COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Neural Machine Translation: directions for improvement

CMSC 470

Marine Carpuat

How can we improve on state-of-the-art machine translation approaches?

- Model
- Training
  - **Data**
  - **Objective**
  - Algorithm

Addressing domain mismatch

Slides adapted from Kevin Duh [Domain Adaptation in Machine Translation, MTMA 2019]

Supervised training data is not always in the domain we want to translate!

- Domain mismatch example:
  - Training data consists of Patent sentences
  - Test sample is Social Media
- Case 1: Test is not in input domain
  - can translate technical words like "NMT"
  - no idea how to translate "OMG"
- Case 2: Input-Output relation changes
  - "CAT" translates to a word that means "Computer Aided Translation" rather than "Cute furry animal"

Example sentences (case 1):

which is Patent, TED, Subtitles, Europarl?

1. We live in a digital world, but we’re fairly analog creatures.
2. The tablets exhibit improved bioavailability of the active ingredient.
3. So, um… she’s kidding.
4. Resumption of the session

Example bitext (case 2)

Medicine (EMEA):

`if you have severe depression, you must not use avonex . / no debe utilizar avonex si padece una depresión grave .`

Parliament (Europarl):

`the economic depression in europe has lasted at least ten years . / europa sufre una crisis económica desde hace , al menos , diez años .`

Domain adaptation is an important practical problem in machine translation

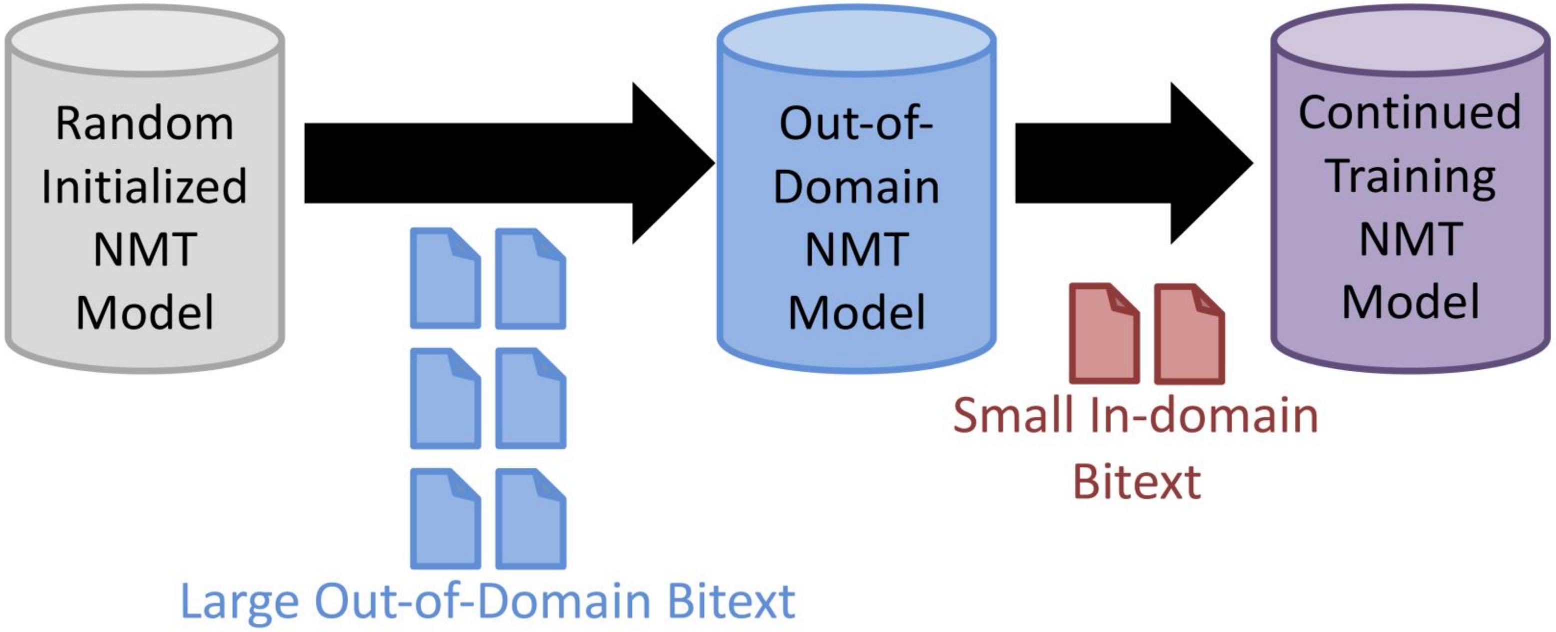
- It may be expensive to obtain training sets that are both large and relevant to test domain
- So we often have to work with whatever we can!

Data Size

Relevance to test domain		Small	Large
	Irrelevant		✔
	Relevant	✔	✔✔

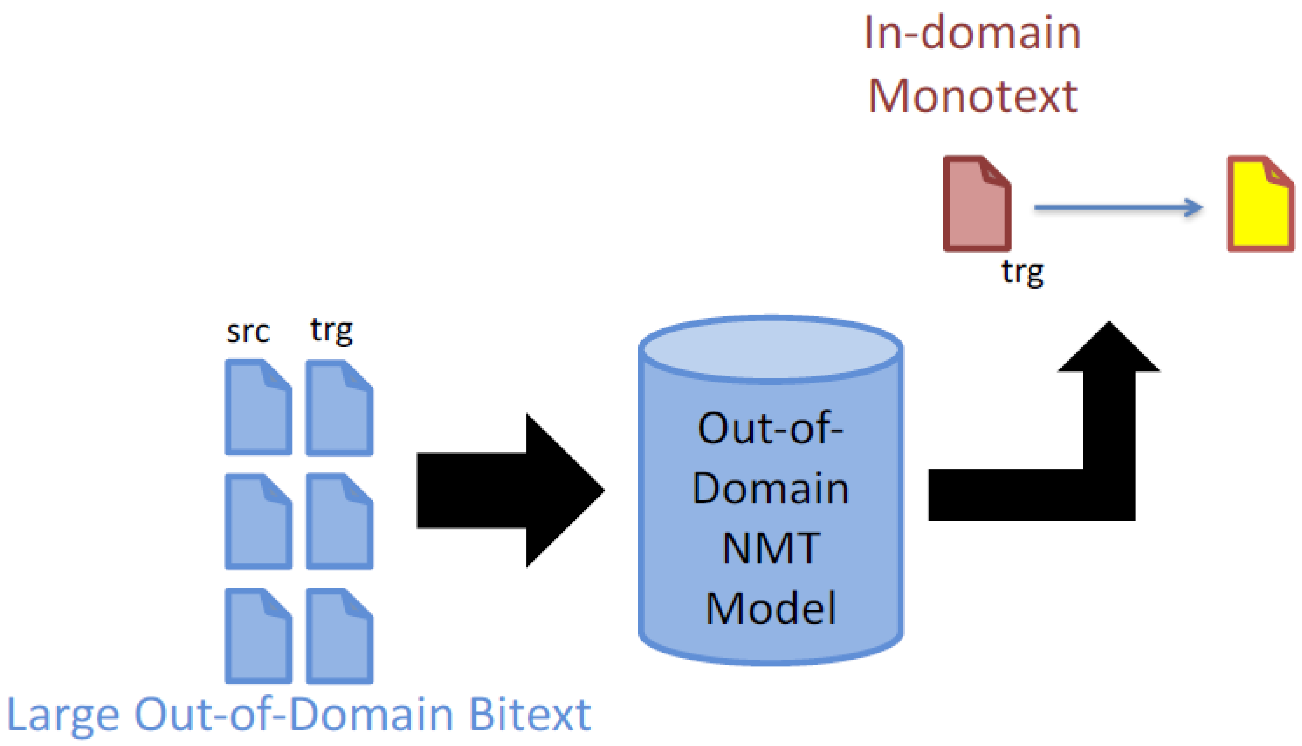
Possible strategies: “Continued Training” or “fine-tuning”

