Natural Language Processing

Decoding Strategies and Evaluations for Natural Language Generation

Outline

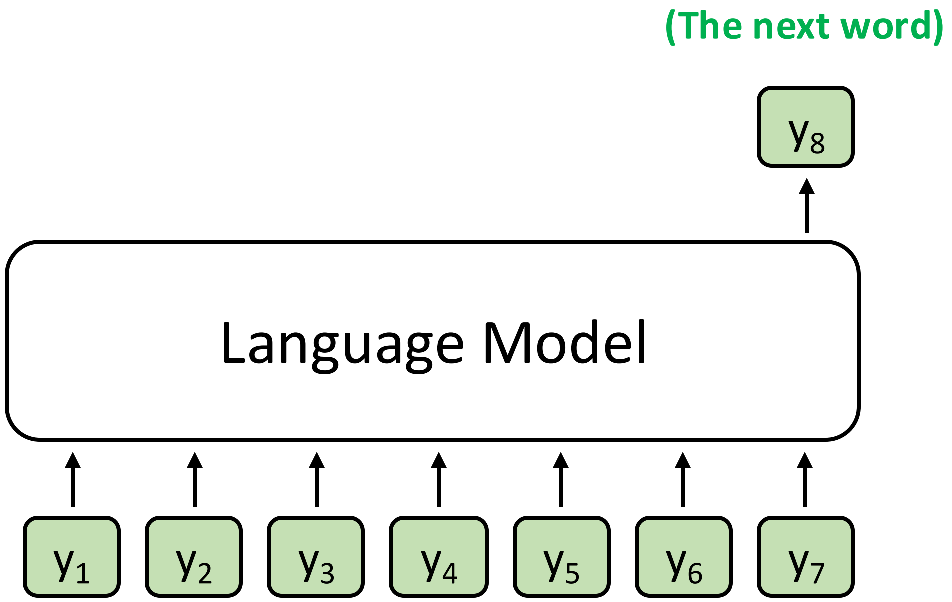
- Recap: Language Generation
- Decoding Strategies
  - Greedy Decoding
  - Beam Search
  - Top-k / Top-p Sampling
- Evaluations

Natural Language Generation (NLG)

- Natural language generation (NLG) is a **process** that **outputs** text.
- NLG includes a wide variety of NLP tasks.

Machine Translation	Abstractive Summarization	Dialogue Generation (e.g., ChatGPT)	Story Generation	Image Captioning	…

Recap: Language Model



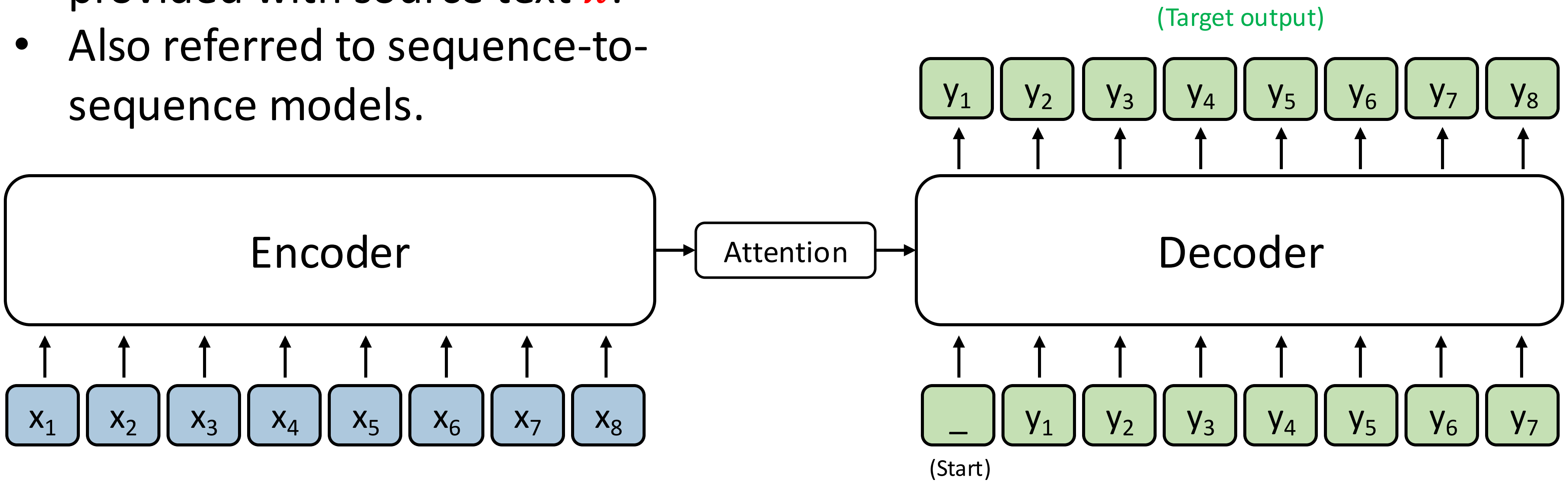
$$P(y_t|y_1, y_2, \dots, y_{t-1})$$

- A model that assigns probabilities to upcoming words is called **a language model**.
- The task involving predictions of upcoming words is **language modeling**.

Recap: Conditional Language Model

- In addition to previous words, a conditional language model is provided with source text x.
- Also referred to sequence-to-sequence models.

$$P(y_t | y_1, y_2, \dots, y_{t-1}, x)$$



Tasks of Conditional Language Model

- In addition to previous words (target), a conditional language model is provided with source text x.

	Source	**Target**
Machine Translation	Language A	Language B
Summarization	Long Text	Concise Text
Dialogue Generation	User Input	Desired User Input
…		

How to train a (Conditional) Language Model?

- First, you need a training (parallel) corpus.

Supervised, Aligned

Example: I love reading books.

Language modeling **(Unsupervised)**

I → love

I love → reading

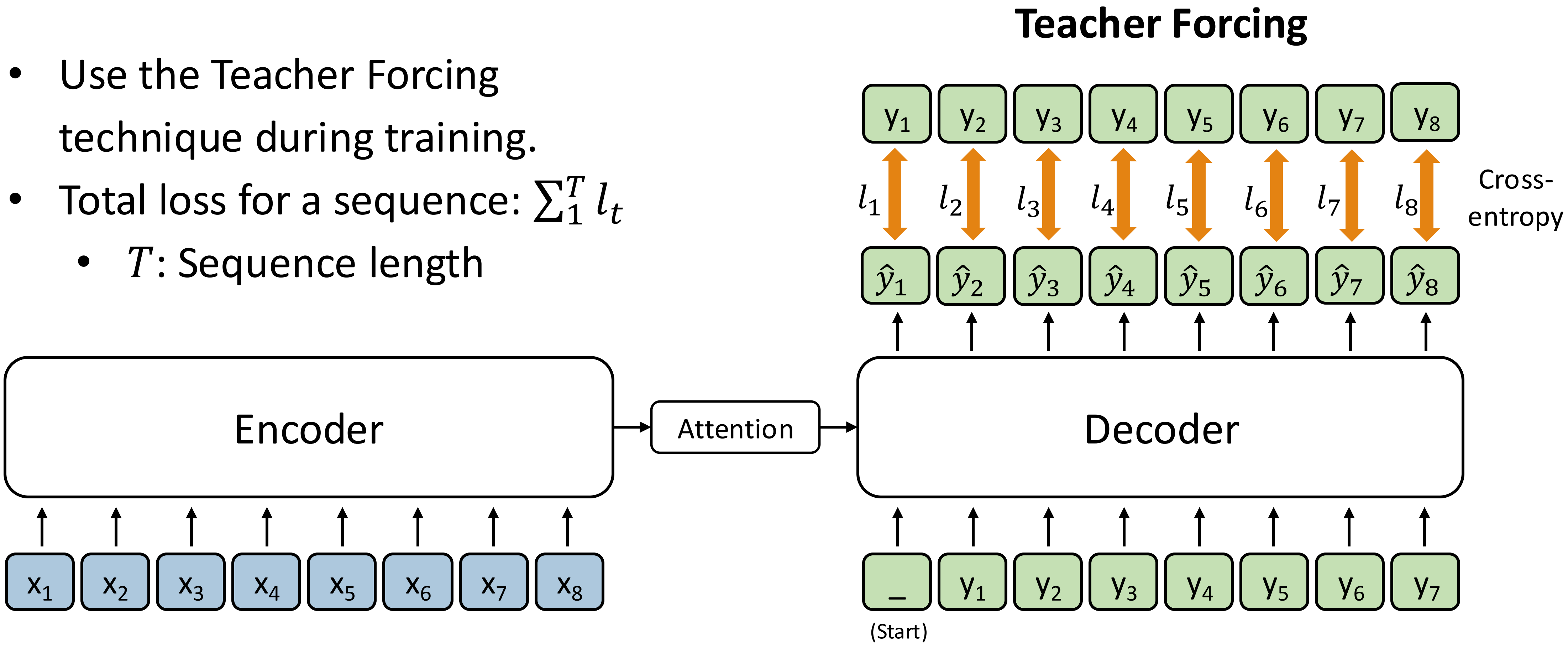
I love reading → books

Machine Translation **(Supervised)**

I	love	reading	books	**English**
我	愛	閱	讀	**Chinese**

How to train a (Conditional) Language Model?

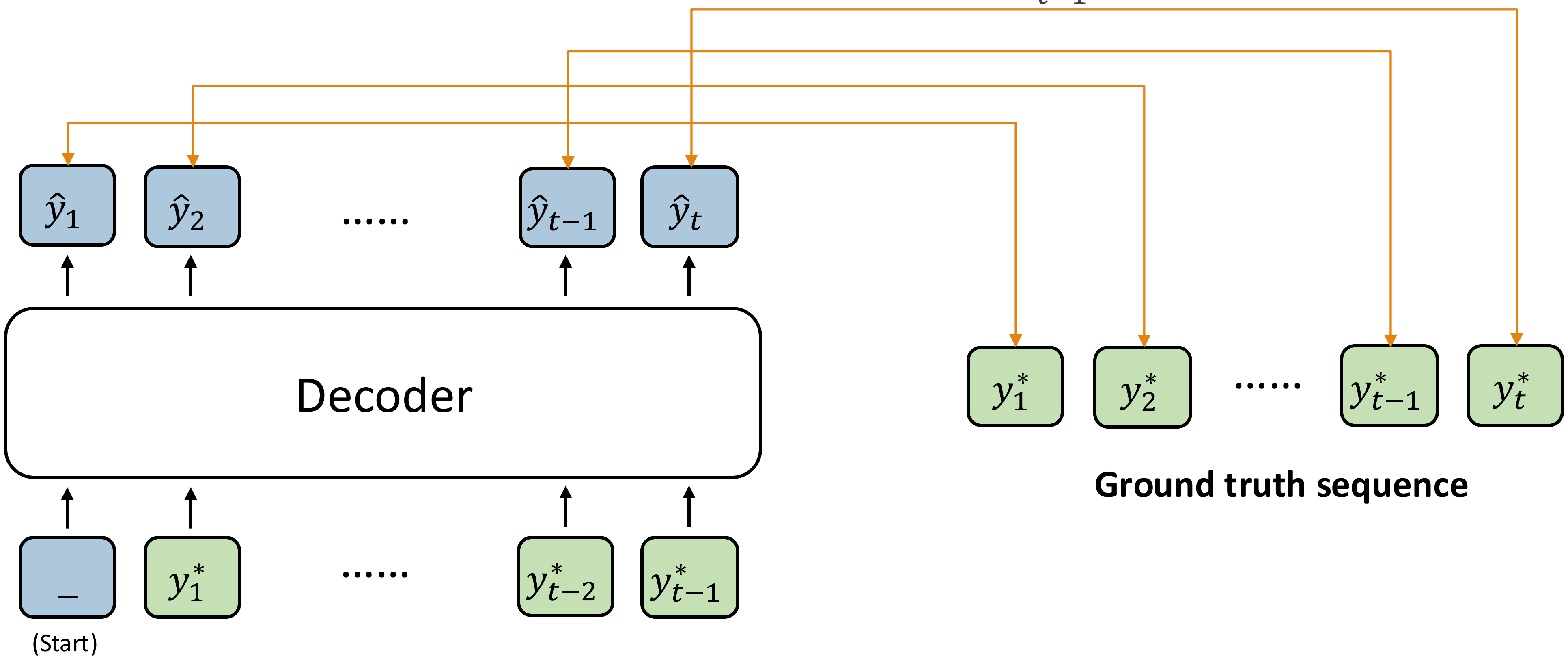
- Use the Teacher Forcing technique during training.
- Total loss for a sequence: $\sum_1^T l_t$
  - T: Sequence length



