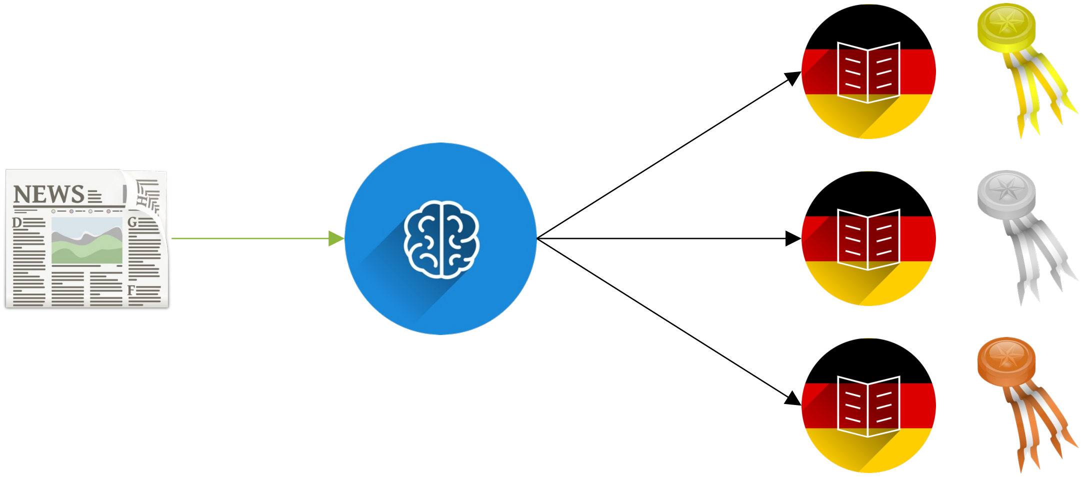
Karlsruher Institut für Technologie

10 - Decoding

Encoder – Decoder: Translation

- How do we generate the translation ?

Machine Translation - Translation



- Search for possible translations
- Model assigns score to each translation
- Find most probable translation

Overview

Karlsruher Institut für Technologie

- Search Problem
- Search Algorithms
- Model/Search errors
- Modeling combination
  - Ensemble
  - Reranking

Encoder – Decoder: Translation

- How do we generate the translation
  - Search for the most accurate translation:

$$y^* = argmin_y E(y, \bar{y})$$

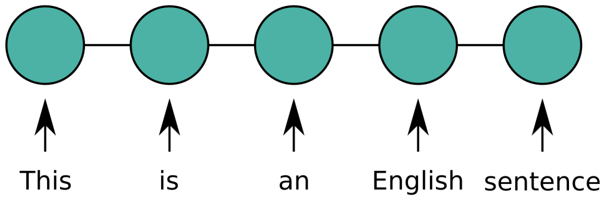
    - At translation time, we don't have the reference $\bar{y}$

  - Search for the most probable translation:

$$y^* = argmax_y P(y|x)$$

Basic search

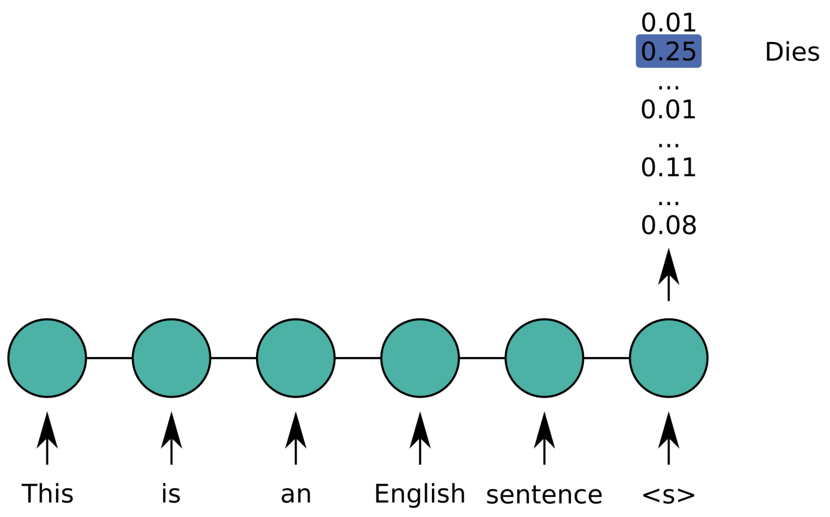
- Input source sentence
  - Forward pass



Basic Search

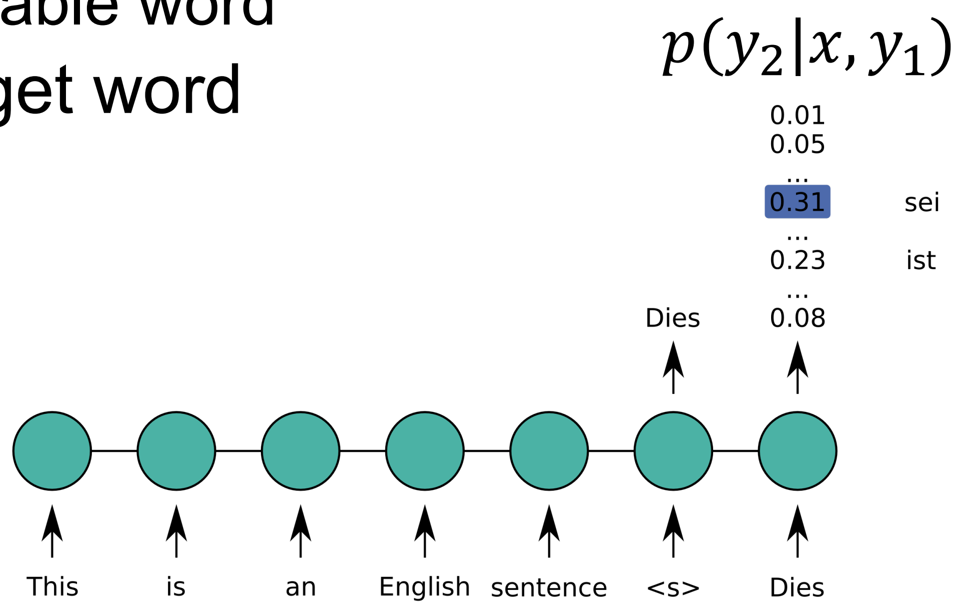
- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word

$$p(y_1|x)$$



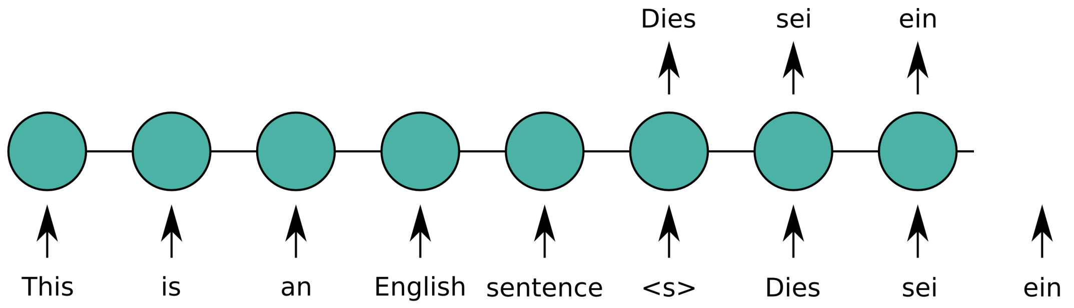
Basic search

- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word
- Input selected target word