- Requires small in-domain parallel data

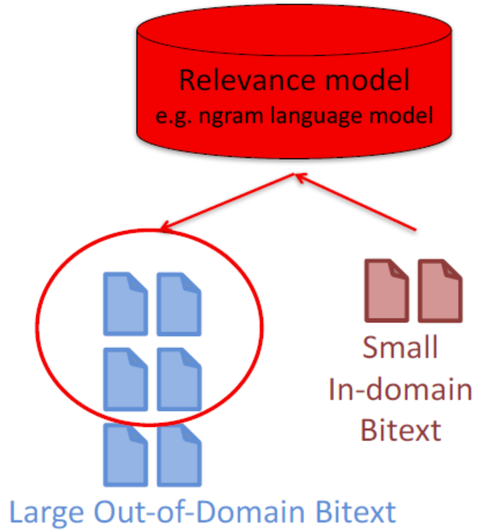


[Luong and Manning 2016]

Possible strategies: back-translation



Possible strategies: data selection



- Train a language model on data representative of test domain
  - N-gram count based model [Moore & Lewis 2010]
  - Neural model [Duh et al. 2013]
  - Neural MT model [Junczys-Dowmunt 2018]

- Use perplexity of LM on new data to measure distance from test domain

Possible strategies: different weights for different training samples

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, Eiichiro Sumita. Instance Weighting for Neural Machine Translation Domain Adaptation. EMNLP 2017

$$J_{dw} = \lambda_{in} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{in}} logp(\mathbf{y}|\mathbf{x}) + \sum_{(\mathbf{x}',\mathbf{y}') \in \mathcal{D}_{out}} logp(\mathbf{y}'|\mathbf{x}').$$

Corpus level weight

Boxing Chen, Colin Cherry, George Foster, Samuel Larkin. Cost Weighting for Neural Machine Translation Domain Adaptation. WNMT 2017

$$\theta^\star = \arg\max_{\theta} \sum_{(x,y) \in D} (1 + p_d(x)) \log p(y|x;\theta)$$

$$p_d(x) = \sigma\left(\tanh\left(W^d r_x + b^d\right)^\top w^d\right)$$

$$\text{where } \sigma(x) = \frac{1}{1+\exp(-x)}$$

Instance level weight
Based on classifier that measures similarity of samples with in domain data

How can we improve on state-of-the-art machine translation approaches?

- Model
- Training
  - **Data**
  - **Objective**
  - Algorithm

Beyond Maximum Likelihood Training

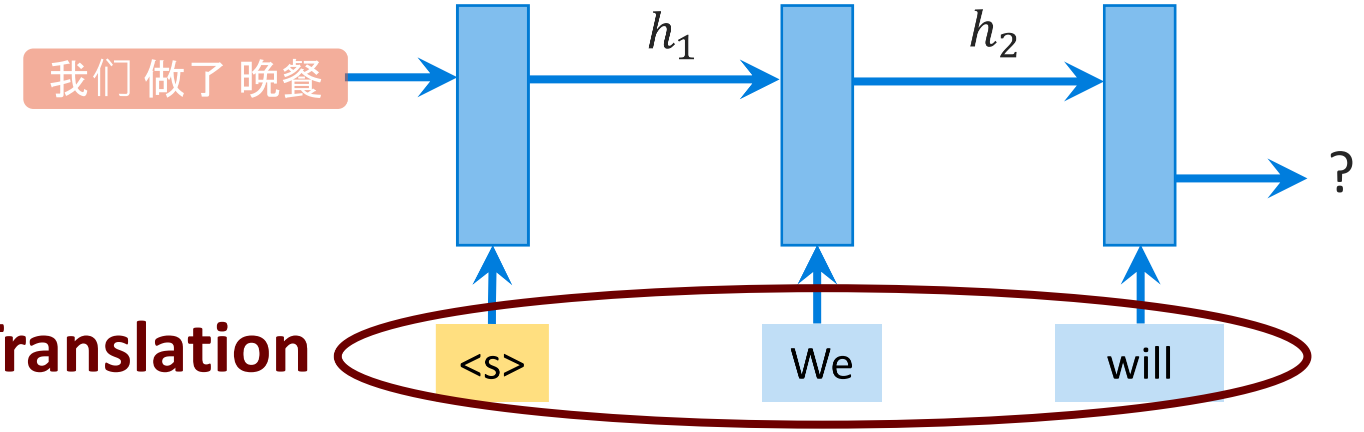
How can we improve NMT training?

- Assumption: References can substitute for predicted translations during training
- Our hypothesis: Modeling divergences between references and predictions improves NMT

Based on paper by [Weijia Xu [NAACL 2019]]