Teacher Forcing – Training stage

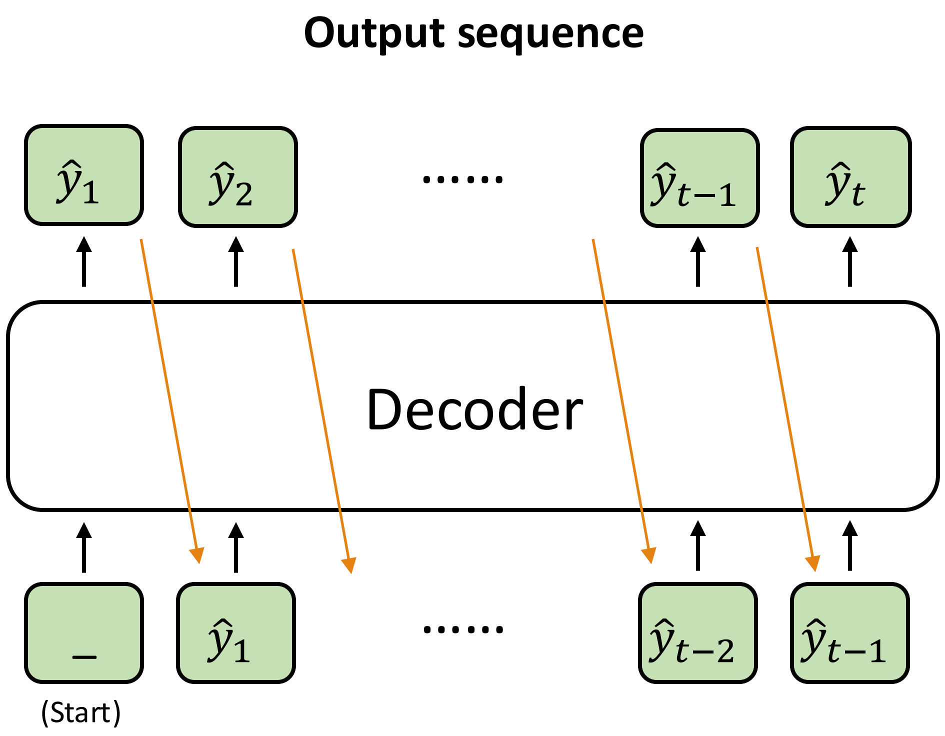
During training:

$$L_{ml} = -\sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \dots, y_{t-1}^*, x)$$



Teacher Forcing – Testing stage

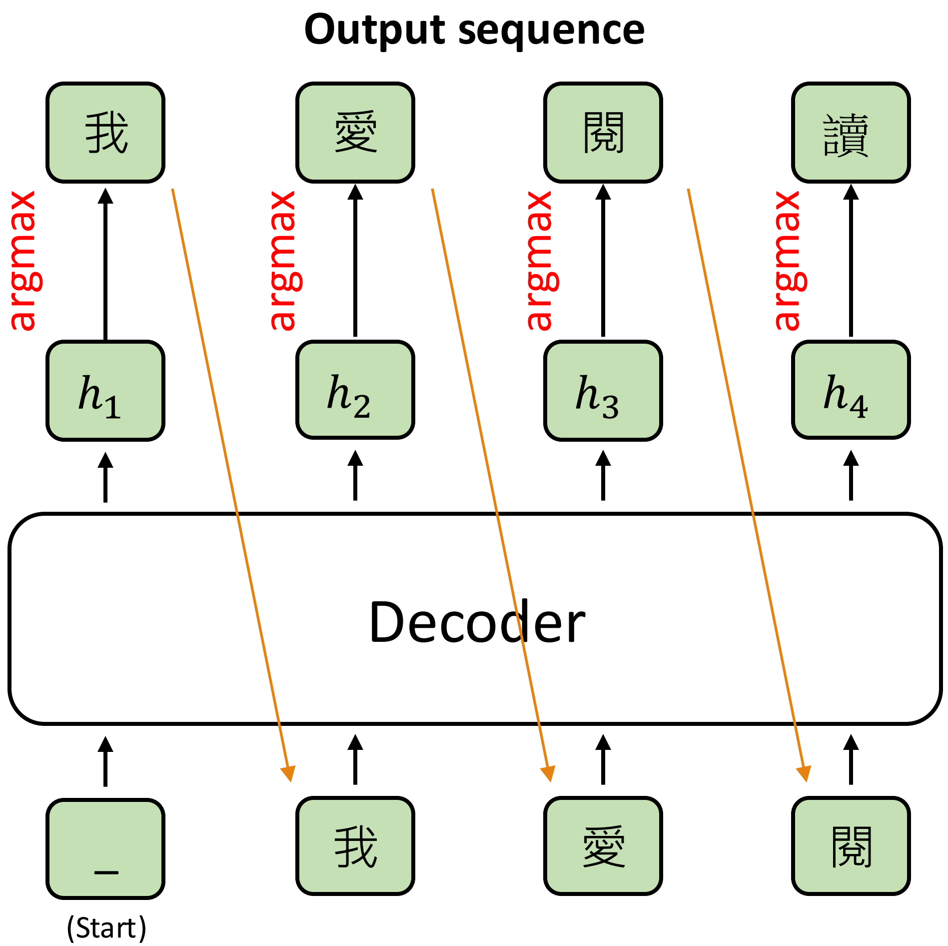
During testing:



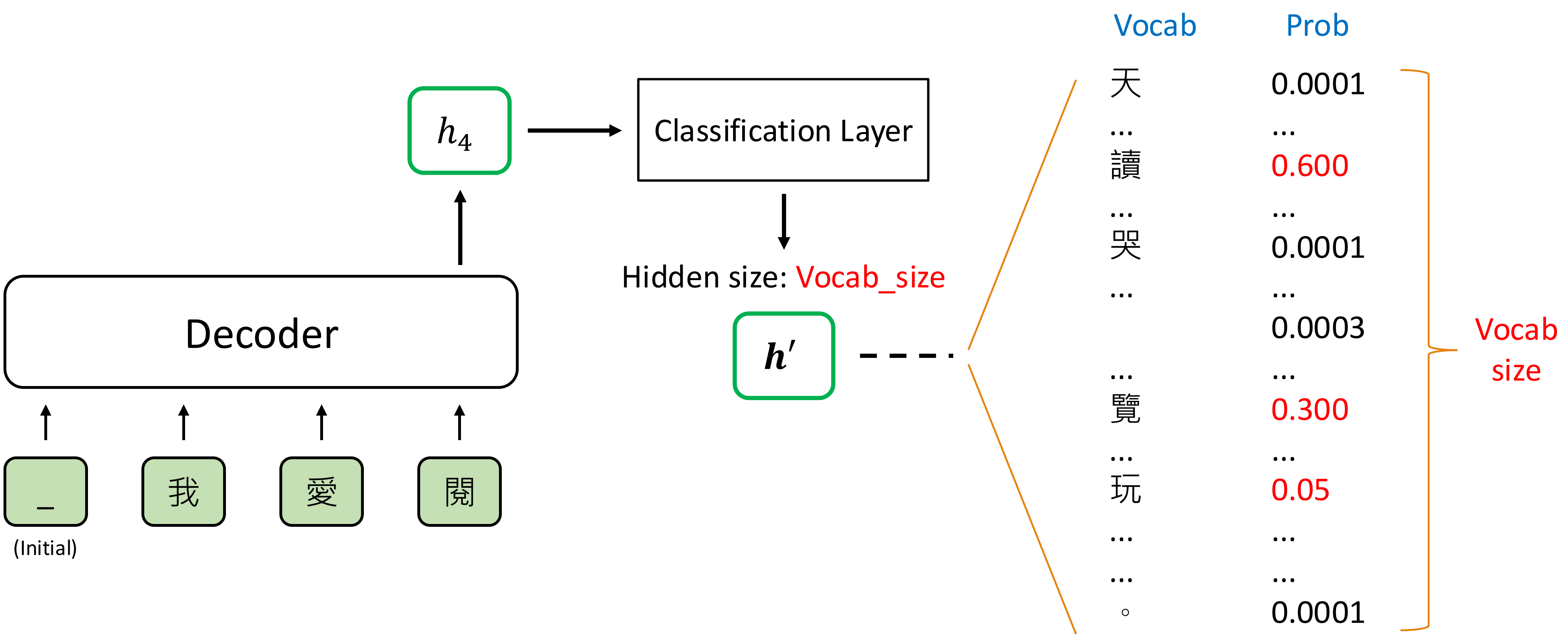
- Advantage: stabilize training and increase performance
- Question: How does the next word be determined?

Greedy Decoding

Example: I love reading books.



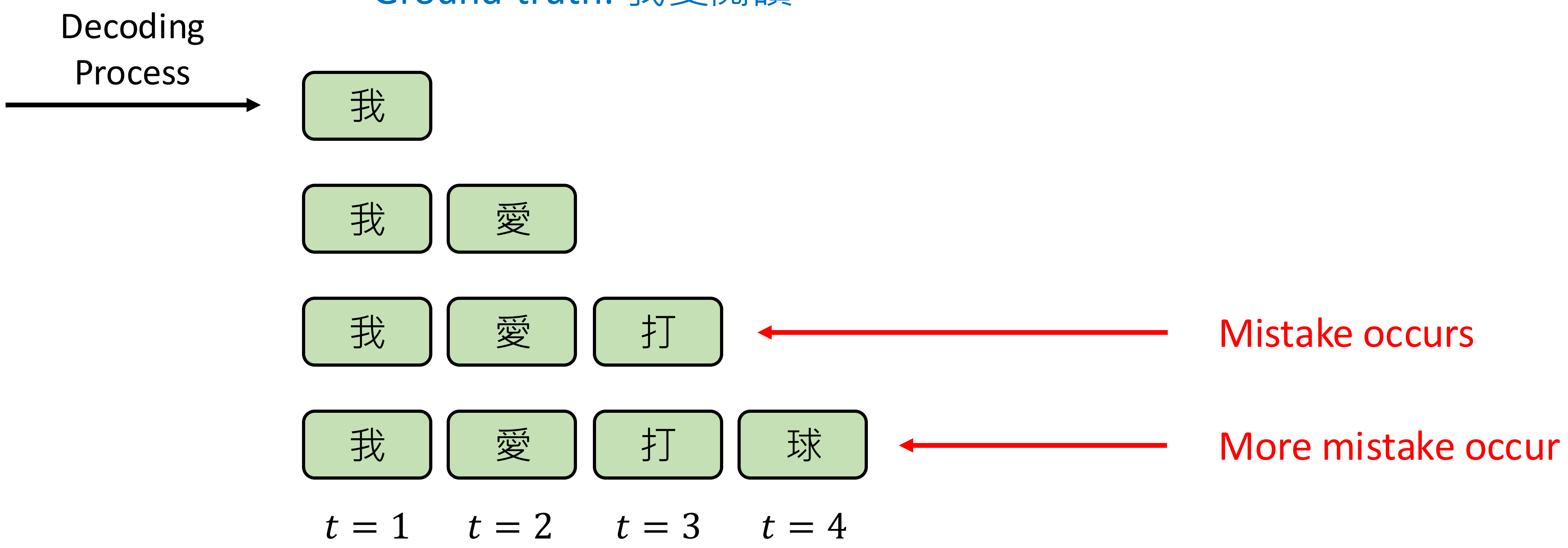
Greedy Decoding – Best Selection Process



Problem of Greedy Decoding

- Greedy decoding cannot undo!

Ground-truth: 我愛閱讀



Re-thinking Greedy Decoding

- Greedy decoding cannot undo!
- Greedy decoding only provides one best choice at each time step.
- How about providing more than one choice at each time step?
  - Stochastic Sampling (Top-k, Top-p sampling)
  - Deterministic Search → **Beam Search**
  - Top-p + Beam?

