Generating Alignments using Target Foresight in Attention-Based Neural Machine Translation

Jan-Thorsten Peter, Arne Nix, Hermann Ney

peter@cs.rwth-aachen.de

Mai 29, 2017

EAMT 2017, Prag

Human Language Technology and Pattern Recognition

Computer Science Department

RWTH Aachen University

Outline

Motivation

Neural Machine Translation

Target Foresight

Guided Alignment Training

Target Foresight with Guided Alignment Training

Conclusion

Motivation

- Alignment use to be important for SMT
- Neural Machine Translation (NMT) uses attention
- There are still application for alignments:
  - Guided alignment training [Chen & Matusov$^+$ 16]
  - Transread [1]
  - Linguee [2]
- Using the attention as alignment produces bad results
- Can we use NMT to create alignments?

[1] https://transread.limsi.fr
[2] http://www.linguee.com

Related Work

D. Bahdanau, K. Cho, Y. Bengio [Bahdanau & Cho^{+} 15]:
Neural machine translation by jointly learning to align and translate. *ICLR, May 2015*.

- Introducing an attention mechanism to neural machine translation

W. Chen, E. Matusov, S. Khadivi, J.-T. Peter [Chen & Matusov^{+} 16]:
Guided alignment training for topic-aware neural machine translation. *AMTA, October 2016*.

- Introduces guided alignment training

Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li [Tu & Lu^{+} 16]:
Modeling coverage for neural machine translation. *ACL, August 2016*.

- Analysing attention of neural machine translation using SAER

Outline

Motivation

Neural Machine Translation

Target Foresight

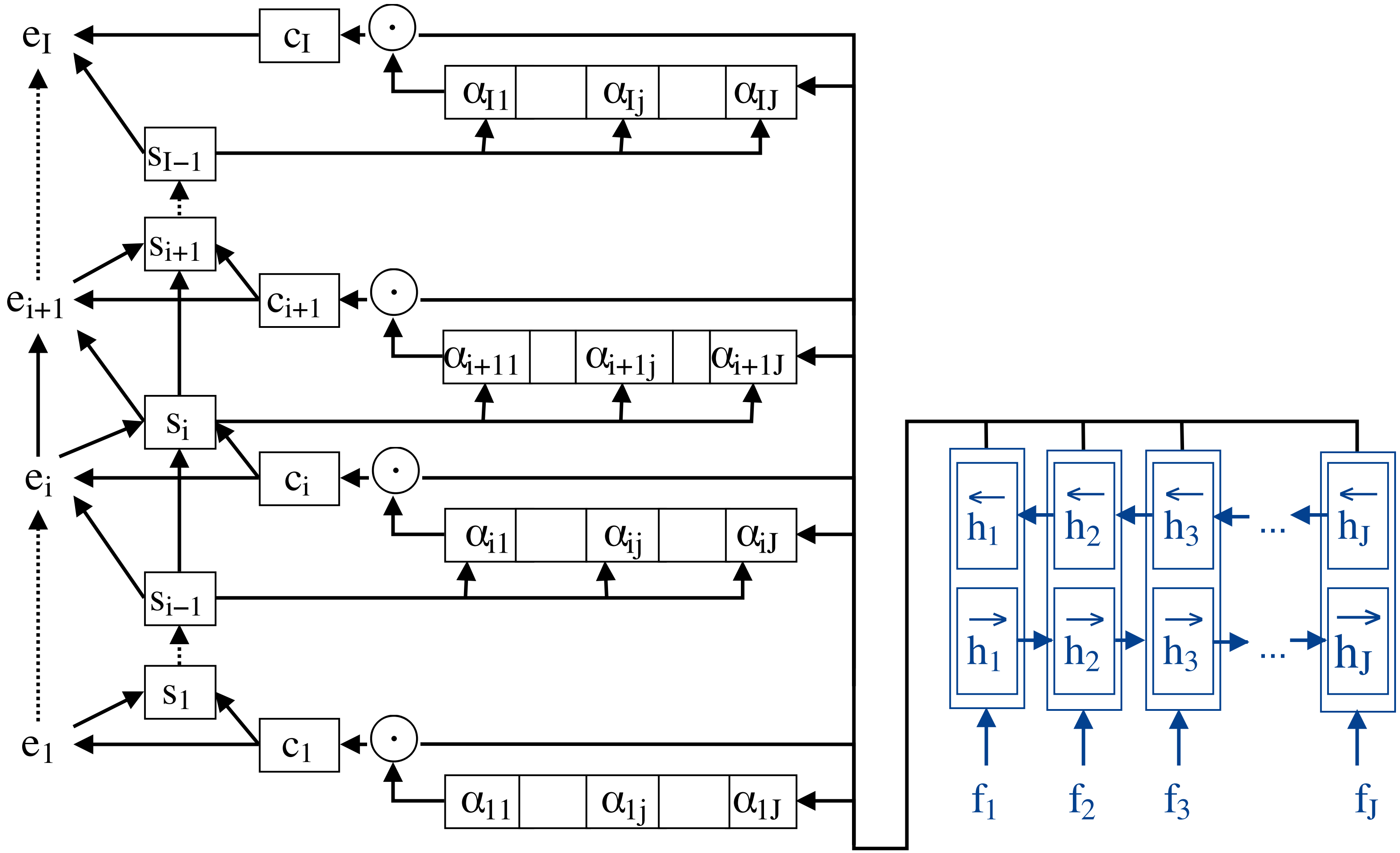
Guided Alignment Training

Target Foresight with Guided Alignment Training

Conclusion

Attention Based NMT

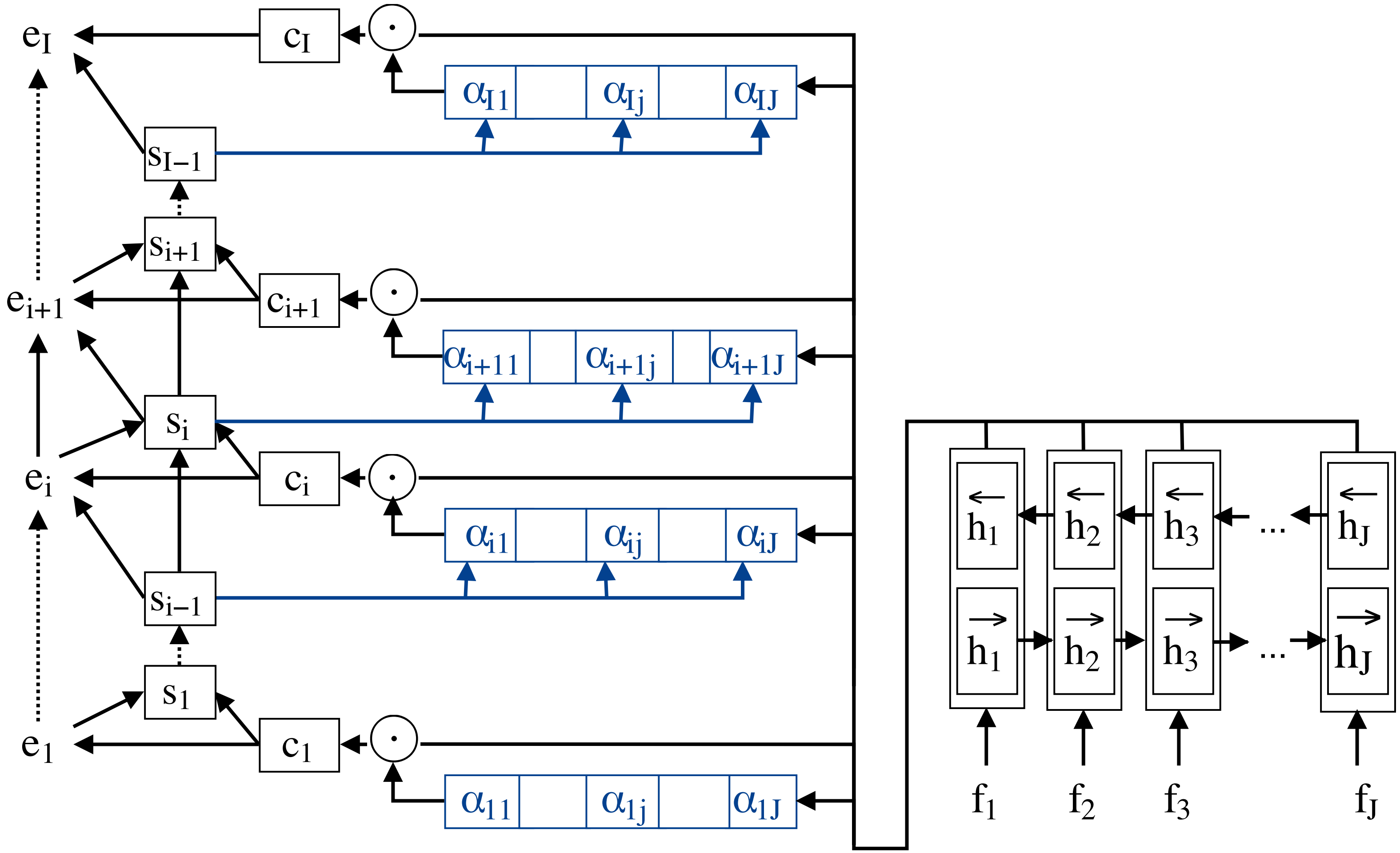
- **Bidirectional RNN encodes source sentence f_1^J into $\overrightarrow{h}_1^J$ and $\overleftarrow{h}_1^J$**
- $h_j := [\overrightarrow{h}_j^T; \overleftarrow{h}_j^T]^T$