Exposure Bias: Gap Between Training and Inference

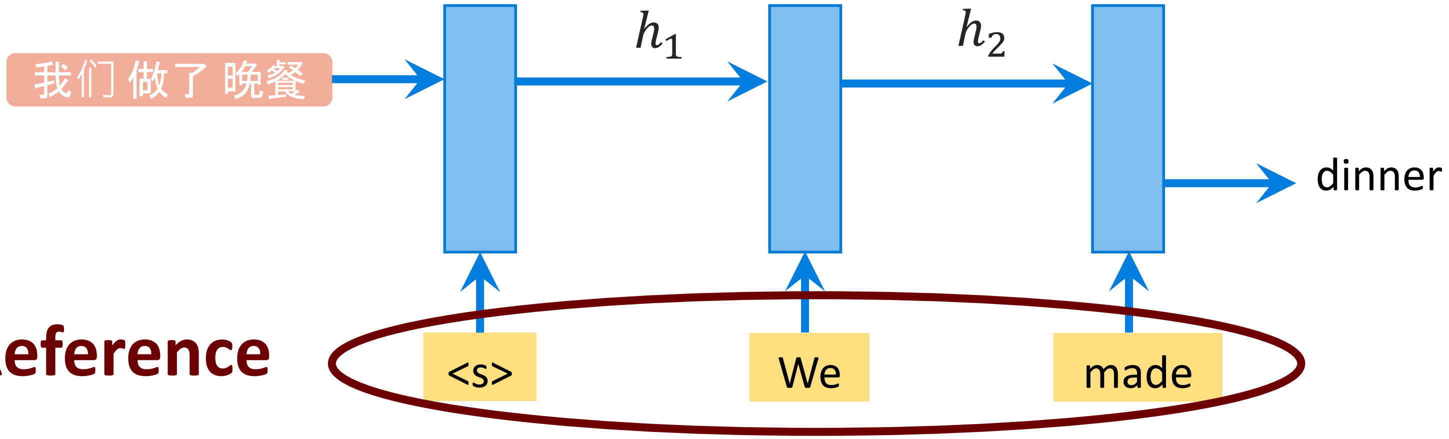
Inference



Model Translation

$$P(y|x) = \prod_{t=1}^{T} p(y_t|y_{<t}, x)$$

Maximum Likelihood Training



Reference

$$\textbf{Loss} = \sum_{t=1}^{T} \log p(y_t|y_{<t}, x)$$

How to Address Exposure Bias?

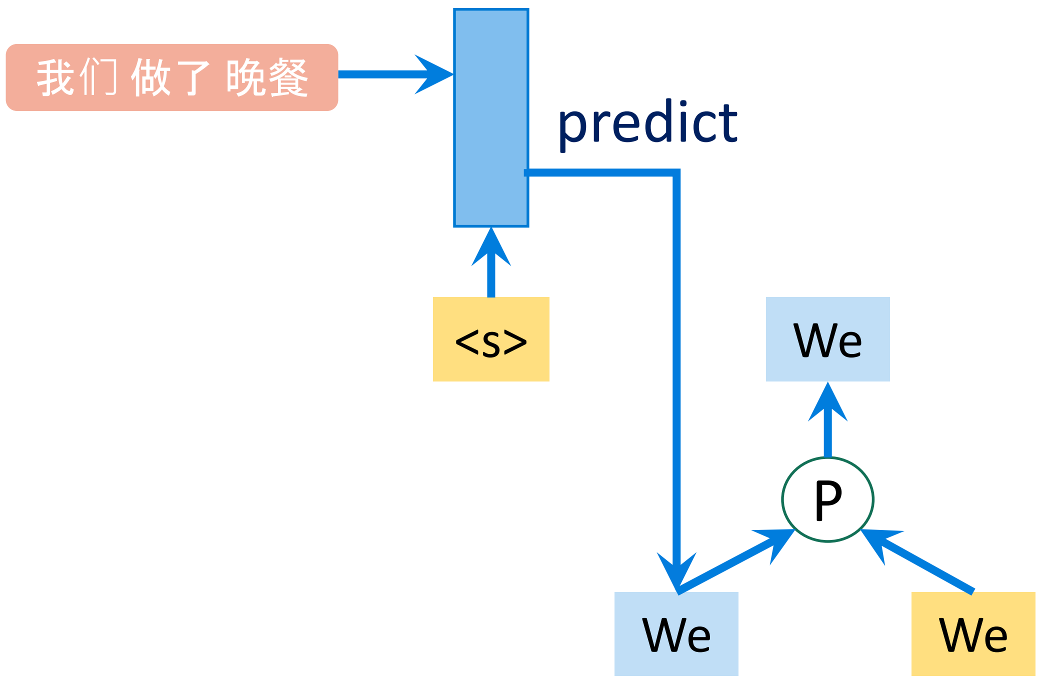
- Because of exposure bias
  - Models don’t learn to recover from their errors
  - Cascading errors at test time

- Solution:
  - Expose models to their own predictions during training
  - But how to compute the loss when the partial translation diverges from the reference?

