Natural Language Processing

Decoding Strategies and Evaluations for Natural Language Generation

Outline

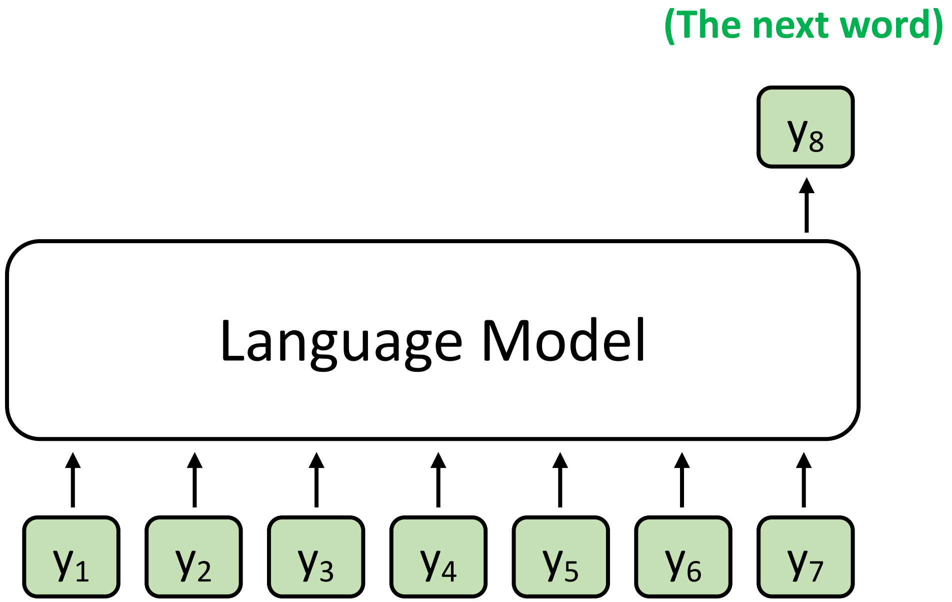
- Recap: Language Generation
- Decoding Strategies
  - Greedy Decoding
  - Beam Search
  - Top-k / Top-p Sampling
- Evaluations

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



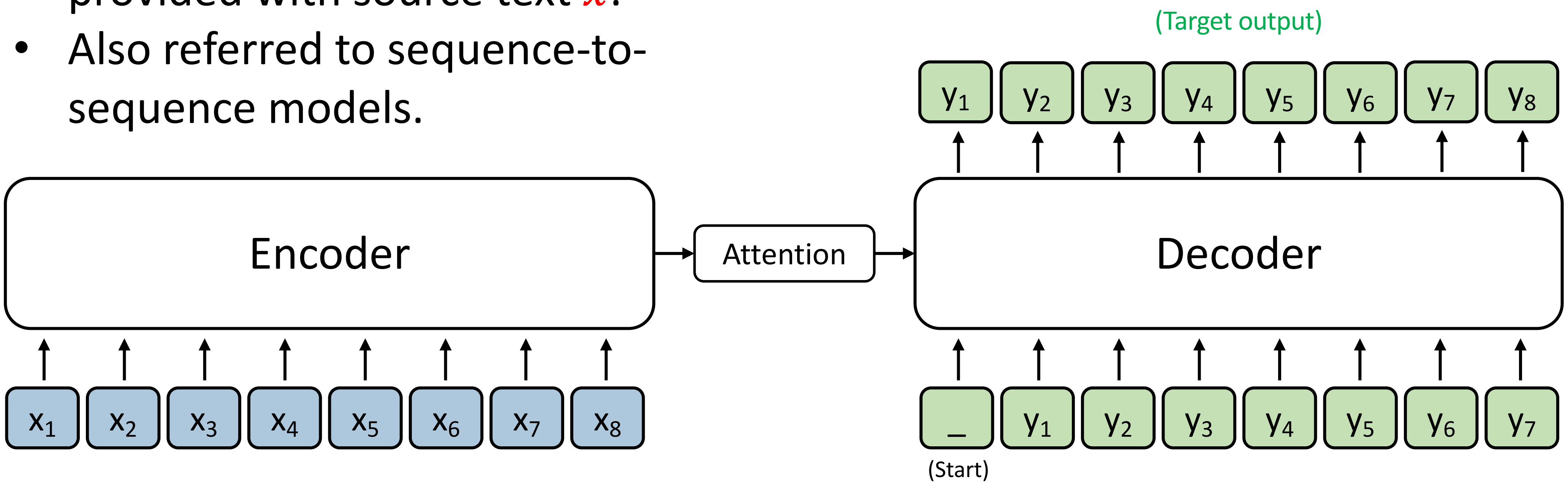
$$P(y_t | y_1, y_2, \dots, y_{t-1})$$

- 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.

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



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

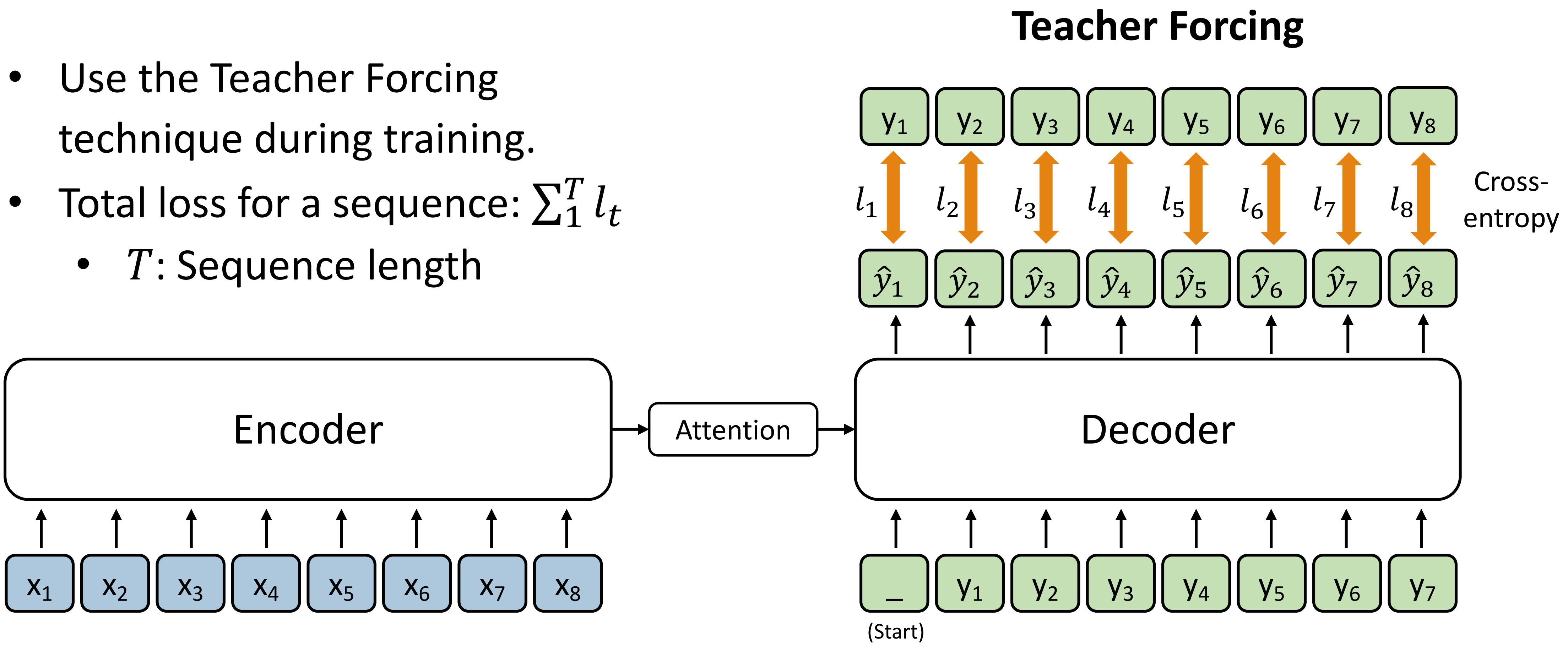
I love reading → books

Machine Translation (Supervised)

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

How to train a (Conditional) Language Model?