Attention Based NMT

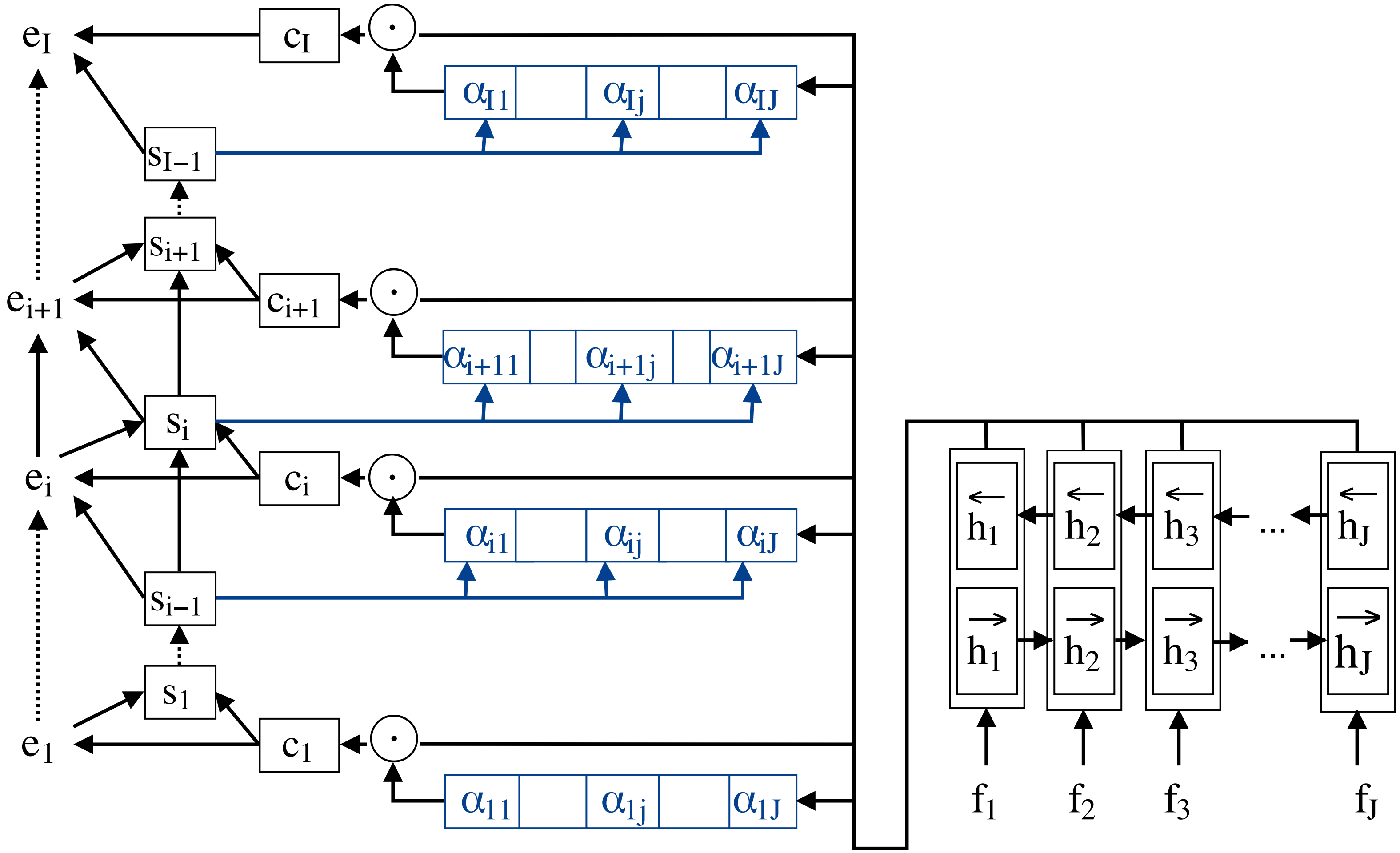
- **Energies computed through MLP:** $\tilde{\alpha}_{ij} = v_a^T \tanh(W_a s_{i-1} + U_a h_j)$

$W_a \in \mathbb{R}^{n \times n}, U_a \in \mathbb{R}^{n \times 2n}, v_a \in \mathbb{R}^{n}$**: weight parameters**



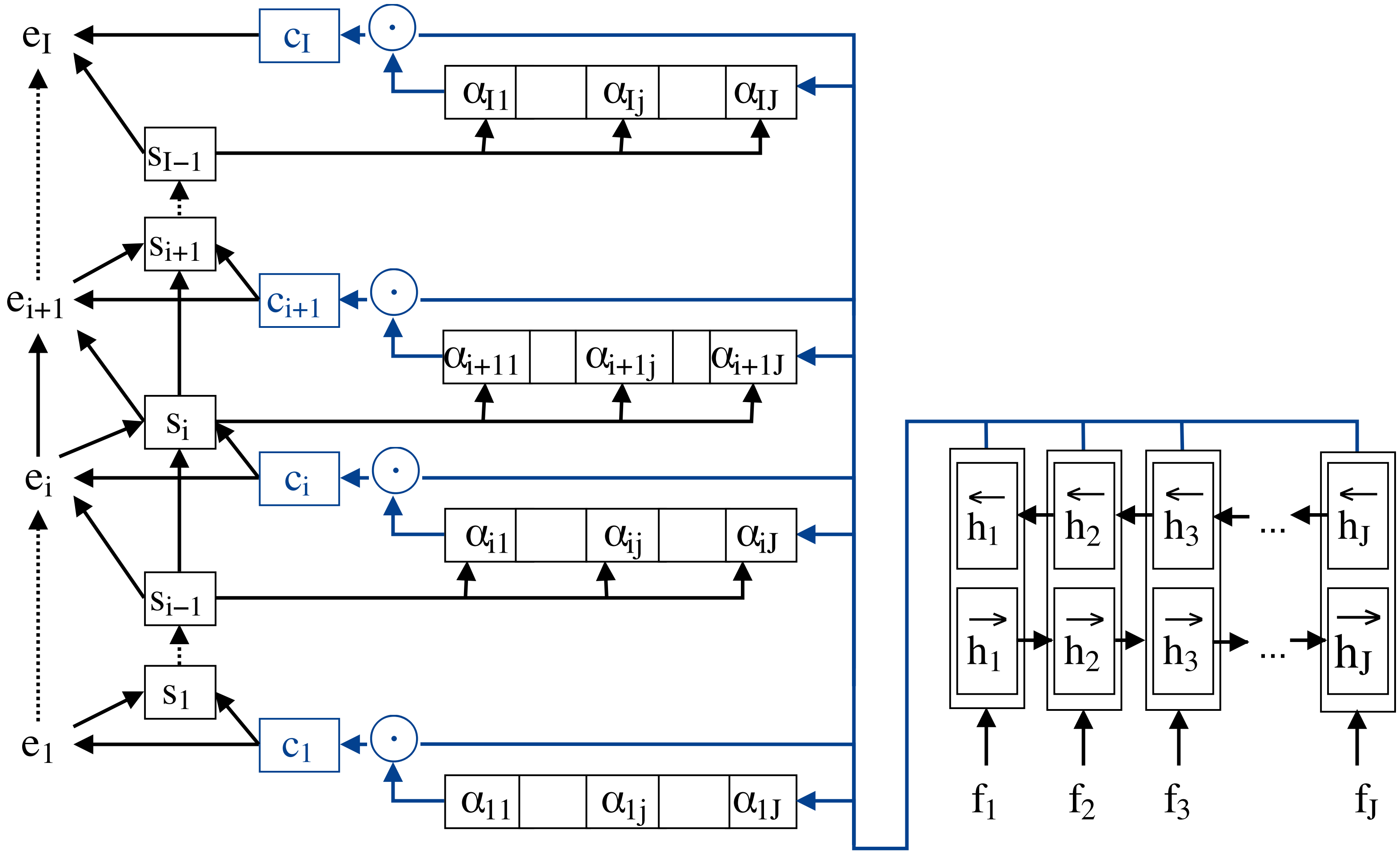
Attention Based NMT

- Attention weights normalized with softmax: $\alpha_{ij} = \frac{\exp(\tilde{\alpha}_{ij})}{\sum_{k=1}^{J} \exp(\tilde{\alpha}_{ik})}$