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

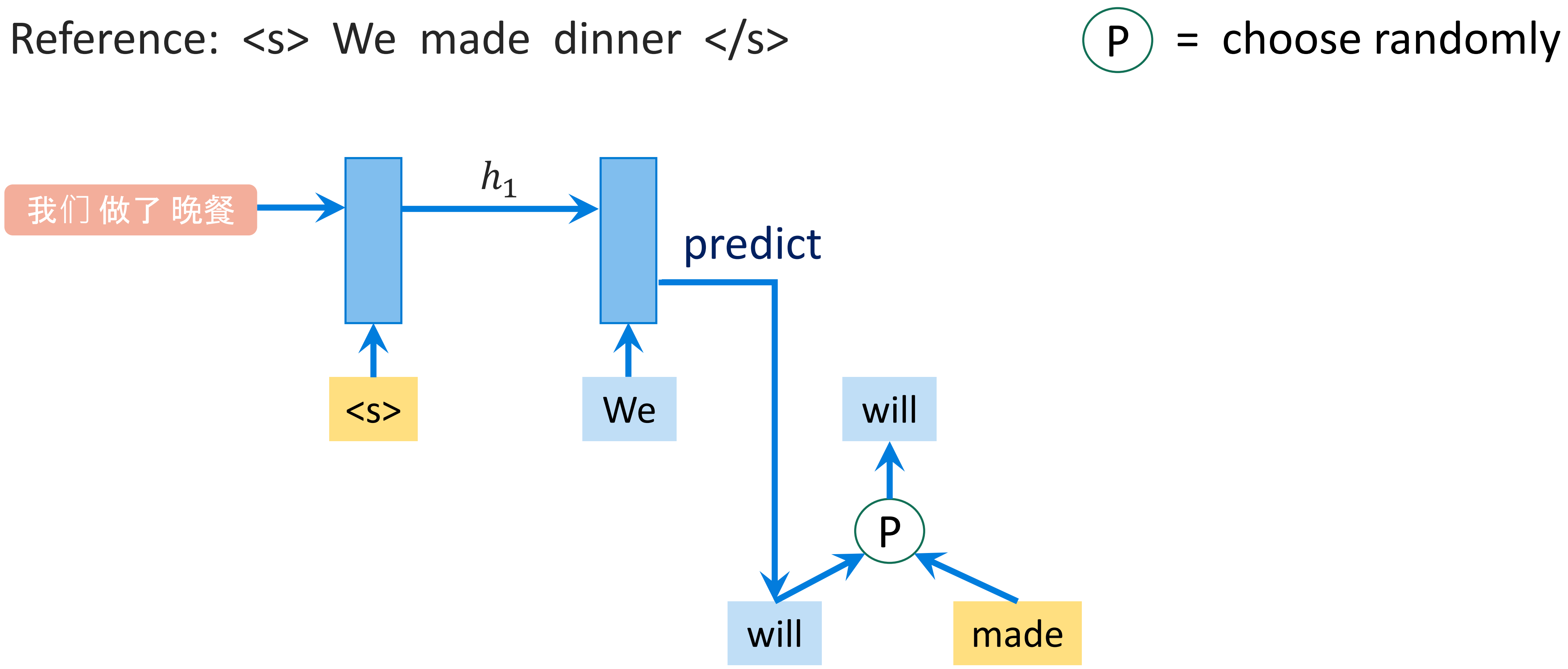
P = choose randomly



[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

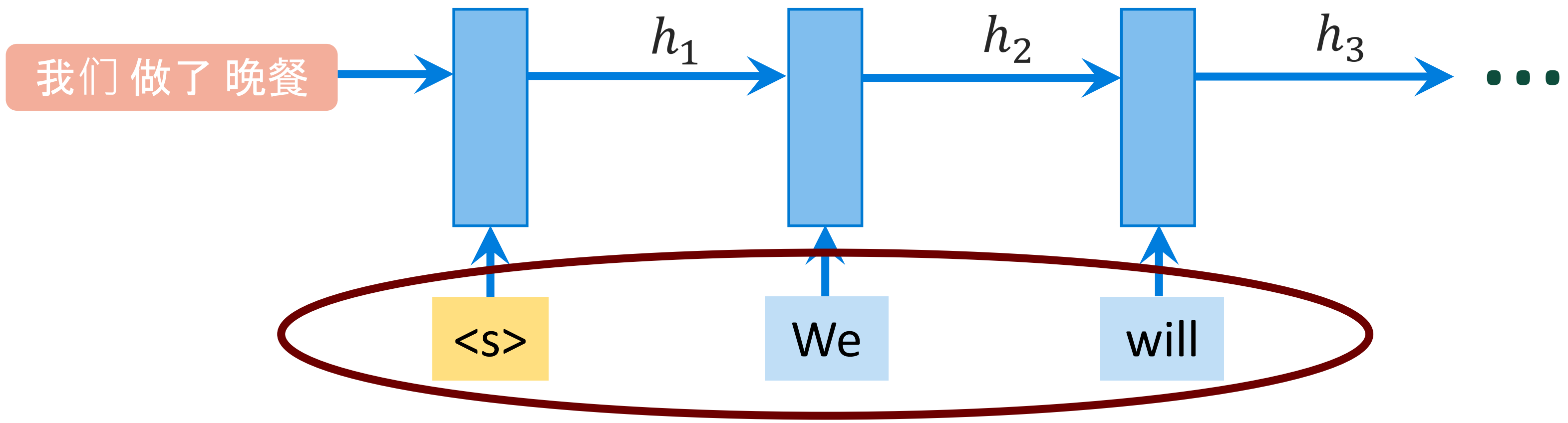
Reference: <s> We made dinner </s>



[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

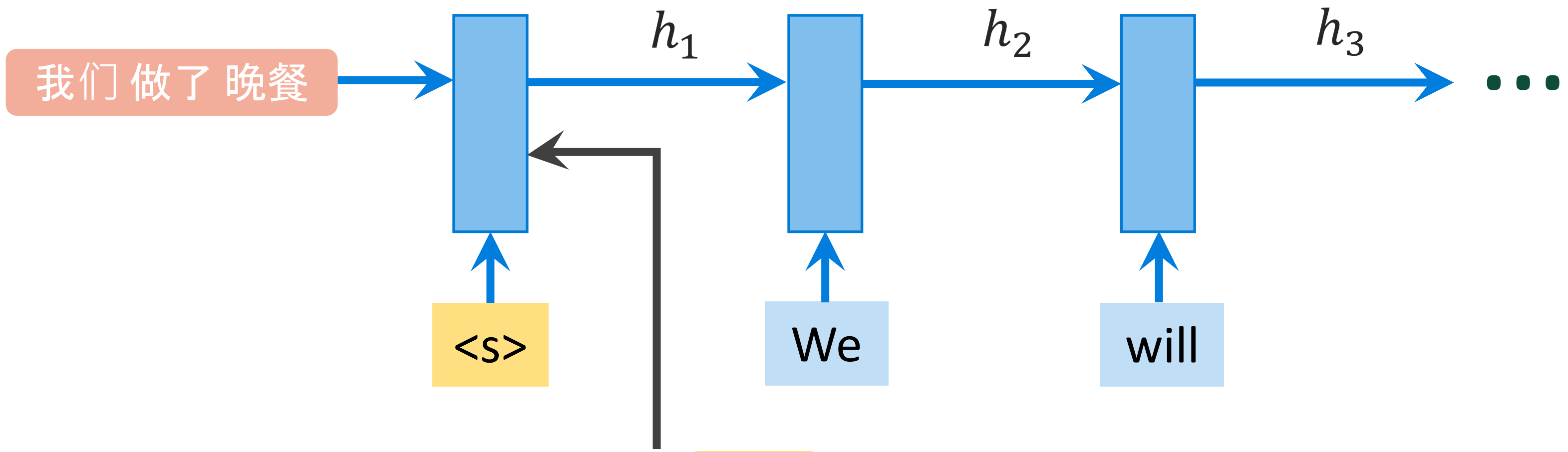
Reference: <s> We made dinner </s>



[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

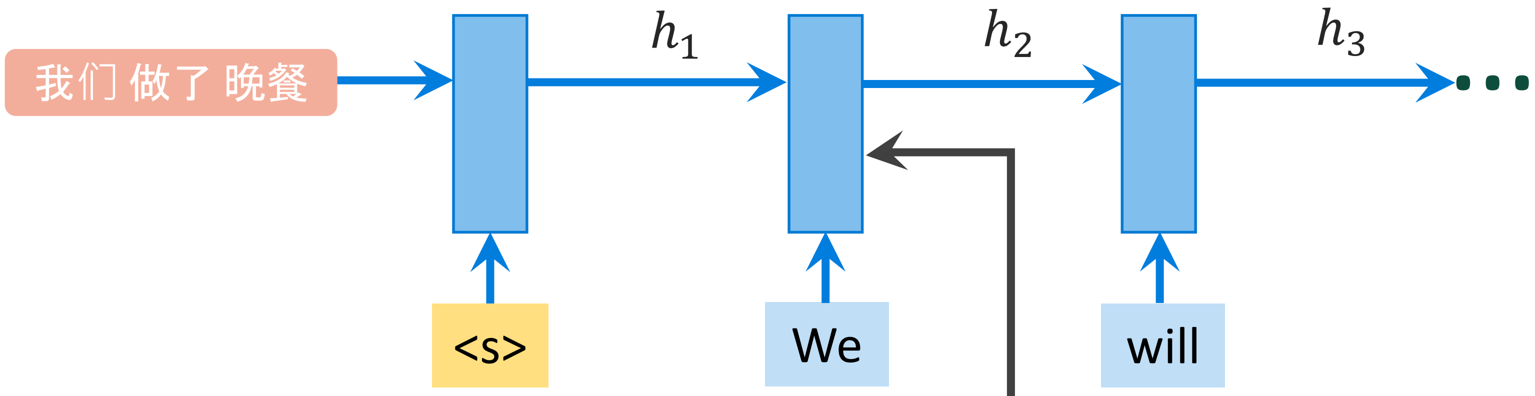


J = log p("We" | "<s>", source)