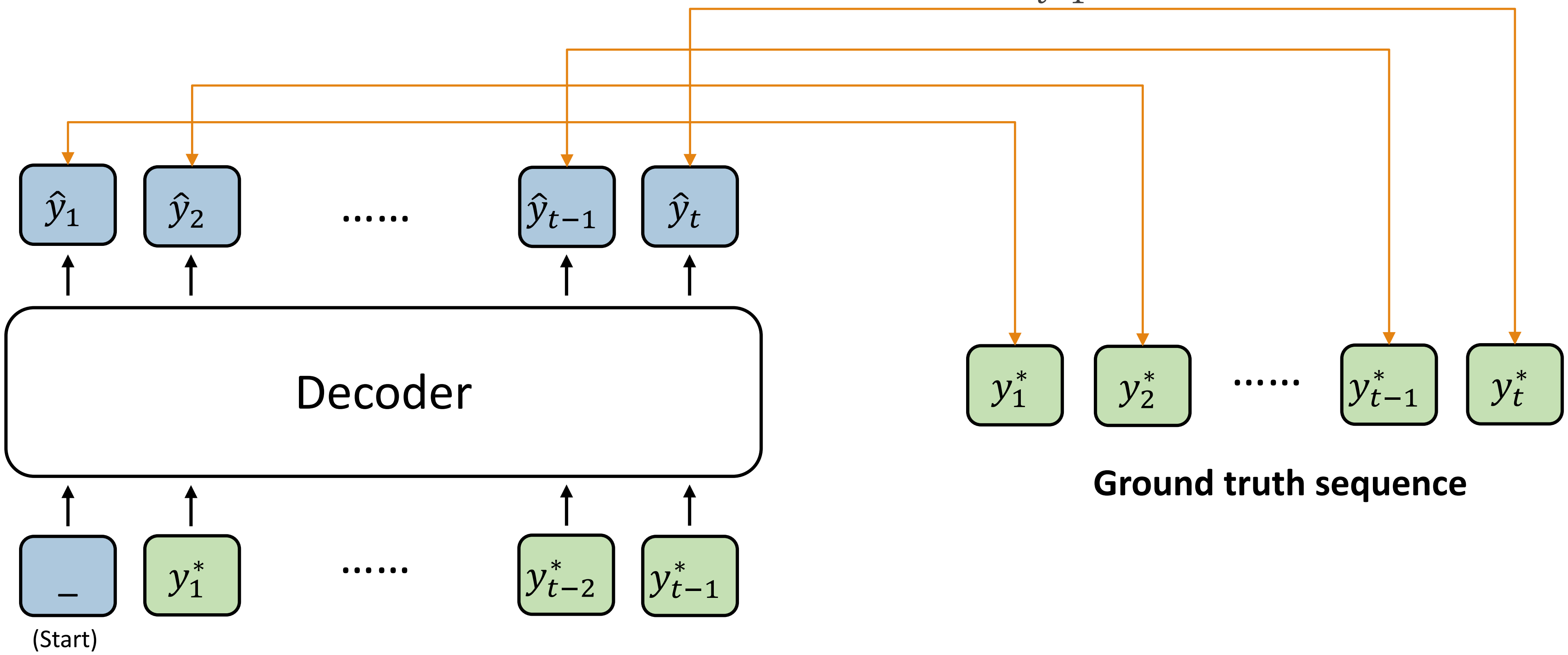
- Use the Teacher Forcing technique during training.
- Total loss for a sequence: $\sum_1^T l_t$
  - T: Sequence length



Teacher Forcing – Training stage

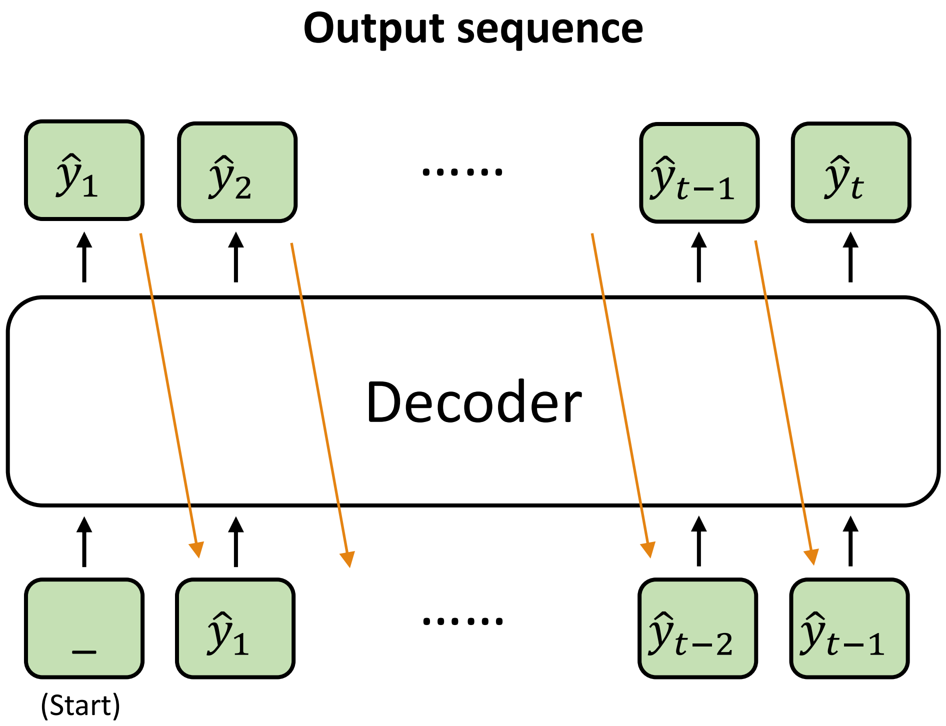
During training:

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



Teacher Forcing – Testing stage

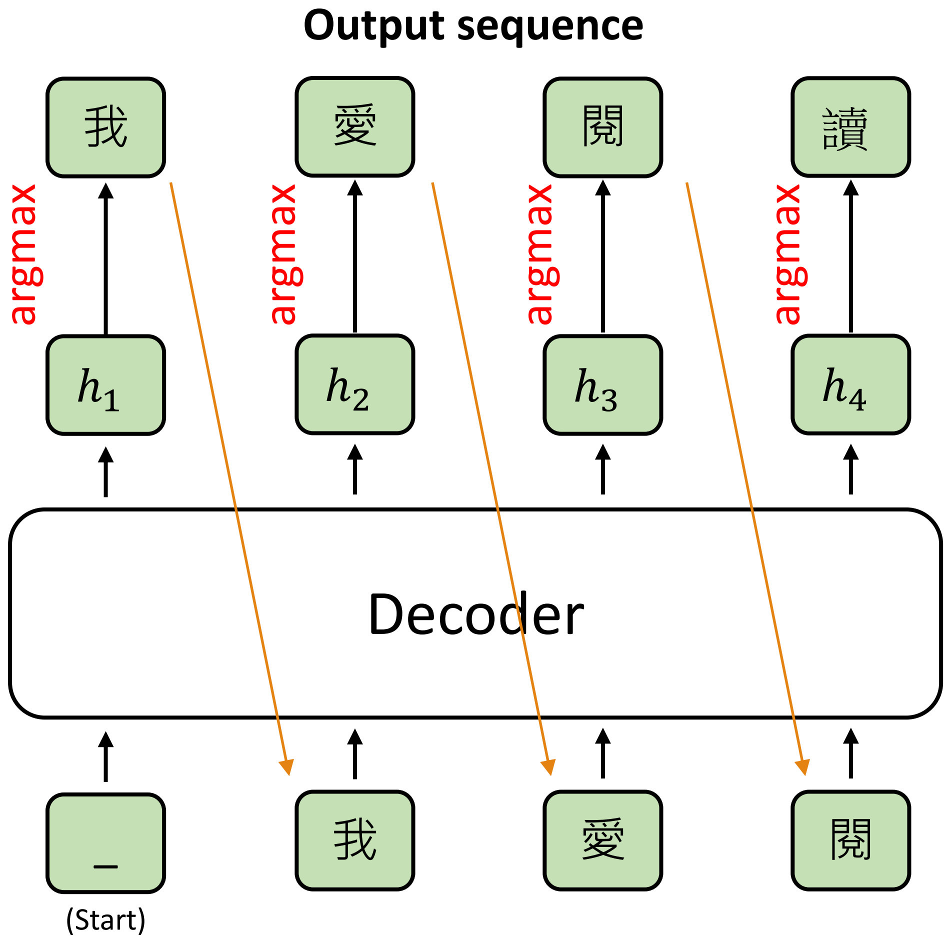
During testing:



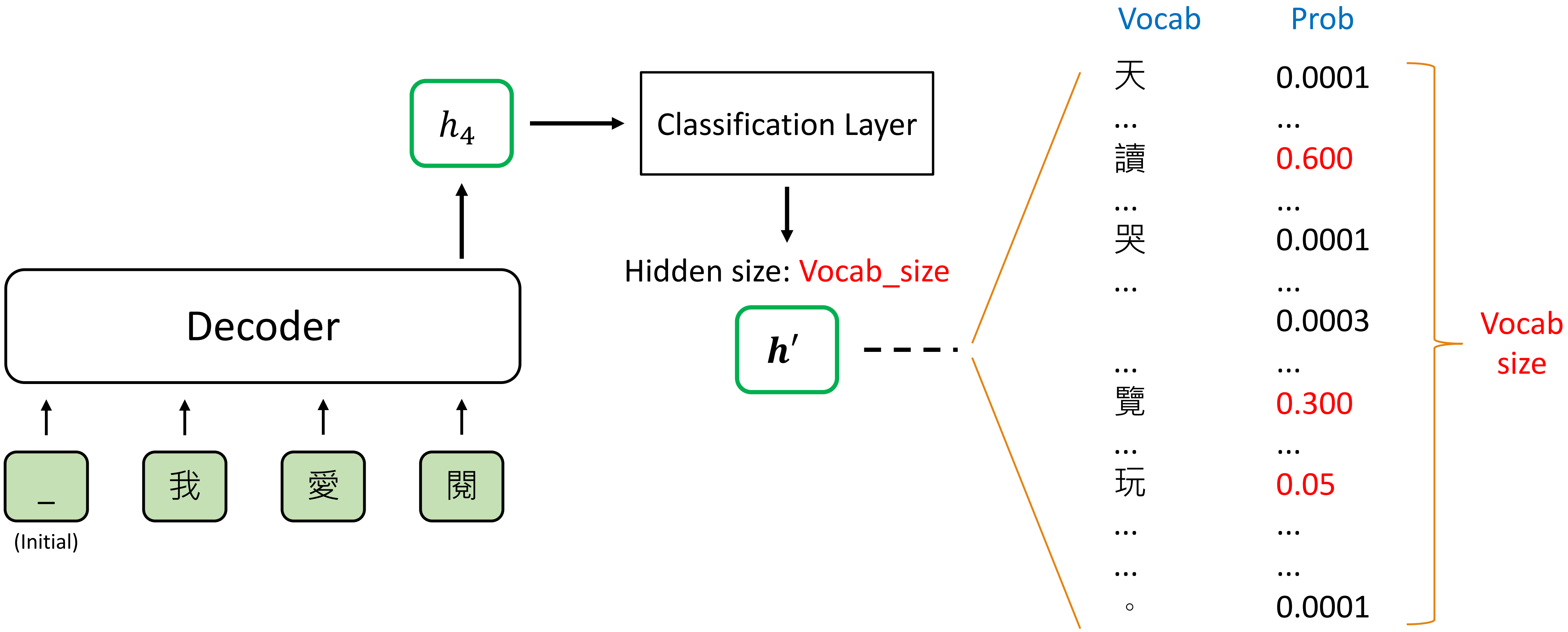
- Advantage: stabilize training and increase performance
- Question: How does the next word be determined?

Greedy Decoding

Example: I love reading books.



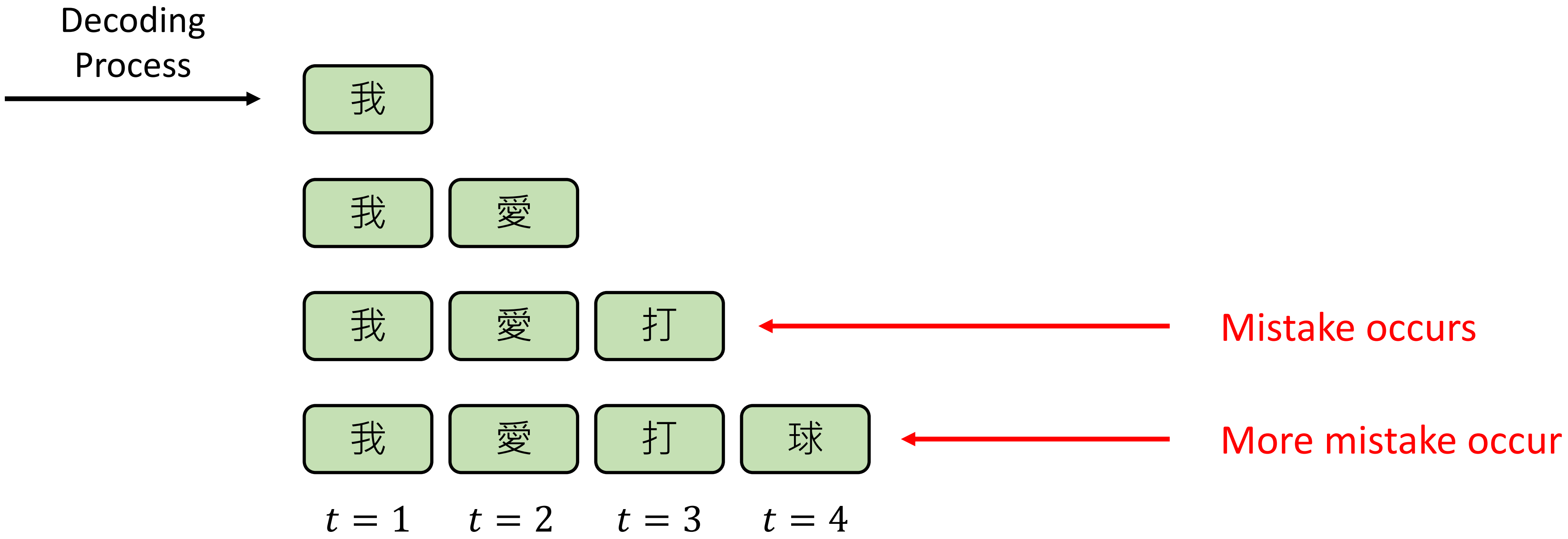
Greedy Decoding – Best Selection Process



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 choices at each time step?

➡ **Beam Search**

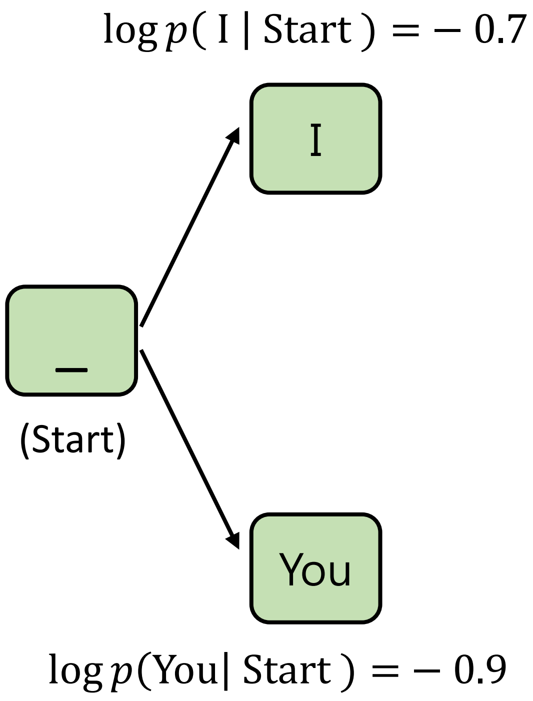
Beam Search

- 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:

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

Beam Search ($t = 1$)

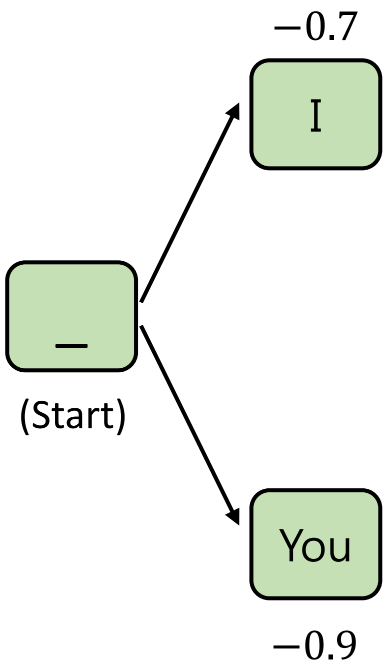
`Beam size` = 2



- At this decoding step, two choices are preserved.

Beam Search ($t = 1$)