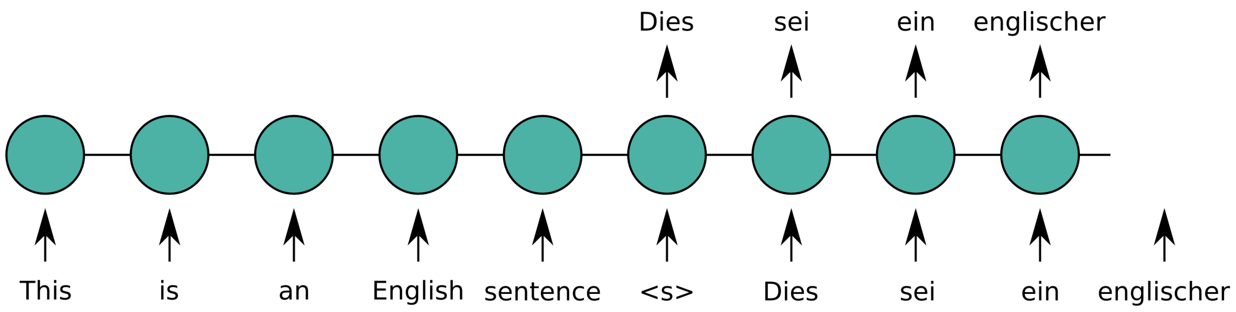
Basic search

- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word
- Input selected target word
- Continue



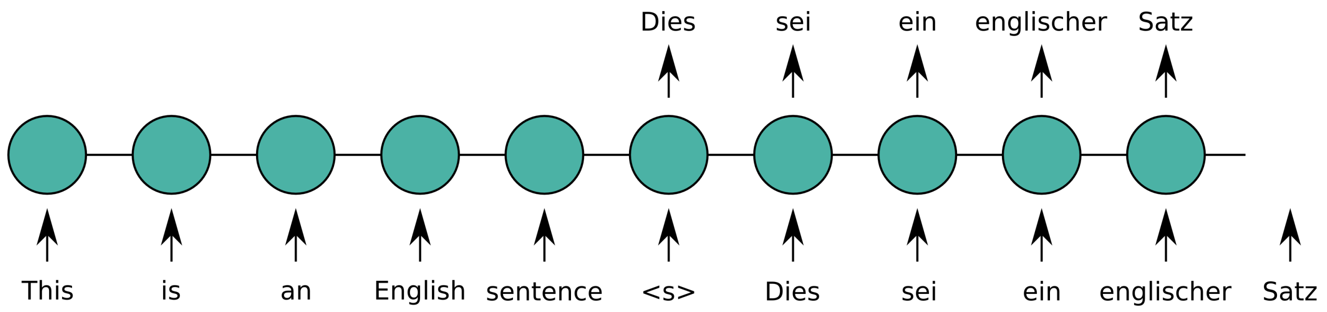
Basic search

- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word
- Input selected target word
- Continue



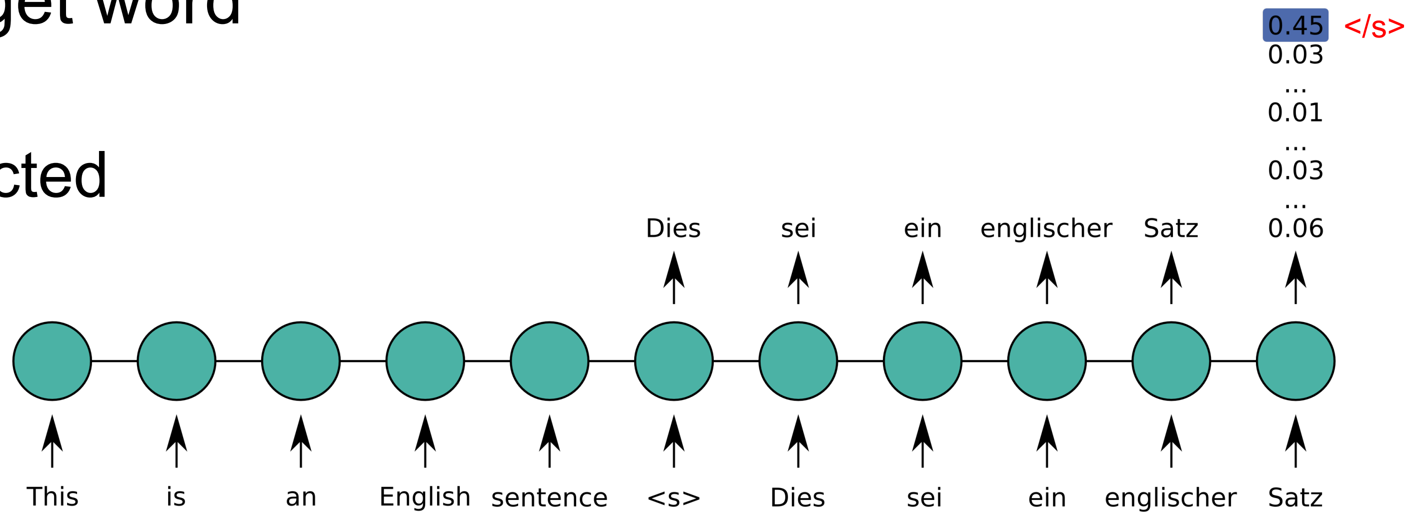
Basic search

- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word
- Input selected target word
- Continue



Basic search

- Input source sentence
  - Forward pass
- Input <s>
  - Calculate output probabilities
  - Select most probable word
- Input selected target word
- Continue
  - Until </s> is selected



Search strategies

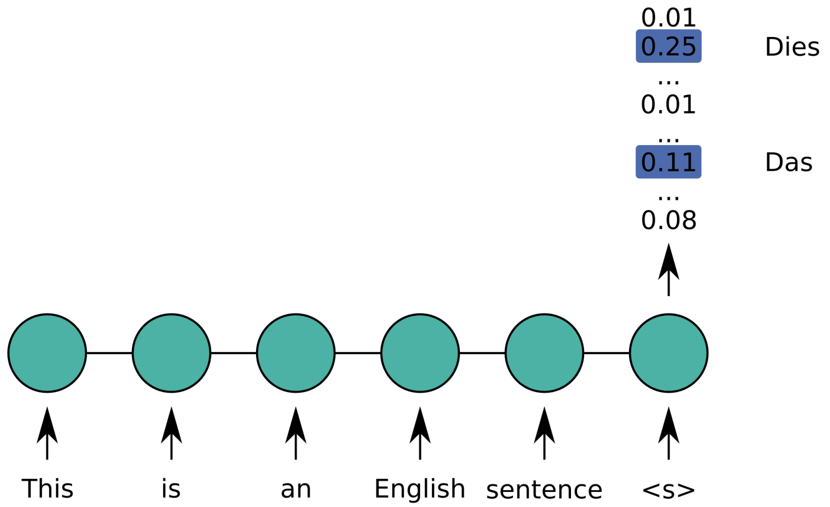
- Greedy search
  - Always select best target word
  - Problem:
    - Autoregressive model: Output influences input