All applied on the last layer output (logits, e.g., 32k token probabilities)

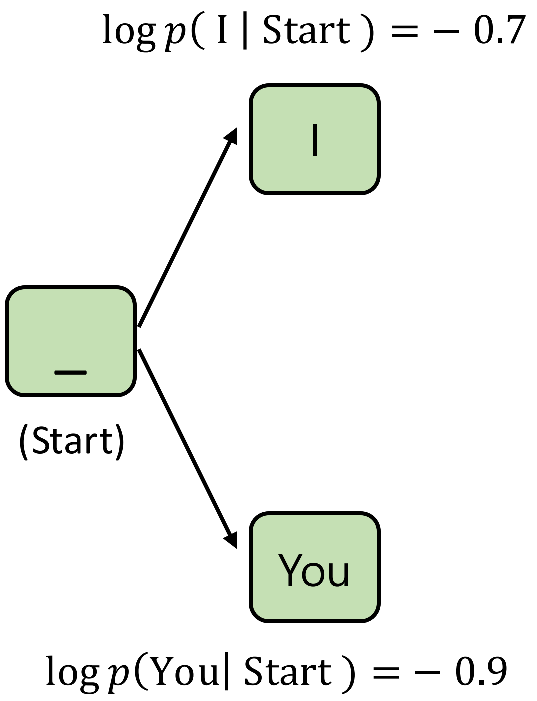
Beam Search

- Set the `Beam size` (or `Beam width`) = 2
  - This means that the number of candidates will be preserved at each decoding time.
  - Beam size is a hyperparameter for beam search decoding.
- At each decoding time step, a score is calculated via the following equation:

$$L_{ml} = \sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*, x)$$

Beam Search ($t = 1$)

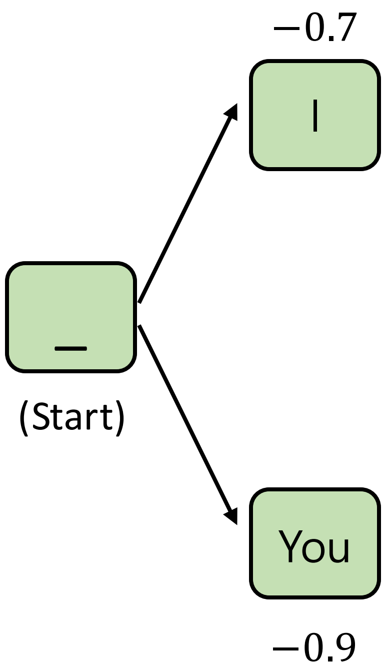
`Beam size` = 2



- At this decoding step, two choices are preserved.

Beam Search ($t = 1$)

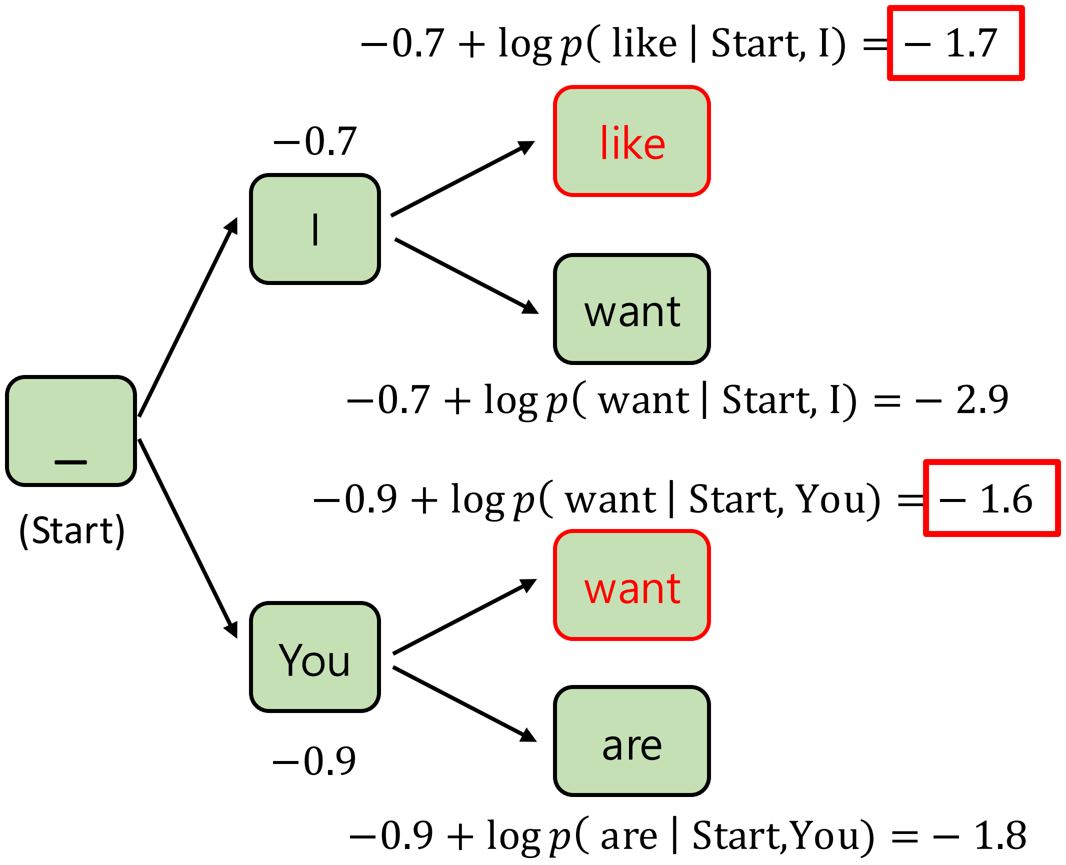
`Beam size` = 2



- At this decoding step, two choices are preserved.

Beam Search ($t = 2$)

`Beam size` = 2

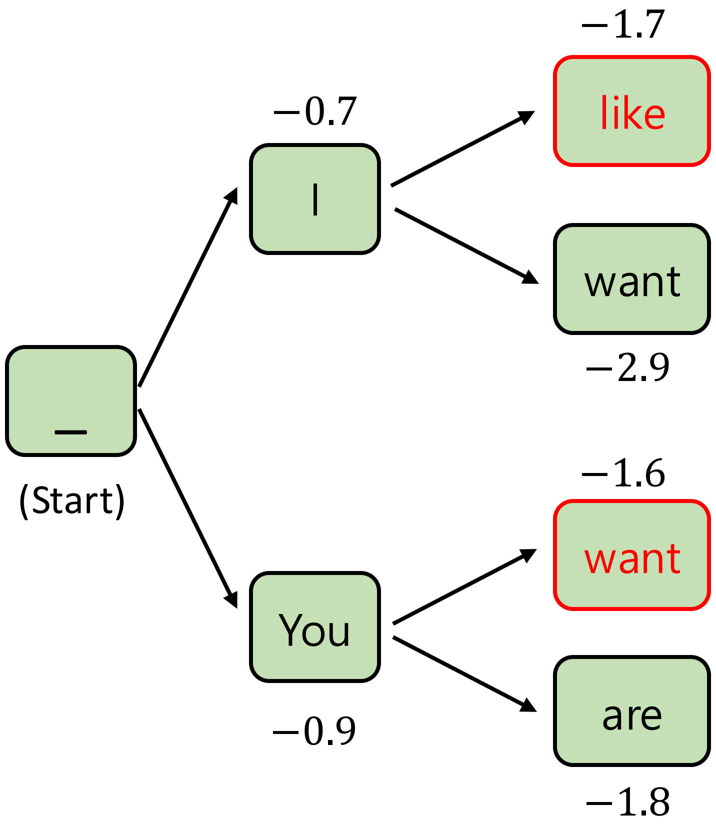


Note the loglikelihood! Being close to zero is better!

- At this decoding step, two choices are preserved, and the other two are discarded.

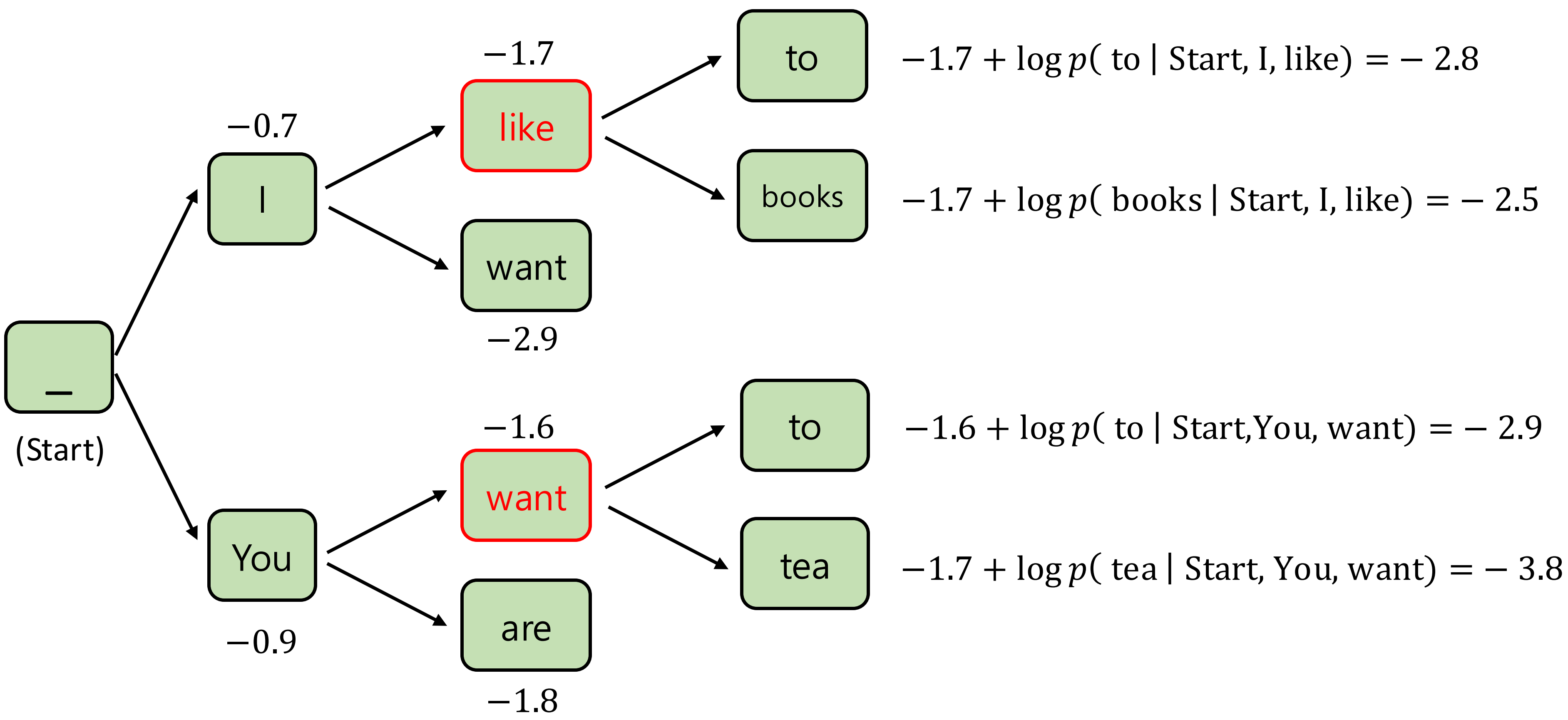
Beam Search ($t = 2$)

