Neural machine translation

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING
UNIVERSITY

Neural Machine Translation (NMT)

- **Neural machine translation (NMT)** models the translation task through a single artificial neural network.

- The first systems for NMT were based on recurrent neural networks; more recent systems typically use Transformers.

- Many practical implementations are based on the OpenNMT ecosystem for neural machine translation.

Link to OpenNMT

The sequence-to-sequence model (seq2seq)

The sequence-to-sequence model consists of two components:

- The **encoder** is a neural network that produces a representation of the source sentence.

  typically implemented as a bidirectional recurrent neural network

- The **decoder** is an autoregressive language model that generates the target sentence, conditioned on the output of the encoder.

  autoregressive = takes its own outputs as new inputs
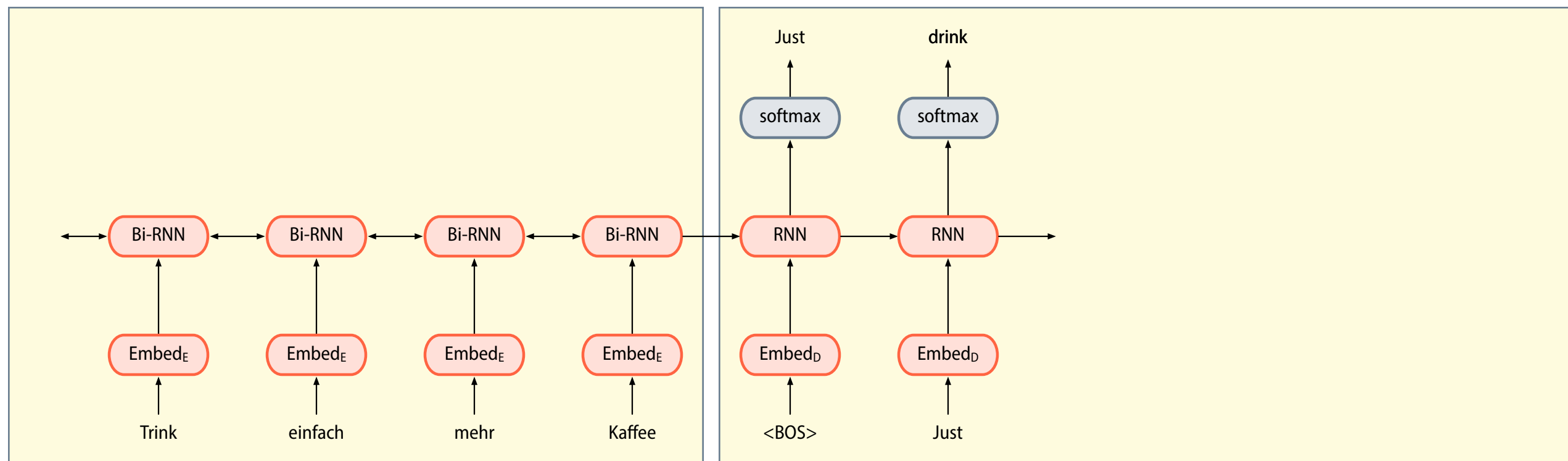
Standard seq2seq architecture
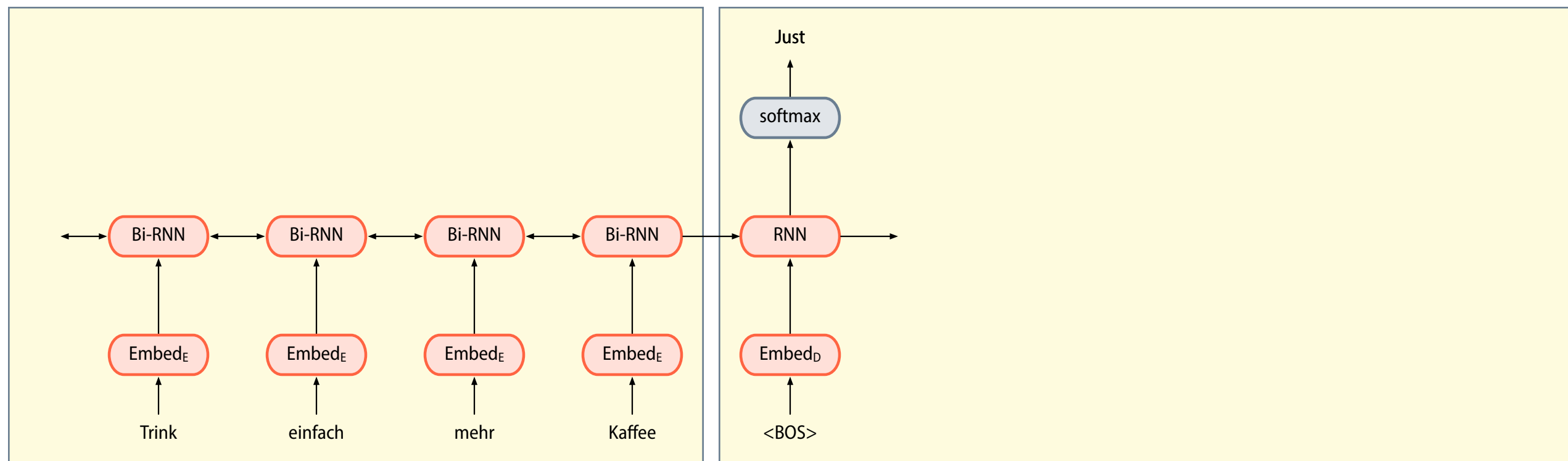
encoder   decoder
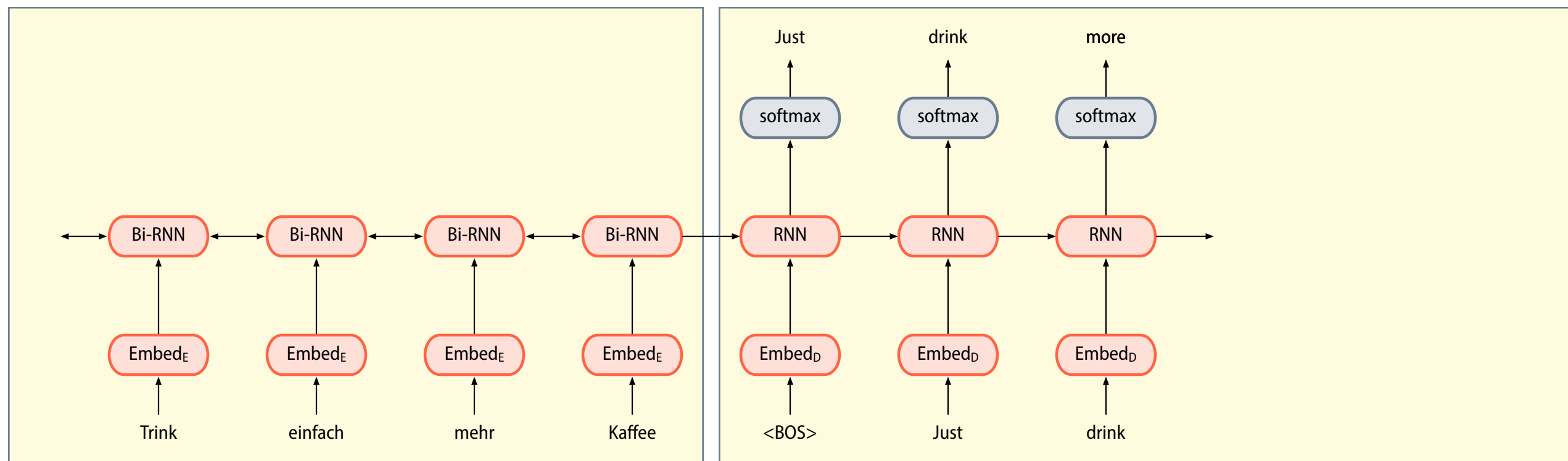
Sutskever et al. (2014)