Greedy search - Challenge

- First word:

$$p(y_1|x)$$

Hypothesis	
Dies	0.25
Das	0.11



Greedy search - Challenge

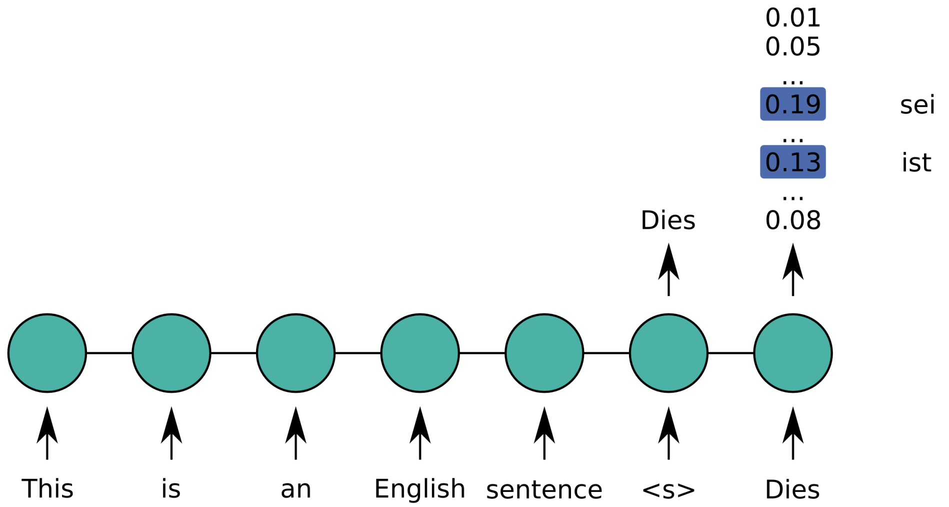
- First word:

$$p(y_1|x)$$

- Second word:

$$p(y_2|x, y_1)$$

Hypothesis	
Dies sei	0.0475
Dies ist	0.0325



Greedy search - Challenge

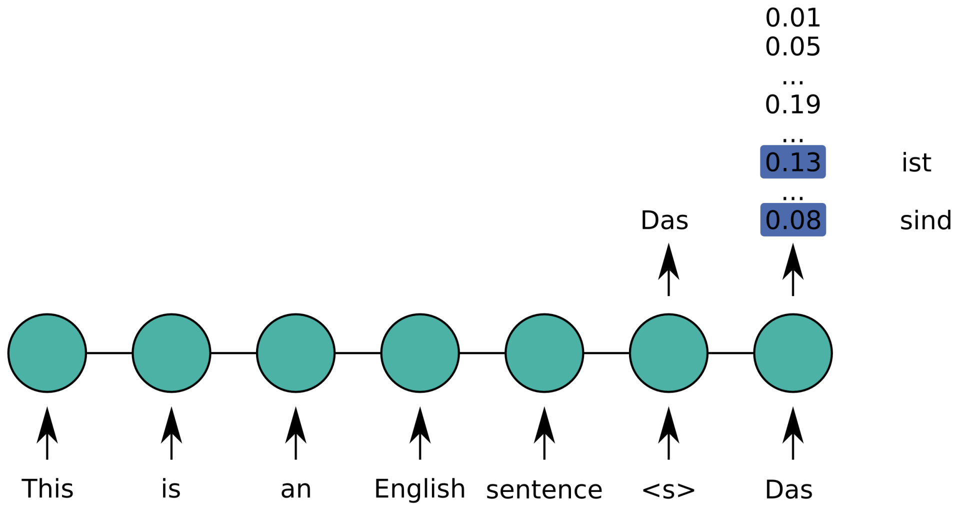
- First word:

$$p(y_1|x)$$

- Second word:

$$p(y_2|x, y_1)$$

Hypothesis	
Das ist	0.0539
Dies sei	0.0475
Dies ist	0.0325
Das sind	0.0187

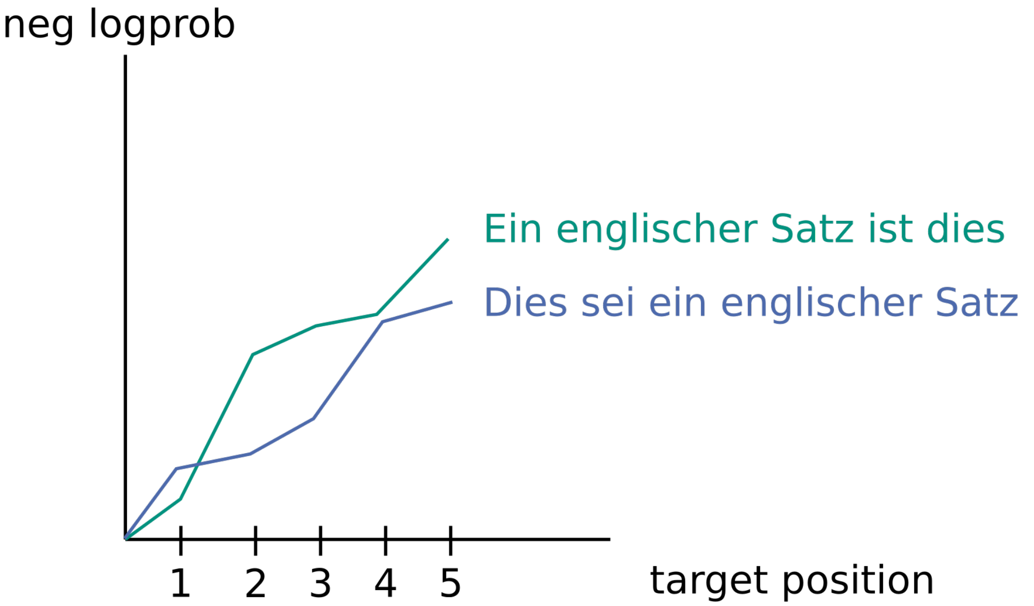


Greedy search - Challenge

- Greedy search
  - Always select best target word
  - Problem:
    - Might not find most probable sentence

- Sentence probability:

$$p(e \mid f) = \prod_{j=1}^{n} p(e_j \mid f, e_1^{j-1})$$



Search strategies

- Greedy search
- Exact search
  - Try all combinations

Search strategies

- Greedy search
- Exact search
  - Try all combinations
  - Maximum a-posterior decoding
  - Challenge:
    - $|V_t|^{|y|}$ combinations
    - Only possible for very short sentences

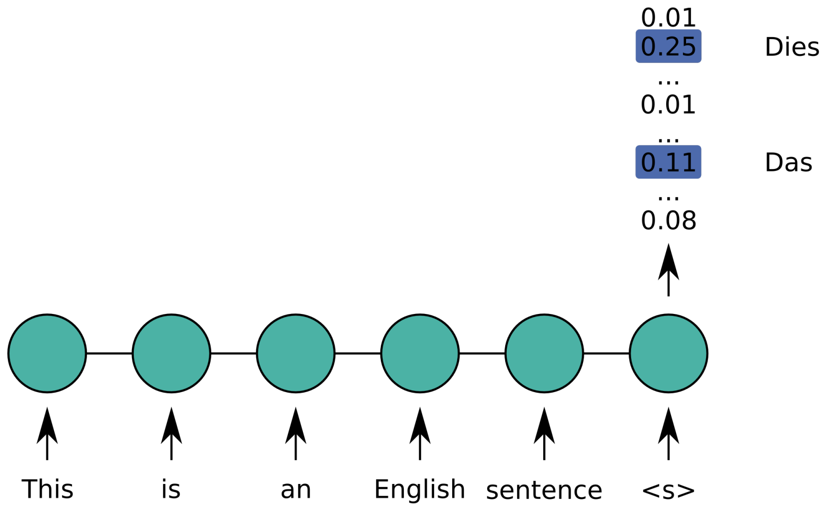
$$y^{\mathrm{MAP}} = \operatorname*{arg\,max}_{h \in \mathcal{Y}} \; p_{Y|X}(h|x, \theta) \; .$$

