m} \underbrace{P(m \mid X, y_1, \ldots, y_{j-1})}_{\text{Probability of model } m}$$

- Second term often set to uniform distribution 1/M

Log-linear Interpolation

- Weighted combination of log probabilities, normalize

$$P(y_j \mid X, y_1, \ldots, y_{j-1}) = \underbrace{\mathrm{softmax}}_{\text{Normalize}} \left(\sum_{m=1}^{M} \underbrace{\lambda_m(X, y_1, \ldots, y_{j-1})}_{\text{Interpolation coefficient for model } m} \underbrace{\log P_m(y_j \mid X, y_1, \ldots, y_{j-1})}_{\text{Log probability of model } m} \right)$$

- Interpolation coefficient often set to uniform distribution 1/M

Linear or Log Linear?

- Think of it in logic!
- **Linear:** “Logical OR”
  - the interpolated model likes any choice that a model gives a high probability
  - use with models that capture different traits
  - necessary when any model can assign zero probability
- **Log Linear:** “Logical AND”
  - interpolated model only likes choices where all models agree
  - use when you want to restrict possible answers

Parameter Averaging

(e.g. Bahar et al. 2017, Wortsman et al. 2022)

- **Problem:** Ensembling means we have to use *M* models at test time, increasing our time/memory complexity
- Parameter averaging is a cheap way to get some good effects of ensembling
- Basically, write out models several times near the end of training, and take the average of parameters to create a single model

Questions?