`Beam size` = 2



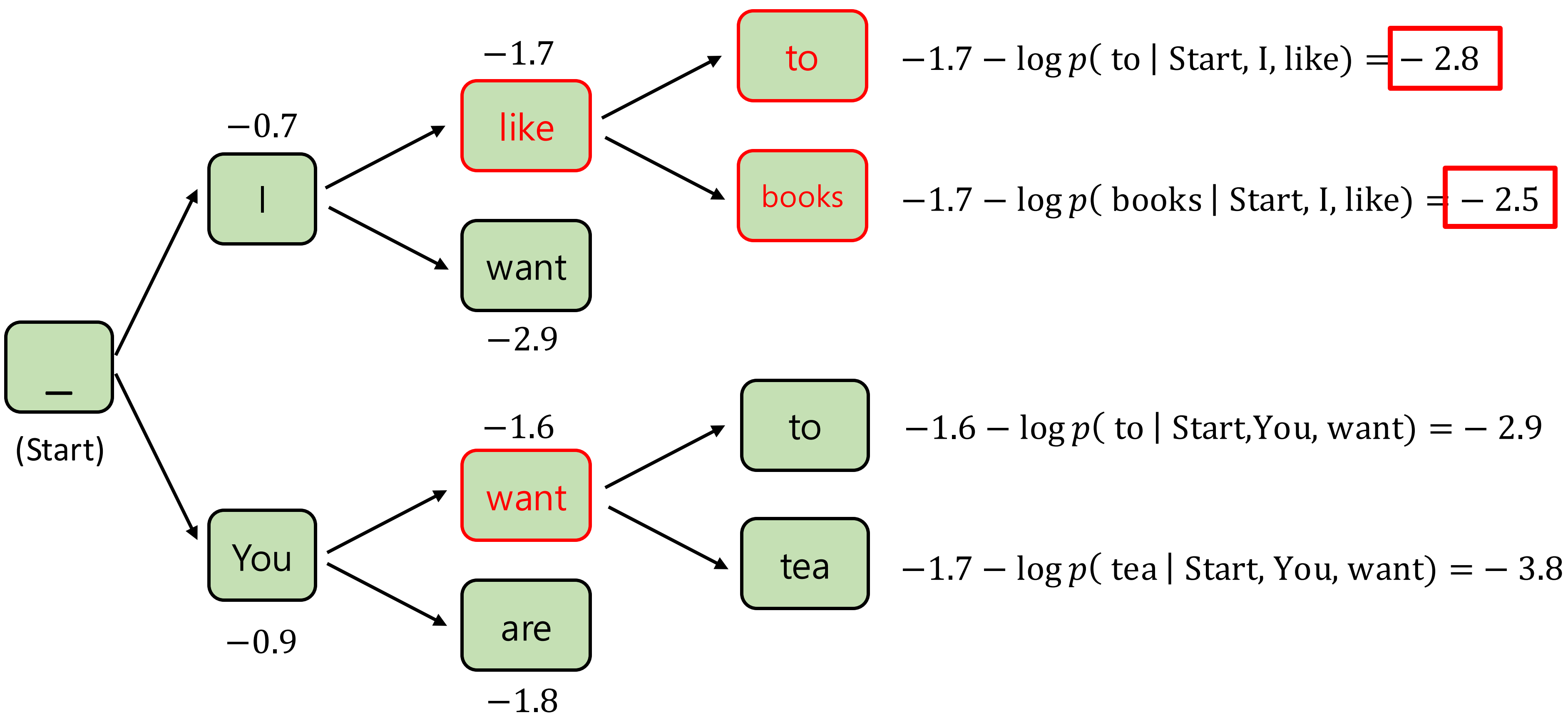
Beam Search ($t = 3$)

`Beam size` = 2



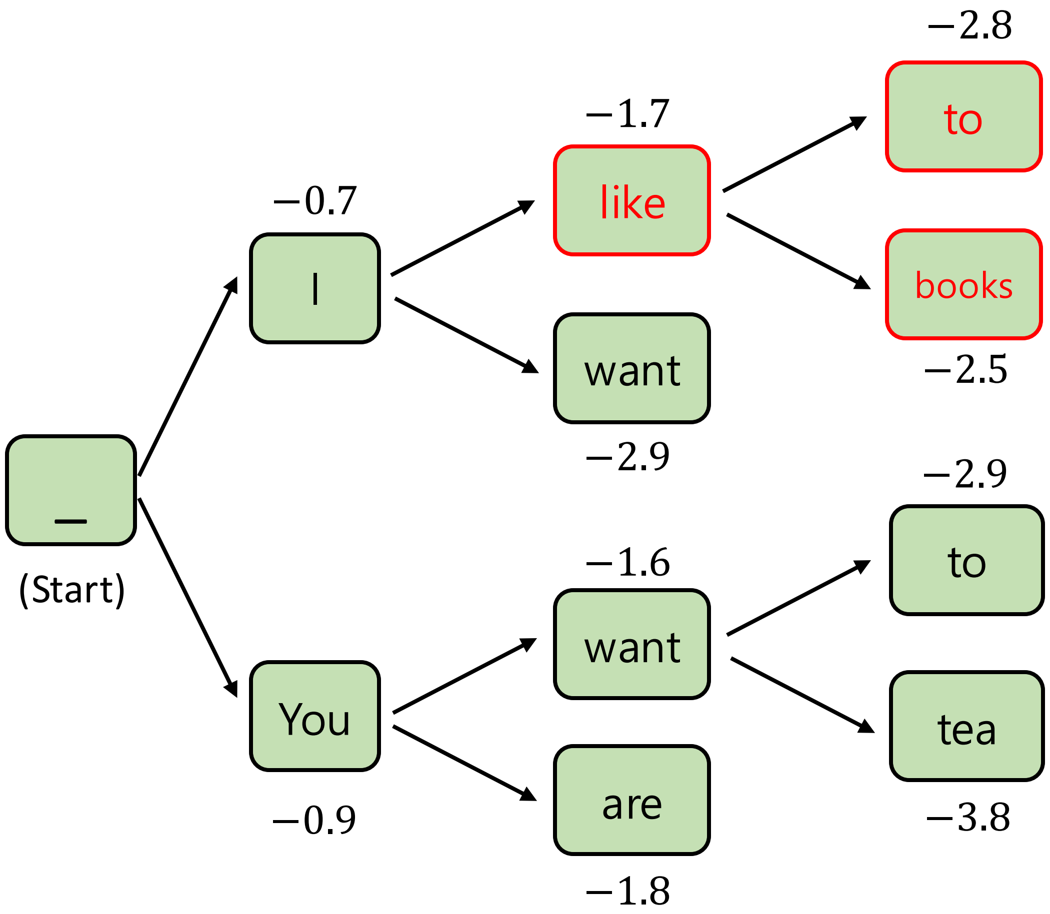
Beam Search ($t = 3$)

`Beam size` = 2



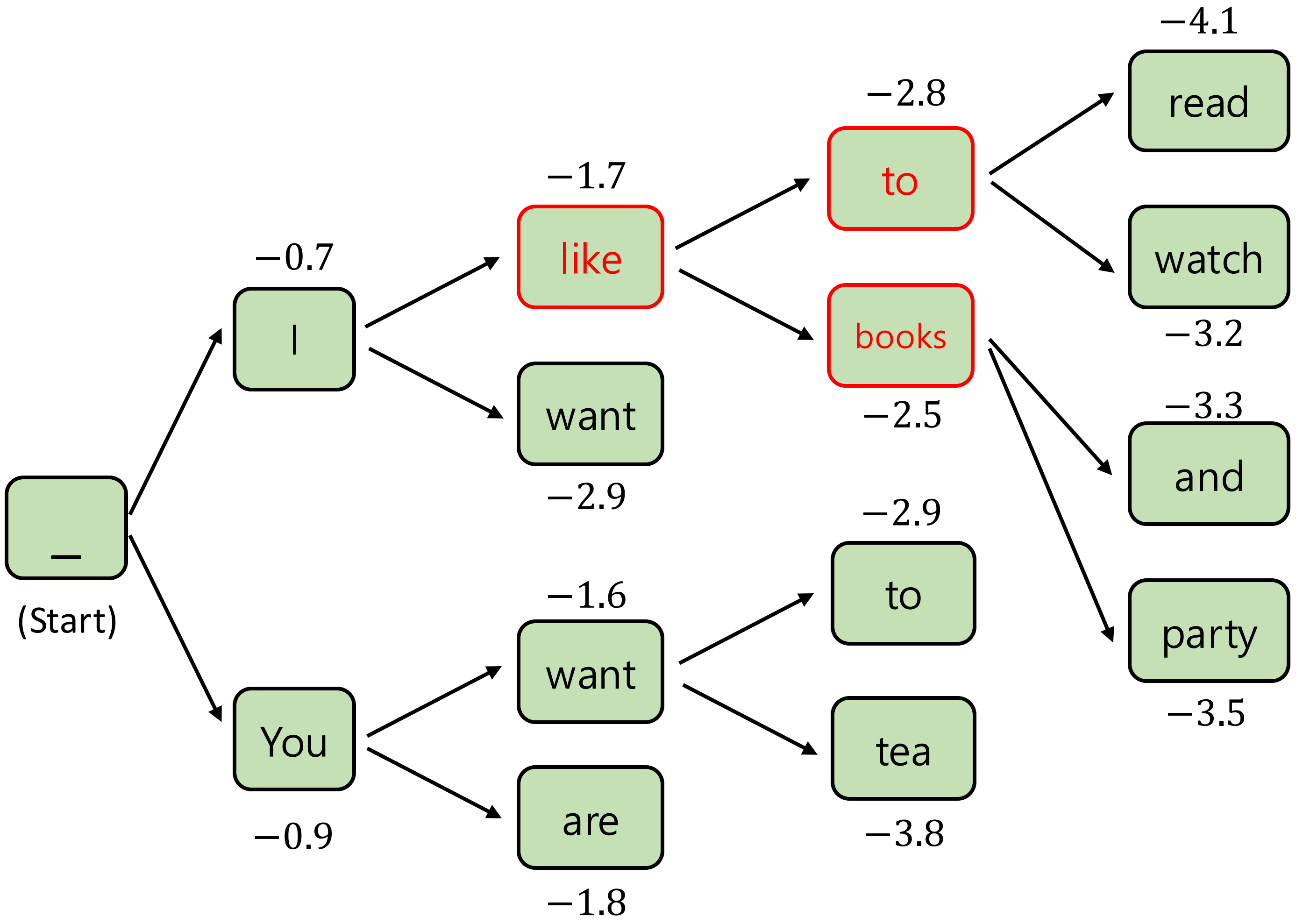
Beam Search ($t = 3$)

`Beam size` = 2



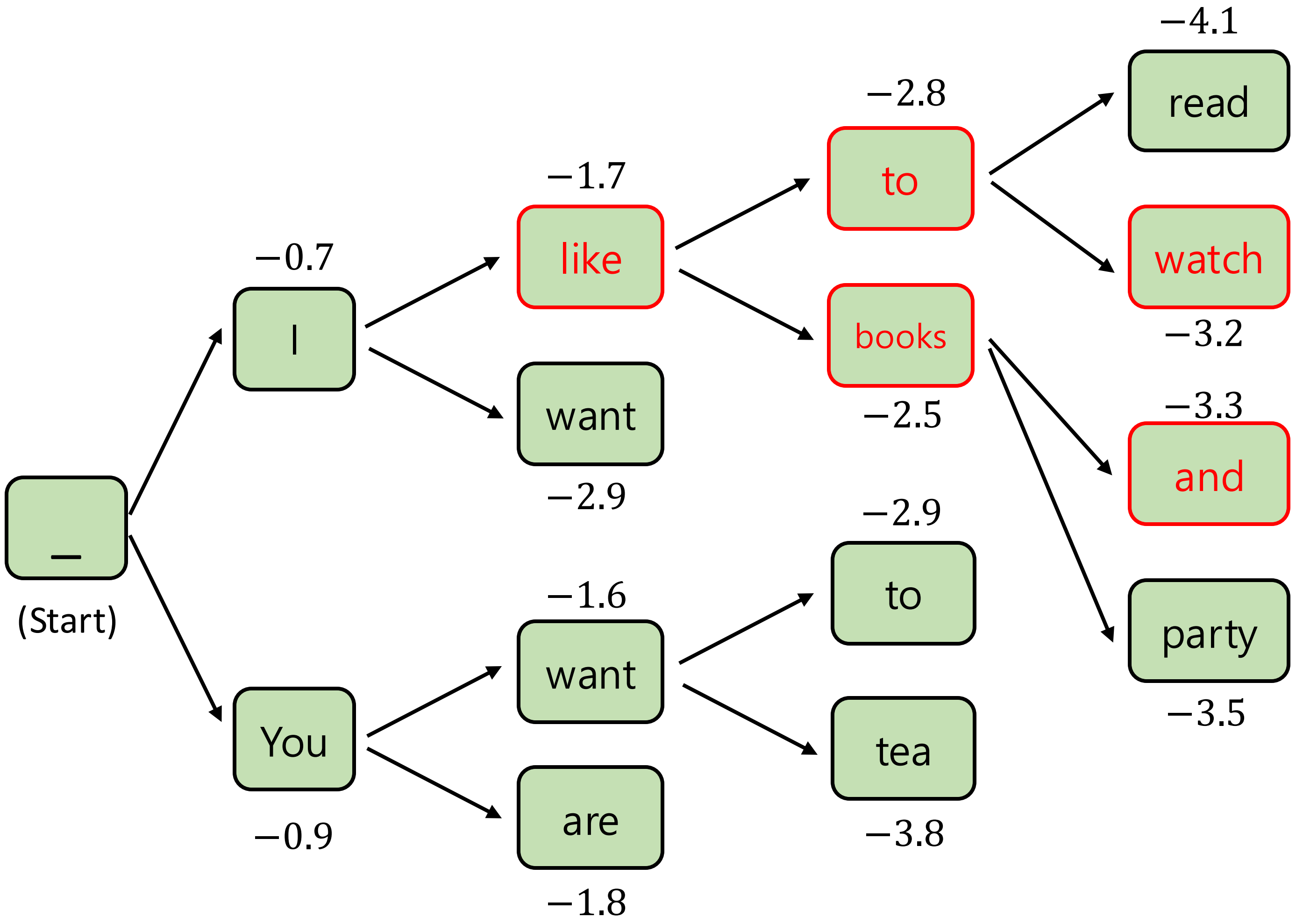
Beam Search ($t = 4$)