Standard seq2seq architecture

encoder   decoder

Sutskever et al. (2014)

Standard seq2seq architecture
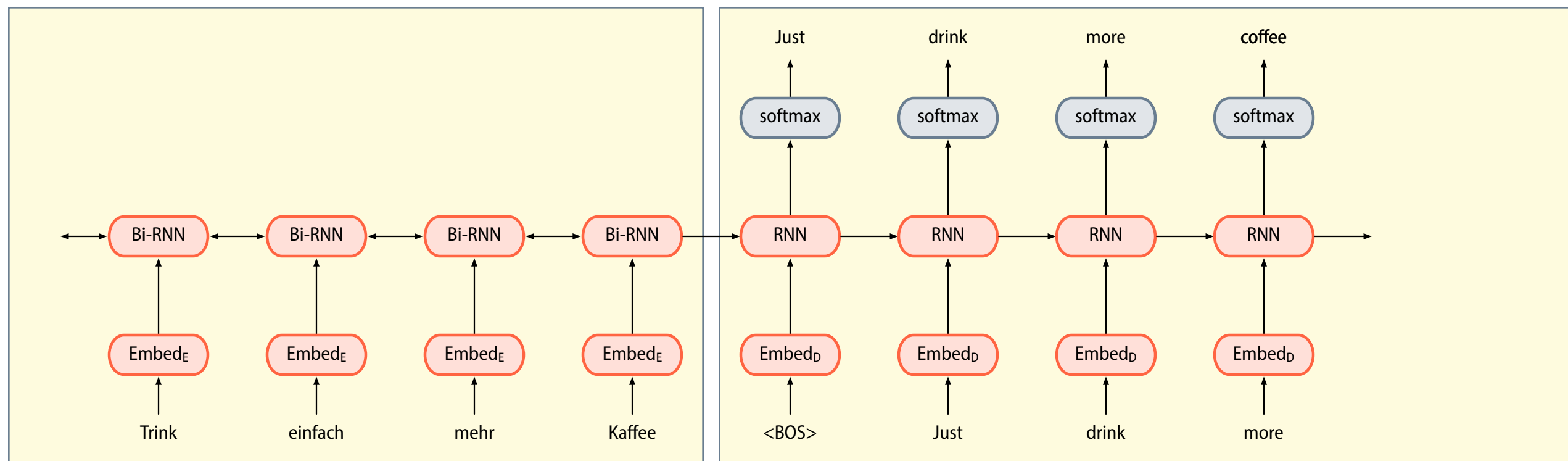

encoder   decoder

Sutskever et al. (2014)

Standard seq2seq architecture

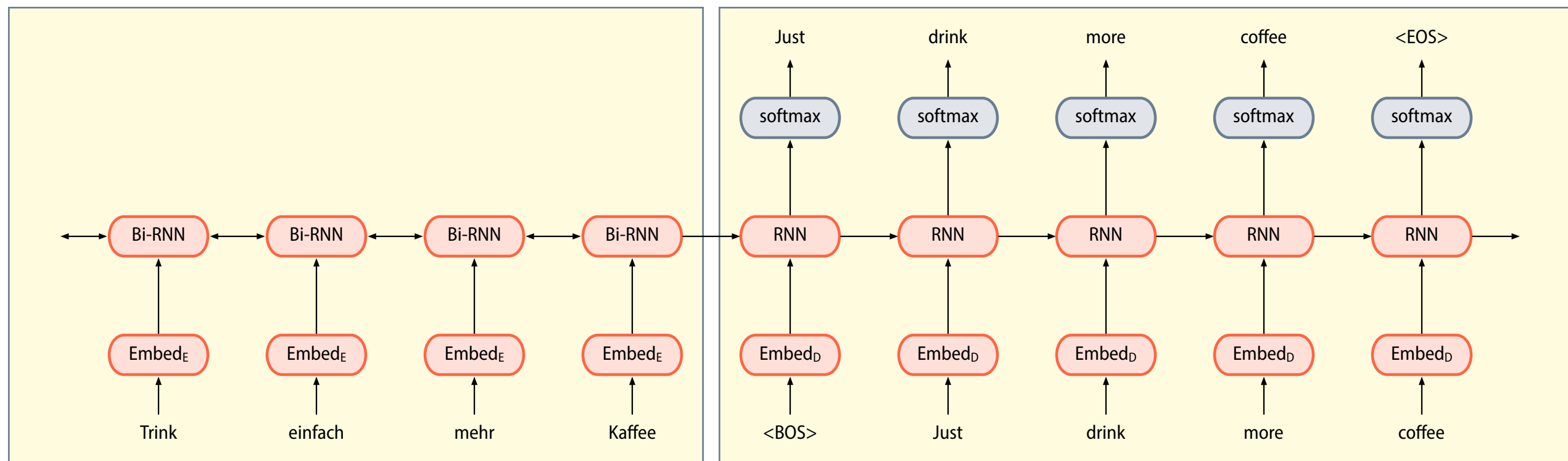

encoder   decoder

Sutskever et al. (2014)