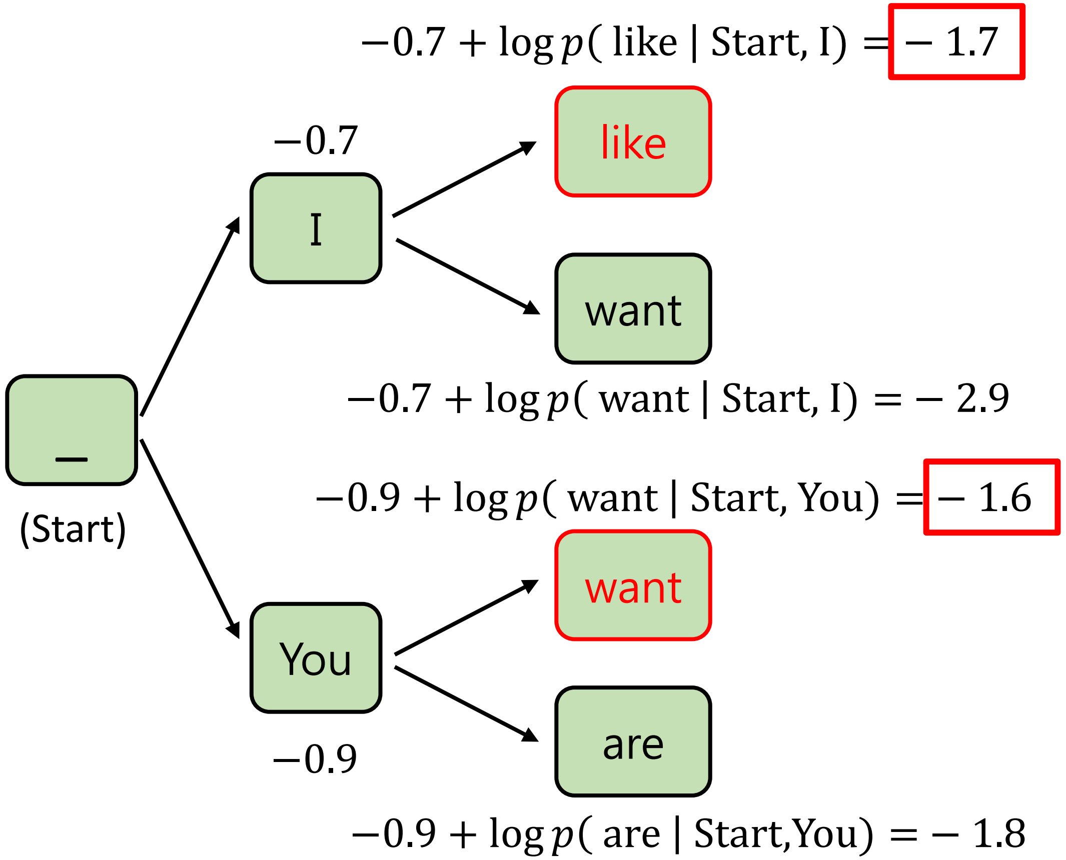
`Beam size` = 2



- At this decoding step, two choices are preserved.

Beam Search ($t = 2$)

`Beam size` = 2

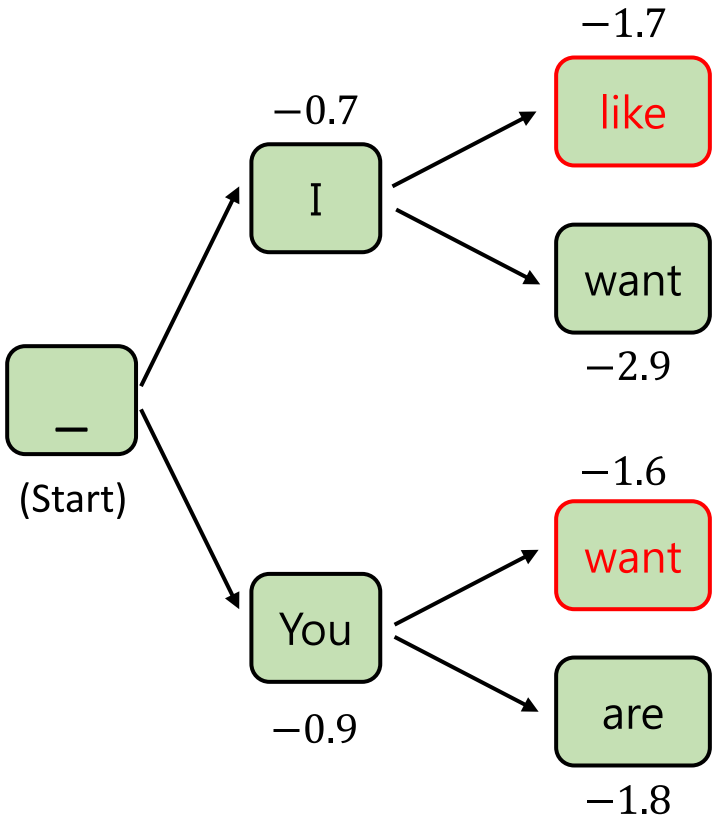


Note the loglikelihood! Being close to zero is better!

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

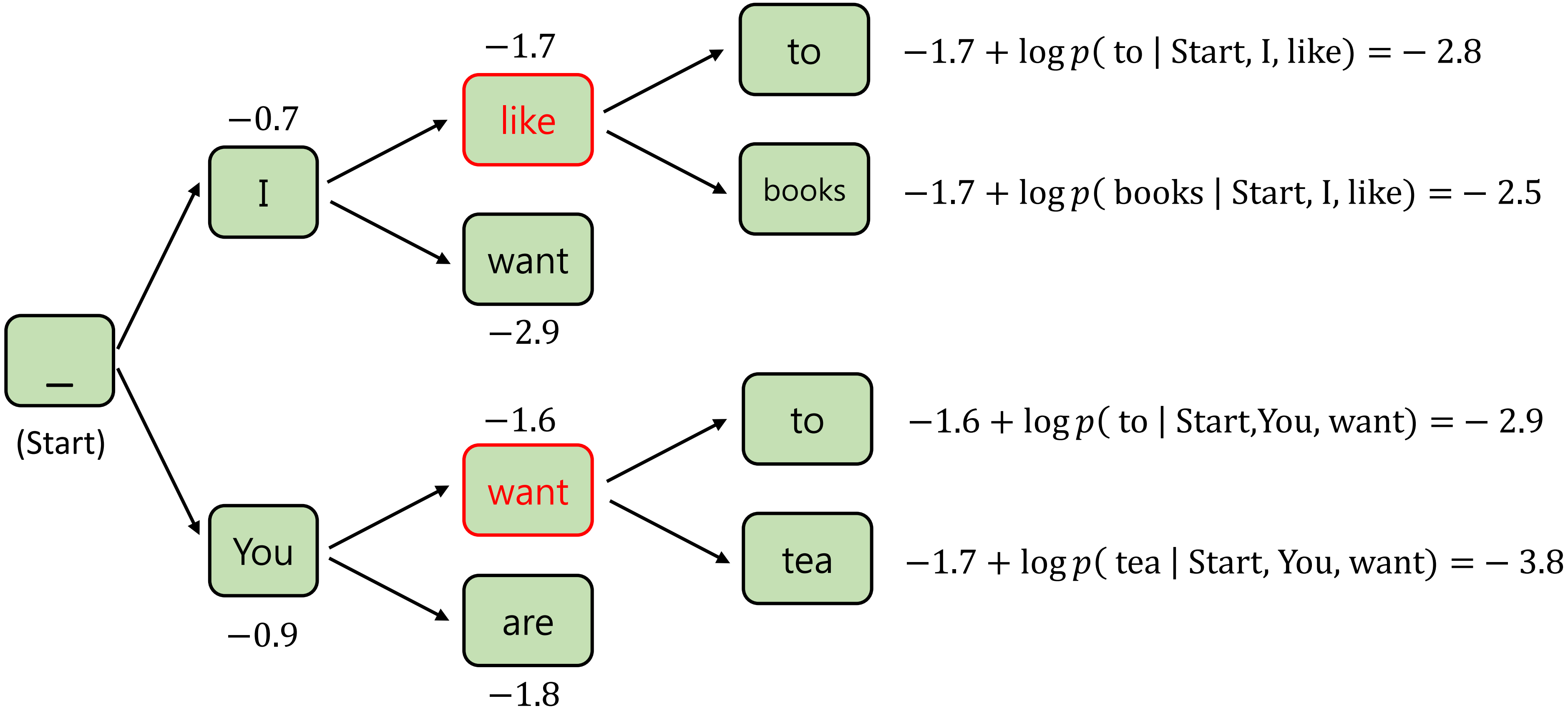
Beam Search ($t = 2$)