Attention Based NMT

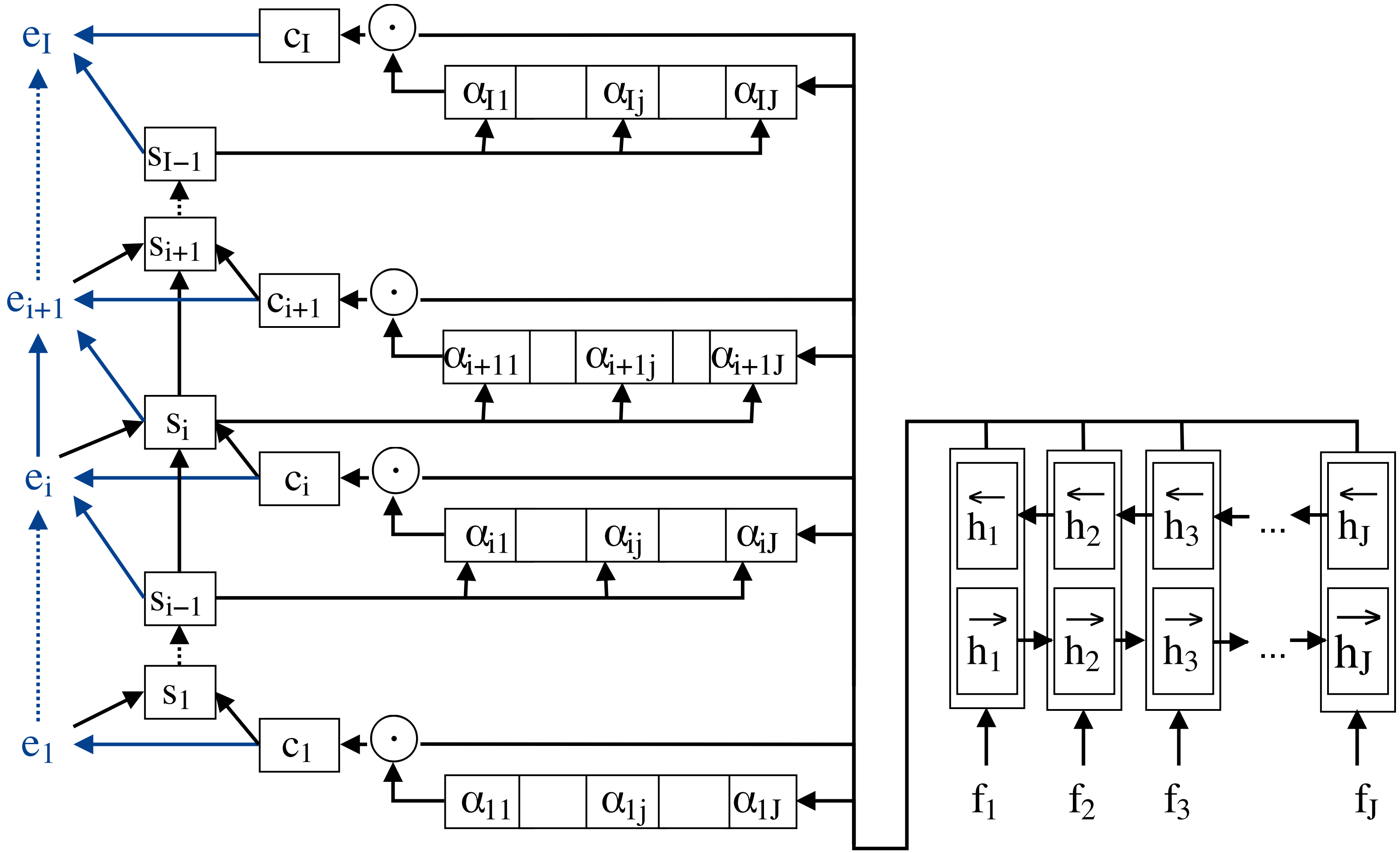
- Context vector as weighted sum: $c_i = \sum_{j=1}^{J} \alpha_{ij} h_j$



Attention Based NMT

- **Neural network output:** $p(e_i|e_1^{i-1}, f_1^J) = g_{\text{out}}(e_{i-1}, s_{i-1}, c_i)$

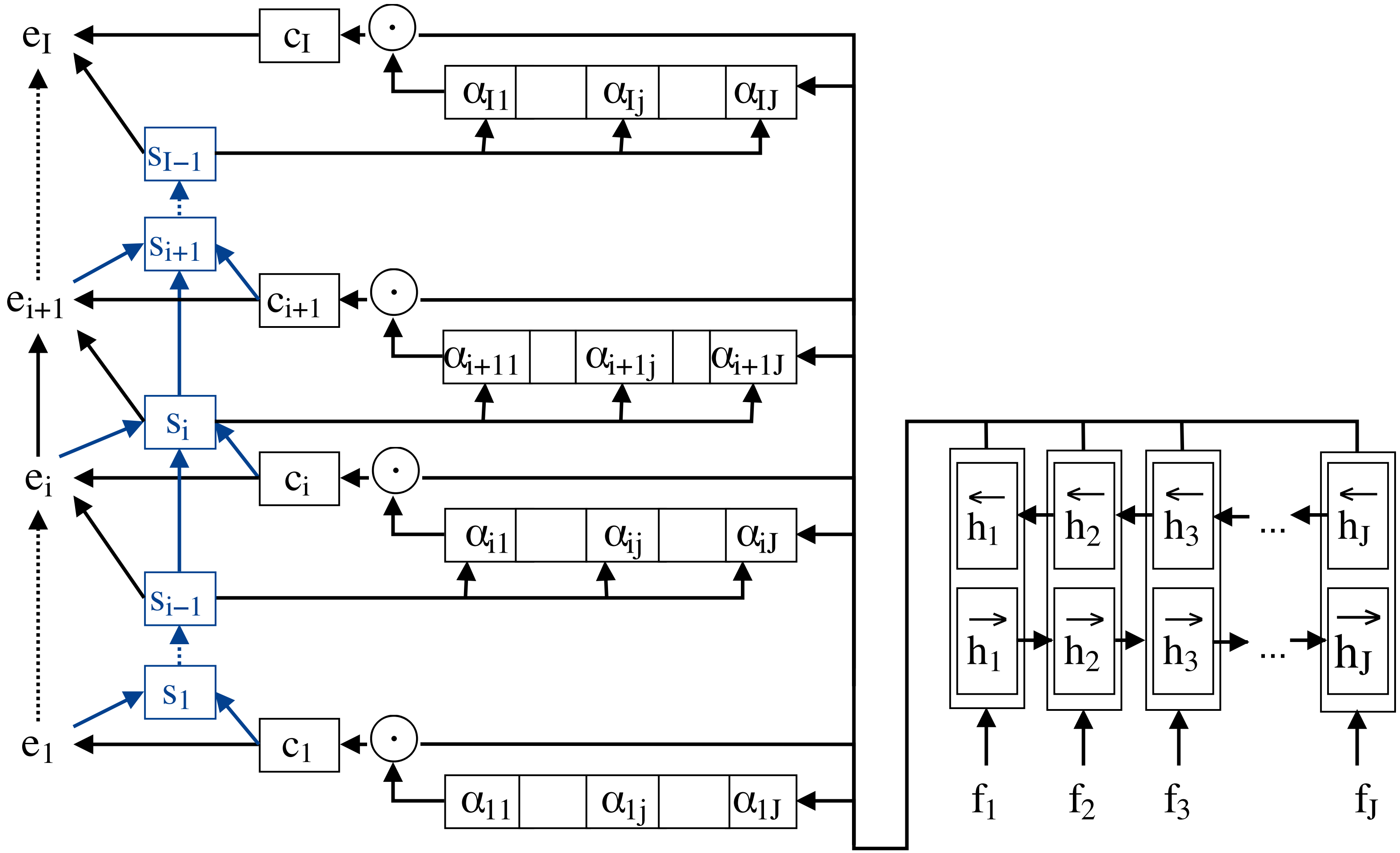
g_{out}: **output function**



Attention Based NMT

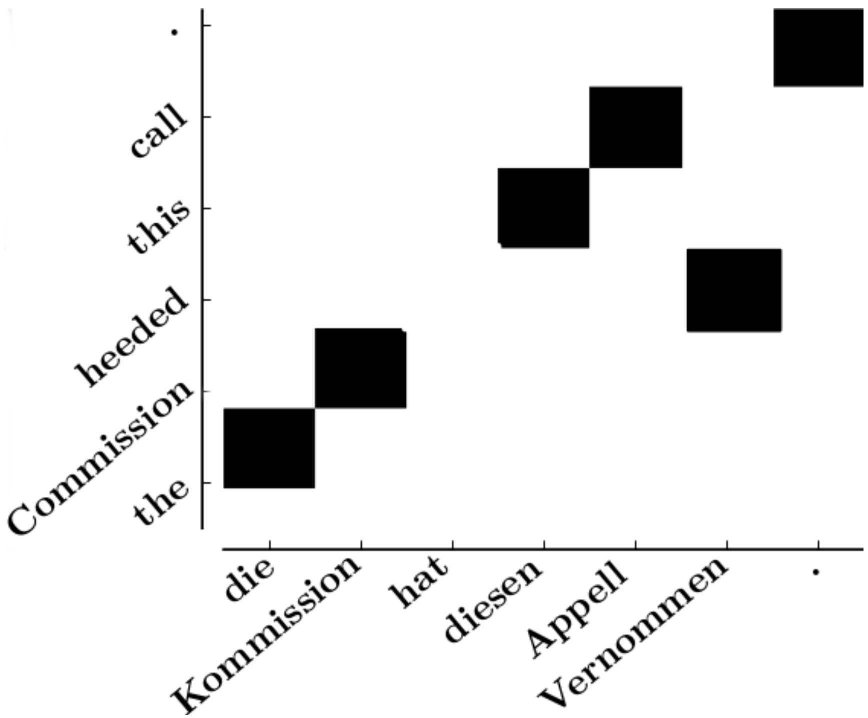
- **Hidden decoder state:** $s_i = g_{\text{dec}}(e_i, c_i; s_{i-1})$