Search strategies

- Greedy search
- Exact search
- Sampling
  - Basic Idea: Randomly select next words based on conditional probability

Sampling

- Randomly choose target word
  - Based on conditional probability
  - Draw uniform random number r between 0 and 1
  - Take word i with:
    - $\sum_{j=0}^{i-1} p(w_j|x,y) < r \le \sum_{j=0}^{i} p(w_j|x,y)$
  - Variation
    - Sample only from most probable k words



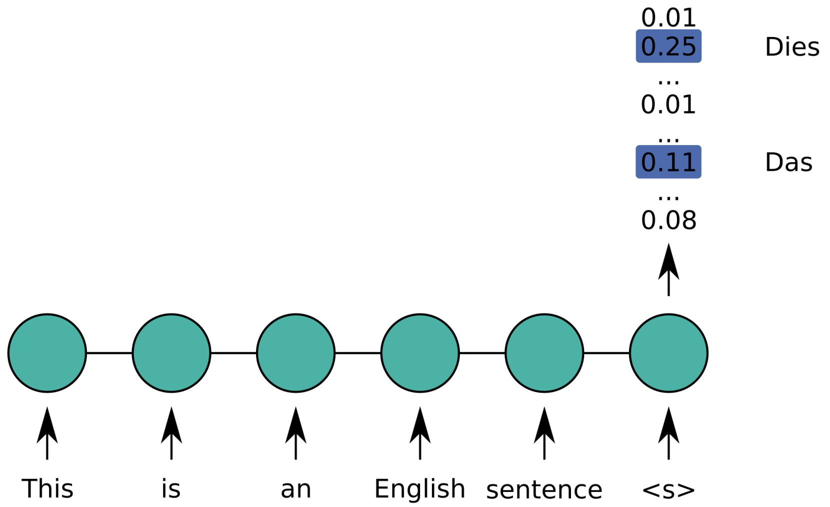
Search stragegies

- Greedy search
- Exact search
- Sampling
- Beam Search
  - Basic Idea: Keep the best n hypotheses
  - In NMT: Only small beam needed

Beam search

- Beam Search:
  - Calculate output probabilities
  - Select best n translations

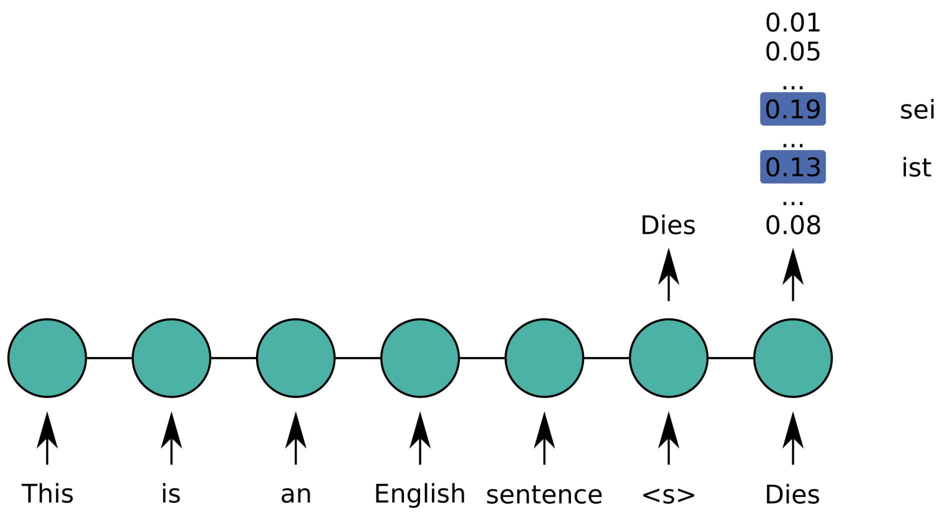
Hypothesis	
Dies	0.25
Das	0.11



Beam search

- Beam Search:
  - Calculate output probabilities
  - Select best n translations
  - Extend all hypothesis in beam

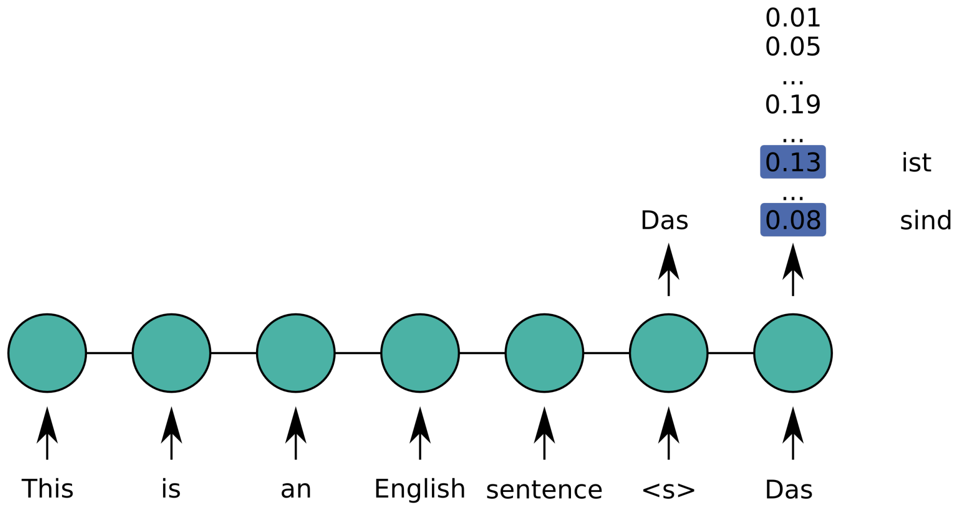
Hypothesis	
Dies sei	0.0475
Dies ist	0.0325



Beam search

- Beam Search:
  - Calculate output probabilities
  - Select best n translations
  - Extend all hypothesis in beam

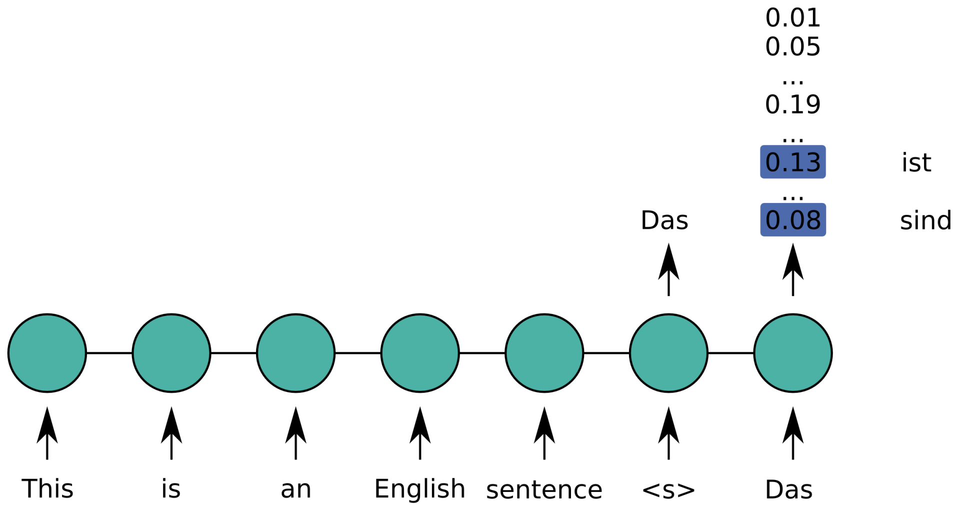
Hypothesis	
Das ist	0.0539
Dies sei	0.0475
Dies ist	0.0325
Das sind	0.0187