`Beam size` = 2



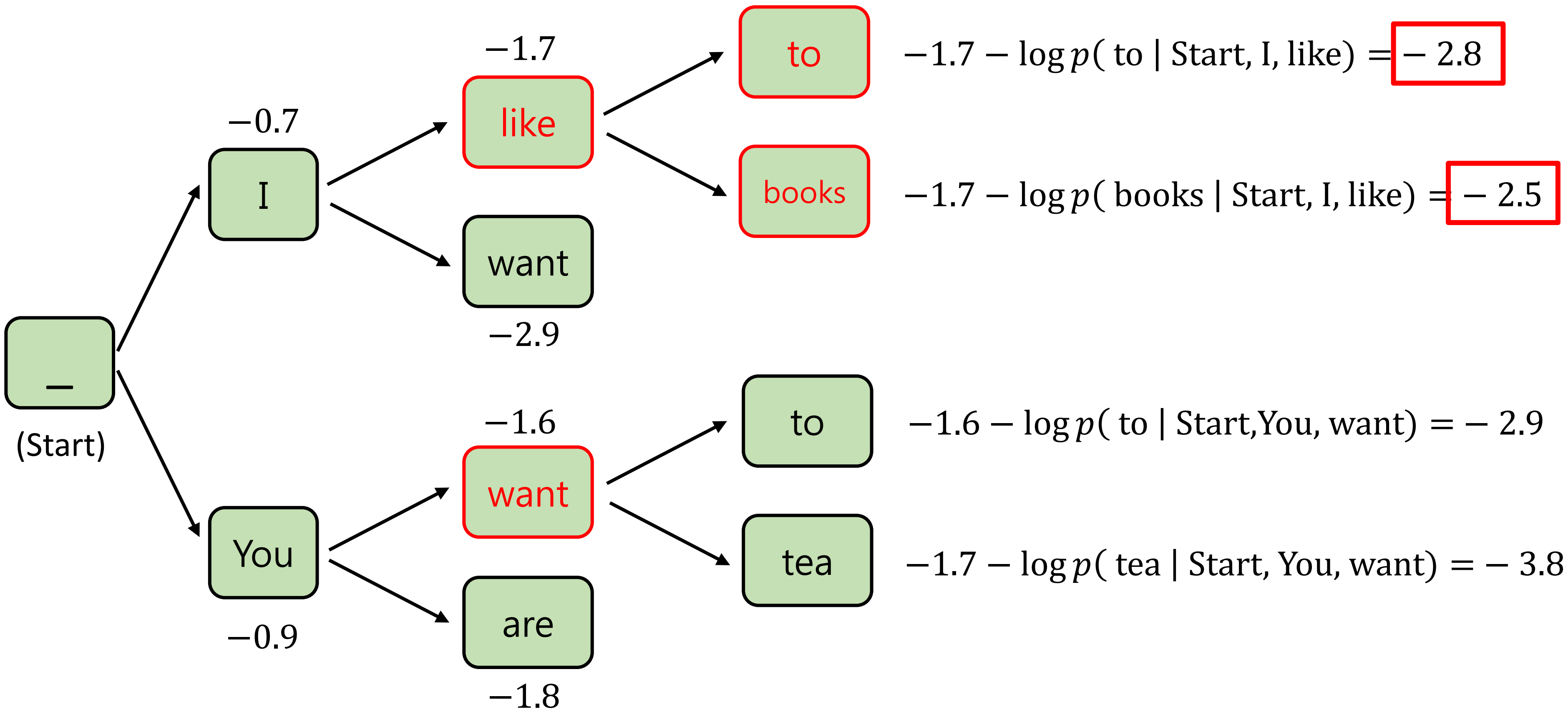
Beam Search ($t = 3$)

`Beam size` = 2



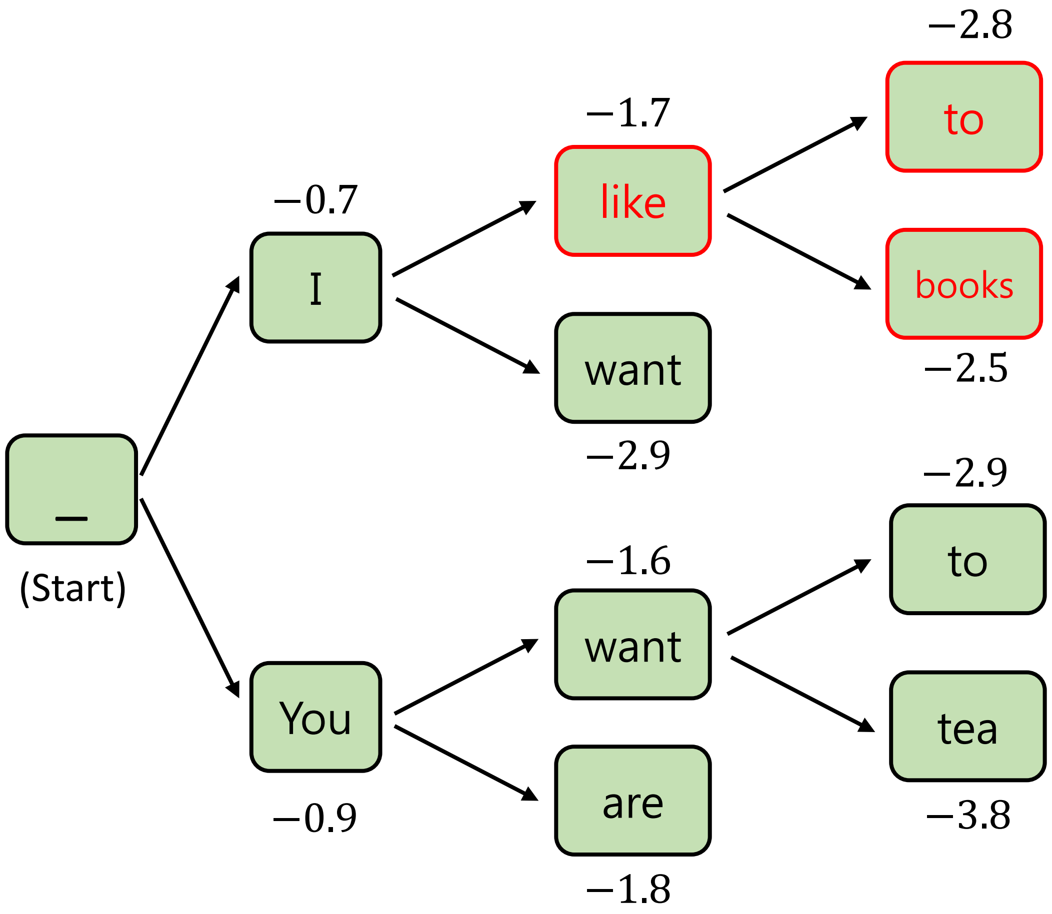
Beam Search ($t = 3$)

`Beam size` = 2



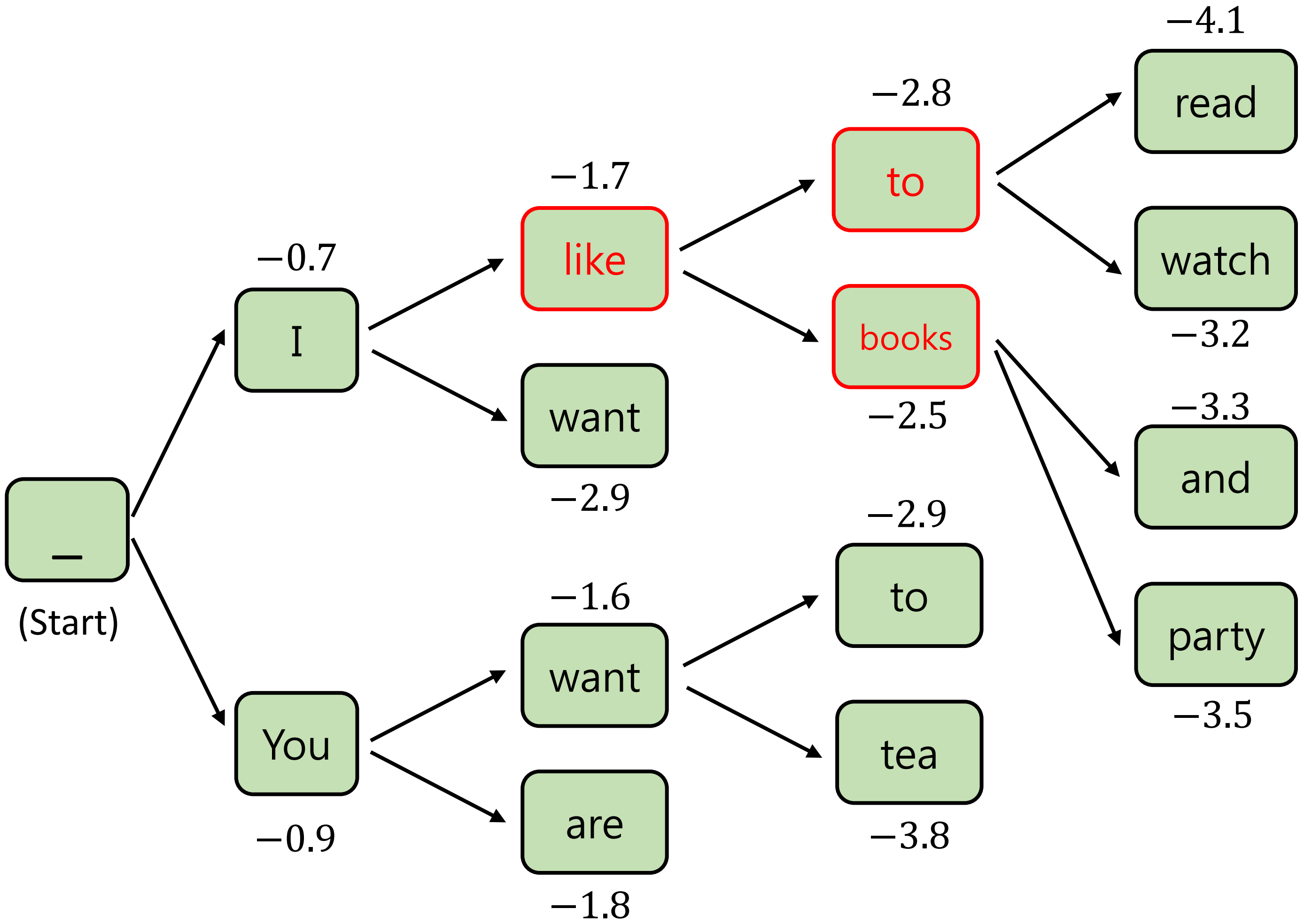
Beam Search ($t = 3$)