g_{dec}: **gated recurrent unit**

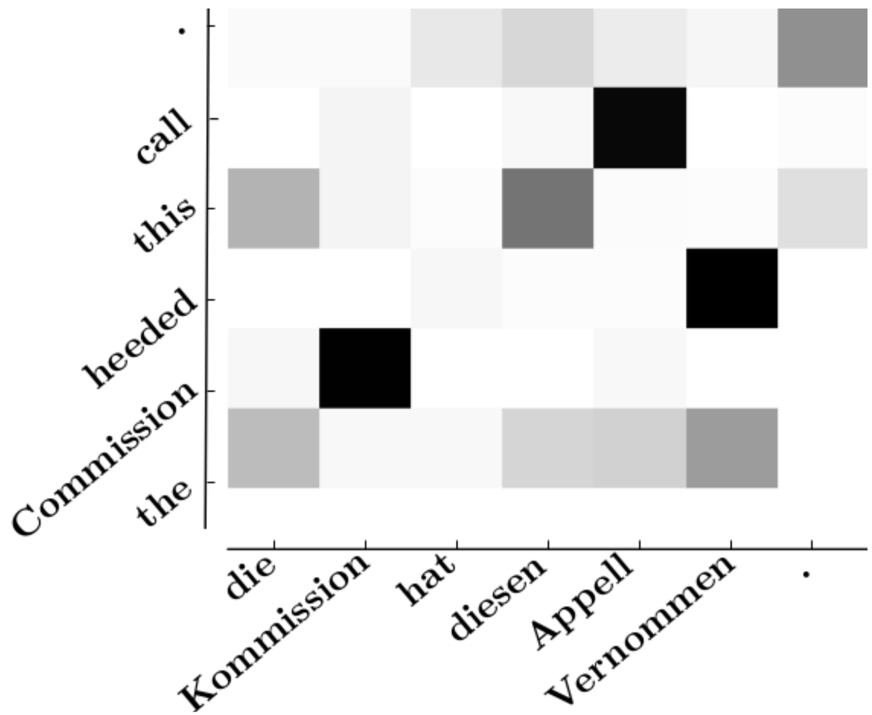


GIZA++ vs. NMT Alignment

GIZA++



NMT



- **GIZA++ creates a clean alignment**
- **Noise NMT alignment**

Alignment Error Rate

- **Alignment Evaluation:**

$$\text{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$ **[Och & Ney 03]**

$$\text{SAER}(M_S, M_P; M_A) = 1 - \frac{|M_A \odot M_S| + |M_A \odot M_P|}{|M_A| + |M_S|}$$ **[Tu & Lu+ 16]**

Europarl De-En	Alignment Test	
Model	AER%	SAER %
GIZA++	21.0	26.8
Attention-Based	38.1	63.6

- Attention is converted into hard alignment in both directions
- Merged using Och's refined method [Och & Ney 03].

Alignment Error Rate

- **Alignment Evaluation:**

$$\text{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \quad \text{[Och \& Ney 03]}$$

$$\text{SAER}(M_S, M_P; M_A) = 1 - \frac{|M_A \odot M_S| + |M_A \odot M_P|}{|M_A| + |M_S|} \quad \text{[Tu \& Lu}^+ \text{16]}$$

Europarl De-En	Alignment Test	
Model	AER%	SAER %
GIZA++	21.0	26.8
Attention-Based	38.1	63.6

- **Attention is converted into hard alignment in both directions**
- **Merged using Och's refined method [Och & Ney 03].**

Outline

Motivation

Neural Machine Translation

Target Foresight

Guided Alignment Training

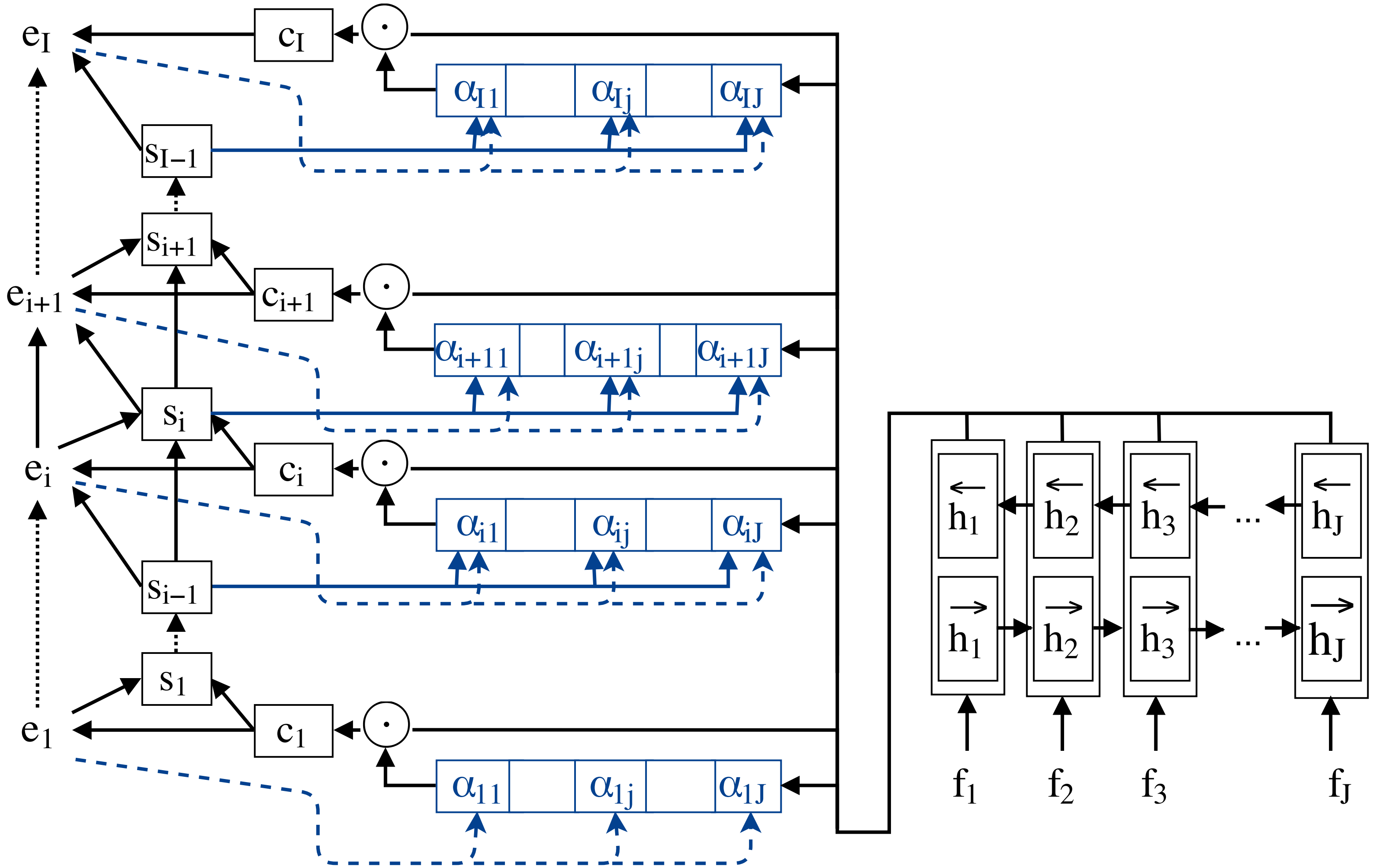
Target Foresight with Guided Alignment Training