[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

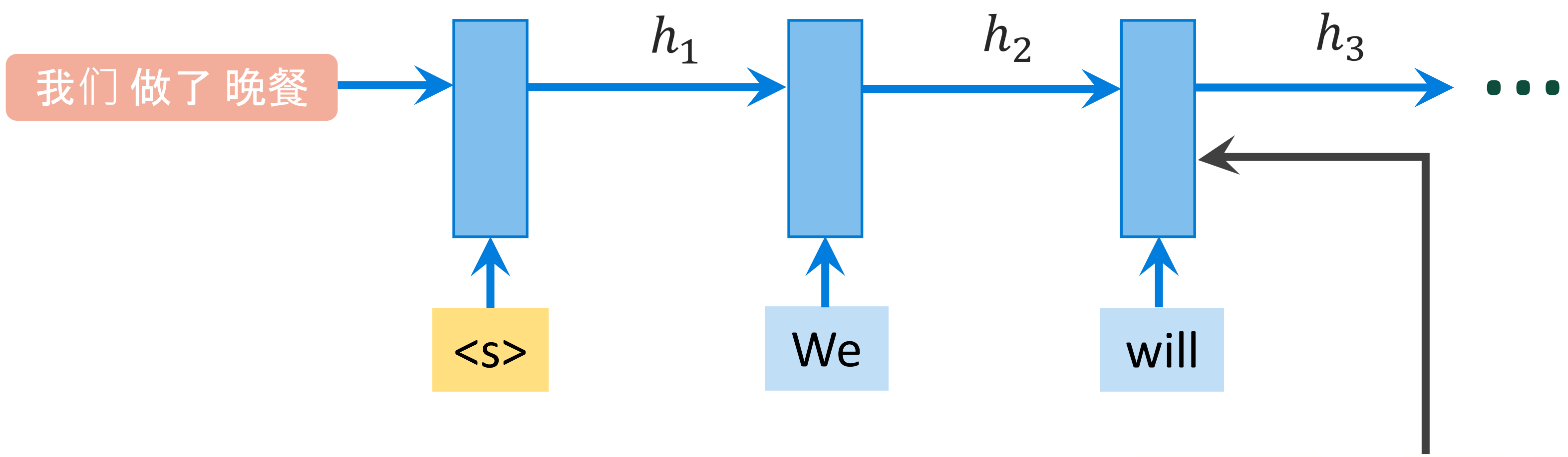


J = log p("made" | "<s> We", source)

[Bengio et al., NeurIPS 2015]

Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

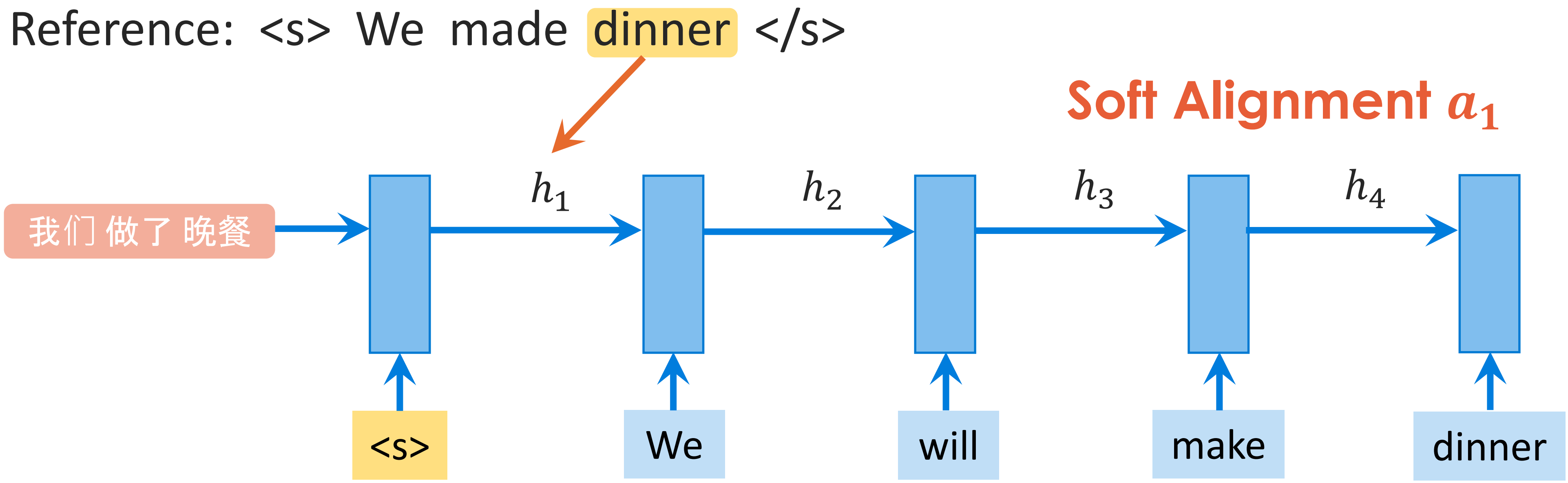


Incorrect synthetic reference:
“We will dinner”

J = log p(“dinner” | “<s> We will”, source)

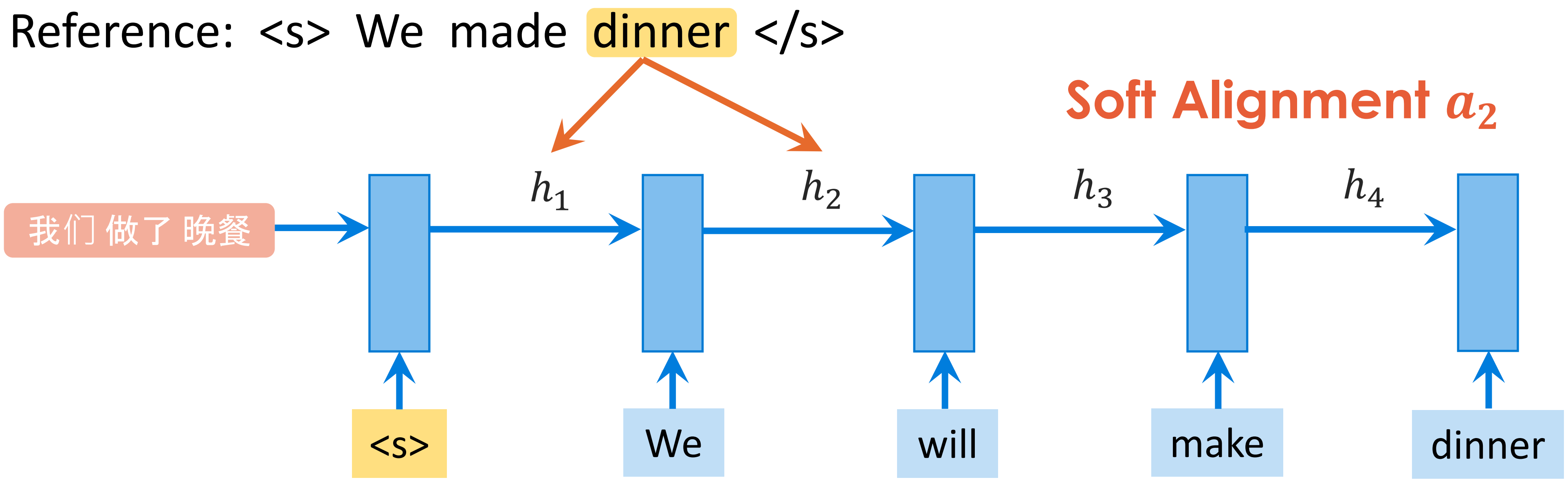
[Bengio et al., NeurIPS 2015]