`Beam size` = 2



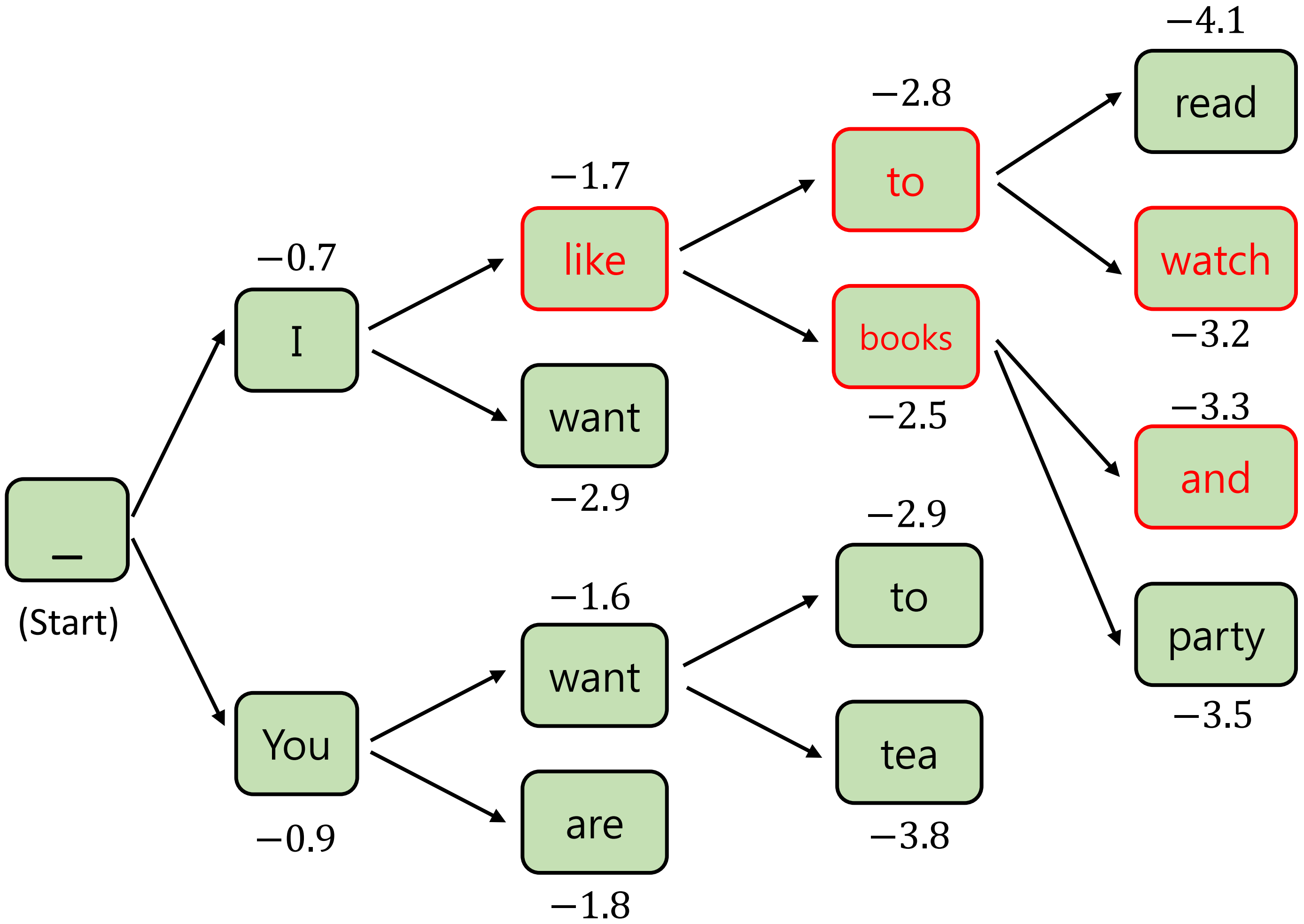
Beam Search ($t = 4$)

`Beam size` = 2



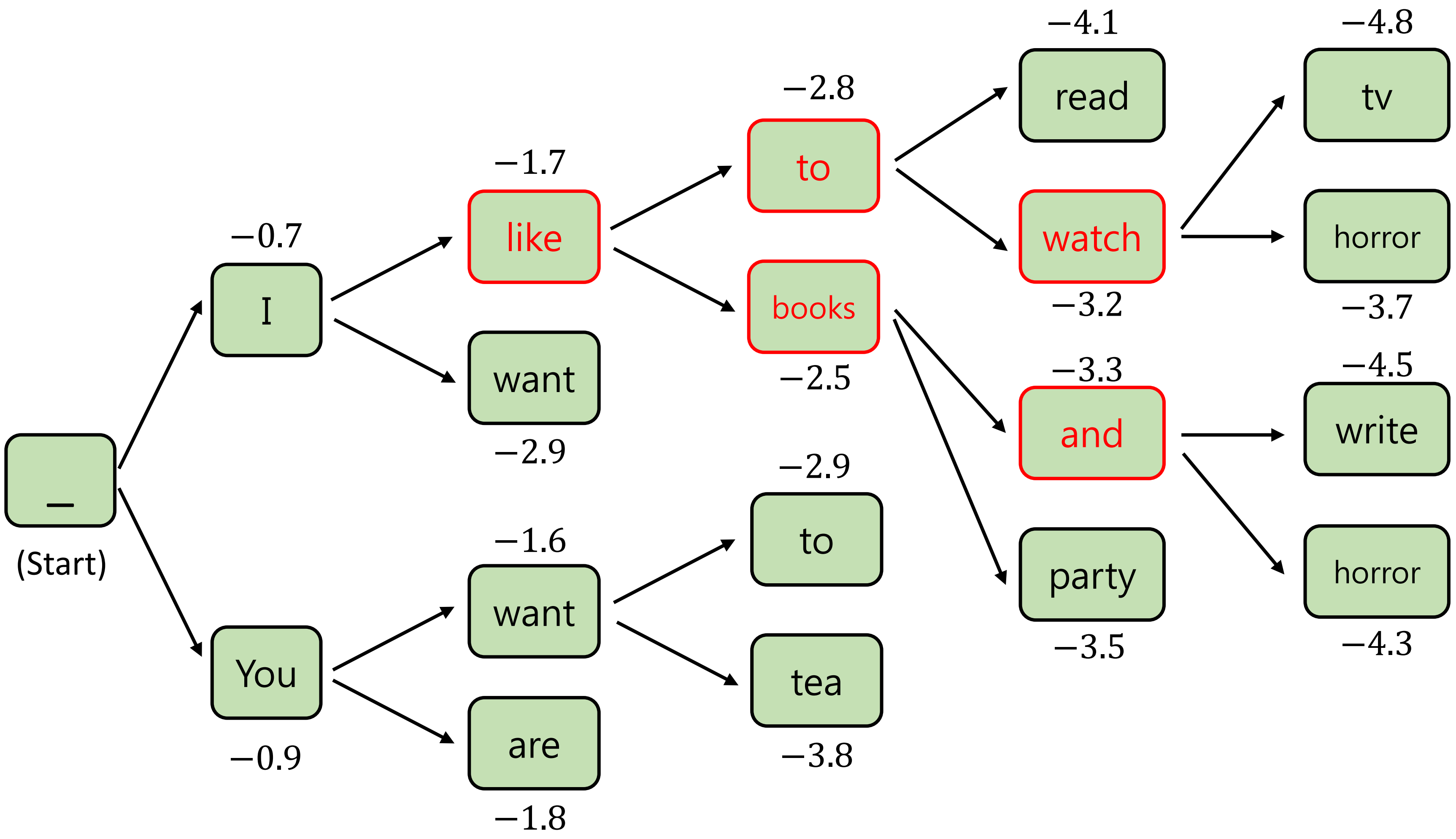
Beam Search ($t = 4$)