Standard seq2seq architecture

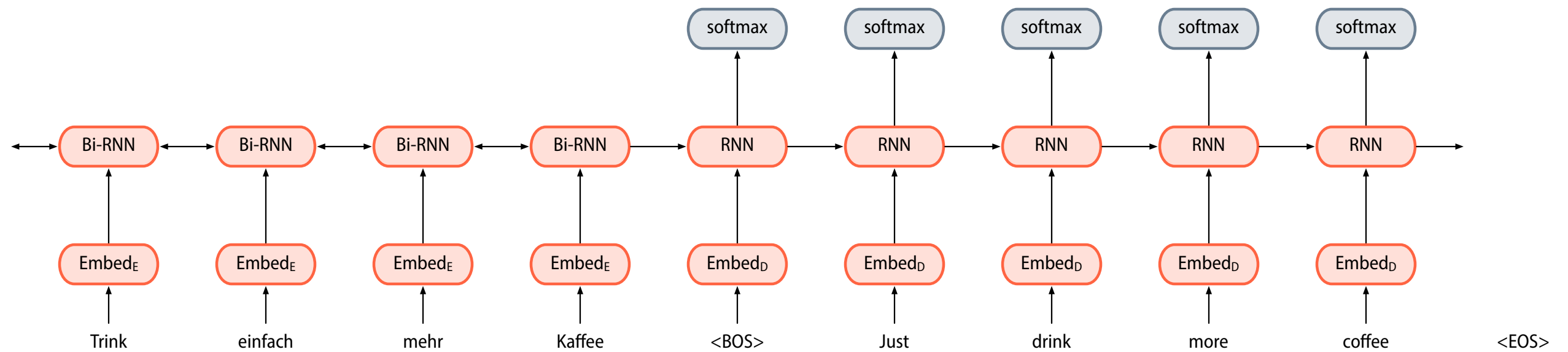encoder   decoder

Sutskever et al. (2014)

Properties of the seq2seq model

- The seq2seq model directly learns and uses $P(y|x)$, rather than decomposing it into $P(x|y)$ and $P(y)$ as in SMT.

- The model can be trained trained end-to-end using backpropagation, without alignments or auxiliary models.
  only needs parallel data

- The seq2seq model is useful for a range of other tasks, including text summarisation, dialogue, and code generation.

Training an encoder–decoder model

Training an encoder–decoder model
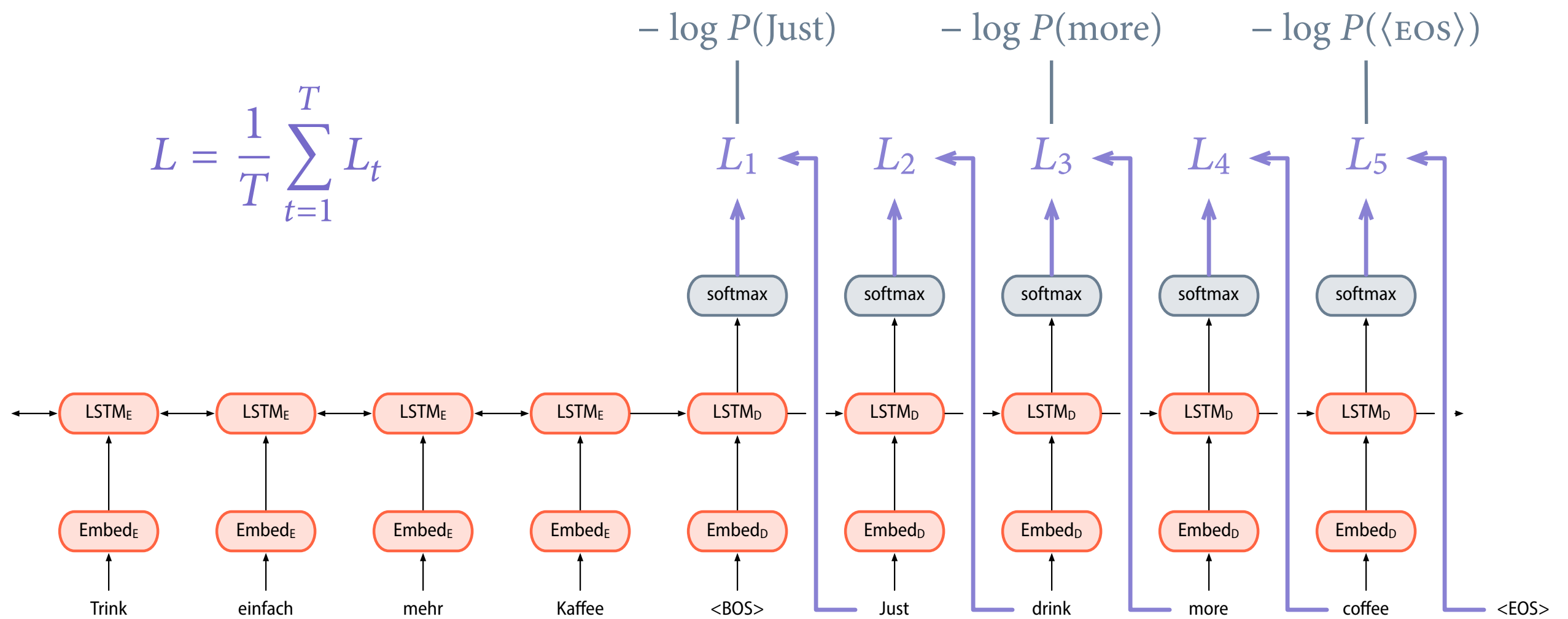
Decoding algorithms

- **Greedy decoding**

  At each step, predict the highest-probability word. Stop when the end-of-sentence marker is predicted.