Conclusion

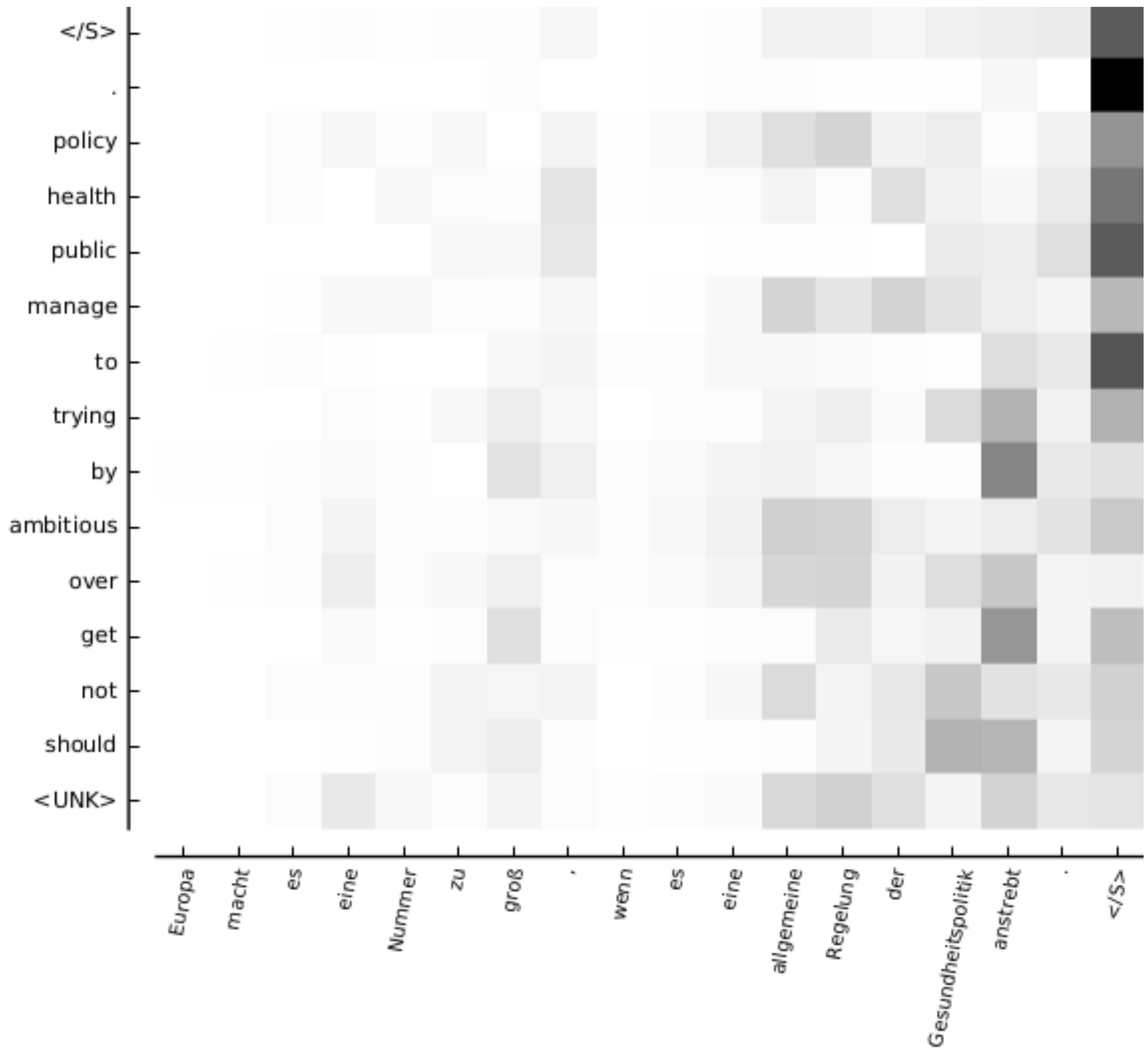
Target Foresight

- **Idea: Use knowledge of the target sentence e_1^I to improve the attention**

$$\tilde{\alpha}_{ij} = v_a^T \tanh(W_a s_{i-1} + U_a h_j + V_a \tilde{e}_i) \qquad V_a \in \mathbb{R}^{n \times p}$$



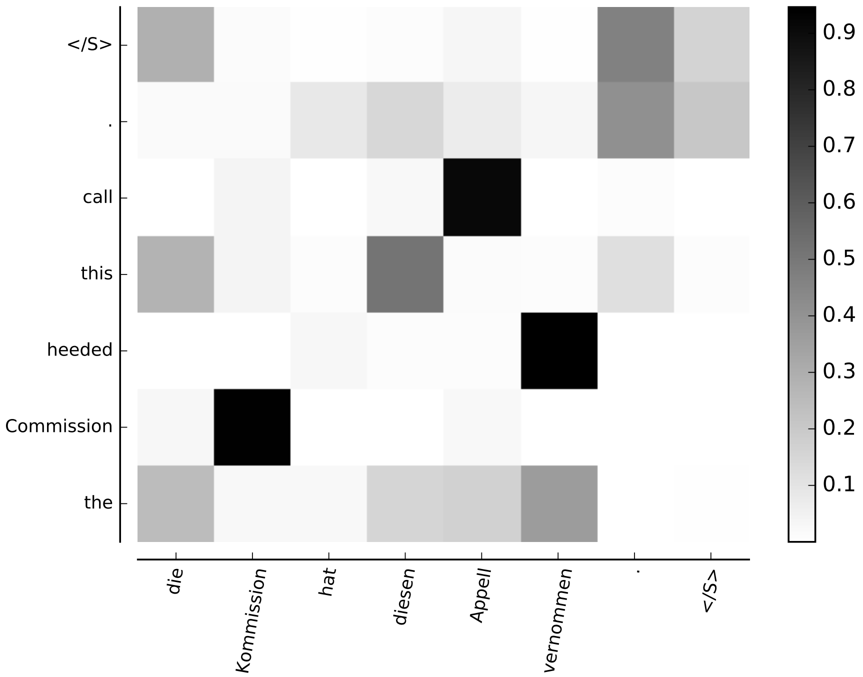
Raw Target Foresight



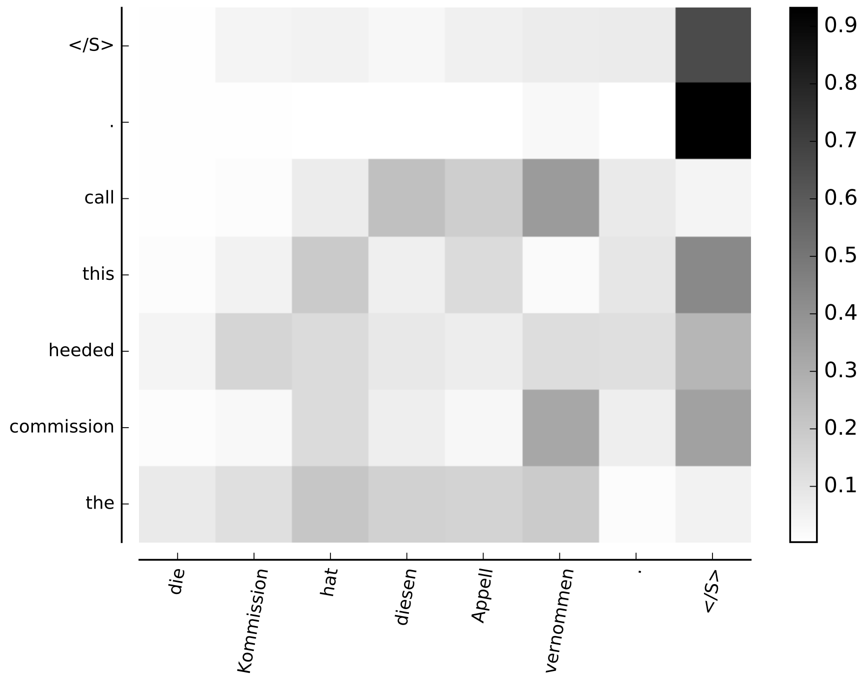
- **Target word encoded in source embedding and attention weights**

