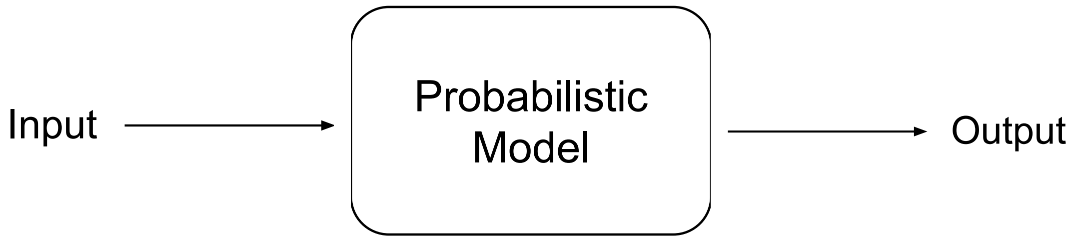
Neural Machine Translation for Low-Resource Scenarios

Yunsu Kim

Promotionsvortrag
RWTH Aachen
07.02.2022

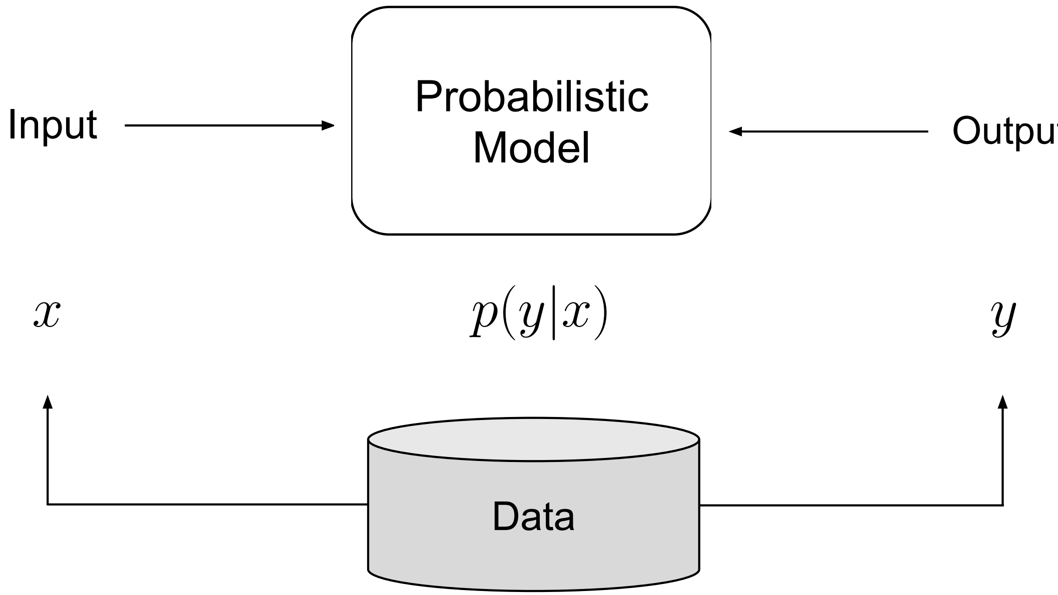
Statistical Machine Learning



x $\qquad$ $p(y|x)$ $\qquad$ $\hat{y} = \underset{y}{\operatorname{argmax}}\ p(y|x)$

Heute ist es sonnig. $\qquad$ Today is sunny.

Supervised Learning



These days: commercializable performance in many tasks

- Model: neural network with attention components [Vaswani & Shazeer+ 17]
- Training: stochastic gradient descent variants [Kingma & Ba 15]
- **Data**: varied among tasks/domains

Low-Resource Scenarios

Why do we lack data?

- Supervised learning: data needs to be *labeled*
- Not feasible to label data at scale: requires *human labor*
- Industry's needs are growing: more tasks, more domains, personalized

Why is the lack of data problematic?

- Model: performance is sensitive to the data size
- Training: hard to generalize to unseen examples

Question: Given a small amount of data, what should we do?

Outline

Preliminaries

To-English Tasks

Non-English Tasks

Conclusion

Outline

Preliminaries

To-English Tasks

Non-English Tasks

Conclusion

Machine Translation

source					target		
Heute	ist	es	sonnig.	$\longrightarrow$	Today	is	sunny.
f_1	f_2	f_3	f_4		e_1	e_2	e_3
	f_1^J					e_1^I	

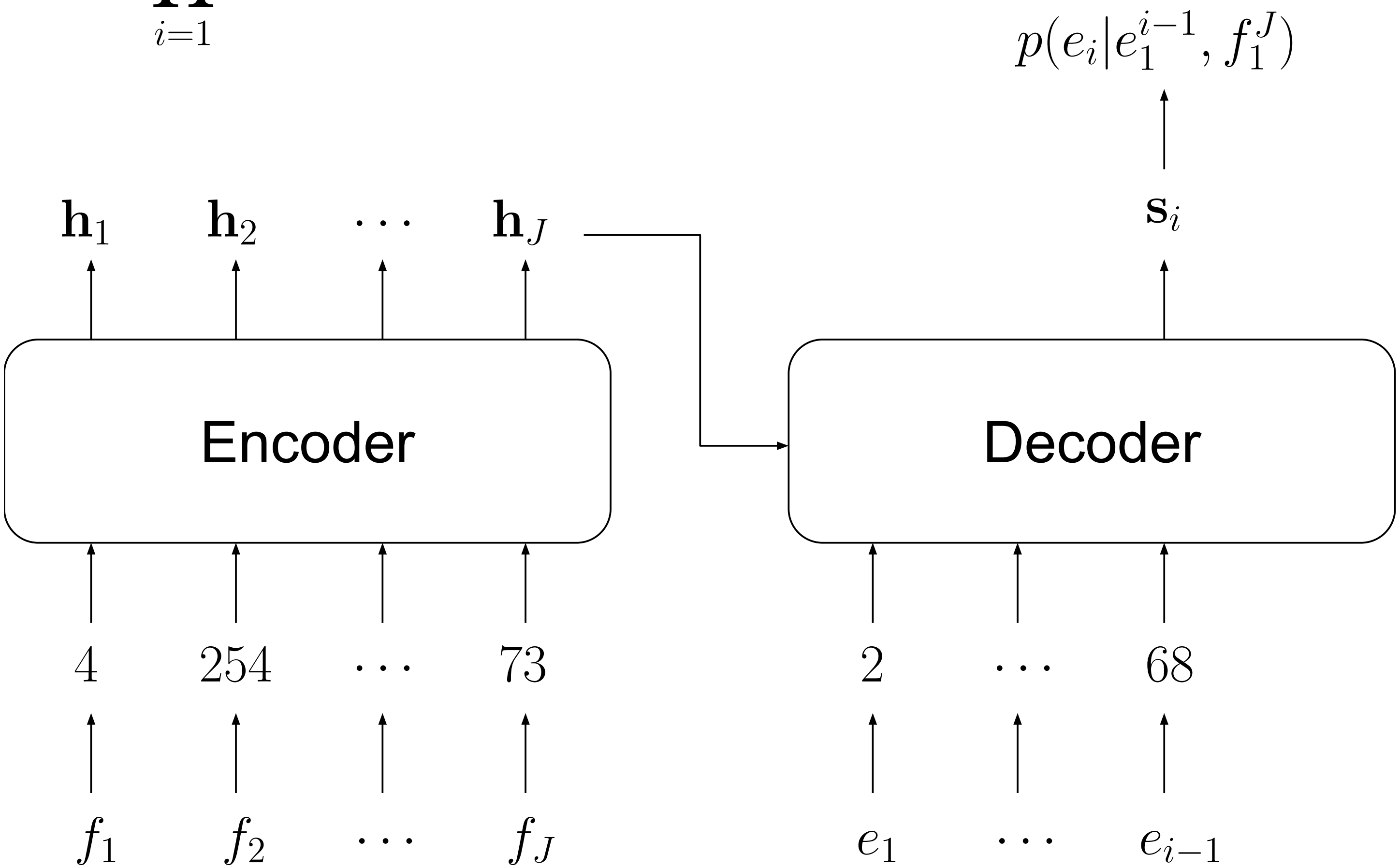
Translation problem

$$f_1^J \mapsto \hat{e}_1^{\hat{I}}(f_1^J) = \operatorname*{argmax}_{I, e_1^I} p(e_1^I | f_1^J)$$