Beam search

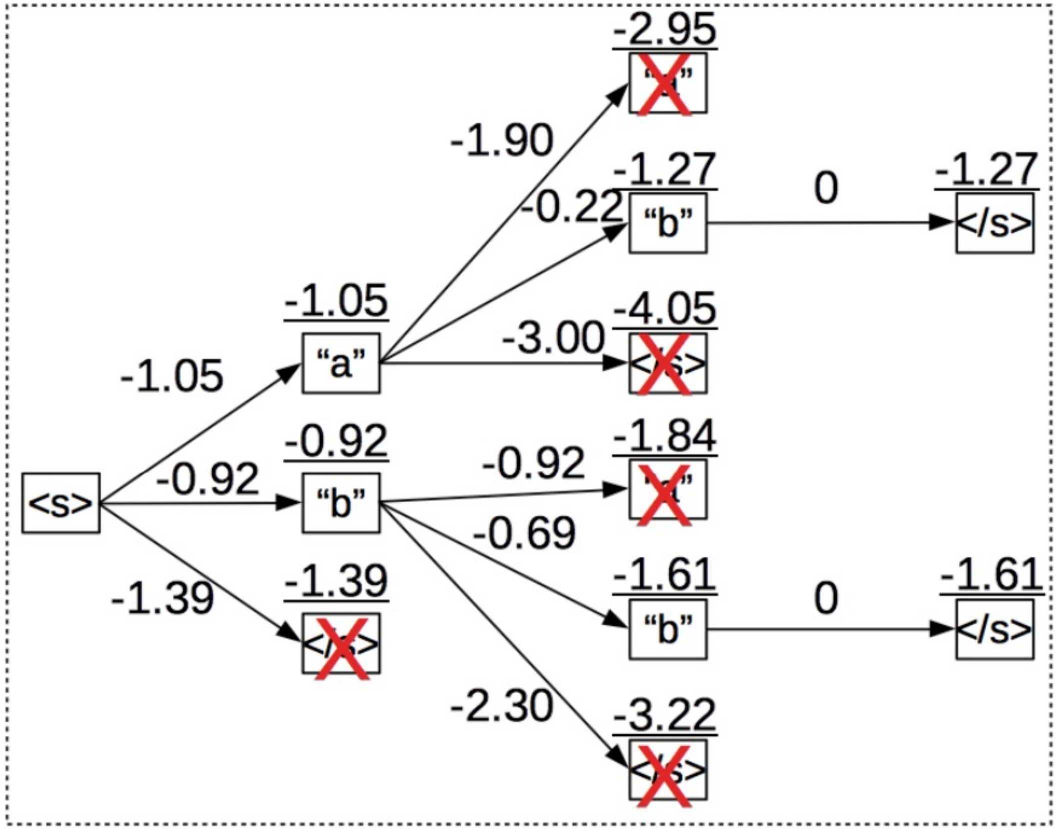
- Beam Search:
  - Calculate output probabilities
  - Select best n translations
  - Extend all hypothesis in beam
  - Prune hypothesis not in beam

Hypothesis	
Das ist	0.0539
Dies sei	0.0475



Beam search

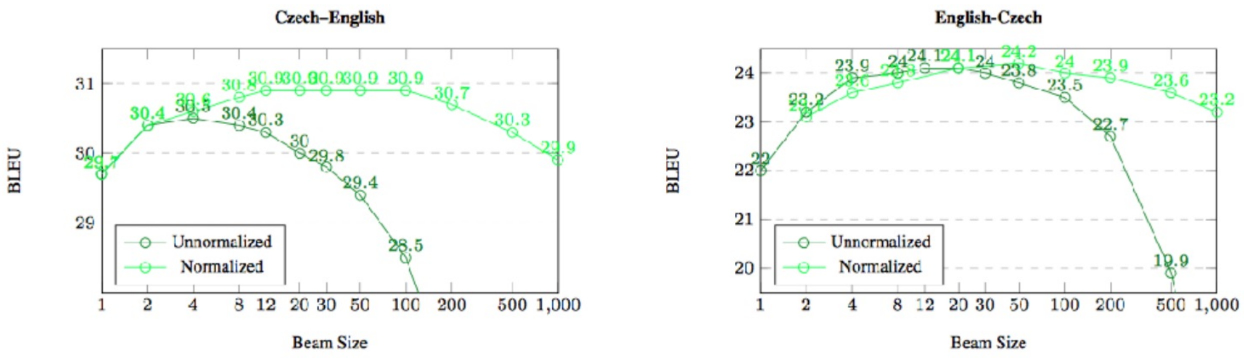
- Beam Search:



(from Neubig 2019)

Beam search

- Beam Search:
  - SMT:
    - Larger beam => larger search space => better score
    - Trade-off between quality and speed (n=300)
  - NMT:
    - Larger beam than an optimal number => more confusions
    - Beyond that optimal beam size (5-12), quality decreases
    - (Niehues et al., 2017): n<50 is sufficient, focus on modeling



(from Koehn and Knowles 2017)

Search

- Model error:
  - The model does not assign the highest probability to the correct translation

- Search error:
  - The search does not find the translation with the highest probability

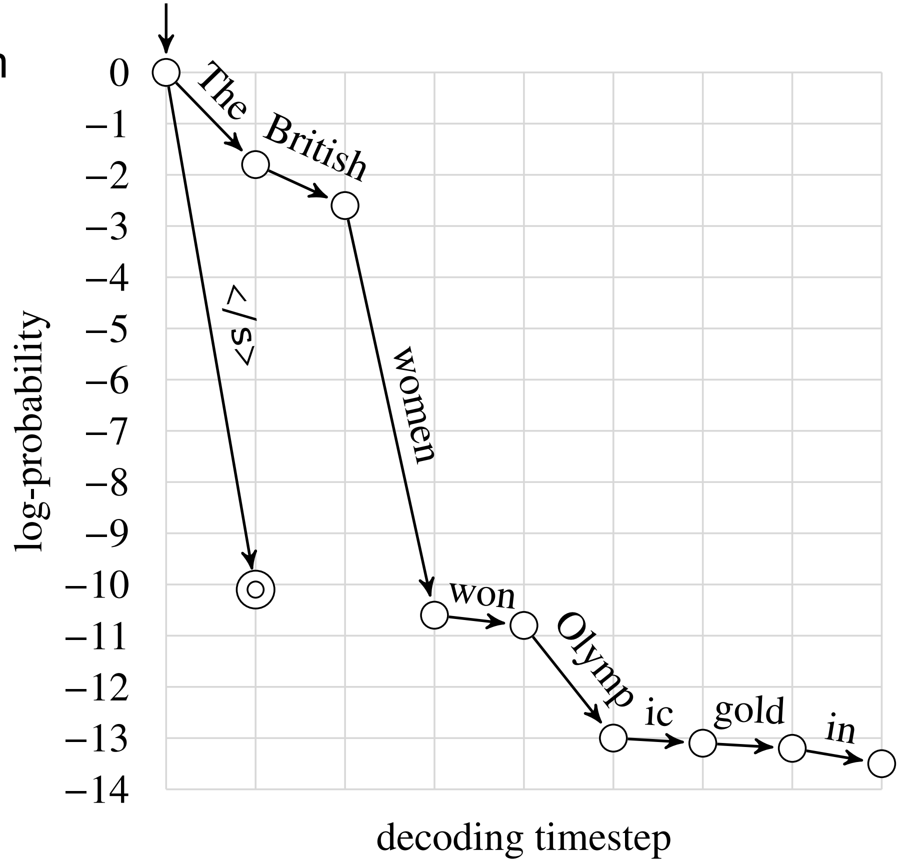
Label / Length bias

- Over-estimate probability of a prefix $y_1,..,y_m$
  - Multiply with conditional probabilities