Target Foresight with Noise

NMT

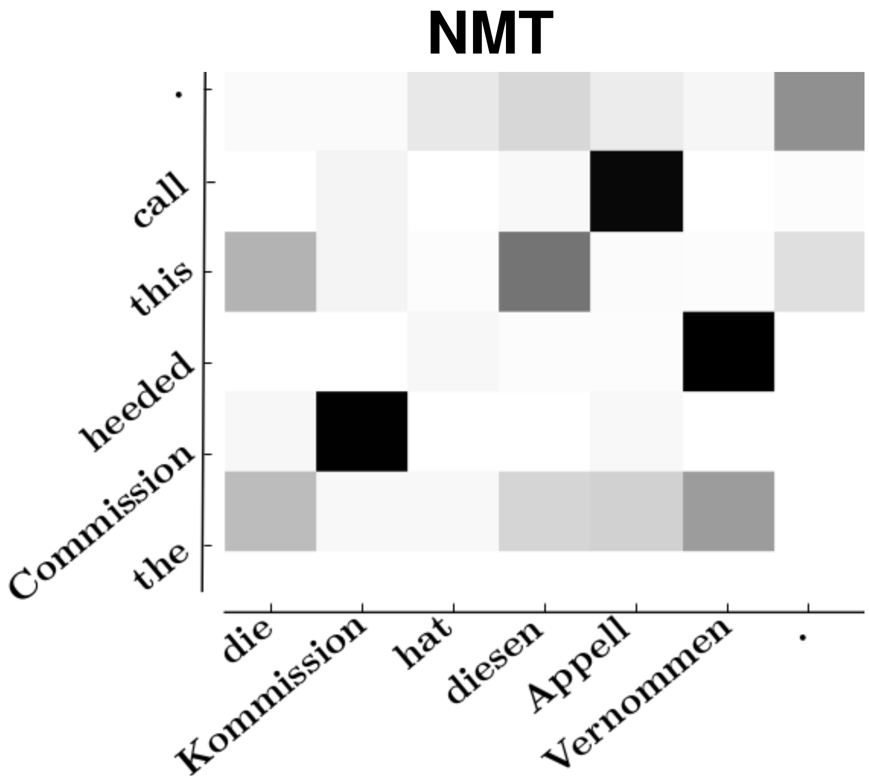


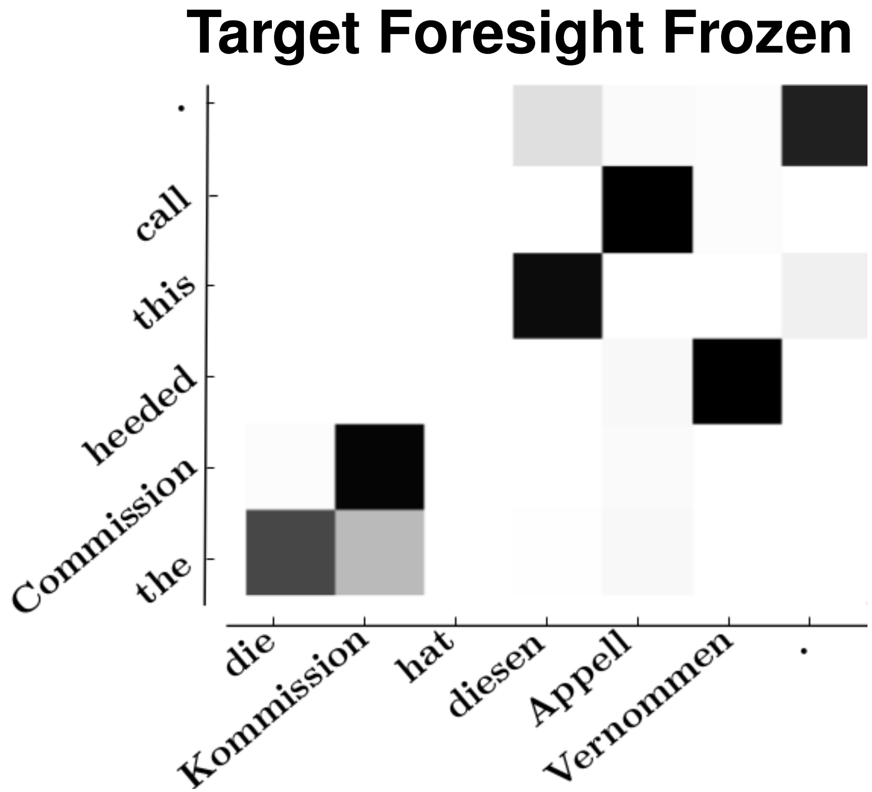
Target Foresight with Noise



- **Adding noise on attention does not help**

Freeze Encoder and Decoder

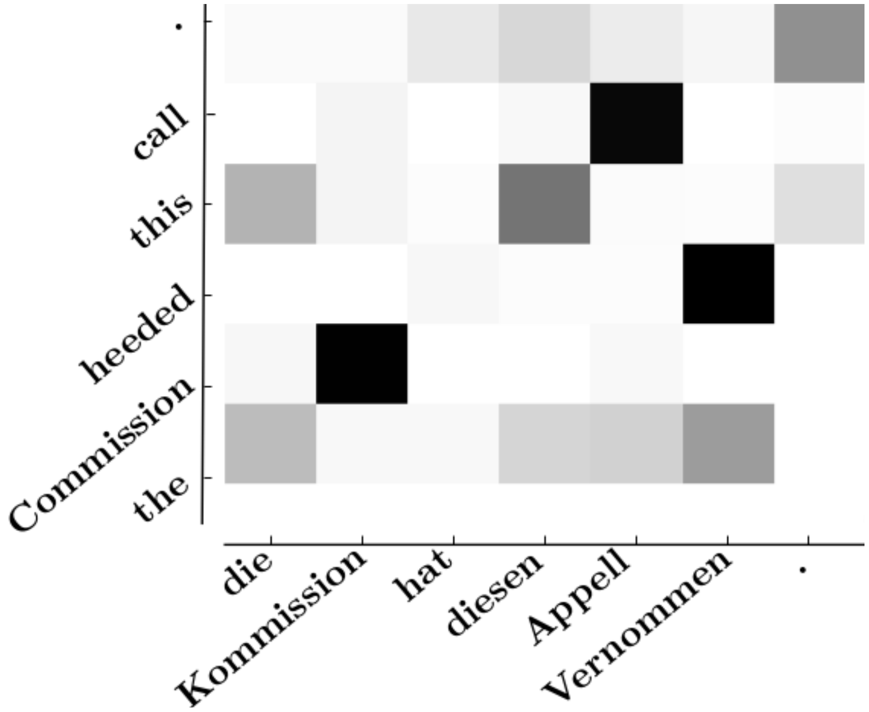




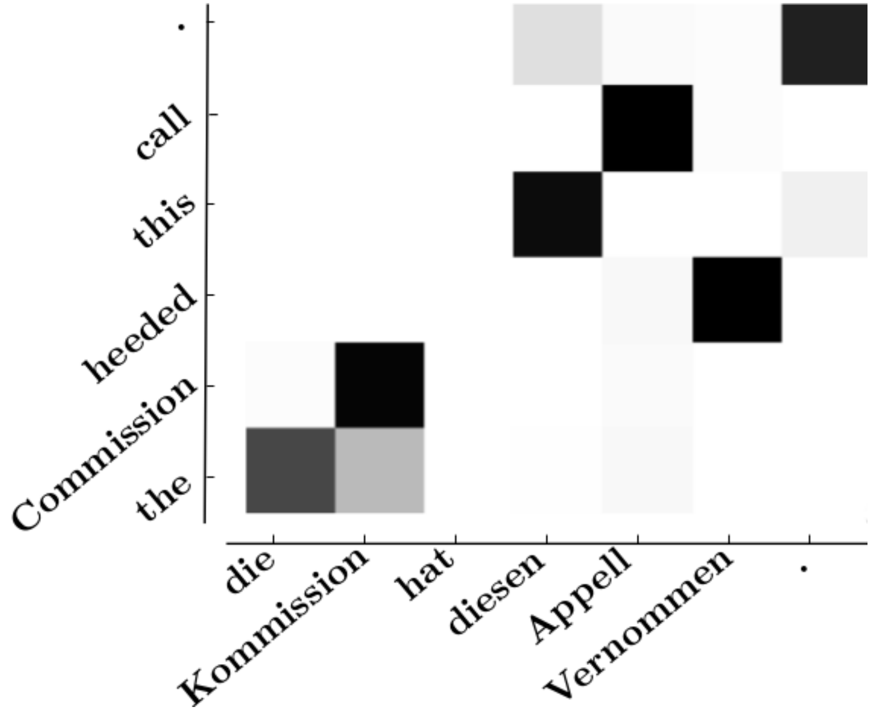
- Train baseline system
- Freeze encoder and decoder weights
- Continue training with target foresight

Freeze Encoder and Decoder

NMT



Target Foresight Frozen



	Alignment Test	
Model	AER %	SAER %
GIZA++	21.0	26.8
Attention-Based	38.1	63.6
+ Target foresight with frozen en-/decoder	33.9	55.6

Outline

Motivation

Neural Machine Translation

Target Foresight

Guided Alignment Training

Target Foresight with Guided Alignment Training

Conclusion

Guided Alignment Training

- **Idea: Introducing target alignment A as a second objective [Chen & Matusov$^+$ 16]**
- **Cross-Entropy cost $\mathcal{L}_{\text{align}}$ between the attention weights α and target alignment A**

$$\mathcal{L}_{\text{align}}(A, \alpha) := -\frac{1}{N} \sum_{n} \sum_{i=1}^{I_n} \sum_{j=1}^{J_n} A_{n,ij} \log \alpha_{n,ij}$$

- **Optimize w.r.t. $\mathcal{L}(A, \alpha, e_1^I, f_1^J) := \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}} + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}}$**
  - **$\mathcal{L}_{\text{CE}}$: standard decoder cost function (cross-entropy)**
  - **$\lambda_{\text{align}}, \lambda_{\text{CE}}$: weights determined through experiments**

Guided Alignment Training

IWSLT De-En	Test	Alignment Test	
Model	BLEU%	AER%	SAER%
Attention-Based	29.3	41.8	66.3
+ GA	30.3	35.4	44.2

- Improves translation by 1.0 BLEU on IWSLT2013 Test
- Great improvement in AER and SAER and Alignment Test
- Trained an all IWSLT 2013 data