- Generate a sequence: large search space
- Language dissimilarity: variable length, reordering

Neural Machine Translation

$$p(e_1^I|f_1^J) = \prod_{i=1}^{I} p(e_i|e_1^{i-1}, f_1^J)$$



Low-Resource Scenarios in Machine Translation

Bilingual data for machine translation

- Requires bilingual speakers to generate: English-centric
- Biased to languages with good research infrastructure
- e.g. German$\rightarrow$English, Chinese$\rightarrow$English

Low-resource scenarios

- English and a non-popular language: e.g. Turkish$\rightarrow$English
- Non-English language pair: e.g. French$\rightarrow$German

Outline

Preliminaries

To-English Tasks

Non-English Tasks

Conclusion

To-English Tasks

Data condition

- Bilingual data: small, limited domains
- Monolingual data: large, available in many domains
- e.g. Turkish→English

Data	#sentences	
	Turkish	English
Bilingual	208k	
Monolingual	4.8M	100M

How can we utilize unlabeled (monolingual) data to compensate for the lack of labeled (bilingual) data?

- **Semi-supervised Learning**

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training
- Data

Non-English Tasks

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- **Training**
- Data

Non-English Tasks

Conclusion

Semi-supervised Learning: Training

How can we exploit monolingual data in training a translation model?

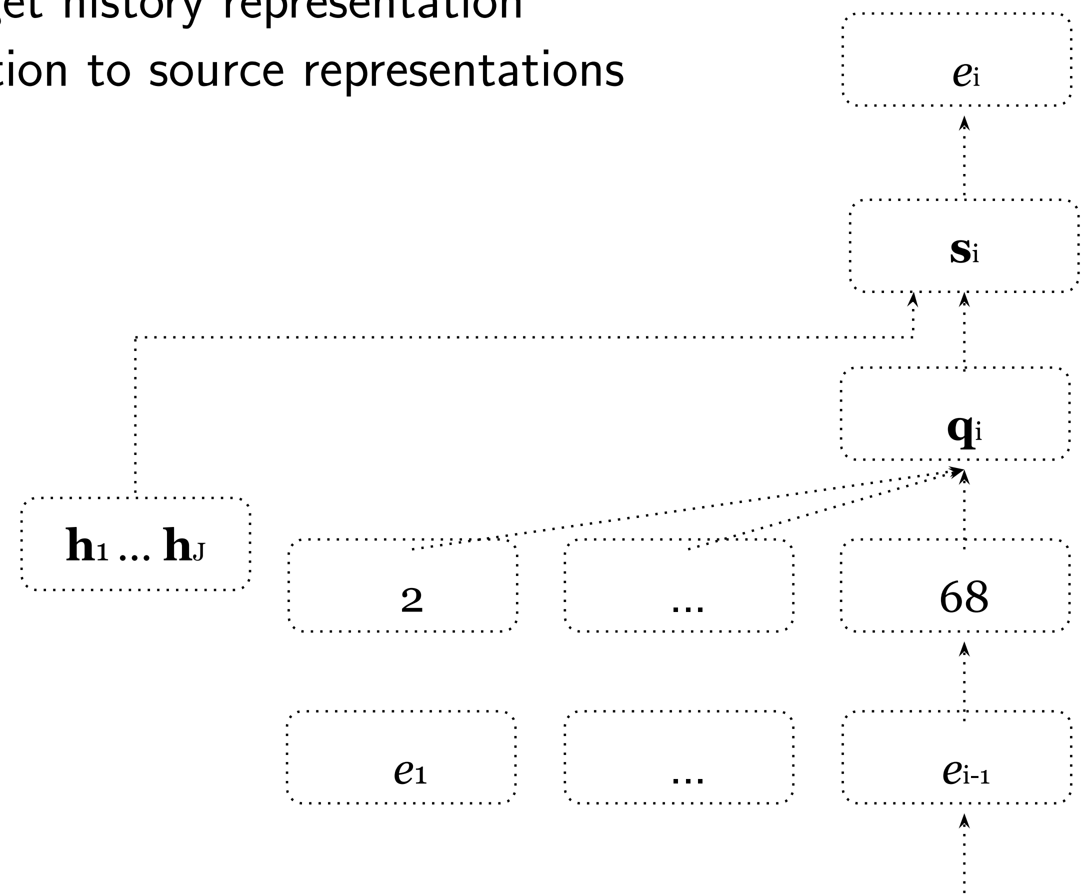
- A part of a translation model resembles a monolingual model
- Train that part as a monolingual model with monolingual data
- Much larger data than bilingual: learn to understand each language better

Question: Which part of a translation model resembles which monolingual model?

Decoder

Decoder (closer look)

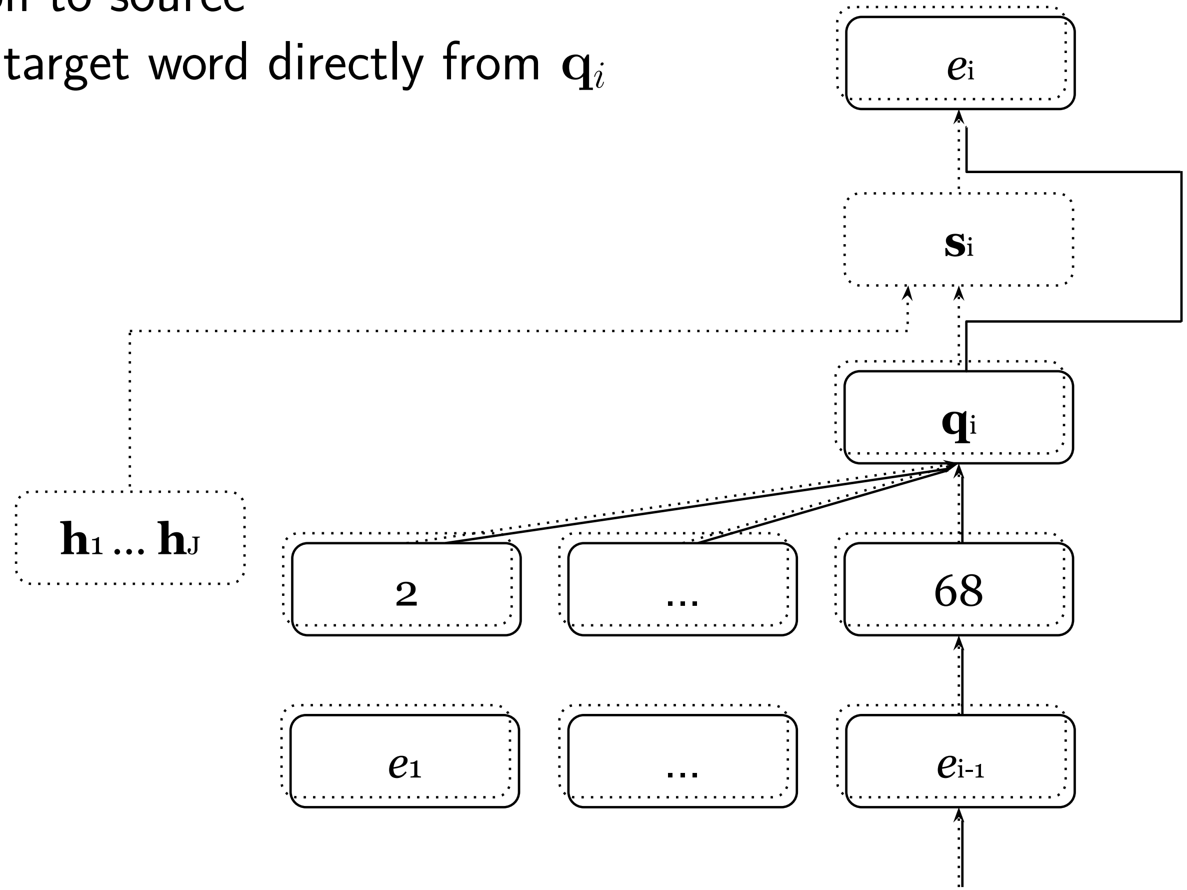
- $\mathbf{q}_i$ = target history representation
- $\mathbf{s}_i$ = relation to source representations