$$p(y_{m+1}|x, y_{1,..,m})$$

  - No possibility to recover

- Prefer short translation



Murray and Chiang, 2018

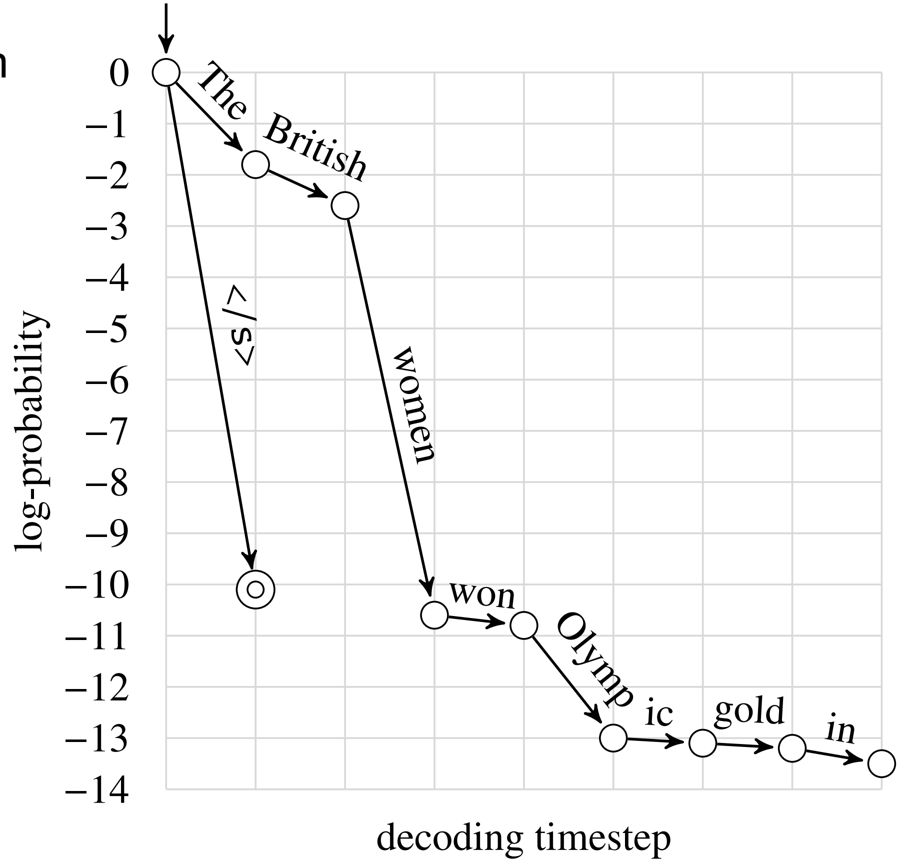
Label / Length bias

- Over-estimate probability of a prefix $y_1,..,y_m$
  - Multiply with conditional probabilities

$$p(y_{m+1}|x, y_{1,..,m})$$

  - No possibility to recover

- Prefer short translation
- Model error
  - Search error in greedy/beam search with small beam does prevent



Murray and Chiang, 2018

Modeling sentence length

- Length normalization

$$s'(e) = s(e) \,/\, m.$$

$$s'(e) = s(e) \Big/ \frac{(5+m)^\alpha}{(5+1)^\alpha}.$$

- Word reward

$$s'(e) = s(e) + \gamma m.$$

Search strategies

- Greedy search
- Exact search
- Sampling
- Beam Search
- Minimum Bayes Risk Decoding
  - Basic Idea:
    - Probability mass distributed over many good translations
    - Find a good representative

Minimum Bayes Risk decoding

- Several good translation distributed accoding to P_{human}
- u(h,r)
  - Utility of the hypothesis accoring to reference r
- Idea:
  - Find translation with highest expected utility

$$
\begin{aligned}
h^{\text{best}} &= \underset{h \in \mathcal{H}}{\arg\max} \underset{r \sim P_{\text{human}}(\cdot|x)}{\mathbb{E}} \{u(h,r)\} \qquad (1) \\
&= \underset{h \in \mathcal{H}}{\arg\max} \sum_{r \in \Omega} u(h,r) P_{\text{human}}(r|x).
\end{aligned}
$$

Minimum Bayes Risk decoding

- Challenge:
  - Reference translations unkown

- Idea:
  - Rely on model

$$h^{\text{model}} = \arg\max_{h \in \mathcal{H}} \sum_{y \in \Omega} u(h, y) P_{\text{model}}(y|x)$$

Minimum Bayes Risk decoding

- Challenge:
  - Sum over all hypothesis
- Idea:
  - Rely on finit sample

$$h^{\mathrm{MBR}} = \arg\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}_{\mathrm{model}}|} \sum_{y \in \mathcal{H}_{\mathrm{model}}} u(h, y).$$

- Candidate pool H
- Pseudo-references H_{model}
  - Use same pool

Minimum Bayes Risk decoding

- Sampling:
  - Independent sampling
    - Uniform probability distribution on the set

$$h^{\mathrm{MBR}} = \arg\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}_{\mathrm{model}}|} \sum_{y \in \mathcal{H}_{\mathrm{model}}} u(h, y).$$

Minimum Bayes Risk decoding

- Utility function:
  - Compare hyothesis and pseudo-reference
  - Related to automatic evaluation
  - Examples:
    - Sentence-level BLEU
    - Neural Evaluation metrics

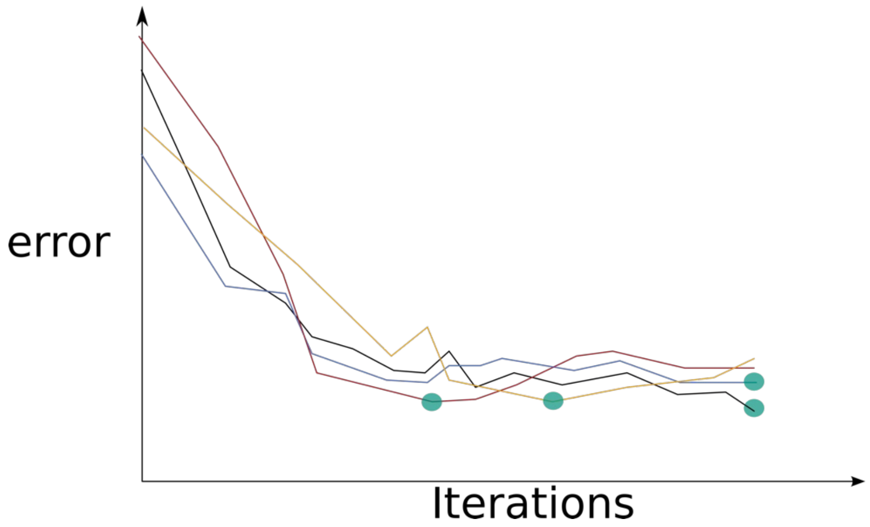
$$h^{\text{MBR}} = \arg\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{H}_{\text{model}}|} \sum_{y \in \mathcal{H}_{\text{model}}} u(h, y).$$