`Beam size` = 2



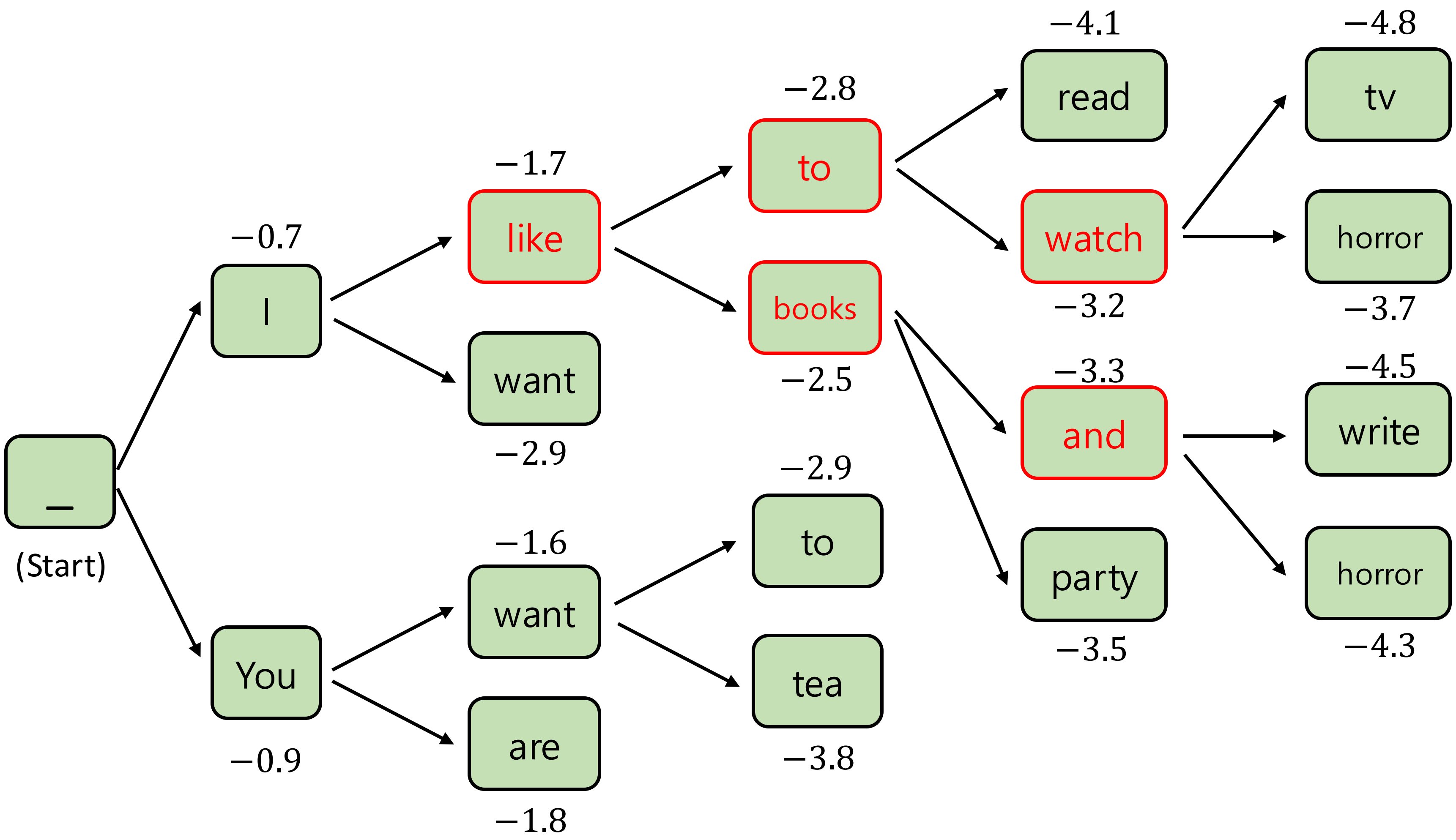
Beam Search ($t = 4$)

`Beam size` = 2



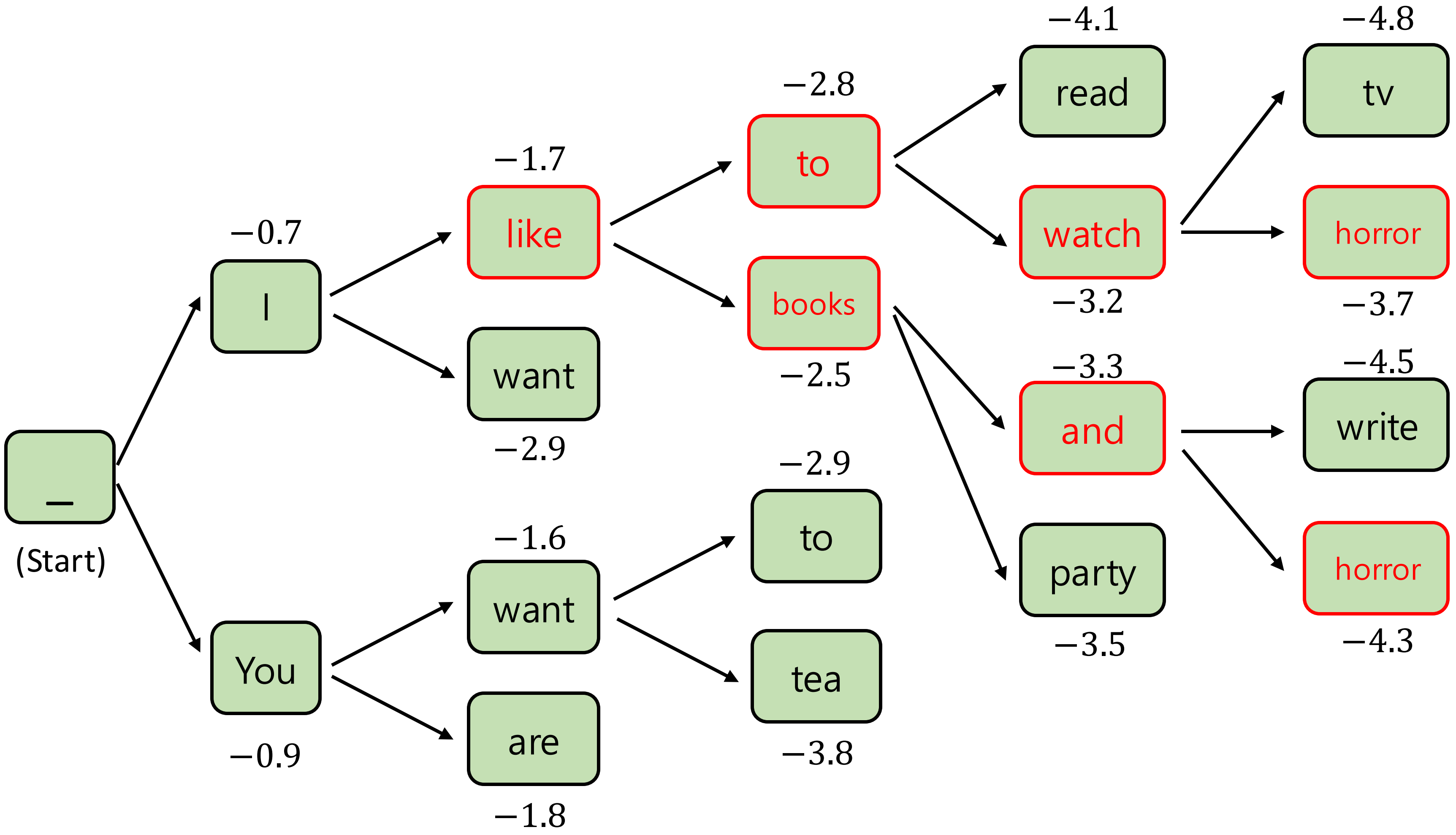
Beam Search ($t = 5$)

`Beam size` = 2



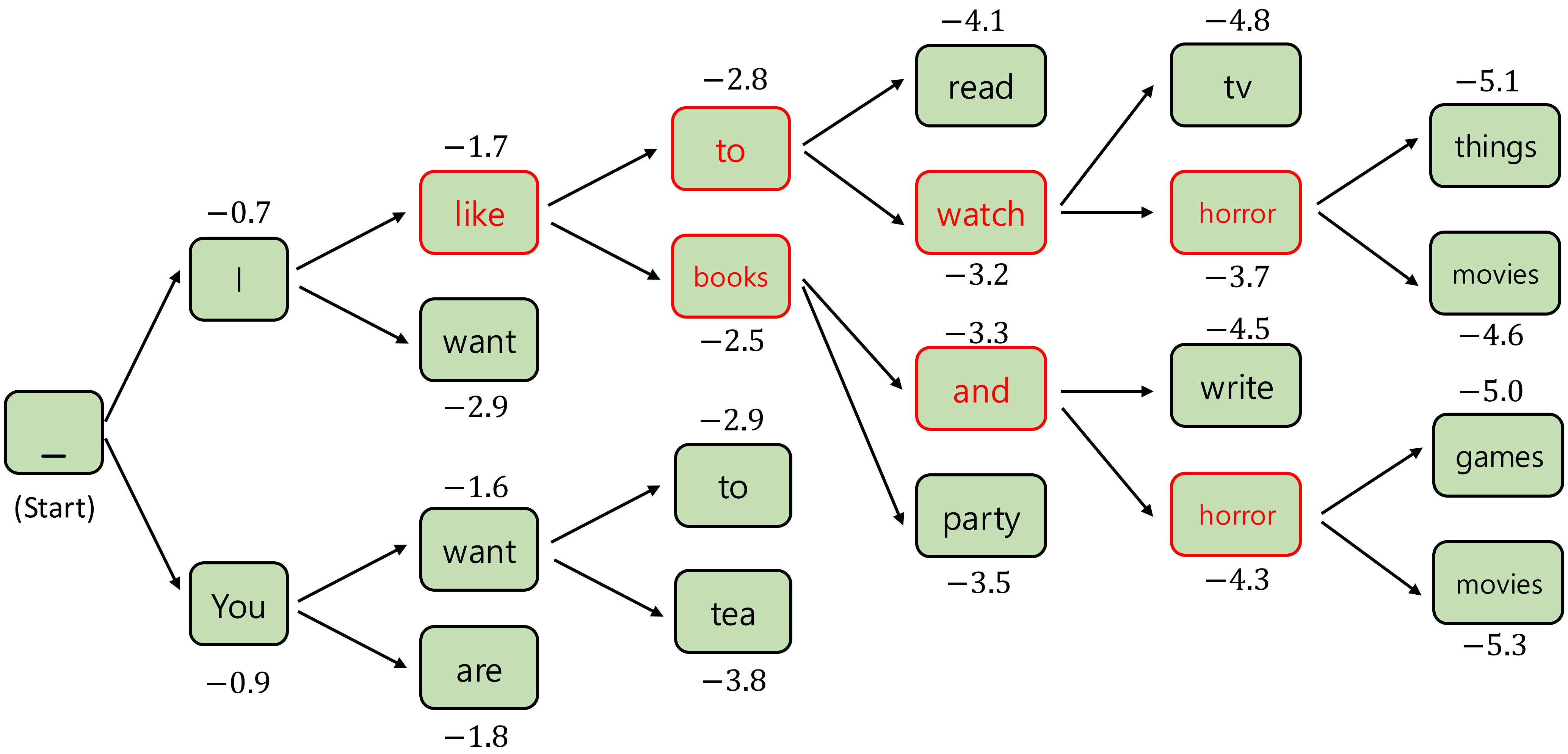
Beam Search ($t = 5$)