- **Beam search**

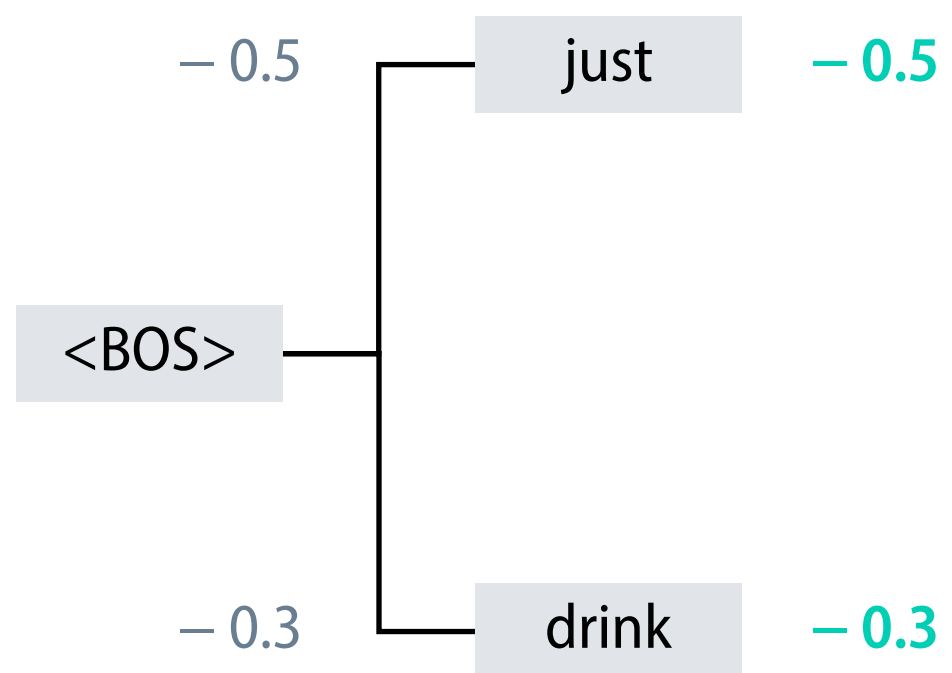 Keep a limited number of highest-scoring partial translations. Expand the items on the beam, score the new items, and prune.