Combination of NMT models

- Randomly initialize models
  - Easy to create many different models
- Design decisions lead to different models
- Each model has strengths and weaknesses
- Methods:
  - Ensemble
  - Reranking

Model Ensemble

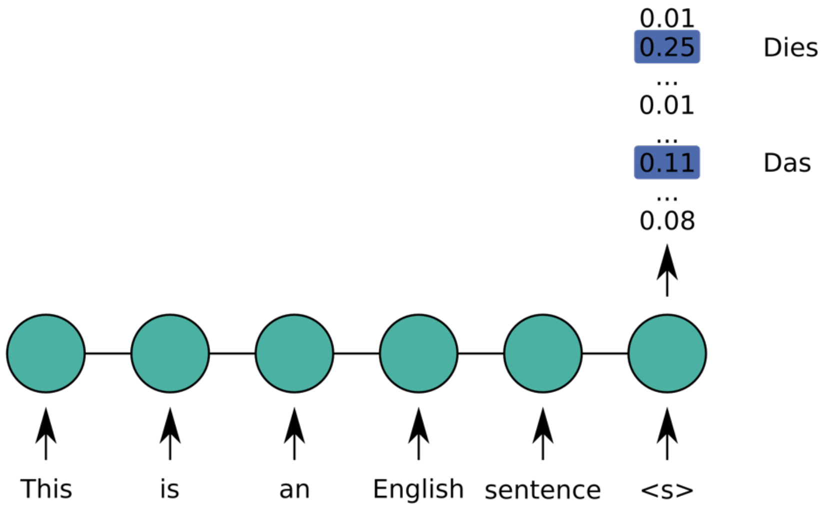
- Combine different models
  - E.g. different initialization



Model Ensemble

- Combine different models
  - E.g. different initialization
- Combine output layer of different models
- Word probability:

$$P\left(e_i = j \middle| e_1^{i-1}, F\right) = \frac{o_j}{\sum_{i=1}^{N} o_i}$$



Ensemble

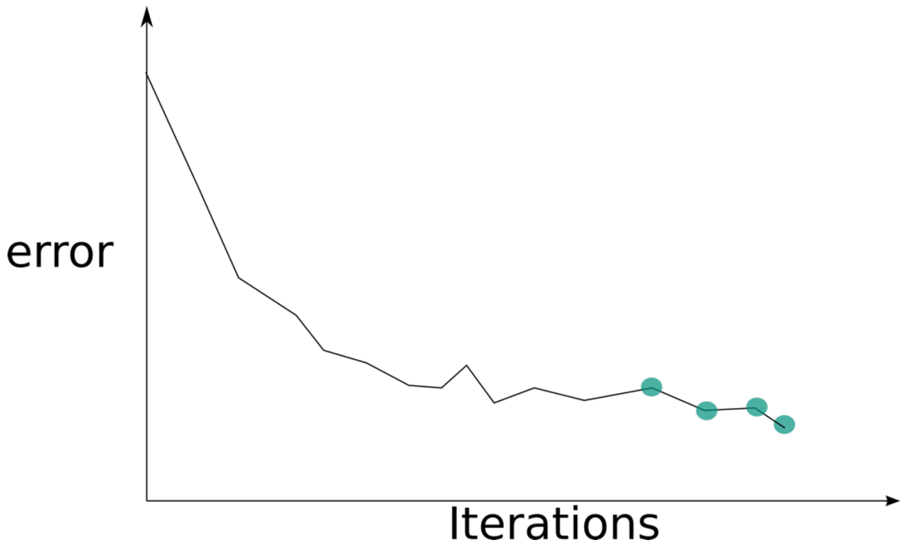
- Combine different models
  - E.g. different initialization
- Combine output layer of different models
- Word probability:

$$P\left(e_i = j \middle| e_1^{i-1}, F\right) = \frac{\sum_{k=1}^{K} o_j^k}{\sum_{i=1}^{N} \sum_{k=1}^{K} o_i^k}$$

- Performance 
- Training speed 
- Decoding speed 

Check-point ensemble

- Train one model
  - Save checkpoints



Check-point ensemble

- Train one model
  - Save checkpoints

- Ensemble models from different checkpoints

- Performance 
- Training speed 
- Decoding speed 

Rescoring/Reranking

- Other ways of modeling translation probability might be complementary

- Different word representations
  - Byte-pair encoding, Character, …

- Right to left decoding:
  - I go home → . home go I

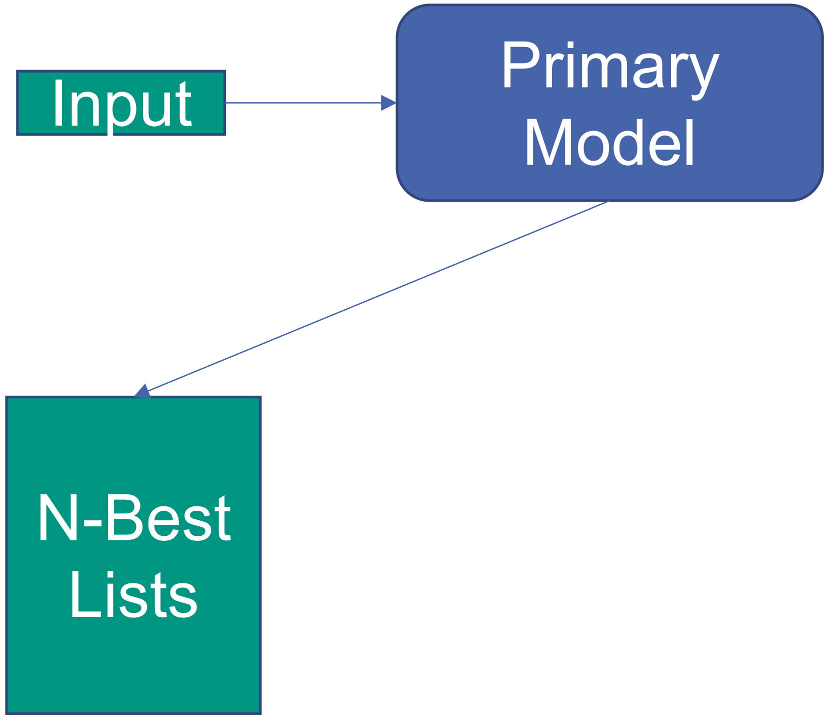
- Inverse translation directions
  - P(I gehe nach Hause | I go home)
  - P(I go home | I gehe nach Hause)

Rescoring/Reranking

- Other ways of modeling translation probability might be complementary
- Challenge:
  - Different search space
  - Cannot score same partical hyptothesis
  - P(I go home | I ??????)
- Idea:
  - Create finit sample of search space
  - Score only finit sample