Our Solution: Align Reference with Partial Translations



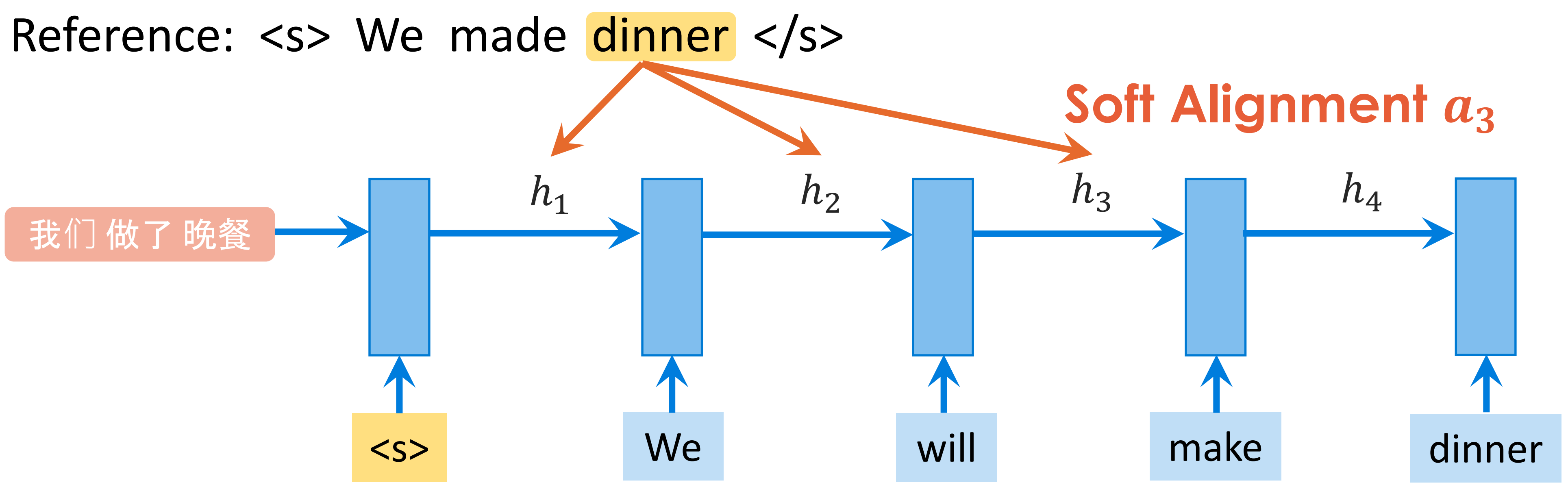
a_1 logp(“dinner” | “<s>”, source)

Our Solution: Align Reference with Partial Translations



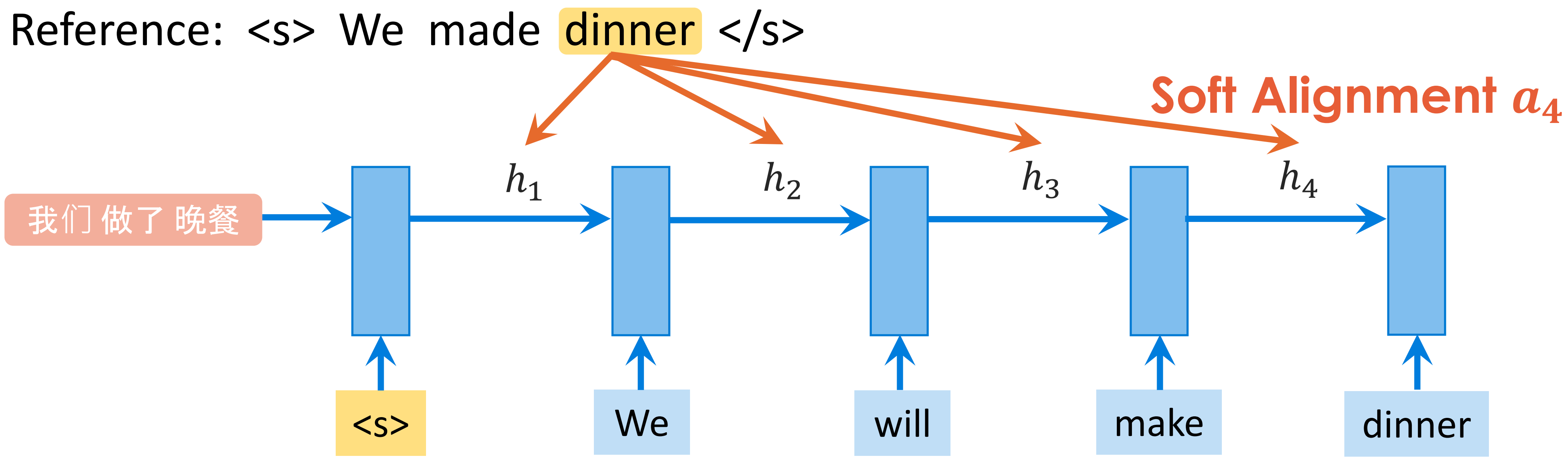
a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source)