Outline

Motivation

Neural Machine Translation

Target Foresight

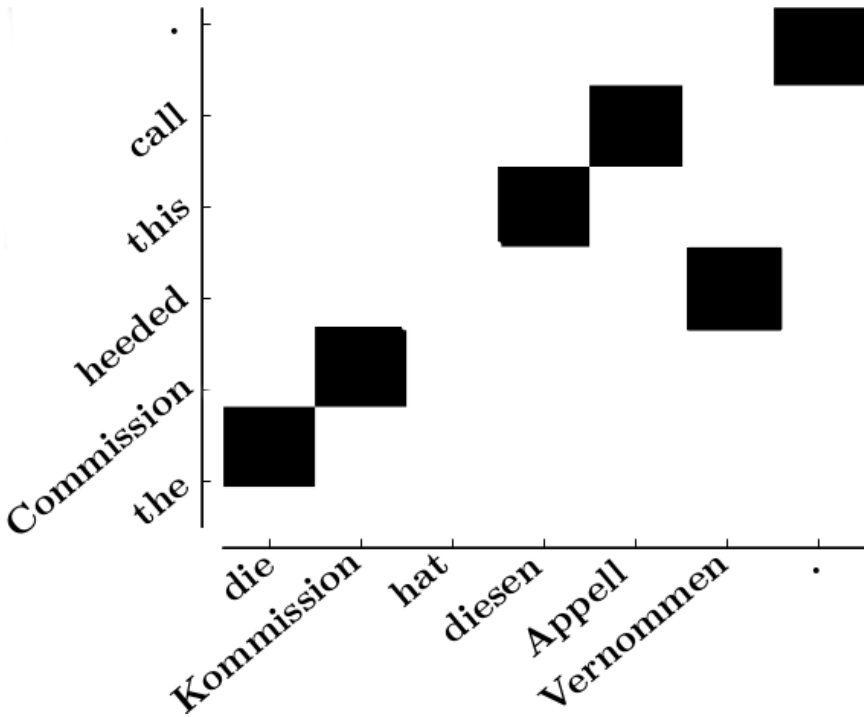
Guided Alignment Training

Target Foresight with Guided Alignment Training

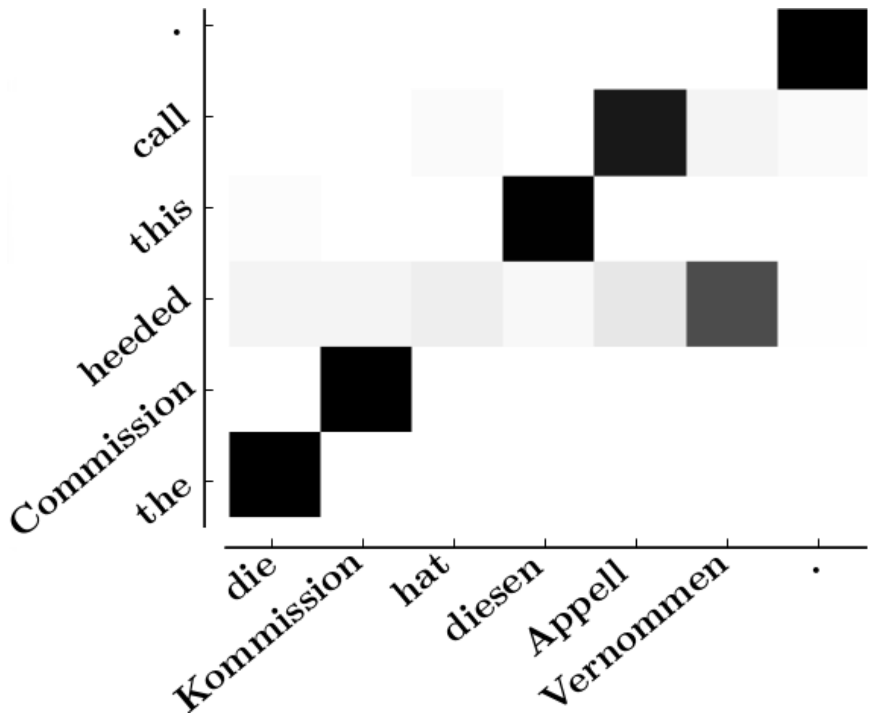
Conclusion

GIZA++ vs. Target Foresight with Guided Alignment

GIZA++



TF + GA



- **Target foresight creates correct alignment**

Results

Model	Alignment Test	
	AER %	SAER %
fast_align	27.9	33.0
GIZA++	21.0	26.8
BerkeleyAligner	20.5	26.4
Attention-Based	38.1	63.6
+ Guided alignment	29.8	38.0
+ Target foresight with frozen en-/decoder	33.9	55.6
+ Target foresight with guided alignment	19.0	34.9
+ converted to hard alignment	19.0	24.6

- Trained on Europal data
- Target foresight improves AER by 2.0% compared to GIZA++
- SAER is biased towards hard alignments

Results

Model	Alignment Test AER %	SAER %
fast_align	27.9	33.0
GIZA++	21.0	26.8
BerkeleyAligner	20.5	26.4
Attention-Based	38.1	63.6
+ Guided alignment	29.8	38.0
+ Target foresight with frozen en-/decoder	33.9	55.6
+ Target foresight with guided alignment	19.0	34.9
+ converted to hard alignment	19.0	24.6

- Trained on Europal data
- Target foresight improves AER by 2.0% compared to GIZA++
- SAER is biased towards hard alignments

Retrain Guided Alignment

- Use improved alignment for guided alignment training
- Test data: IWSLT 2013
- Train data: Europarl corpus

Model	Test BLEU	Alignment Test AER %	Alignment Test SAER %
Attention-Based	16.0	38.1	63.6
+ GA using GIZA++	18.4	29.8	38.0
+ GA using target-foresight alignments	18.8	28.5	36.7

Outline

Motivation

Neural Machine Translation

Target Foresight

Guided Alignment Training

Target Foresight with Guided Alignment Training

Conclusion

Conclusion

- **Improvement of AER by 2.0% compared to GIZA++**
- **Can easily be used to align unseen data**
- **Aligned data can again be used for guided alignment training**
- **Neural networks will cheat if it is possible**
- **Guided alignment keeps it from cheating**

Thank you for your attention!

Jan-Thorsten Peter, Arne Nix, Hermann Ney

peter@cs.rwth-aachen.de

Experiment Setup

System configuration:

- 30000 most frequent words as source and target vocabulary
- Out-of-vocabulary words are mapped to unknown tokens
- Bi-directional encoder with 1000 GRU nodes each
- GRU based decoder with 1000 nodes
- Alignment computation also has an internal dimension of 1000

Training:

- 250000 iterations
- Evaluation after each 10000 iterations on corresponding dev set

Europarl De-En

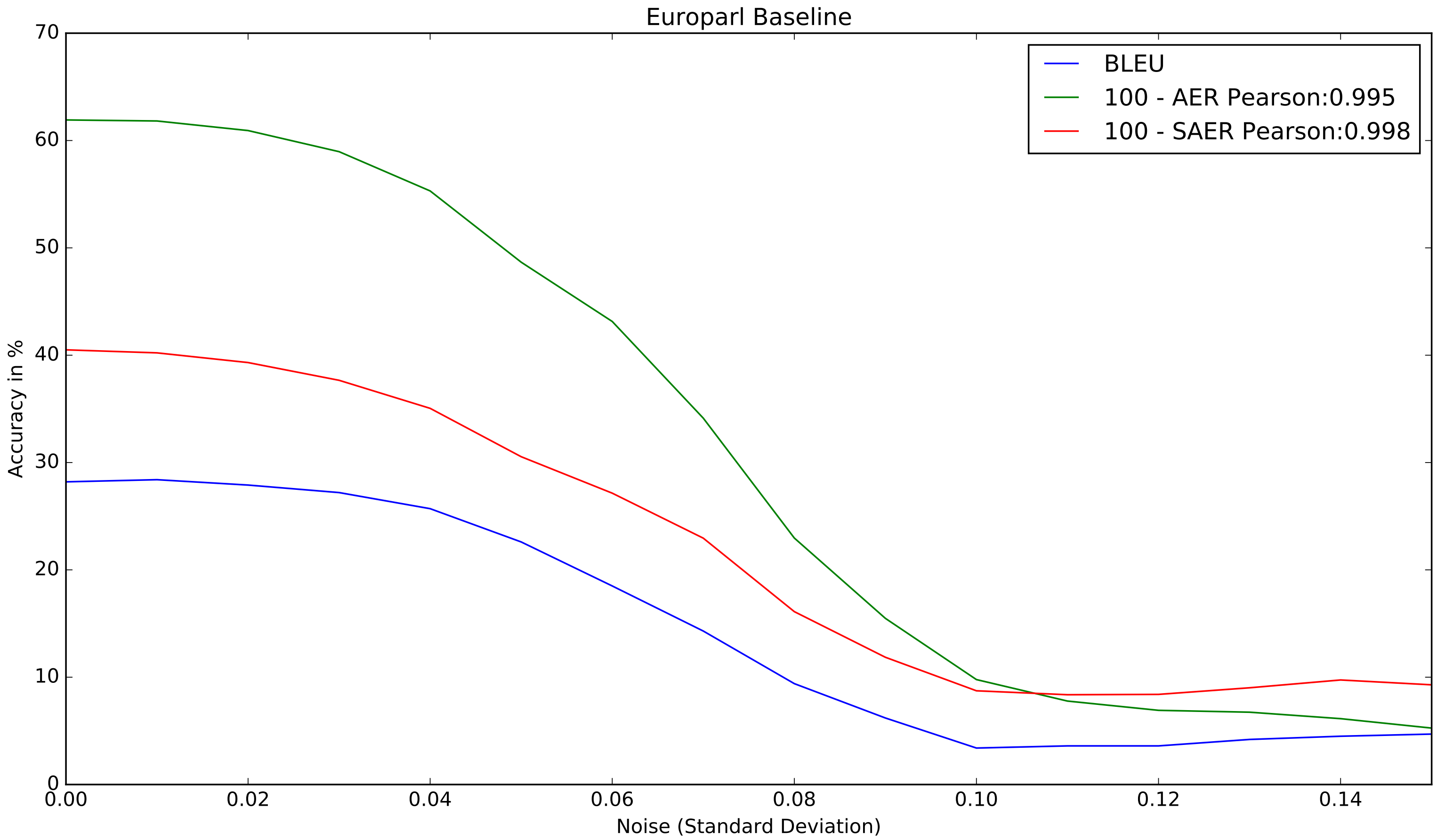
		German	English
`train (full data)`	Sentences	1.2M	
	Running Words	32M	34M
	Vocabulary	305K	100K
`align-test`	Sentences	504	
	Running Words	9.9K	10.3K
	Vocabulary	2.8K	2.4K
	OOVs with full vocabulary	6 (0.1%)	1 (0.0%)
	OOVs with 30K shortlist (Rate)	276 (2.8%)	50 (0.5%)

Analyzing Attention-based Alignments

- How good is the alignment quality of attention-based NMT?
- How can we evaluate attention-based alignments?
- How important are attention-based alignments for translation?