Decoder As a Language Model

Decoder resembles a language model

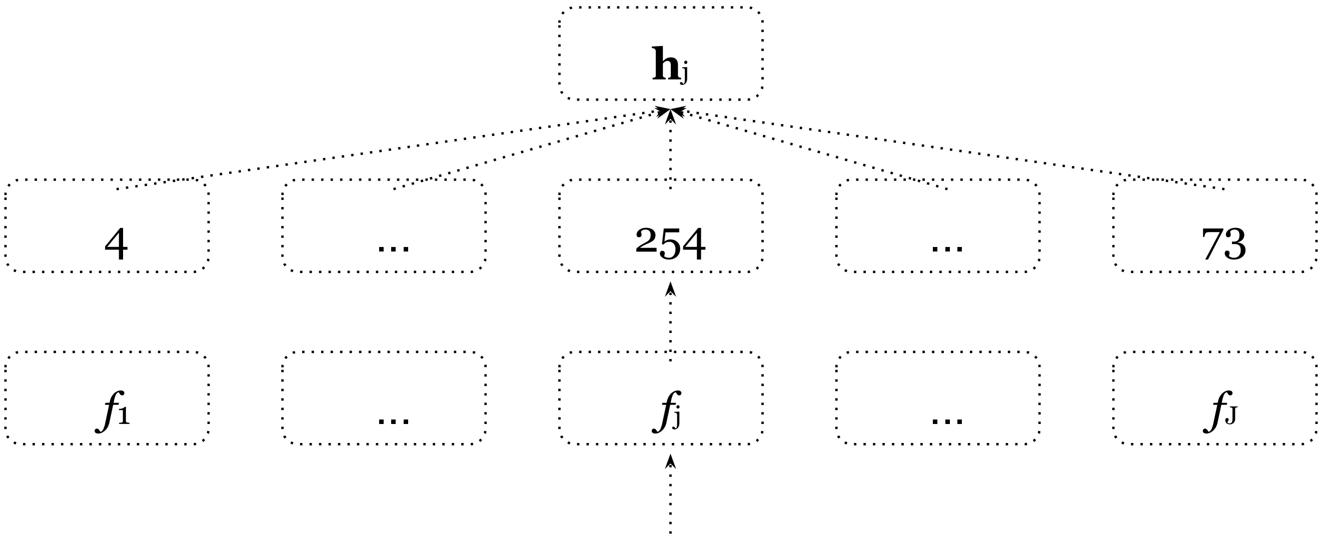
- No relation to source
- Predict a target word directly from $\mathbf{q}_i$



Encoder

Encoder (closer look)

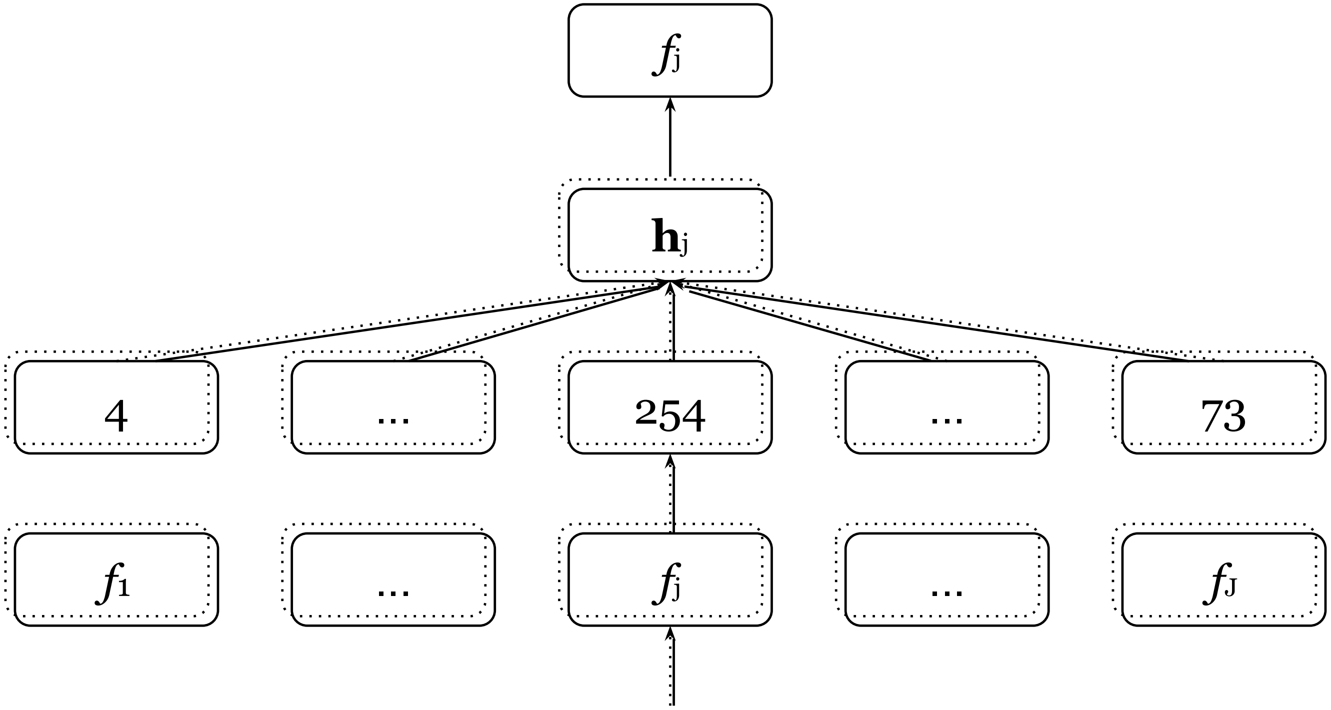
- Learns source representations
- No prediction by itself



Encoder As a Cloze Task Model

Encoder resembles a Cloze task model

- Predict a source word given its surrounding context [Taylor 53]
- Basis of the groundbreaking BERT [Devlin & Chang+ 19]



Monolingual Pre-Training

Pre-train for monolingual tasks $\longrightarrow$ Train for a translation task

- [Ramachandran & Liu^{+} 17]: LM pre-training for RNN translation model
- **This work**: re-evaluate in Transformer, also test Cloze task pre-training

Turkish→English	Monolingual		newstest2016		newstest2017	
	Encoder	Decoder	Bleu [%]	Ter [%]	Bleu [%]	Ter [%]
Bilingual	-	-	19.0	70.5	18.9	71.1
Monolingual→Bilingual	Cloze	LM	19.9	68.8	19.6	69.7
	LM	LM	19.6	70.1	19.4	69.8
	Cloze	Cloze	**20.0**	**68.5**	**19.8**	**69.2**

Monolingual pre-training helps the translation

- Cloze task is more suitable for both encoder/decoder
- Richer context is more important than the exact parameter overlap
- In the thesis: multi-task learning, cross-lingual pre-training

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data

Non-English Tasks

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- **Data**

Non-English Tasks

Conclusion

Semi-supervised Learning: Data

How can we augment bilingual training data?