`Beam size` = 2



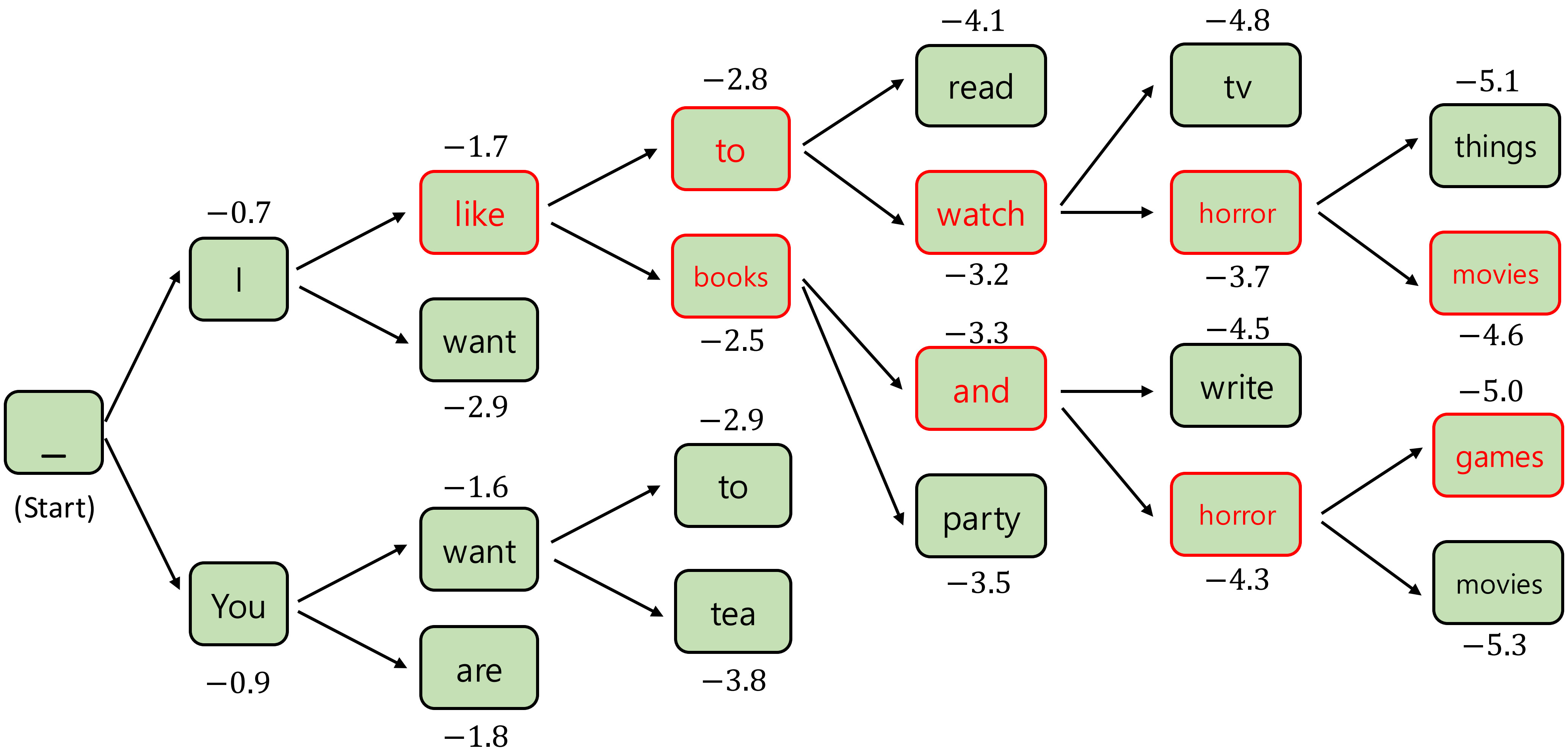
Beam Search ($t = 6$)

`Beam size` = 2



Beam Search ($t = 6$)

`Beam size` = 2



Stop Criterion

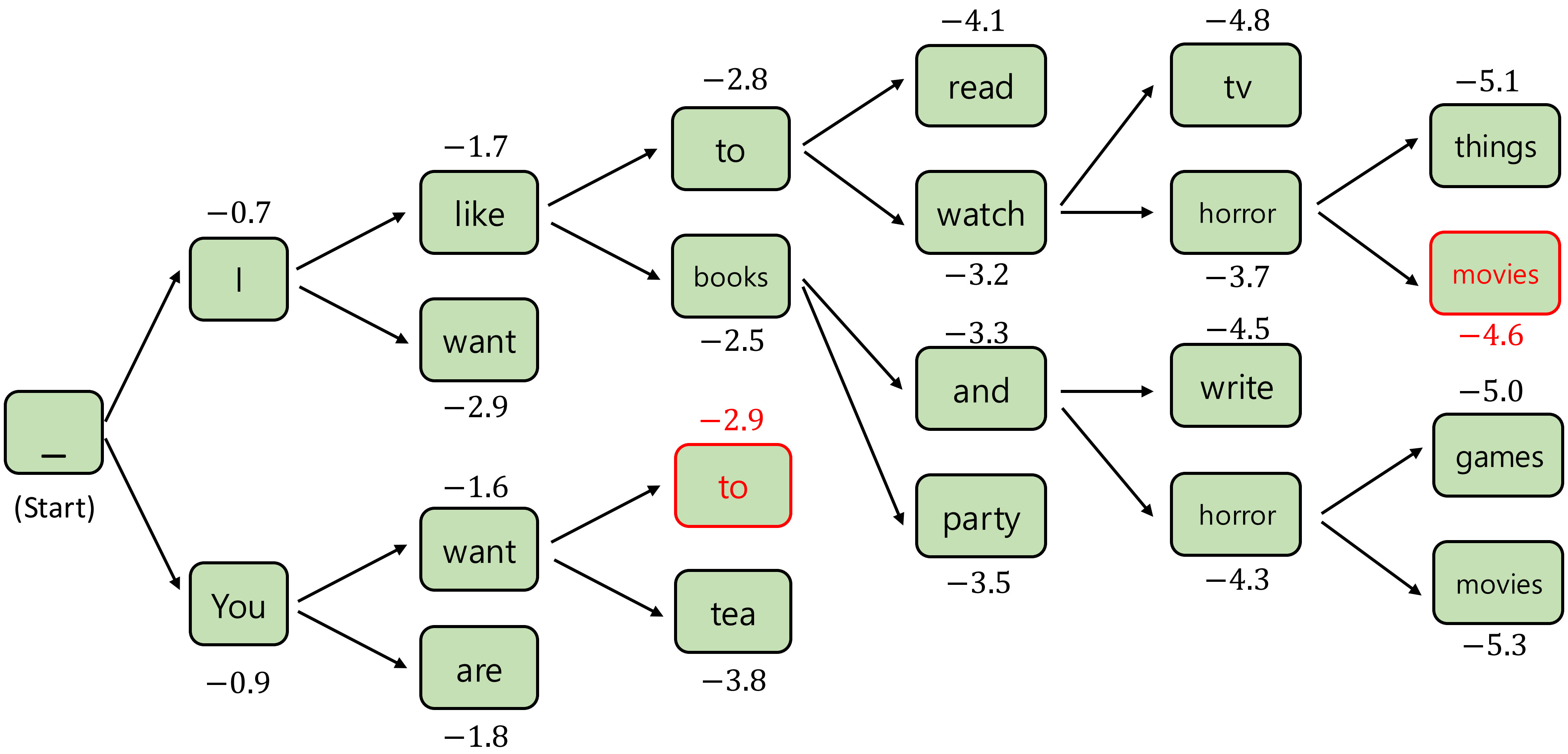
- There are two common stop criteria, either for greedy decoding or beam search decoding:
  - We consider a sequence of generation complete when the <EOS> token is produced by a model. *<EOS>: End of sequence
    - E.g., <Start> I like to watch horror movies <EOS>
  - A generated sequence reaches a pre-defined maximal length.

Problem of Beam Search

- Longer candidates will have lower scores.
- (Let's see again the 6th time step)

Beam Search ($t = 6$)

`Beam size` = 2



Problem of Beam Search

- Longer candidates will have lower scores.
- Solution: Perform normalization to penalize on length

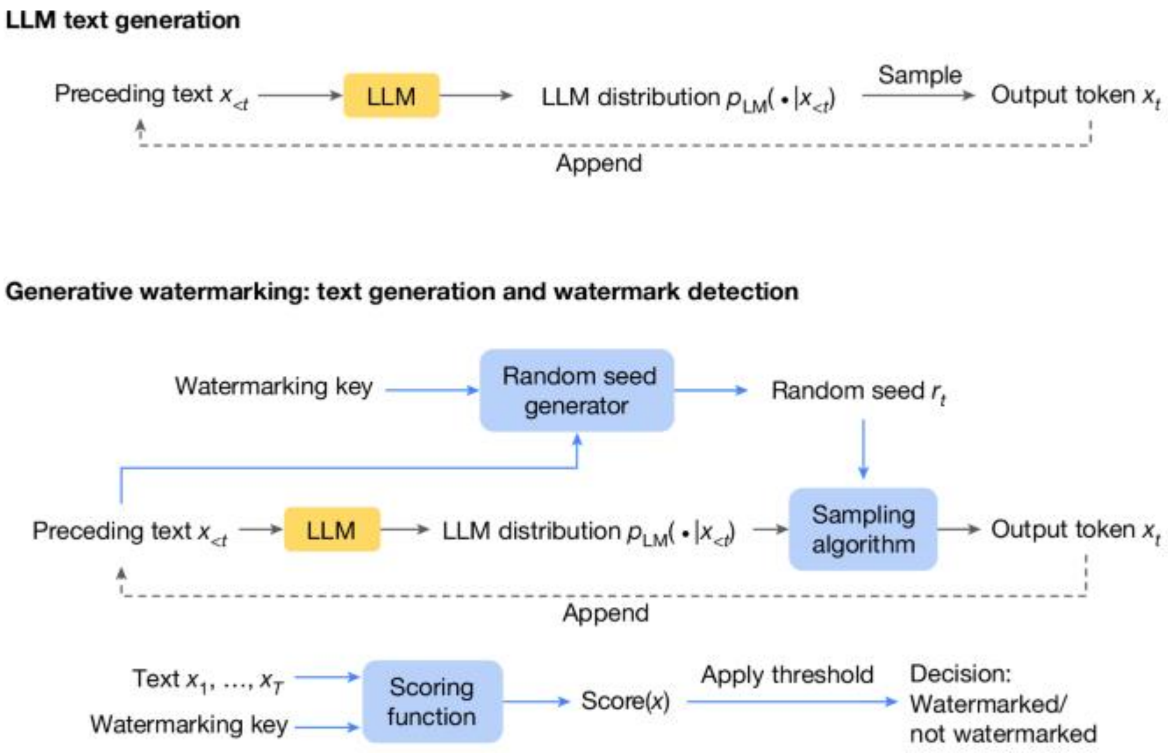
$$L_{ml} = \frac{1}{T}\sum_{t=1}^{T} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*, x)$$