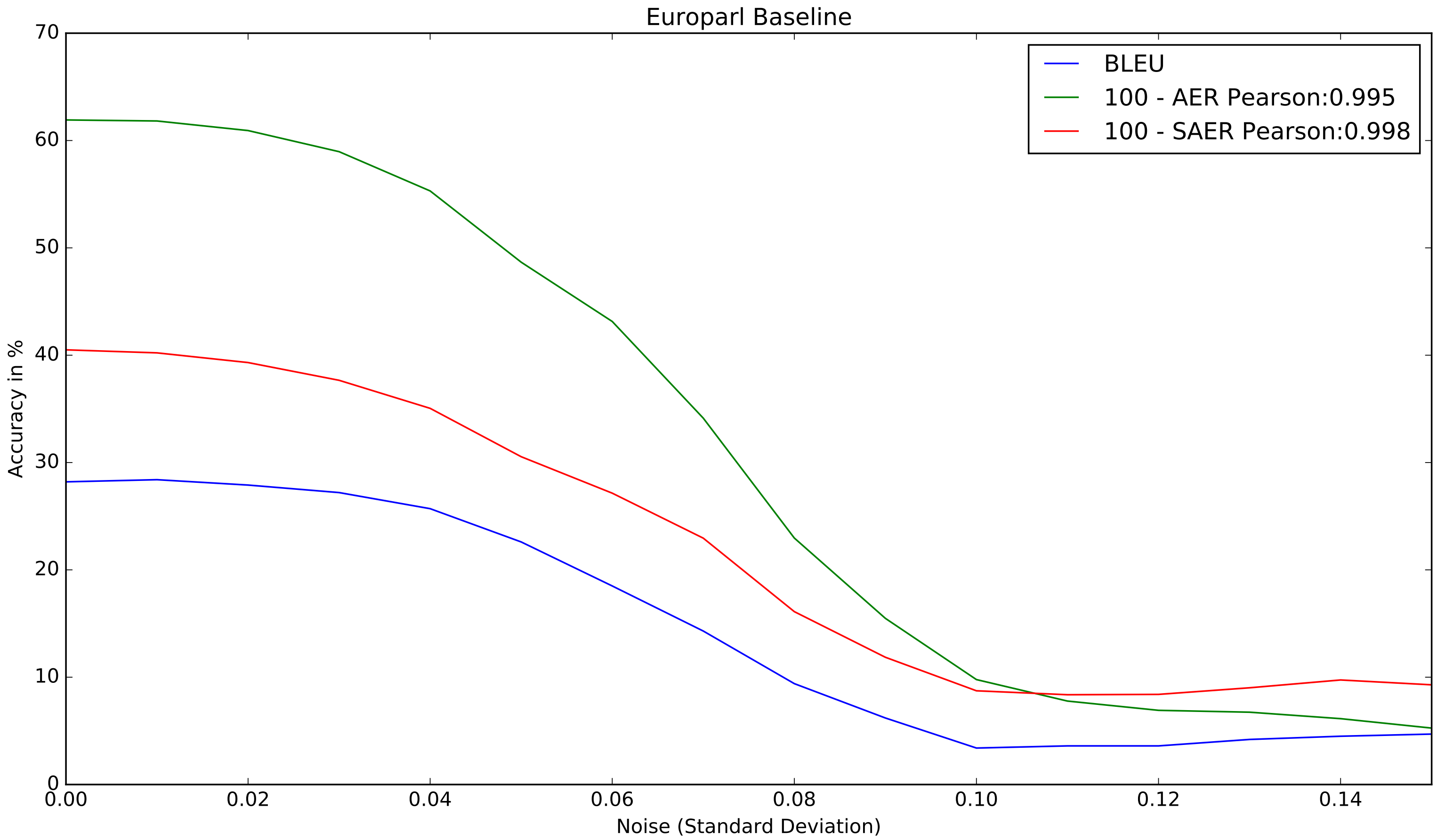
Analysing Attention-based Alignments (Europarl)

- **Compare BLEU to AER, SAER on model with increasing noise on alignment parameters**



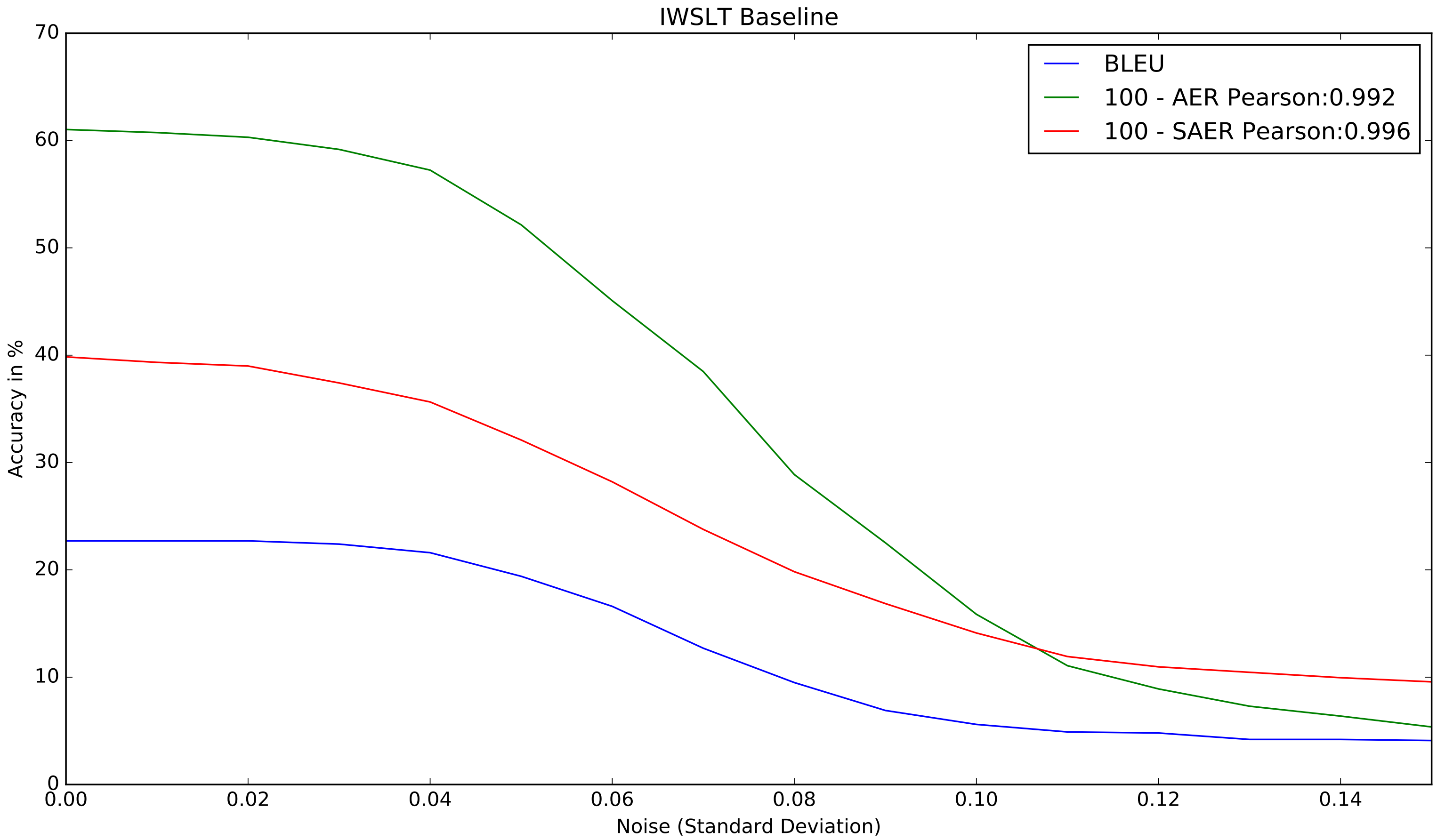
Analysing Attention-based Alignments (Europarl)

- Attention parameters are robust to noise up to a certain degree
- Alignment quality correlates with translation quality for all evaluation methods



Analysing Attention-based Alignments (IWSLT2013)

- Attention parameters are robust to noise up to a certain degree
- Alignment quality correlates with translation quality for all evaluation methods



- D. Bahdanau, K. Cho, Y. Bengio.
  Neural machine translation by jointly learning to align and translate.
  In *ICLR*, May 2015.
- W. Chen, E. Matusov, S. Khadivi, J.-T. Peter.
  Guided alignment training for topic-aware neural machine translation.
  Austion, Texas, October 2016. Association for Machine Translation in the Americas.
- F. J. Och, H. Ney.
  A systematic comparison of various statistical alignment models.
  Computational Linguistics, Vol. 29, No. 1, pp. 19–51, 2003.
- Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li.
  Modeling coverage for neural machine translation.
  In *54th Annual Meeting of the Association for Computational Linguistics*, August 2016.