Our Solution: Align Reference with Partial Translations



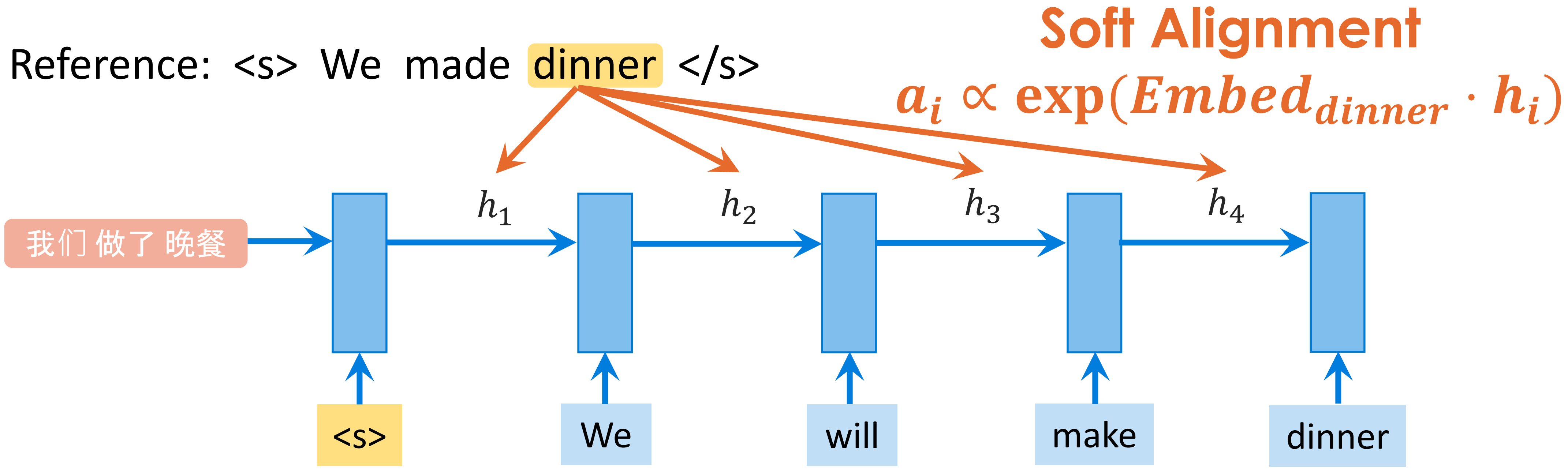
a_1 logp("dinner" | "<s>", source) + a_2 logp("dinner" | "<s> We", source) +
a_3 logp("dinner" | "<s> We will", source)

Our Solution: Align Reference with Partial Translations



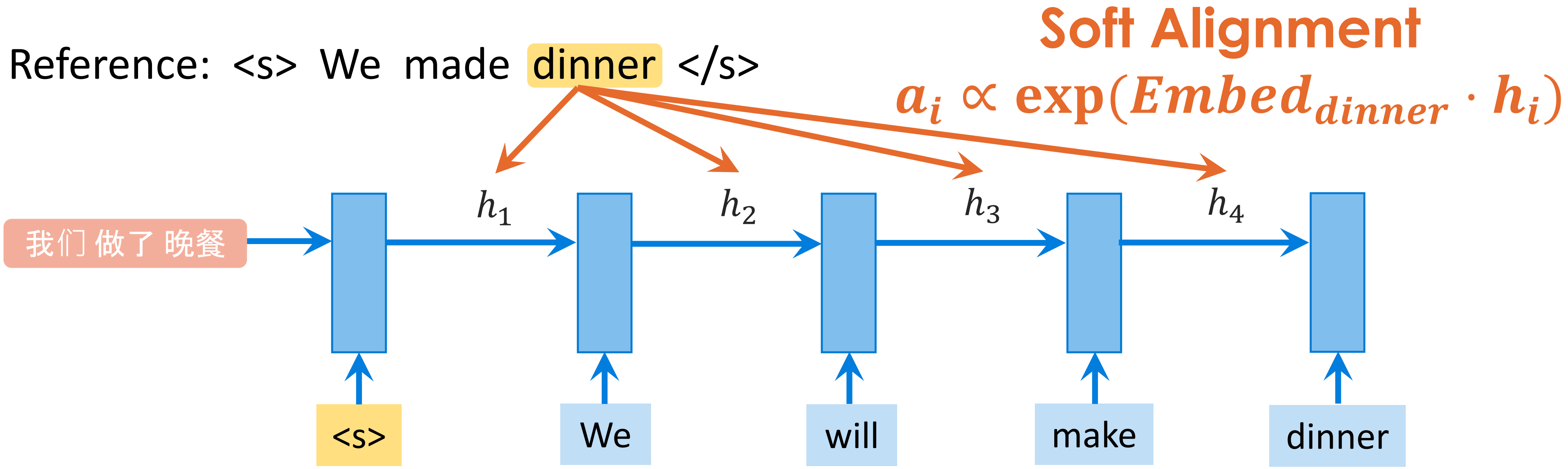
a_1 logp("dinner" | "<s>", source) + a_2 logp("dinner" | "<s> We", source) +
a_3 logp("dinner" | "<s> We will", source) + a_4 logp("dinner" | "<s> We will make", source)

Our Solution: Align Reference with Partial Translations



a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source) +
a_3 logp(“dinner” | “<s> We will”, source) + a_4 logp(“dinner” | “<s> We will make”, source)

Our Solution: Align Reference with Partial Translations



a_1 logp(“dinner” | “<s>”, source) + a_2 logp(“dinner” | “<s> We”, source) +
a_3 logp(“dinner” | “<s> We will”, source) + a_4 **logp(“dinner” | “<s> We will make”, source)**

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(\boxed{y_t} \mid \tilde{y}_{<j}, x)$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(\boxed{y_t} \mid \tilde{y}_{<t}, x)$$

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(y_t \mid \tilde{y}_{<j}, x)$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(y_t \mid \tilde{y}_{<t}, x)$$

Training Objective

Ours:

Soft alignment between y_t and $\tilde{y}_{<j}$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log \sum_{j=1}^{T'} a_{tj}\, p(y_t \mid \tilde{y}_{<j}, x)$$

Combined with maximum likelihood:

$$J = J_{SA} + J_{ML}$$

Scheduled Sampling:

Hard alignment by time index t

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} log\, p(y_t \mid \tilde{y}_{<t}, x)$$