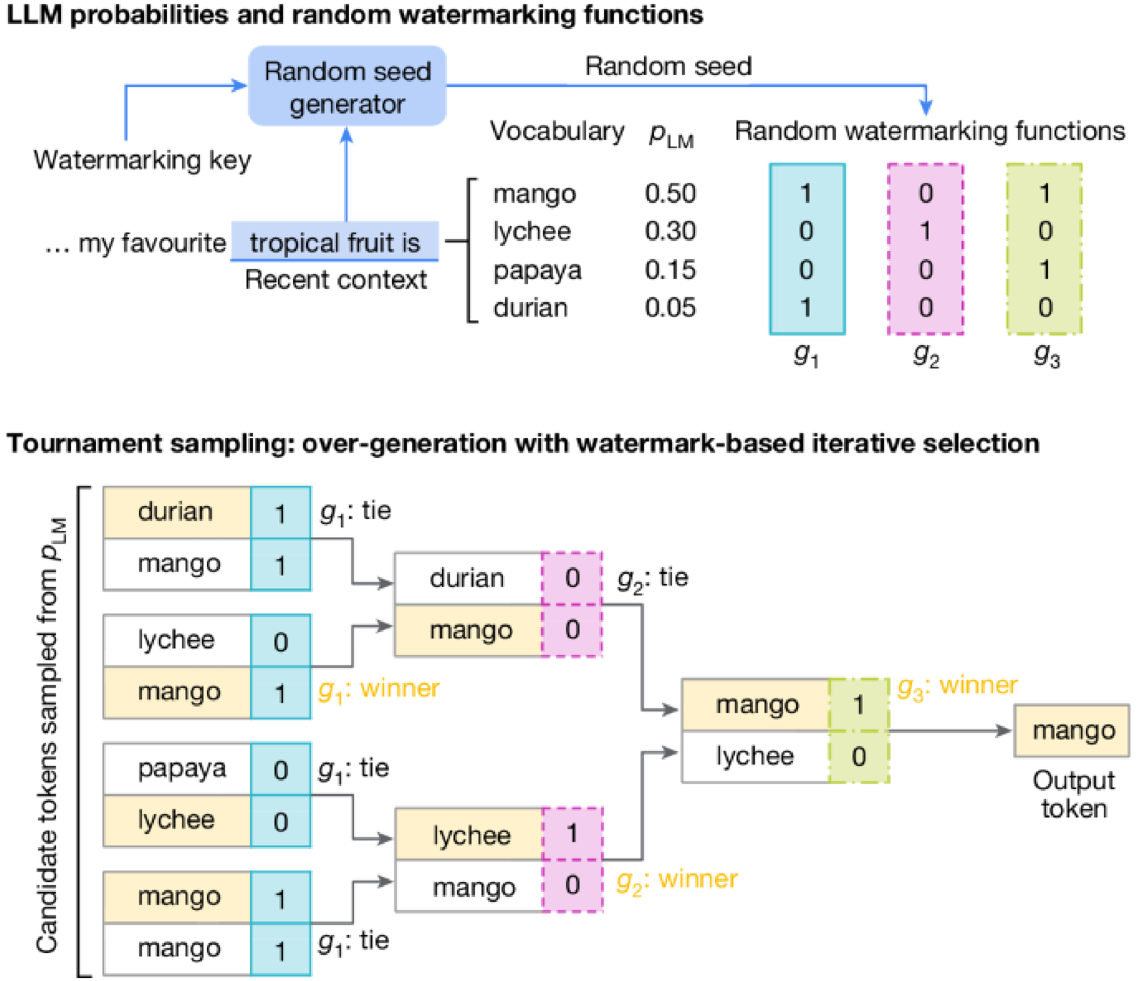
https://www.nature.com/articles/s41586-024-08025-4

Watermark for text

Scalable watermarking for identifying large language model outputs,

Nature 2024 Oct.





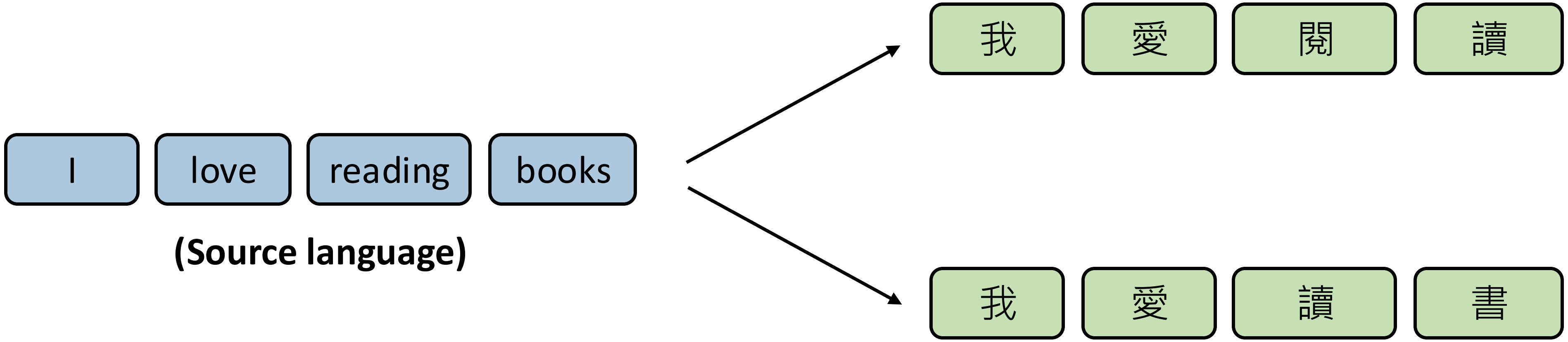
preserves text quality and enables high detection accuracy

Evaluation

How to evaluate natural language generation?

- Natural language is hard to evaluate due to subjectivity and language diversity.

For example: Machine Translation



(Source language)

(Target language)

- Human evaluations
- Automatic evaluations (We will focus on this topic.)

BLEU (Bilingual Evaluation Understudy)

- A word-based metric.
  - It is very sensitive to word tokenization
- Core concept: Compute precision for n-grams:
  - Unigrams -> BLEU-1
  - Bigrams -> BLEU-2
  - Trigrams -> BLEU-3
  - 4-grams -> BLEU-4

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Precision and Recall

$$\text{Precision}=\frac{\text{Relevant and retrieved instances}}{\text{All retrieved instances}}$$

Predicted by a model

$$\text{Recall}=\frac{\text{Relevant and retrieved instances}}{\text{All relevant instances}}$$

Ground-truths

Relevant and retrieved instances: Intersection between predictions and ground-truths

Calculation of BLEU Score (Example)

Assume we now translate from Chinese to English.

Calculate BLEU-1 score

Chinese: 我想要讀那本書

Reference1: I want to read the book.

Reference2: I want to read that book.

Model output: the the the the the the.


```
"id": 391895,
"file_name": "000000391895.jpg",
"width": 640,
"height": 480,
"captions": [
  "A man riding a bicycle down a street next to parked cars.",
  "A person on a bike rides past cars on a city street.",
  "A man rides a bicycle past a row of parked cars.",
  "A bicyclist in a blue shirt rides by parked vehicles on the road.",
  "A person riding a bike on a street with parked cars."
]
```


An example of COCO dataset

Calculation of BLEU Score (Example)

Assume we now translate from Chinese to English.

Calculate BLEU-1 score

Chinese: 我想要讀那本書

Reference1: I want to read the book.

Reference2: I want to read that book.

Model output: the the the the the the.

Precision: $\frac{6}{6}$

100%! Can this be true?

Calculation of BLEU Score (Example)

Assume we now translate from Chinese to English.

Calculate BLEU-1 score

Chinese: 我想要讀那本書

Reference1: I want to read the book.

Reference2: I want to read that book.

Model output: the the the the the the.

$\cancel{\text{Precision: } \frac{6}{6}}$

Modified Precision: $\frac{1}{6}$

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Why should we use modified precision?

- The output sequences can be total mistakes.
  - E.g., the the the the the the
- Original precision is in favor of longer output sequences.
- Therefore, we should use modified precision to prevent bad evaluations.

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

More than one references can be provided for machine translation!

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

1 2 3 4 5 6

	Count	
the dog	2	(duplicated)
dog the	1	
dog on	1	
on the	1	
the bed	1	

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

Clips to the reference

	Count	Count_{clip}
the dog	2	1
dog the	1	
dog on	1	
on the	1	
the bed	1	

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

	Count	Count_{clip}
the dog	2	1
dog the	1	0
dog on	1	
on the	1	
the bed	1	

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

	Count	$Count_{clip}$
the dog	2	1
dog the	1	0
dog on	1	1
on the	1	
the bed	1	

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

Count only one time even mapped to both references.

	Count	Count_{clip}
the dog	2	1
dog the	1	0
dog on	1	1
on the	1	1
the bed	1	

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Calculation of BLEU Score (Example)

For each unique ngram
- count its maximum frequency in each of the reference sentences.
- The clipped count = minimum (this special count, the original count)

Calculate BLEU-2 score

Reference1: The dog is on the bed.

Reference2: There is a dog on the bed.

Model output: The dog the dog on the bed.

Count only one time even mapped to both references.

	Count	$\text{Count}_{\text{clip}}$
the dog	2	1
dog the	1	0
dog on	1	1
on the	1	1
the bed	1	1

Modified Precision: $\frac{4}{6}$

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Formula of BLEU Score

Summation for unigram, bigram, tri-gram, and 4-gram

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Summation for all candidates (model outputs) of each translation

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

What we've learned BLEU so far

- The BLEU score is calculated from the summation of 1-gram to 4-gram.
  - You can also measure n-gram individually.
- We use modified precision to prevent bad evaluations.
- What will happen if a model tends to generate really short sentences?

➡ **More penalty for calculating BLEU score!**

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Brevity Penalty (BP)

- BP is used to penalize short candidates.

c: The length of a candidate sequence
r: The length of a reference sequence that is closest to c (shorter one)

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Then,

$$\text{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

N=4 to include 1-gram to 4-gram

Weight for each n-gram (was set 1/4 in the original paper)

Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

ROUGE: A Package for Automatic Evaluation of Summaries (Recall-Oriented Understudy for Gisting Evaluation)

- ROUGE-N: N-gram Co-Occurrence Statistics (recall base)
- ROUGE-L: count LCS

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

S1. police killed the gunman (reference)
S2. police kill the gunman (summary candidate 1)
S3. the gunman kill police (summary candidate 2)

ROUGE-2_{S2} = ROUGE-2_{S3}
ROUGE-L_{S2} = ¾ > ROUGE-L_{S3} = ½
police the gunman vs the gunman

S4. the gunman police killed

ROUGE-2_{S4} > ROUGE-2_{S2}
ROUGE-L_{S4} < ROUGE-L_{S2}

(Recap) Perplexity

Perplexity (PPL) is a quantitative criterion used to evaluate the capacities of language modeling models.

- Given the sequence of words $W = w_1 w_2 \ldots w_N$ and an N-gram model. The PPL of the model was computed by:

$$Perplexity(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} = \sqrt[N]{\prod_{k=1}^{n} \frac{1}{P(w_k | w_{k-N+1:k})}}$$

The lower the value of perplexity, the better the language modeling capability of the model.

Comparison for Human and Automatic Evaluations

- Human evaluations
  - Pros: More accurate for subjectivity, flexibility for any desired comparison
  - Cons: Less objective, time-consuming, expensive
- Automatic evaluations
  - Pros: Objective enough to serve as common evaluation metrics, fast
  - Cons: Cannot meet language diversity
    - Take machine translation for instance, there are always other valid ways to translate the source sentence.