- Synthesize bilingual data from monolingual data
- e.g. Generate a source sentence from a target monolingual sentence

Back-Translation: Use target$\rightarrow$source translation model

$$e_1^I \xrightarrow{p(f_1^J|e_1^I)} \tilde{f}_1^{\tilde{J}}$$

- Use synthetic bilingual sentences $(\tilde{f}_1^{\tilde{J}}, e_1^I)$ along with real bilingual data
- In this work, real:synthetic = 1:2

Generation Strategy: Decoding

How should we generate a source sentence $\tilde{f}_1^{\tilde{J}}$ using $p(f_1^J|e_1^I)$?

Decoding: Do the usual translation using beam search [Sennrich & Haddow+ 16]

source		**target**
Heute ist es sonnig.	$\longleftarrow$	Today is sunny.

- Biased to use frequent words [Ott & Auli+ 18]
- Tends to perform less reordering
- Does not reflect the variability of human translations

Generation Strategy: N-best List

How can we generate a potentially human-like source sentence $\tilde{f}_1^{\tilde{J}}$?

N-**best List**: Randomly choose one of the N-best hypotheses from beam search

source		target
Heute ist es sonnig.	$\longleftarrow$	Today is sunny.
Heute ist es sonnig!		
Heute ist sonnig.		

- Mechanical variations, e.g., different punctuation, dropping one word
- Computational complexity increases linearly with N

Generation Strategy: Restricted Sampling

How can we generate a potentially human-like source sentence $\tilde{f}_1^{\tilde{J}}$?

Restricted Sampling: Randomly sample a token from left to right only if $p(f_j | \tilde{f}_1^{j-1}, e_1^I) > \tau$ [Graça & Kim⁺ 19]

source					**target**
Heute	ist	es	sonnig.	$\longleftarrow$	Today is sunny.
Heute	ist		sonnig.		
Heute	scheint	die	Sonne.		

- Allow less probable tokens: more variability
- τ prohibits nonsense tokens in sampling
- Much faster than beam search: $O(\log_2 V) \ll O(NV)$ per position

Comparison of Generation Strategies

Turkish→English	Generation Strategy	newstest2016 BLEU [%]	newstest2016 TER [%]	newstest2017 BLEU [%]	newstest2017 TER [%]
Real bilingual data	-	19.0	70.5	18.9	71.1
+ Synthetic data	Beam search	24.5	65.7	23.1	67.2
	N-best list	24.7	65.7	23.1	67.1
	Restricted sampling	**25.0**	**64.7**	**23.6**	**66.2**

Restricted sampling is the best strategy to synthesize bilingual data

- More realistic variability in syntax and semantics

Translated by	Proportion [%] Delete	Insert	Reorder
Human	11.1	10.2	22.4
Beam Search	7.6	7.7	20.1
Restricted Sampling	8.1	8.6	20.7

- More suitable for large-scale synthesis

Monolingual Pre-Training vs. Synthetic Data

	newstest2016		newstest2017	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Baseline	19.0	70.5	18.9	71.1
+ Pre-training	20.0	68.5	19.8	69.2
+ Synthetic data	**25.0**	**64.7**	**23.6**	**66.2**
+ Pre-training + synthetic data	25.1	64.2	23.7	66.2

Synthesizing data is a more effective semi-supervised method

- Provides additional training data in the exact form expected by the model
- Up to +6.0% BLEU and -4.9% TER
- Combination with monolingual pre-training yields no significant difference

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks

Conclusion

Non-English Tasks

Data condition

- source-target: small
- source-English: large
- English-target: large

Language Pair	#sentences
French-German	2.5M
French-English	35M
English-German	9.1M

Language Pair	#sentences
German-Czech	226k
German-English	10M
English-Czech	49M

How can we utilize large bilingual data of related language pairs?

- **Cross-lingual Learning** [Kim & Petrov$^+$ 19]

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- Training
- Data

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- **Training**
- Data

Conclusion

Baseline: Pivoting

Two-step translation with English as a **pivot** language

source		**pivot**		**target**
Aujourd'hui est ensoleillé.	$\longrightarrow$	Today is sunny.	$\longrightarrow$	Heute ist es sonnig.
f_1 f_2 f_3		g_1 g_2 g_3		e_1 e_2 e_3 e_4
f_1^J		g_1^K		e_1^I

- Slow: doubled decoding time
- Translation errors are propagated or expanded from pivot to target
- Cannot utilize source-target bilingual data

Cross-lingual Learning: Training

Can we avoid pivoting and build a better single-size model?

- Faster translation: perform decoding only once
- No propagation of errors in the middle
- Utilize all three data sources: source-target, source-English, English-target

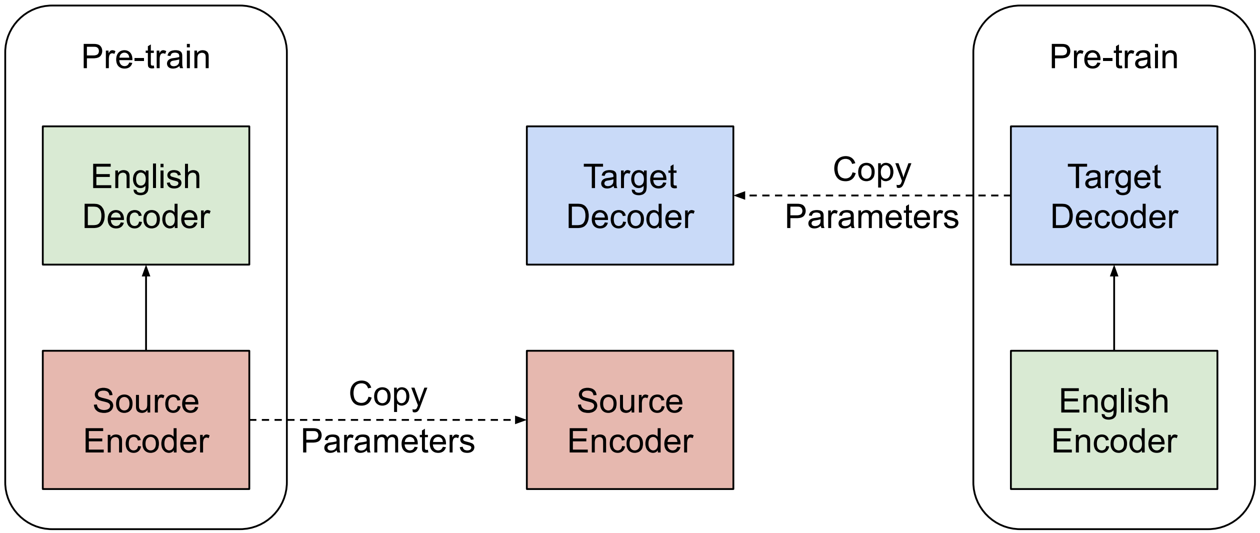
Sequential Transfer: Pre-Training for base tasks $\rightarrow$ Training for the main task