  Typical beam widths are between 2 and 16.
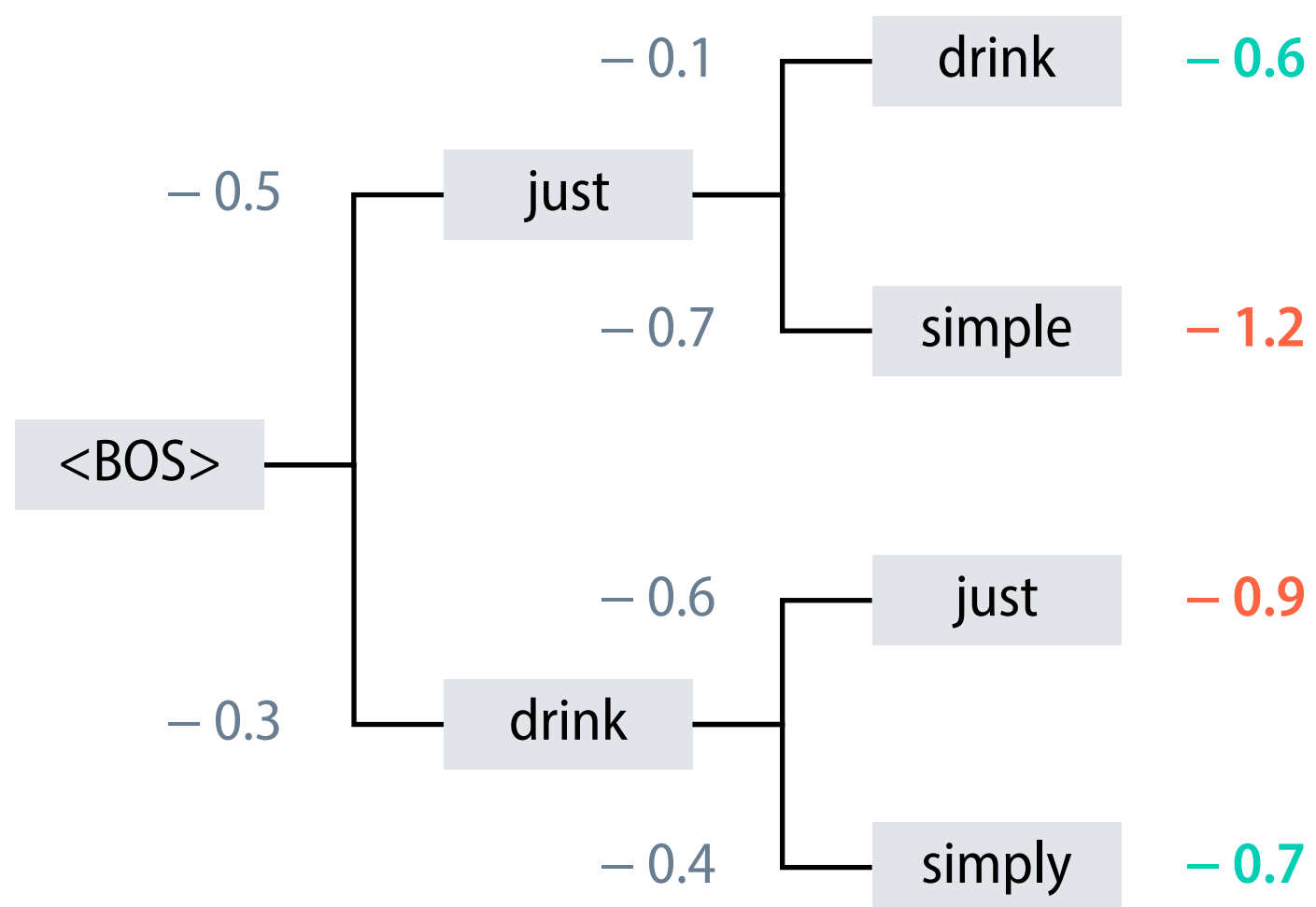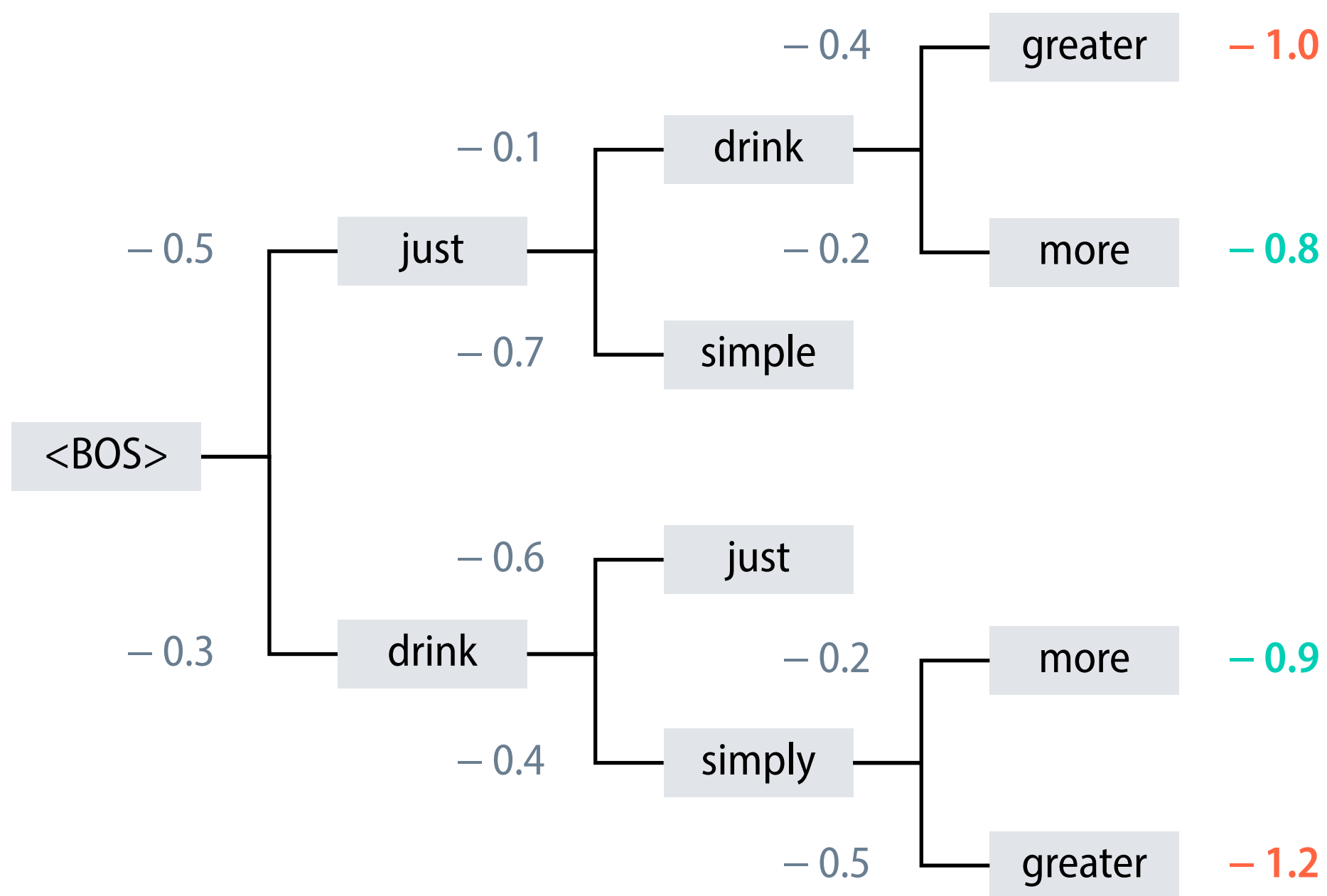
Beam search example