NLP Benchmark – GLUE ICLR 2019, https://openreview.net/pdf?id=rJ4km2R5t7

GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING

GLUE: General Language Understanding Evaluation

- Consist of **9** task
  - Using one model to fit all tasks
  - Prove that model have **general language understanding**
- Task can be divided into **three** categories
  - **Single** Sentence Classification
  - **Pairwise** Text Classification
  - Text **Similarity**
- Average score
  - Average 9 task performance
  - **Higher** means better
  - **Human** baseline: **87.6** on average

NLP Benchmark - GLUE

Single Sentence Classification

- Sentence **Acceptability**
  - Given a sentence **s**, is **s** a valid sentence, i.e., have **correct grammar** and **meaningful content**?
  - Dataset: **CoLA**
- **Sentiment** Analysis
  - Given a sentence **s**, is **s** expressing **postive** sentiment or **negative** sentiment?
  - Dataset: **SST-2**

Pairwise Text Classification

- Natural Language **Inference** (NLI)
  - Given premise **p** and hypothesis **h**, is p **implies**, **contradict** or **not related** to **h**?
  - Dataset: **MNLI-(m/mm)**, **RTE**, **WNLI**, **QNLI**
- Paraphrase
  - Given two sentences **s1** and **s2**, is **s1 means s2**, i.e., **s1** is "**in other words**" **s2**?
  - Dataset: **QQP**, **MRPC**

Text **Similarity**

- Given two sentences **s1** and **s2**, from **1 to 10**, how much does **s1 similar to s2**?
- Dataset: **STS-B**

GLUE Benchmark

https://paperswithcode.com/dataset/glue

Data source: https://gluebenchmark.com/tasks

Data Name	Full Name	Train / Val Size	Task
CoLA	The Corpus of Linguistic Acceptability	8.6k / 1.0k / 1.1k	Single-sentence Classification
SST-2	The Stanford Sentiment Treebank	67k / 87k / 1.8k	Single-sentence Classification
MRPC	Microsoft Research Paraphrase Corpus	3.7k / 408 /1.7k	Pair-sentence Classification
STS-B	Semantic Textual Similarity Benchmark	5.8k / 1.5k / 1.4k	Semantic Similarity
QQP	Quora Question Pairs	364k / 40.4k / 391k	Question Answering
MNLI-m	Multi-Genre Natural Language Inference (matched)	393k / 9.8k / 9.8k	Pair-sentence Classification
MNLI-mm	Multi-Genre Natural Language Inference (mis-matched)	393k / 9.8k / 9.9k (Same training set as MNLI-m)	Pair-sentence Classification
QNLI	Question NLI	105k / 5.5k / 5.5k	Pair-sentence Classification
RTE	Recognizing Textual Entailment	2.5k / 277 / 3k	Pair-sentence Classification
WNLI	Winograd NLI	635 / 71 / 146	Pair-sentence Classification

CoLA: The Corpus of Linguistic Acceptability

imbalanced dataset (7:3)

Corpus Sample

clc95	0	*	In which way is Sandy very anxious to see if the students will be able to solve the homework problem?
c-05	1		The book was written by John.
c-05	0	*	Books were sent to each other by the students.
swb04	1		She voted for herself.
swb04	1		I saw that gas can explode.

CoLA hard examples

X I can't go because too tired.

X The child seems sleeping.

O The pizza cried all night.

NLP Benchmark - SQuAD

SQuAD: Stanford Question Answering Dataset

- A **reading comprehension** dataset
- Consisting of **questions** posed by crowdworkers on a set of **Wikipedia** atricles
- The **answer** to every question is **a segment of text from the corresponding reading passage**
- Or the question might be **unanswerable**

SQuAD 1.1

- 100000+ question-answer pairs
- 500+ articles

SQuAD 2.0

- SQuAD 1.1 + 50000 **unanswerable** questions
- Unanswerable questions look similar to answerable ones.
- To do well on SQuAD 2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering

SQuAD 2.0

Computational_complexity_theory

The Stanford Question Answering Dataset

Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other. A computational problem is understood to be a task that is in principle amenable to being solved by a computer, which is equivalent to stating that the problem may be solved by mechanical application of mathematical steps, such as an algorithm.

What branch of theoretical computer science deals with broadly classifying computational problems by difficulty and class of relationship?

Ground Truth Answers: Computational complexity theory | Computational complexity theory | Computational complexity theory

By what main attribute are computational problems classified utilizing computational complexity theory?

Ground Truth Answers: inherent difficulty | their inherent difficulty | inherent difficulty

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

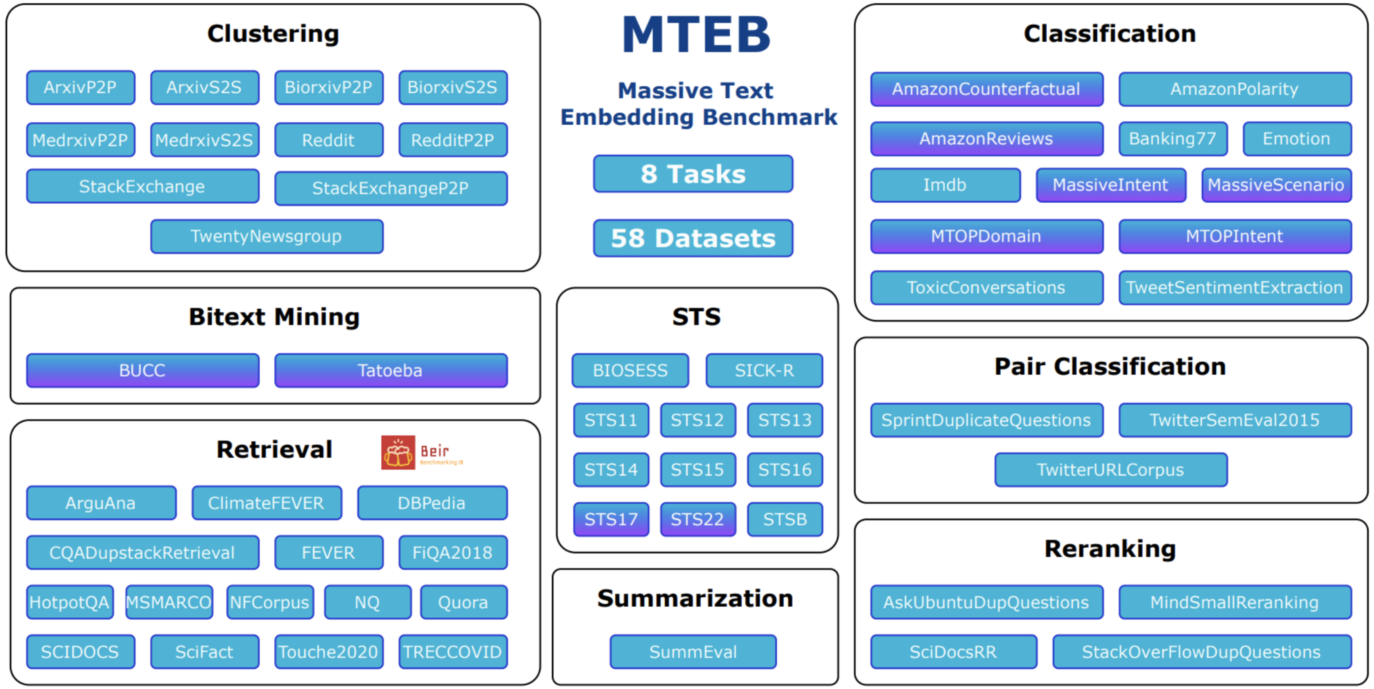
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**

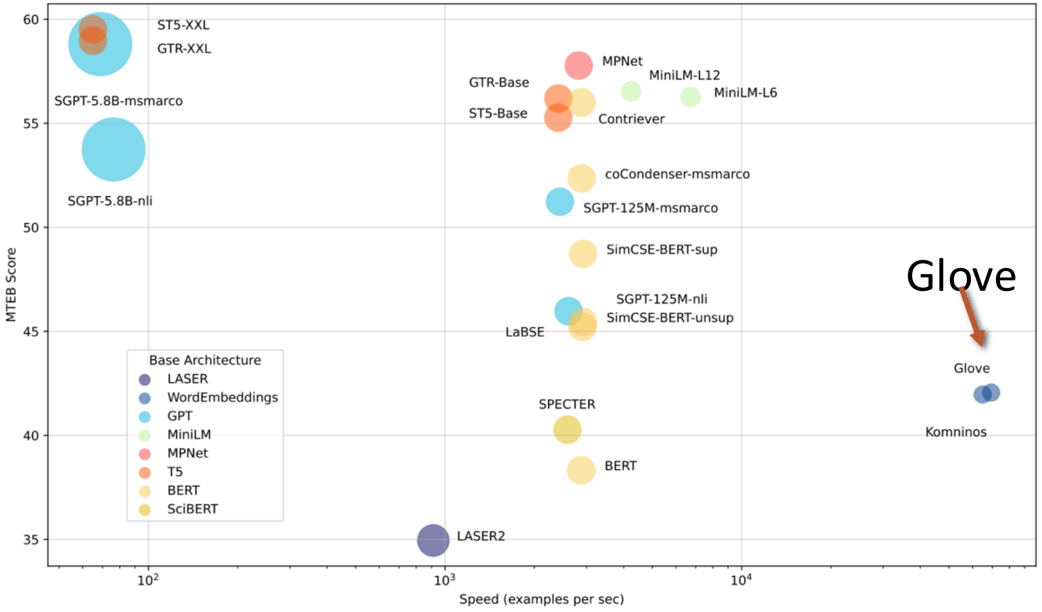
https://arxiv.org/abs/2210.07316
https://github.com/embeddings-benchmark/mteb

MTEB (Massive Text Embedding Benchmark)

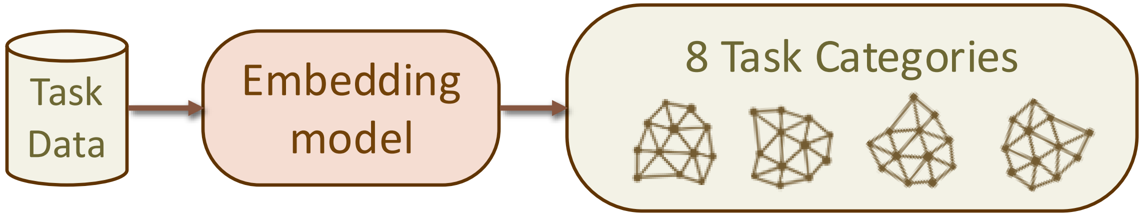
(mMTEB, C-MTEB)

https://huggingface.co/spaces/mteb/leaderboard





SOTA: Qwen-Embedding-8B (70.58)



Task models

(nDCG@10, Recall@k, Accuracy, F1, V-measure, ARI, Spearman, Pearson correlation)

In-domain, out-of-domain, and open-domain data

In-Domain Data

- **Definition**: In-domain data refers to data that is **similar** to the data used to train a model. It typically comes from the same distribution, context, or domain. (Aligned with the training data)

- **Examples**: If a model is trained on news articles about technology, in-domain data would also consist of technology-related news articles.

Out-of-Domain Data

- **Definition**: Out-of-domain data consists of data that is **significantly different** from the training data, either in terms of content, style, or context.

- **Examples**: Continuing with the previous example, if the model is tested on health-related articles, this would be considered out-of-domain data.

Open-Domain Data

- **Definition**: Open-domain data refers to data that covers a wide range of topics and is not restricted to a specific domain. It is designed to be general and comprehensive.

- **Examples**: A chatbot that can answer questions about history, science, sports, and technology would work with open-domain data.

MMLU: Massive Multitask Language Understanding

1 measure knowledge acquired during pretraining by evaluating models exclusively in **zero-shot** and **few-shot** settings.

2 covers 57 subjects across STEM, the humanities, the social sciences, and more.

3 multiple-choice QA

Also TMMLU, MediaTek
https://arxiv.org/pdf/2309.08448

Q: 「臺灣原住民的布只有形制屬傳統或較現代的分別，像圓領的剪裁、鈕扣和棉布的使用等，都是受漢人的影響而來。泰雅族的貝珠鈴衣，是貝珠串底下加銅鈴裝飾，銅鈴也是和漢人交易而來。日治時代的原住民服裝，還出現以漢人棉布做底、日本布做袖口、原住民圖案做主要裝飾的混搭法。」這段文字的主旨最可能是下列何者？
(A)不同文化的碰撞，可融合並產生新的火花
(B)外來文化的入侵，讓在地的傳統文化日漸消失
(C)臺灣原住民的文化，影響了漢人與日本人的穿著
(D)觀察不同族群的服飾，就能了解不同文化的差異

A: (A)