- Shared model parameters throughout several training stages
- Optimized to the main task at the end

Sequential Transfer: Individual Pre-Training

1. Pre-train source→English and English→target models
   - The two models do not depend on each other (parallelizable)

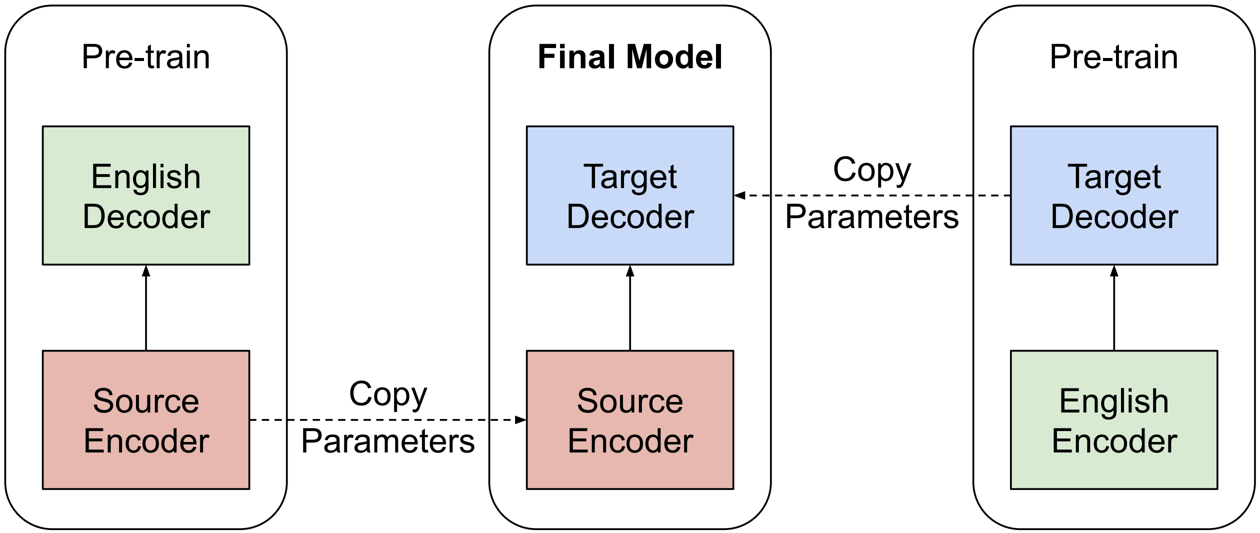
Sequential Transfer: Individual Pre-Training



2. Take components from pre-trained models
   - Source encoder from source→English model
   - Target decoder from English→target model

Problem: Individually pre-trained components are not compatible with each other

Sequential Transfer: Individual Pre-Training



3. Combine the pre-trained components and train with source-target data
   - Learn to connect encoder representations with decoder computations
   - What if source-target data is small? (**low-resource**)

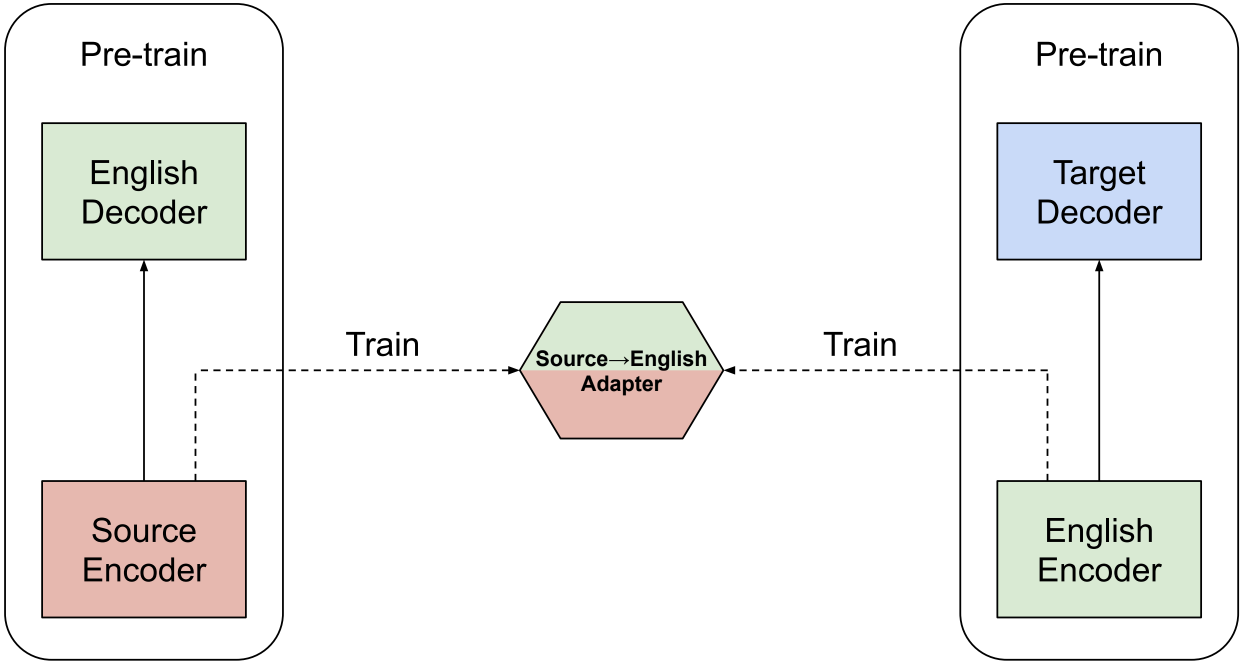
Sequential Transfer: Pivot Adapter

How can we mitigate the mismatch between components after pre-training?

Pivot Adapter: Transform source encoder outputs like English encoder outputs

- Target decoder is trained to use English encoder outputs
- Source encoder produces outputs which are familiar to target decoder