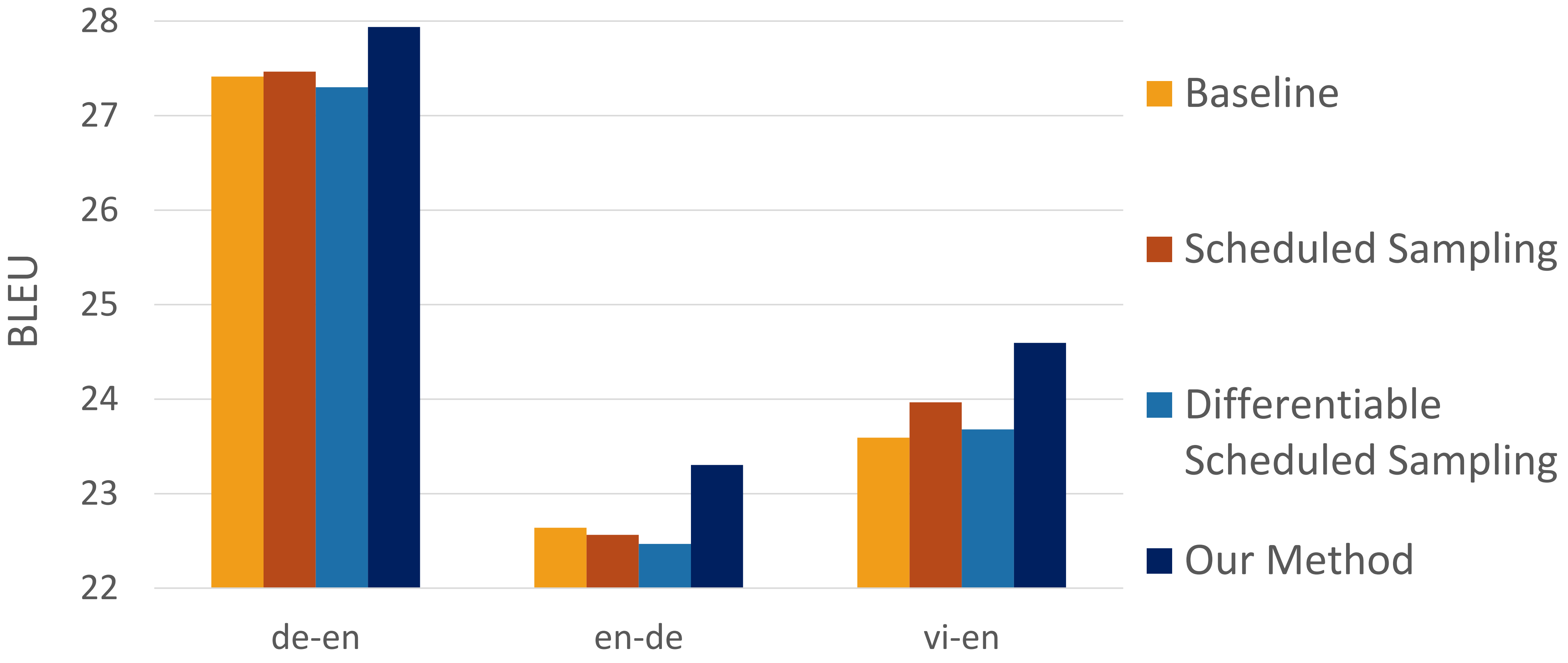
Experiments

- **Data**
- IWSLT14 de-en
- IWSLT15 vi-en

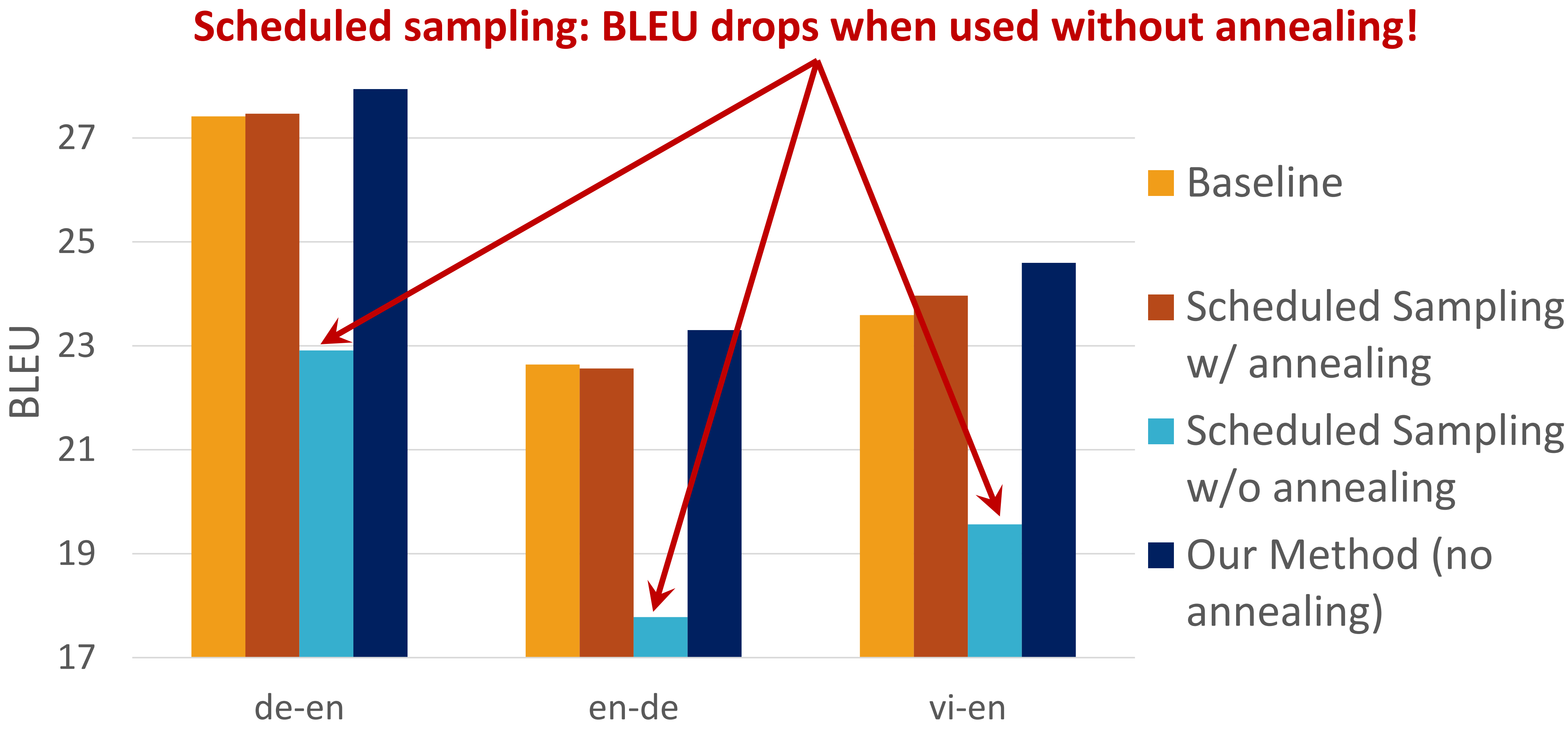
Task	sentences (K)			vocab (K)	
	train	dev	test	src	tgt
de-en	153.3	7.0	6.8	113.5	53.3
vi-en	121.3	1.5	1.3	23.9	50.0

- **Model**
- Bi-LSTM encoder, LSTM decoder, multilayer perceptron attention
- Differentiable sampling with Straight-Through Gumbel Softmax
- Based on AWS sockeye

Our Method Outperforms Maximum Likelihood and Scheduled Sampling



Our Method Needs No Annealing



Summary

Introduced a new training objective

1. **Generate translation prefixes** via differentiable sampling
2. Learn to **align** the reference words with sampled prefixes

Better BLEU than the maximum likelihood and scheduled sampling (de-en, en-de, vi-en)

Simple to train, no annealing schedule required

What you should know

- Lots of things can be done to improve neural MT even without changing the model architecture

- The domain of training data matters
  - Simple techniques can be used to measure distance from test domain
  - And to adapt model to domain of interest

- The standard maximum likelihood objective is suboptimal
  - It does not directly measure translation quality
  - It is based on reference translations only, so the model is not exposed to its own errors during training
  - Developing reliable alternatives is an active area of research