`Beam size` = 2



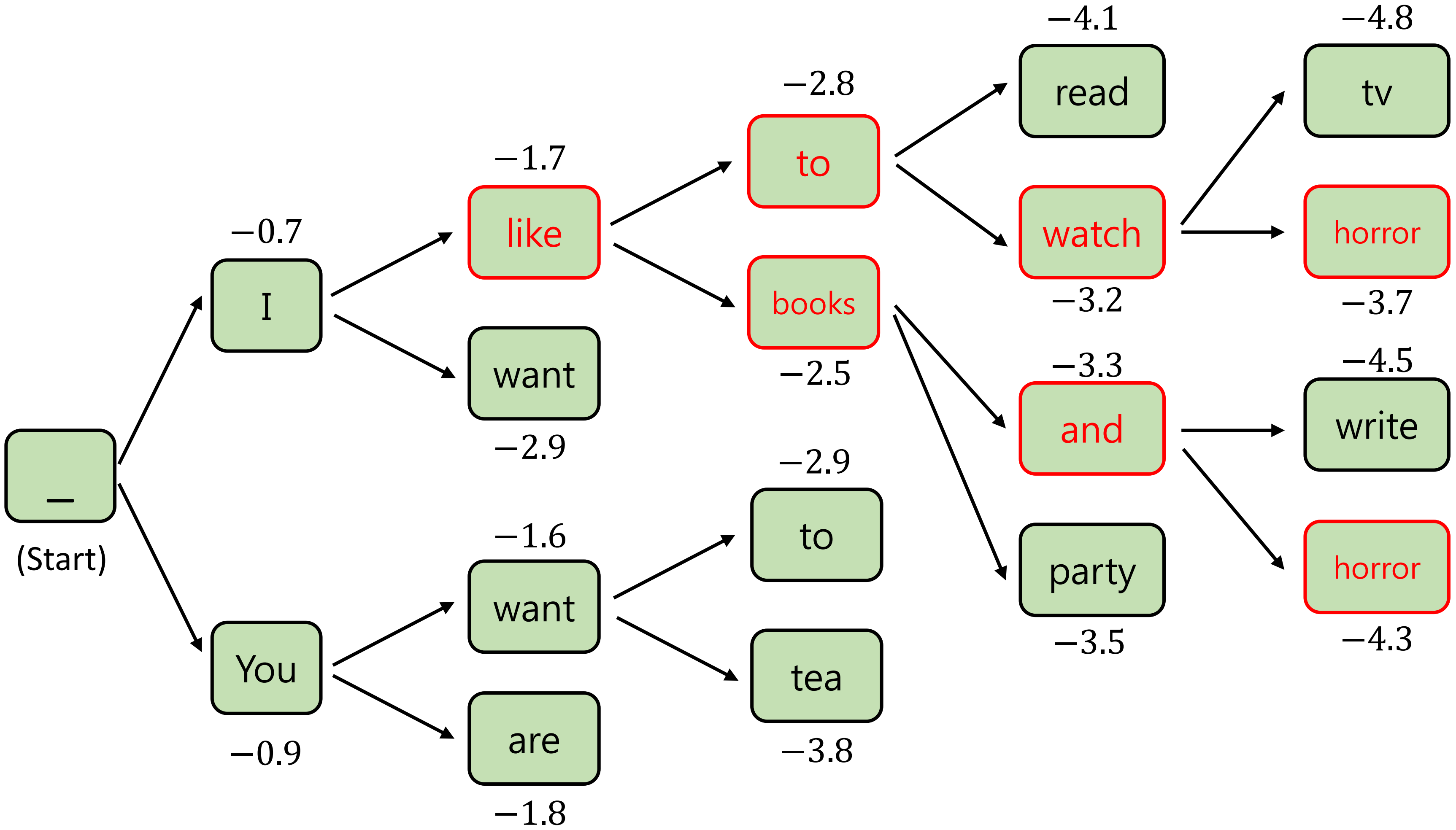
Beam Search ($t = 5$)

`Beam size` = 2



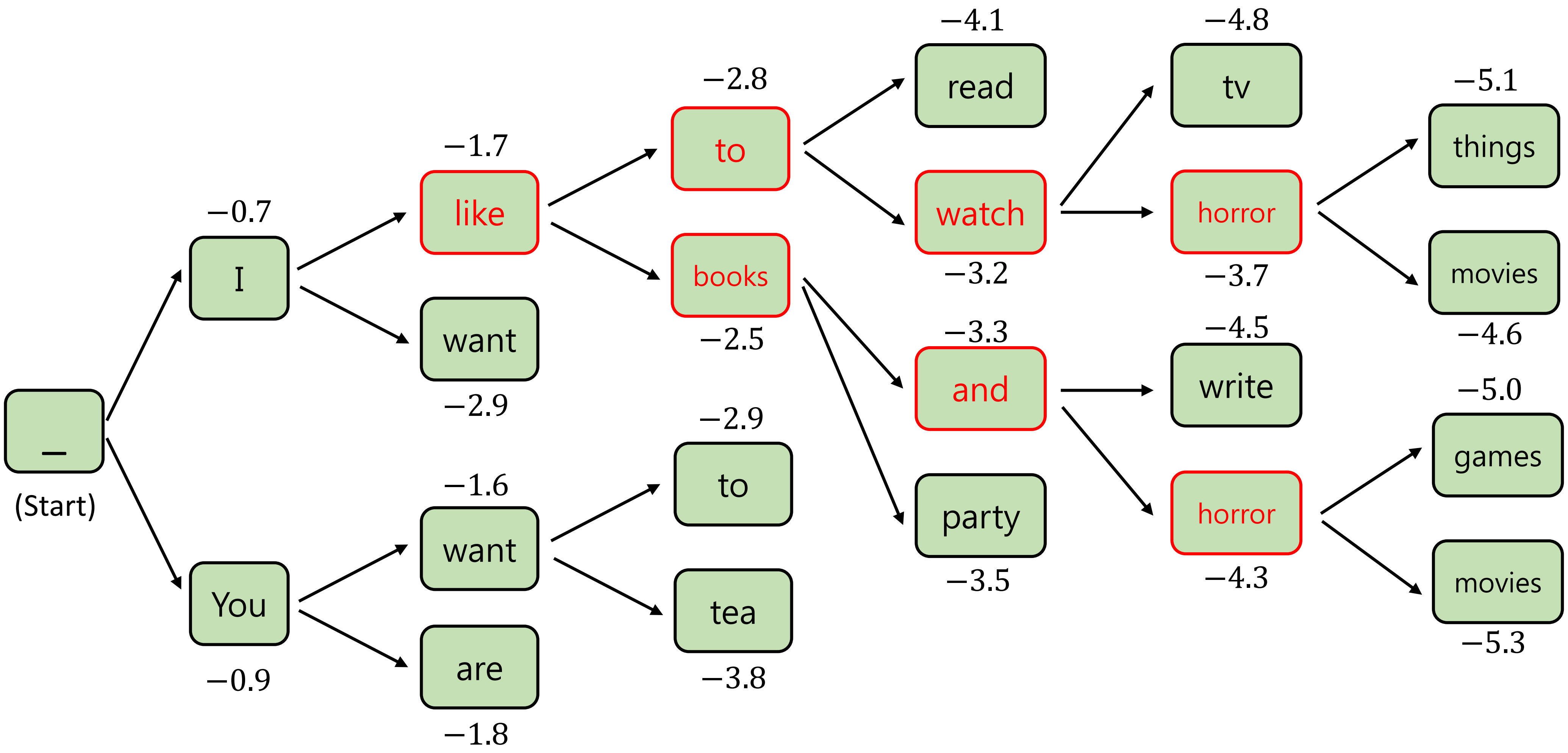
Beam Search ($t = 5$)