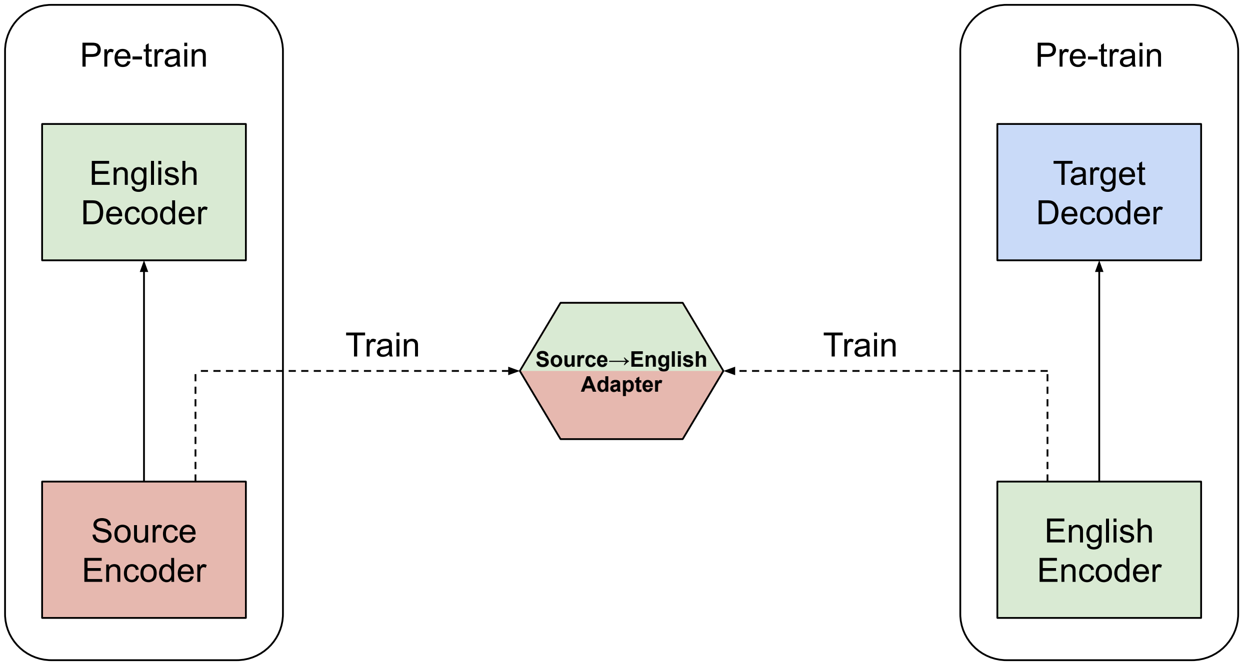
Sequential Transfer: Pivot Adapter



1. Feed source/English encoders with source-English data: get representation pairs

$$f_1^J \xrightarrow{\text{encoder}} \mathbf{h}_{f,1}^J \xrightarrow{\text{pooling}} \mathbf{h}_f$$
$$g_1^K \xrightarrow{\text{encoder}} \mathbf{h}_{g,1}^K \xrightarrow{\text{pooling}} \mathbf{h}_g$$

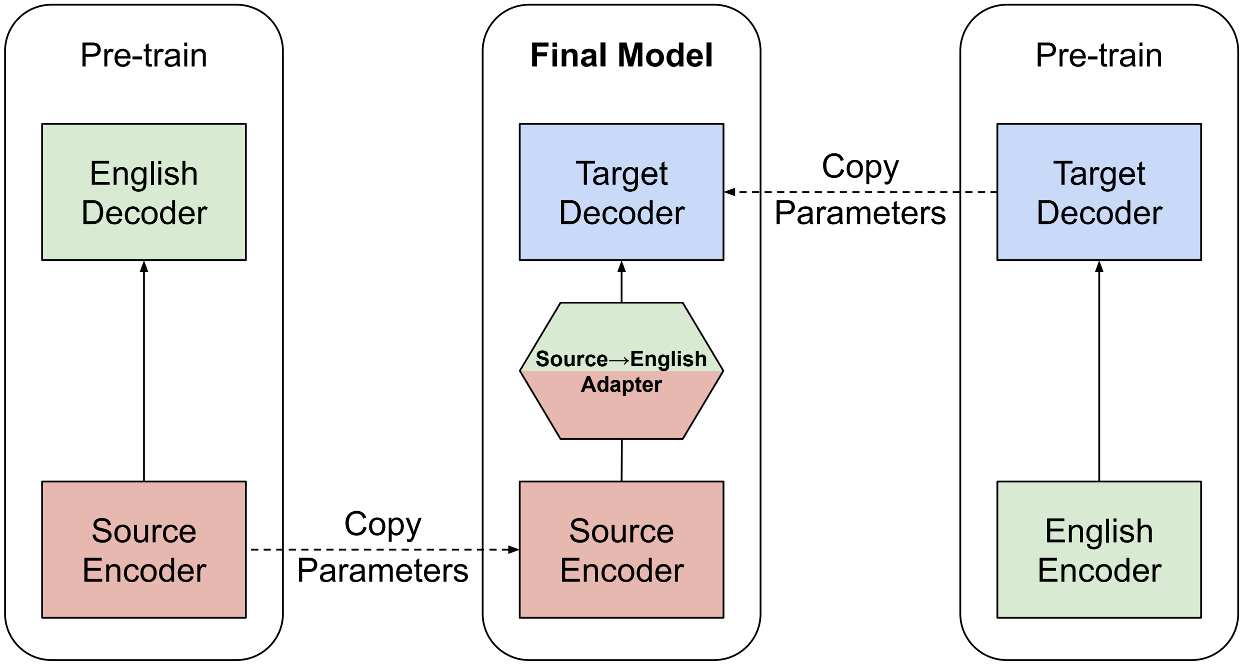
Sequential Transfer: Pivot Adapter



2. Train a linear mapping from source encoder outputs to English encoder outputs

$$\mathbf{W}_{f \to g} = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{(\mathbf{h}_f, \mathbf{h}_g)} \|\mathbf{W} \cdot \mathbf{h}_f - \mathbf{h}_g\|^2$$

Sequential Transfer: Pivot Adapter



3. Insert the adapter layer between source encoder and target encoder
   - Smoother connection of representation spaces
   - Continue training with source-target data

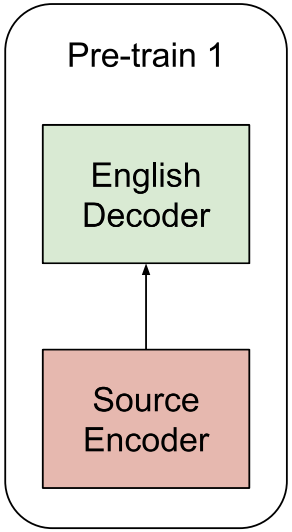
Sequential Transfer: Step-wise Transfer

How can we fundamentally prevent the mismatch between components during pre-training?

Step-wise Pre-Training: Pre-train for source→English and English→target in consecutive steps

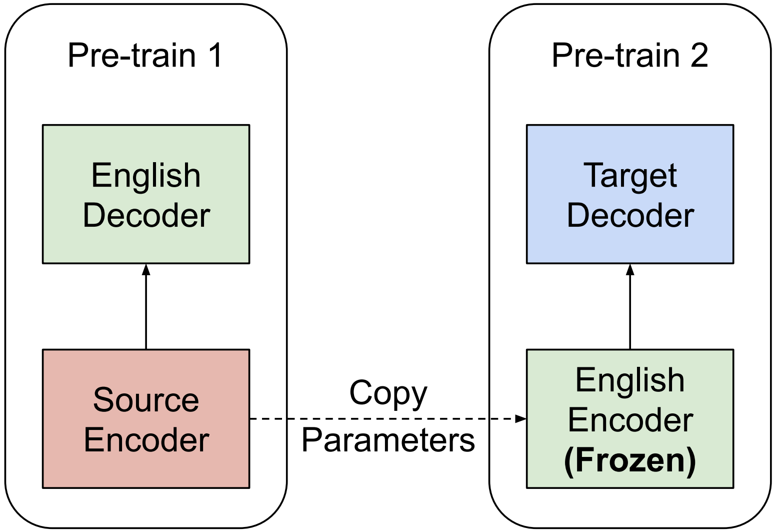
- Same data, different order of training
- Explicitly force target decoder to use source encoder representations