<BOS>

Beam search example
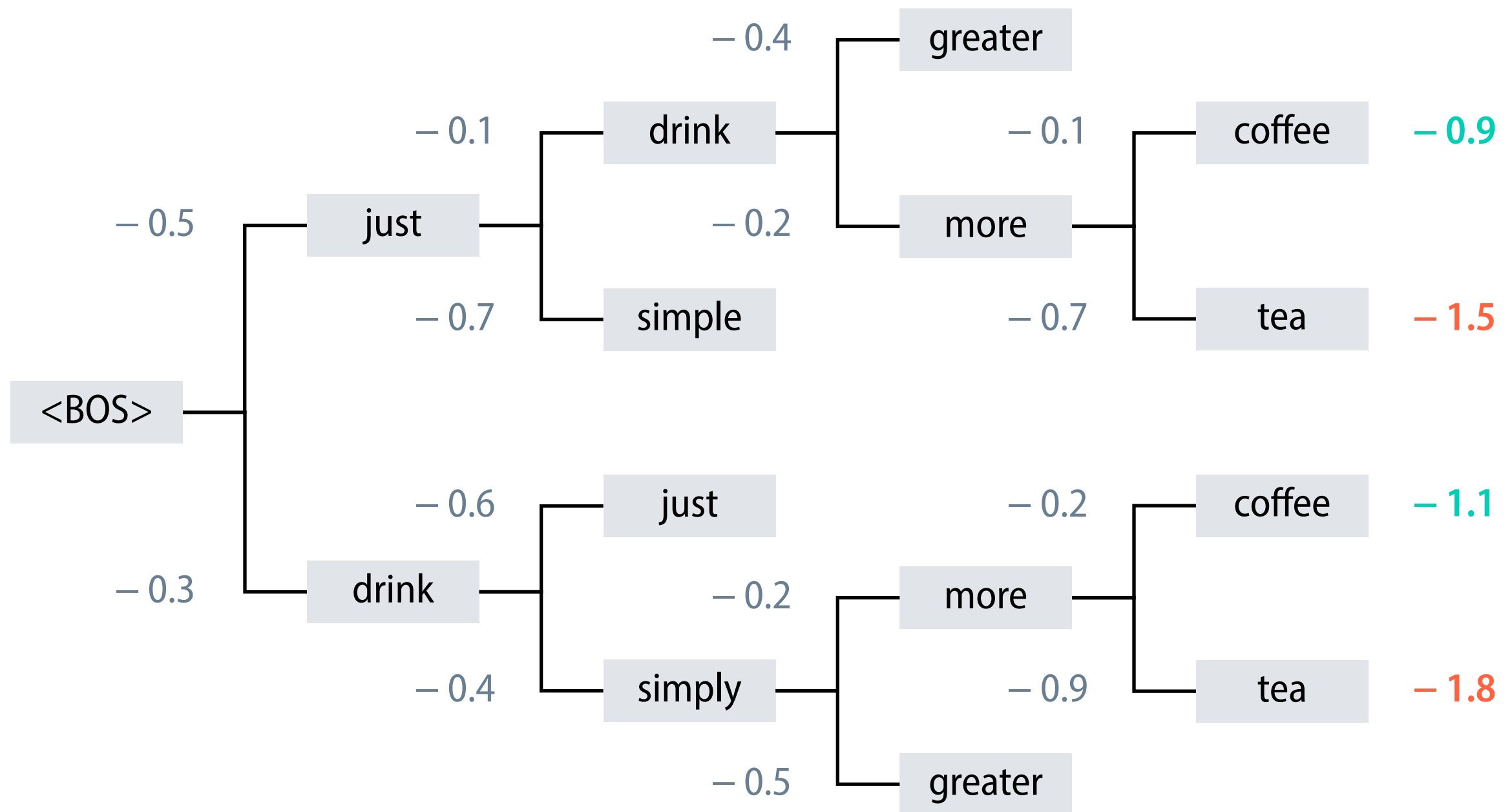
Beam search example

Beam search example
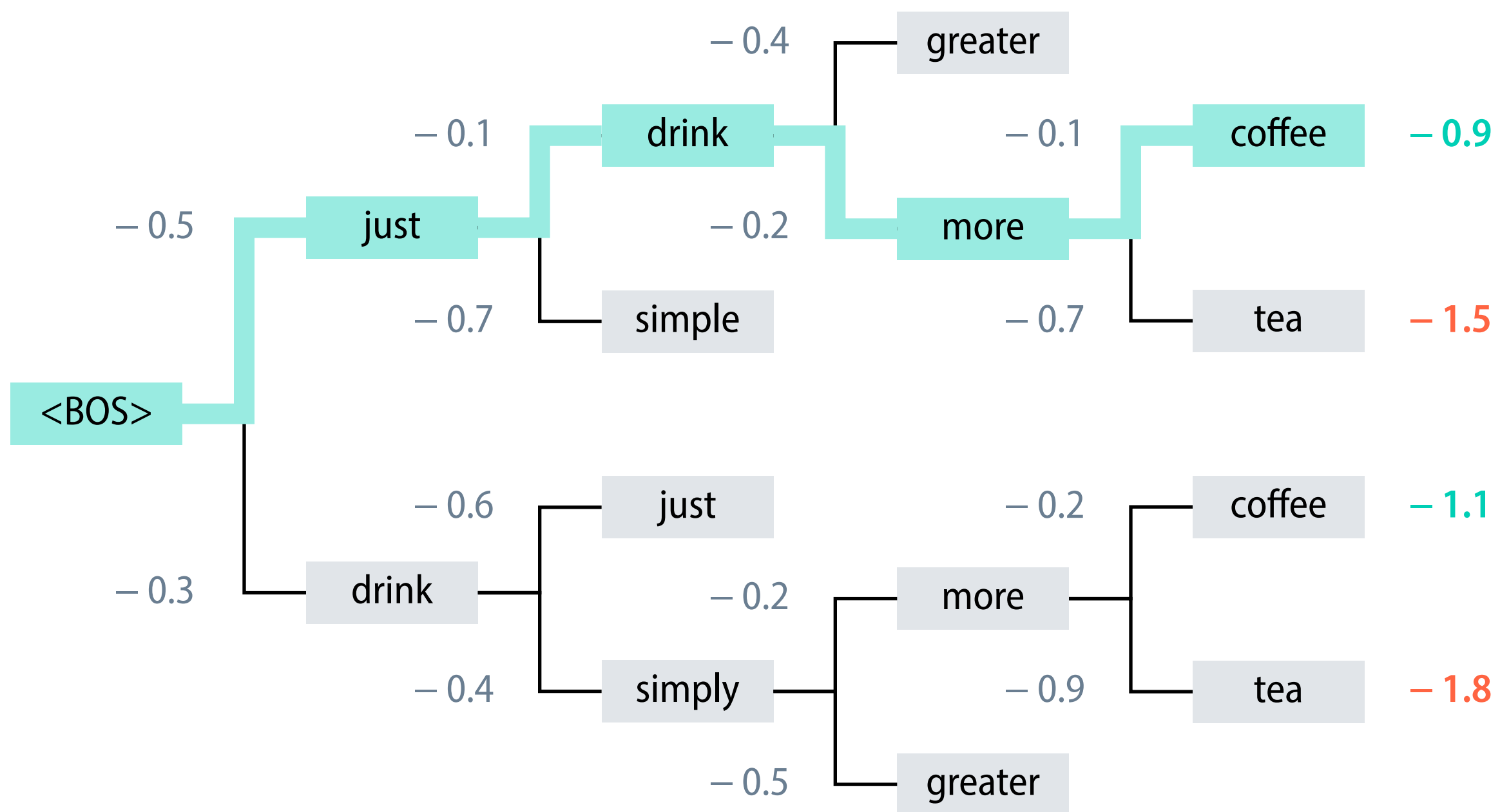
Beam search example

Beam search example

Termination criteria

- When the expansion of a partial translation generates the ⟨EOS⟩ marker, store the result as a complete translation.

- End the search after a fixed number of steps, or when enough complete translations have been generated.

- Evaluate the translations found during search based on their length-normalised scores and return the highest-scoring one.

  different from standard beam search