`Beam size` = 2



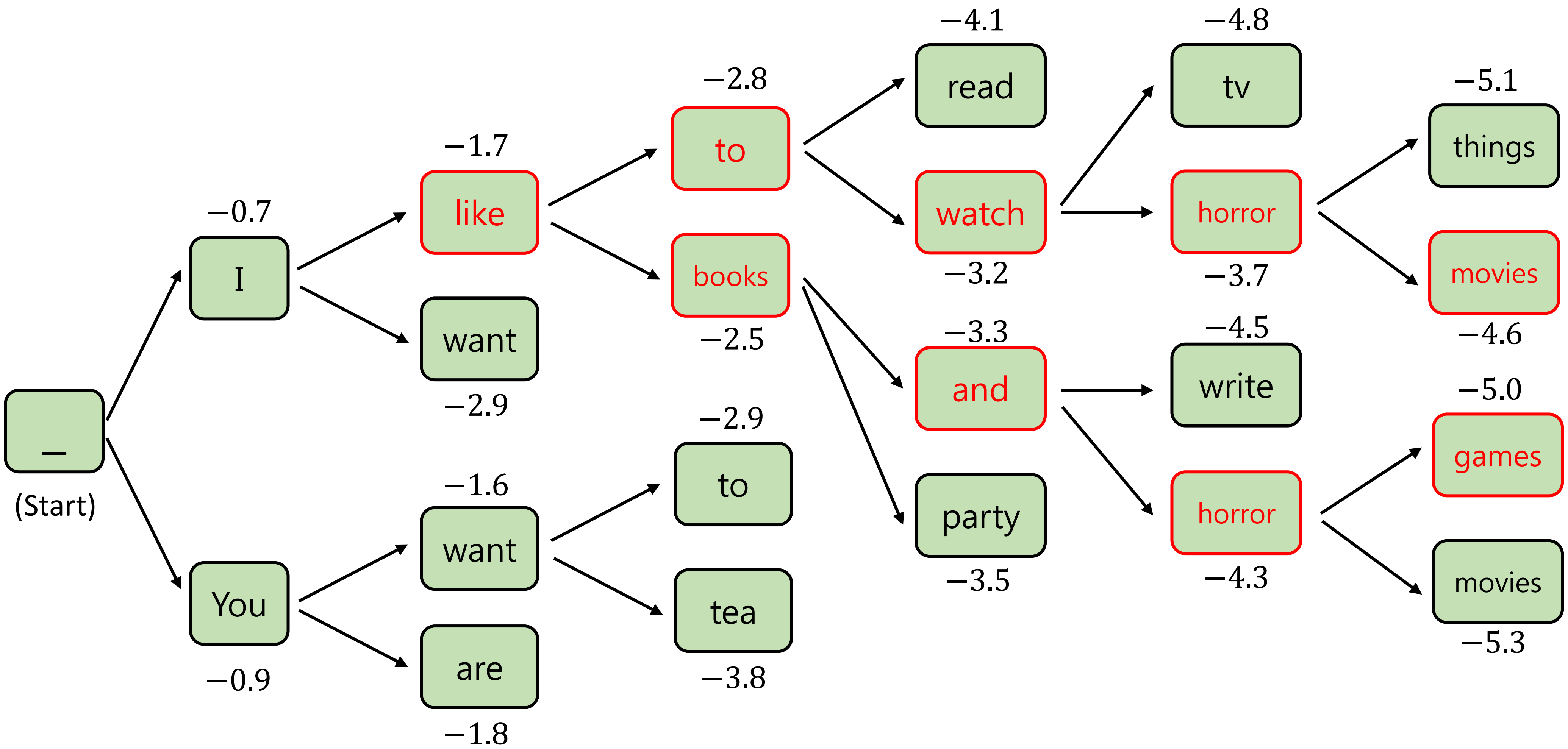
Beam Search ($t = 6$)

`Beam size` = 2



Beam Search ($t = 6$)

`Beam size` = 2



Stop Criterion

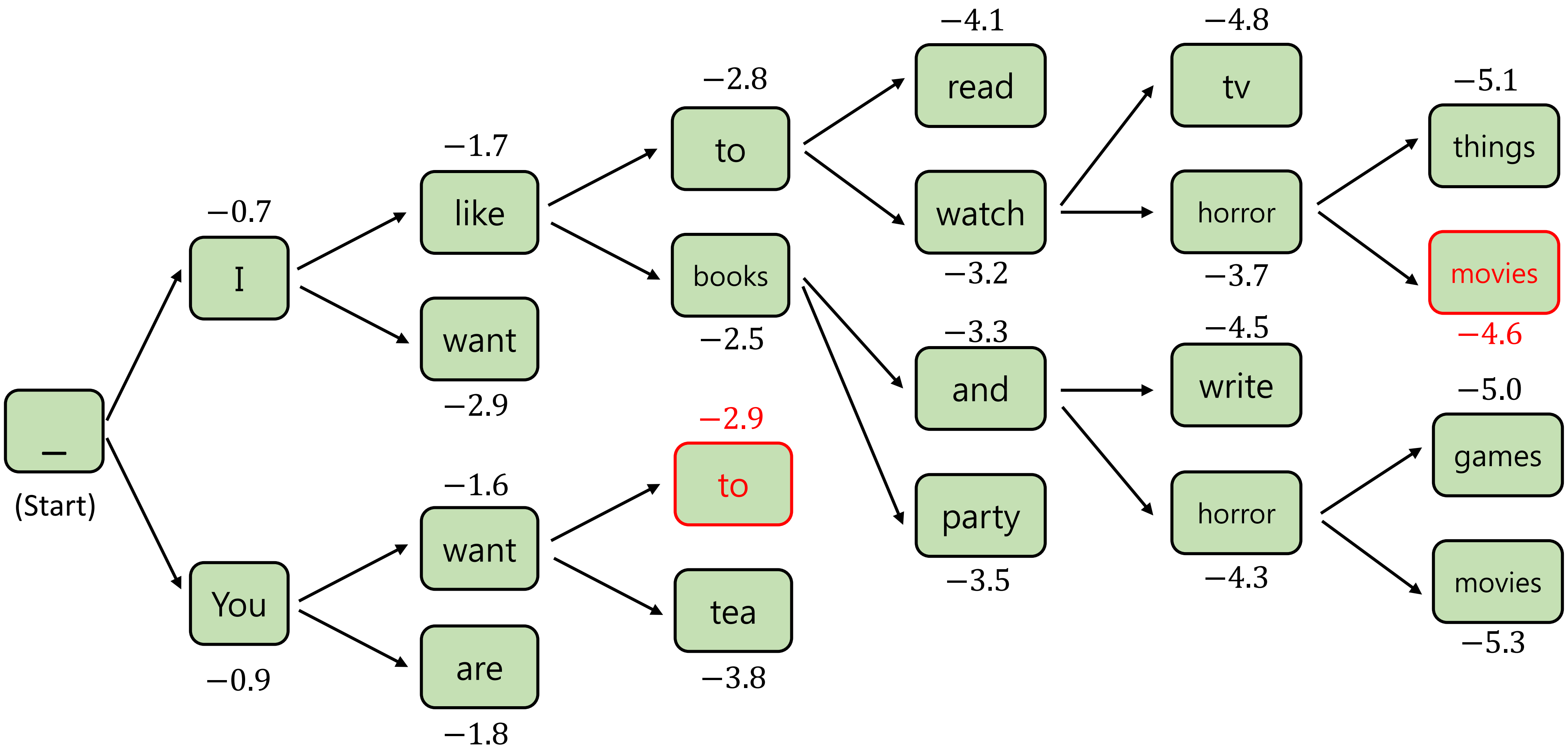
- There are two common stop criteria, either for greedy decoding or beam search decoding:
  - We consider a sequence of generation complete when the <EOS> token is produced by a model. *<EOS>: End of sequence
    - E.g., <Start> I like to watch horror movies <EOS>
  - A generated sequence reaches a pre-defined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

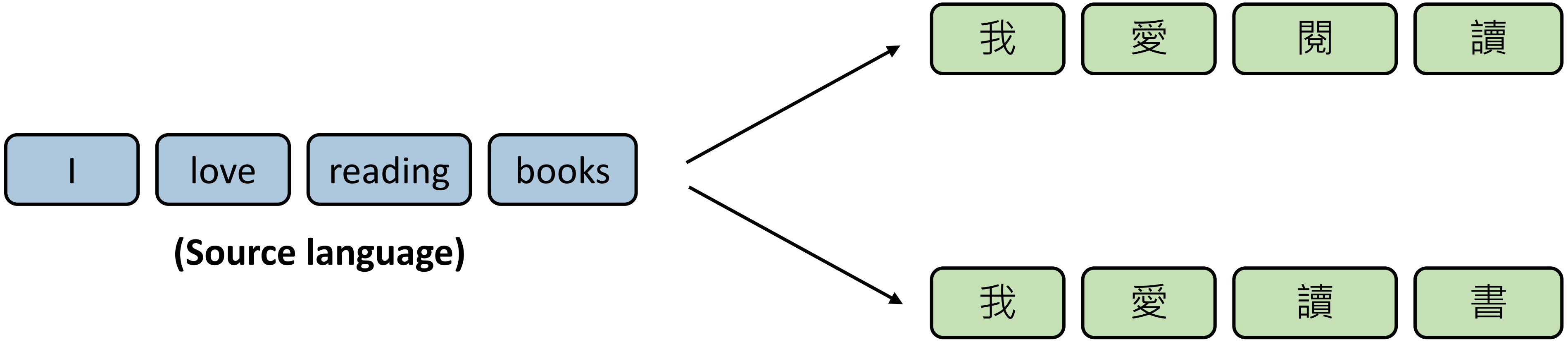
- Longer candidates will have lower scores.
- Solution: Perform normalization to penalize on length

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

How to evaluate natural language generation?

- Natural language is hard to evaluate due to <u>subjectivity</u> and language <u>diversity</u>.

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.

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.

~~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	$\text{Count}_{\text{clip}}$
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)

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,

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

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.

ROGUE: A Package for Automatic Evaluation of Summaries

- ROGUE-N: N-gram Co-Occurrence Statistics (recall base)
- ROGUE-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)

$\text{ROGUE-2}_{S2} = \text{ROGUE-2}_{S3}$
$\text{ROGUE-L}_{S2} = 3/4 > \text{ROGUE-L}_{S3} = 1/2$

S4. the gunman police killed

$\text{ROGUE-2}_{S4} > \text{ROGUE-2}_{S2}$
$\text{ROGUE-L}_{S4} < \text{ROGUE-L}_{S2}$

(Recap) Perplexity

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

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

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

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.