Sequential Transfer: Step-wise Transfer



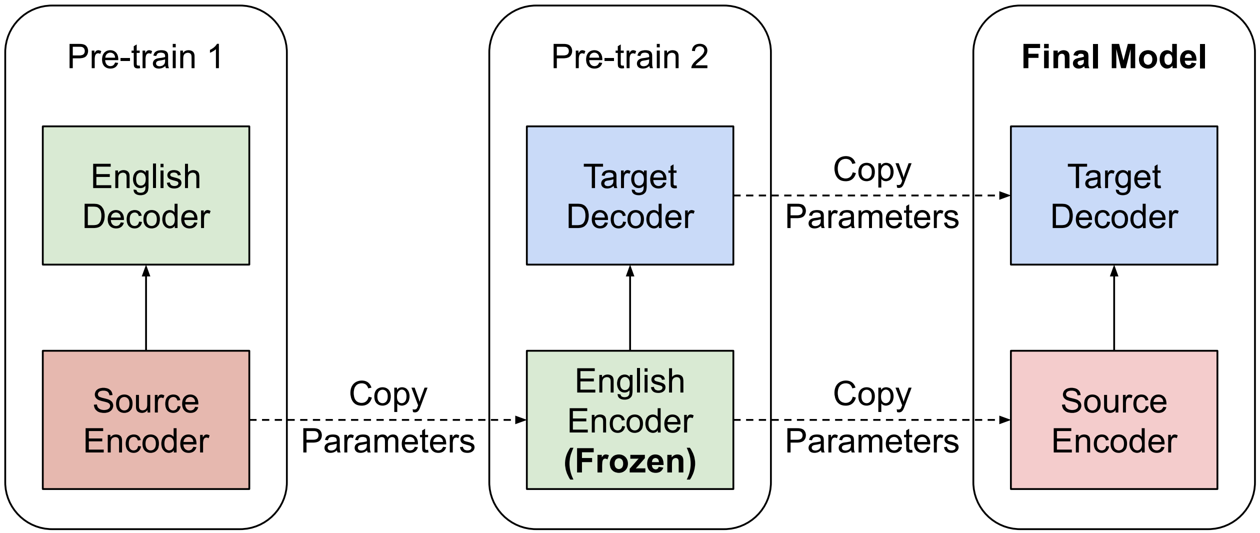
1. Pre-train source→English model

Sequential Transfer: Step-wise Transfer



2. Take source encoder parameters and train English→target model
   - English sentences are fed to (frozen) source encoder: random semantics
   - Target at least learns to use source encoder's representation space

Sequential Transfer: Step-wise Transfer



3. Continue training with source-target data
   - Encoder was frozen: Can still model source sentences well
   - Decoder computations are already connected with encoder representations

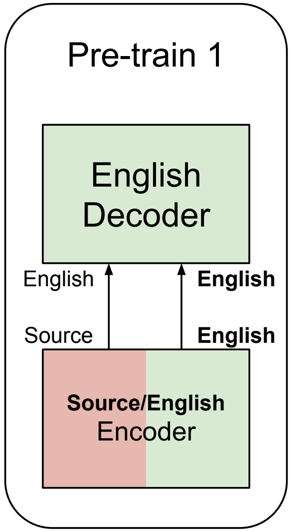
Sequential Transfer: Cross-lingual Encoder

How can we improve the second pre-training step (English$\rightarrow$target)?

Cross-lingual Encoder: Encoder models source and English languages together

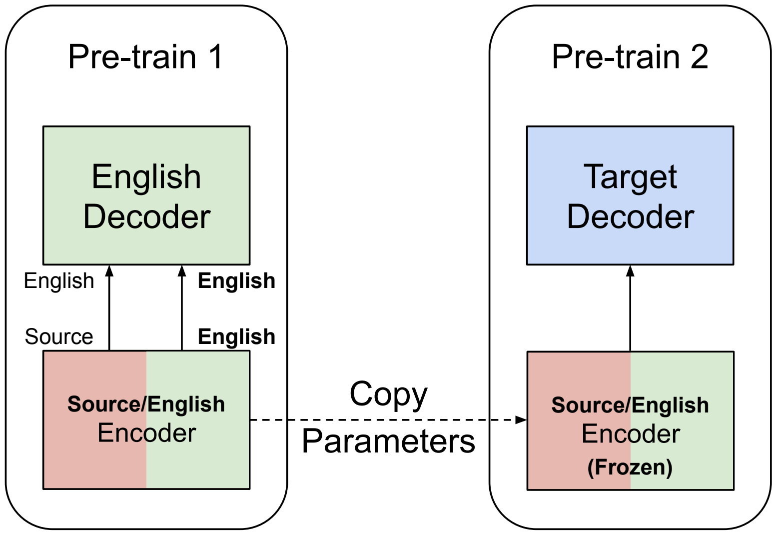
- Encodes source and English sentences in the same mathematical space
- Convey meaningful English representations to target decoder

Sequential Transfer: Cross-lingual Encoder



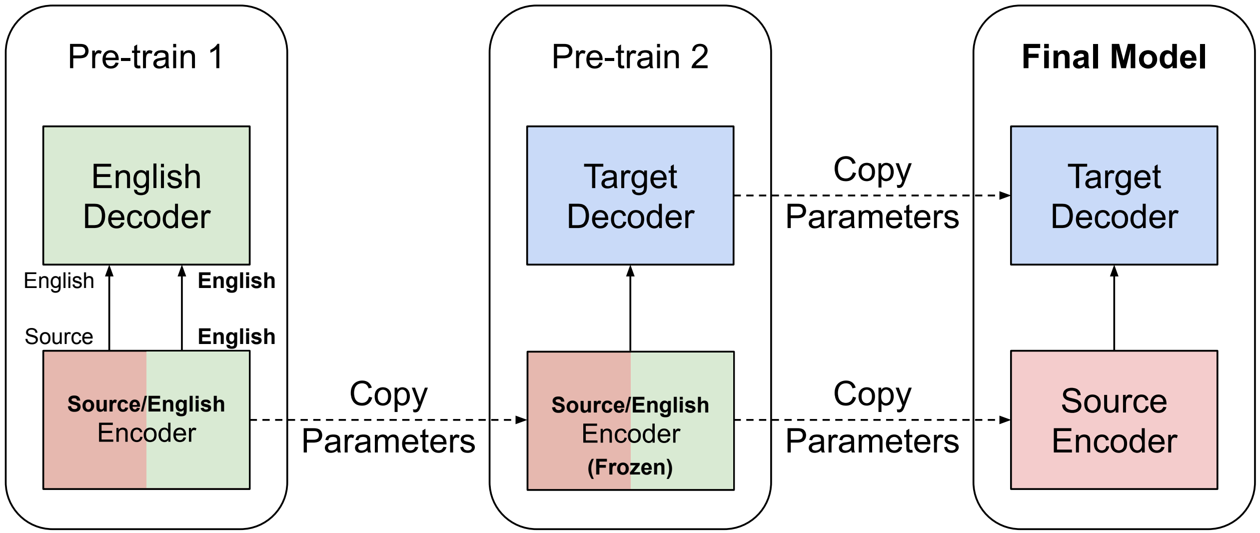
1. Pre-training for source/English→English with source-English data
   - Source→English: source as input, English as output
   - English→English: English as both input and output (autoencoding)
   - Similar encoder output for paired source-English sentences
   - Also used in parallel corpus mining/filtering
     [Rossenbach & Rosendahl$^+$ 18, Kim & Rosendahl$^+$ 19]

Sequential Transfer: Cross-lingual Encoder



2. Take source/English encoder parameters and train with English-target data
   - Encoder produces meaningful semantics for target decoder
   - Decoder learns to work with (shared) source representation space even if the input is in English