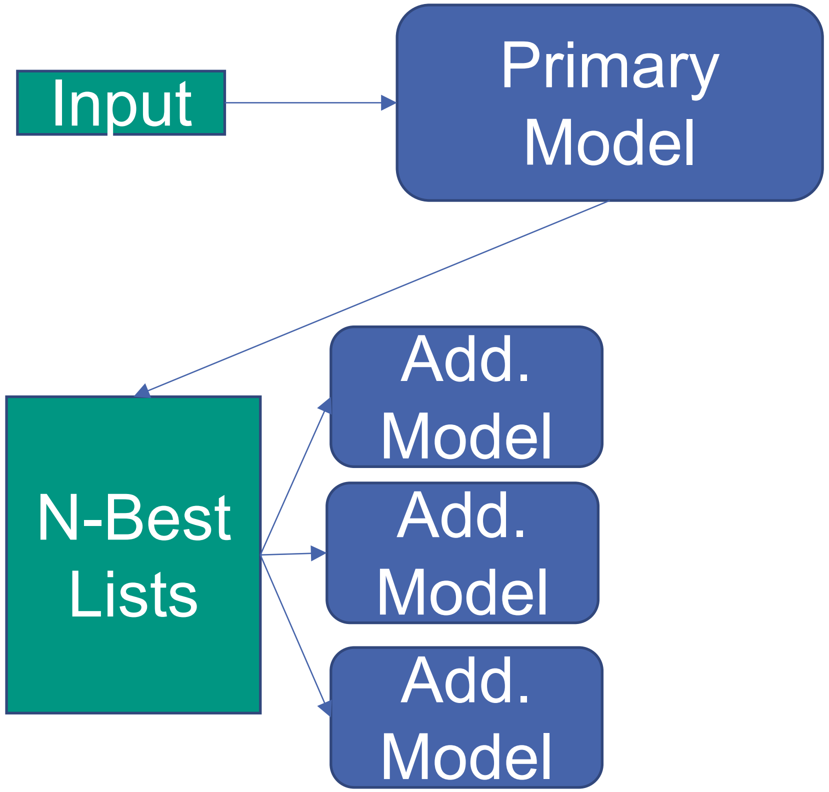
Rescoring/Reranking

- Create N-Best List by primary model



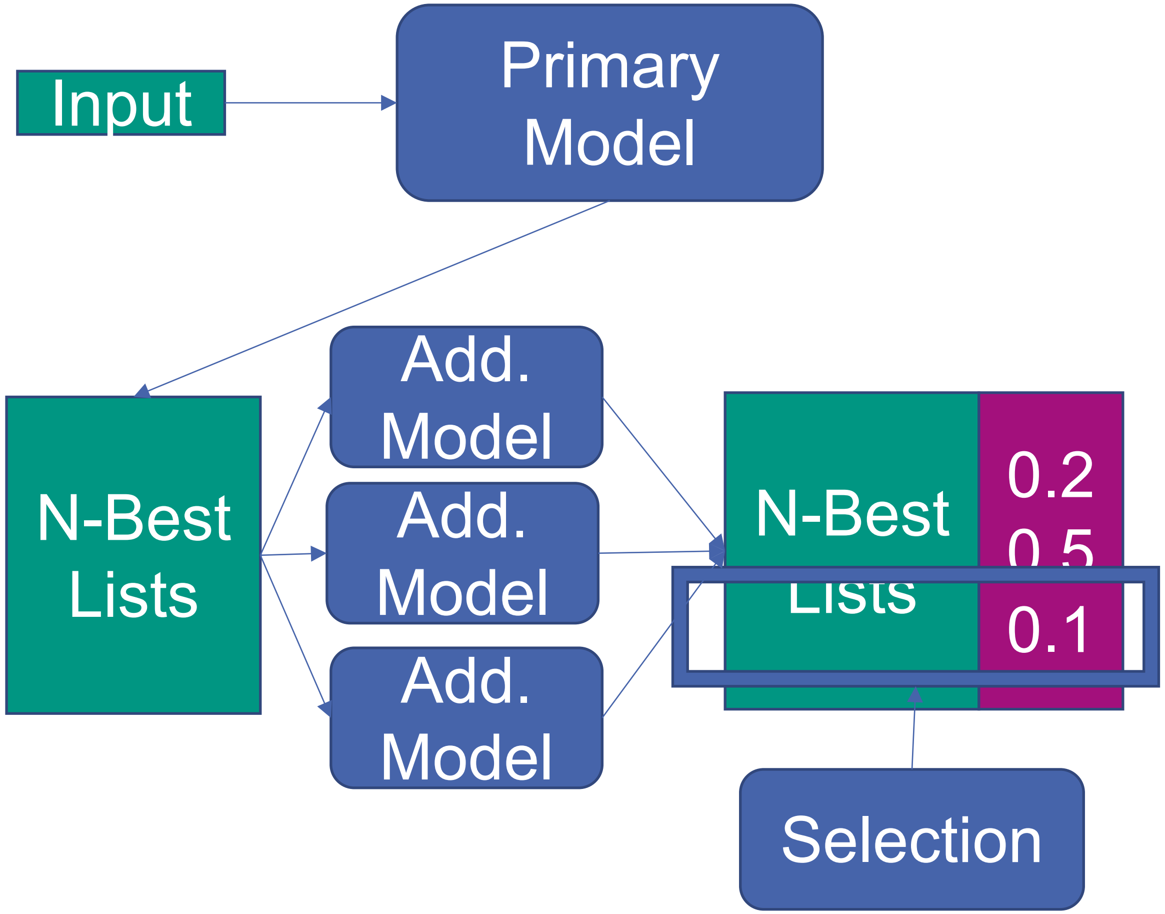
Rescoring/Reranking

- Create N-Best List by primary model
- Rescore using additional models



Rescoring/Reranking

- Create N-Best List by primary model
- Rescore using additional models
- Select best
  - Sum scores
  - Weights sum
    - Training using e.g. MERT



Deversity Bias term

- N-Best List looks very similar
  - He never wanted to participate in any kind of confrontation.
  - He never wanted to take part in any kind of confrontation.
  - He never wanted to participate in any kind of argument.
  - He never wanted to take part in any kind of argument.
  - He never wanted to participate in any sort of confrontation.
  - He never wanted to take part in any sort of confrontation.
  - He never wanted to participate in any sort of argument.
  - He never wanted to take part in any sort of argument.
  - He never wanted to participate in any kind of controversy.
  - He never wanted to take part in any kind of controversy.
  - He never intended to participate in any kind of confrontation.
  - He never intended to take part in any kind of confrontation.
  - He never wanted to take part in some sort of confrontation.
  - He never wanted to take part in any sort of controversy.

Deversity Bias term

- N-Best List looks very similar
- Single hypothesis generates too many of the surviding next hyptothesis
- Add a cost based on rank
  - Most probable: no cost
  - Second most probable: cost c
  - Third most probable: cost 2c

Constrainted Decoding

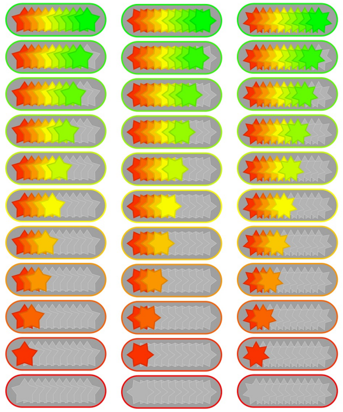
- The translation needs to fullfill additional constraints

- Example:
  - KIT → KIT (not Bausatz)

- Search only for hypothesis where the word KIT occures
  - Alignment difficult

Overview

Search



Ranking