Sequential Transfer: Cross-lingual Encoder



3. Continue training with source-target data
   - Decoder has better initial parameters for the last training step

Sequential Transfer: Experiments

German→Czech	newstest2012		newstest2013	
	Bleu [%]	Ter [%]	Bleu [%]	Ter [%]
Direct source→target	12.0	79.7	13.5	76.3
Individual pre-training	15.4	75.4	18.0	70.9
+ Pivot adapter	15.9	75.0	18.7	70.3
Step-wise pre-training	15.6	75.0	18.1	70.9
+ Cross-lingual encoder	**16.2**	**74.6**	**19.1**	**69.9**
Pivoting	18.0	73.6	21.3	68.8

Transfer learning from two high-resource language pairs gives large improvement

- Pivot adapter gives additional performance gain

Sequential Transfer: Experiments

German→Czech	newstest2012		newstest2013	
	Bleu [%]	Ter [%]	Bleu [%]	Ter [%]
Direct source→target	12.0	79.7	13.5	76.3
Plain transfer	15.4	75.4	18.0	70.9
+ Pivot adapter	15.9	75.0	18.7	70.3
Step-wise pre-training	15.6	75.0	18.1	70.9
+ Cross-lingual encoder	**16.2**	**74.6**	**19.1**	**69.9**
Pivoting	18.0	73.6	21.3	68.8

Best combination = step-wise pre-training + cross-lingual encoder

- Direct connection between pre-trained components + fully utilizing high-resource data in all pre-training stages
- +1.1% Bleu, -1.0% Ter against simple sequential transfer
- Still behind pivoting

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- Training: Sequential Transfer
- Data

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- Training: Sequential Transfer
- **Data**

Conclusion

Cross-lingual Learning: Synthetic Data

How can we synthesize bilingual data from source-English and English-target data?

Pivot-based Back-Translation: Translate pivot side of pivot-target data into source language [Bertoldi & Barbaiani$^+$ 08]

source		pivot		target
Aujourd'hui est ensoleillé.	$\longleftarrow$	Today is sunny.	—	Heute ist es sonnig.
$\tilde{f}_1$ $\tilde{f}_2$ $\tilde{f}_3$		g_1 g_2 g_3		e_1 e_2 e_3 e_4
$\tilde{f}_1^{\tilde{J}}$		g_1^K		e_1^I
(generated source)		(pivot-target bilingual data)		

- Use high-resource pivot$\rightarrow$source model: high-quality generations (c.f. low-resource target$\rightarrow$source model for semi-supervised learning)

Sequential Transfer with Pivot-based Synthetic Data

Add pivot-based synthetic data in source→target training step

German→Czech	newstest2012		newstest2013	
	Bleu [%]	Ter [%]	Bleu [%]	Ter [%]
Direct source→target	12.0	79.7	13.5	76.3
+ Synthetic data	15.7	76.5	18.5	72.0
Sequential transfer	16.2	74.6	19.1	69.9
+ Synthetic data	**18.0**	**72.7**	**21.3**	**68.0**
Pivoting	18.0	73.6	21.3	68.8

Low-resource: Synthetic data gives large additional gain to single-size models

- Sequential transfer <u>reaches</u> the pivoting performance with 2x faster decoding

Sequential Transfer with Pivot-based Synthetic Data

Add pivot-based synthetic data in source→target training step

French→German	newstest2012		newstest2013	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Direct source→target	20.1	69.8	21.9	69.2
+ Synthetic data	21.1	68.2	22.6	68.1
Sequential transfer	20.9	69.4	23.1	68.0
+ Synthetic data	**21.9**	**67.6**	**23.4**	**67.4**
Pivoting	20.6	68.9	22.3	68.5

Mid-resource: Synthetic data gives small additional gain to single-size models

- Sequential transfer outperforms the pivoting with 2x faster decoding

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- Training: Sequential Transfer
- Data: Pivot-based Back-Translation

Conclusion

Outline

Preliminaries

To-English Tasks: Semi-supervised Learning

- Training: Monolingual Pre-Training
- Data: Back-Translation

Non-English Tasks: Cross-lingual Learning

- Training: Sequential Transfer
- Data: Pivot-based Back-Translation

Conclusion

Conclusion

Question: Given a small amount of bilingual data, what should we do to make a good machine translation model?

Answer: Exploit additional data sources

- To-English tasks: monolingual data
- Non-English tasks: source-English and English-target bilingual data

How?

- Most important: **Synthesize** bilingual data with restricted sampling
- **Pre-train** model parameters for related tasks in the right order

In the thesis, you can find also:

- Unsupervised learning for (neural) machine translation

End

Thank you!

`kim@cs.rwth-aachen.de`

References

[Bertoldi & Barbaiani+ 08] N. Bertoldi, M. Barbaiani, M. Federico, R. Cattoni.
Phrase-based statistical machine translation with pivot languages.
In Proceedings of 5th International Workshop on Spoken Language Translation (IWSLT 2008), pp. 143–149, Honolulu, HI, USA, October 2008.

[Devlin & Chang+ 19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp. 4171–4186, Minneapolis, MN, USA, June 2019.

[Graça & Kim+ 19] M. Graça, Y. Kim, J. Schamper, S. Khadivi, H. Ney.
Generalizing back-translation in neural machine translation.
In Proceedings of the 4th Conference on Machine Translation (WMT 2019), pp. 45–52, Florence, Italy, August 2019.

[Kim & Petrov+ 19] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, H. Ney.
Pivot-based transfer learning for neural machine translation between non-English languages.
In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), pp. 866–876, Hong Kong, China, November 2019.

[Kim & Rosendahl+ 19] Y. Kim, H. Rosendahl, N. Rossenbach, J. Rosendahl, S. Khadivi, H. Ney.
Learning bilingual sentence embeddings via autoencoding and computing similarities with a multilayer perceptron.
In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP 2019), pp. 61–71, Florence, Italy, August 2019.

[Kingma & Ba 15] D. P. Kingma, J. Ba.
Adam: A method for stochastic optimization.
In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 2015.

[Ott & Auli+ 18] M. Ott, M. Auli, D. Grangier, M. Ranzato.
Analyzing uncertainty in neural machine translation.
In Proceedings of the 35th International Conference on Machine Learning (ICML 2018), pp. 3956–3965, Stockholm, Sweden, July 2018.

References

[Ramachandran & Liu+ 17] P. Ramachandran, P. Liu, Q. Le.
Unsupervised pretraining for sequence to sequence learning.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pp. 383–391, Copenhagen, Denmark, September 2017.

[Rossenbach & Rosendahl+ 18] N. Rossenbach, J. Rosendahl, Y. Kim, M. Graça, A. Gokrani, H. Ney.
The RWTH aachen university filtering system for the WMT 2018 parallel corpus filtering task.
In Proceedings of the 3rd Conference on Machine Translation (WMT 2018), pp. 946–954, Belgium, Brussels, October 2018.

[Sennrich & Haddow+ 16] R. Sennrich, B. Haddow, A. Birch.
Improving neural machine translation models with monolingual data.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pp. 86–96, Berlin, Germany, July 2016.

[Taylor 53] W. L. Taylor.
Cloze procedure: A new tool for measuring readability.
Journalism Quarterly, Vol. 30, No. 4, pp. 415–433, 1953.

[Vaswani & Shazeer+ 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. ukasz Kaiser, I. Polosukhin.
Attention is all you need.
In Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998–6008, Long Beach, CA, USA, December 2017.