The RWTH Aachen Machine Translation System for IWSLT 2016

**Jan-Thorsten Peter, Andreas Guta,
Nick Rossenbach, Miguel Graça,
and Hermann Ney**

<surname>@i6.informatik.rwth-aachen.de

December 8th, 2016, Seattle

**Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University**

Outline

Overview

Phrase-based System (PBT)

Joint Translation and Reordering System (JTR)

Neural Machine Translation System (NMT)

System Combination

Conclusion

Overview

- German → English
- TED and MSLT task
- Based on system combination using:
  - Phrase-based System
  - JTR Systems
  - NMT Systems
- Specialized systems:
  - TED task, optimized on TED.dev2010
  - MSLT task, optimized on TEDX.dev2012

Phrase-based System (PBT)

- Alignment with GIZA++ [Och and Ney, 2003]
- SCSS decoding using Jane [Wuebker et al., 2012]
- Optimization on BLEU with MERT [Och, 2003]
- Language Models:
  - 5-gram in-domain
  - 5-gram out-domain, with data selection [Moore and Lewis, 2010]
  - 7-gram word-class [Wuebker et al., 2013]
- Hierarchical Reordering Model [Galley and Manning, 2008]
- Reranking on 1000-best lists
  - Recurrent Neural Network Language Model with LSTM [Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2014]
  - Attention-based Neural Network Model [Bahdanau & Cho^{+} 15]

Joint Translation and Reordering System (JTR)

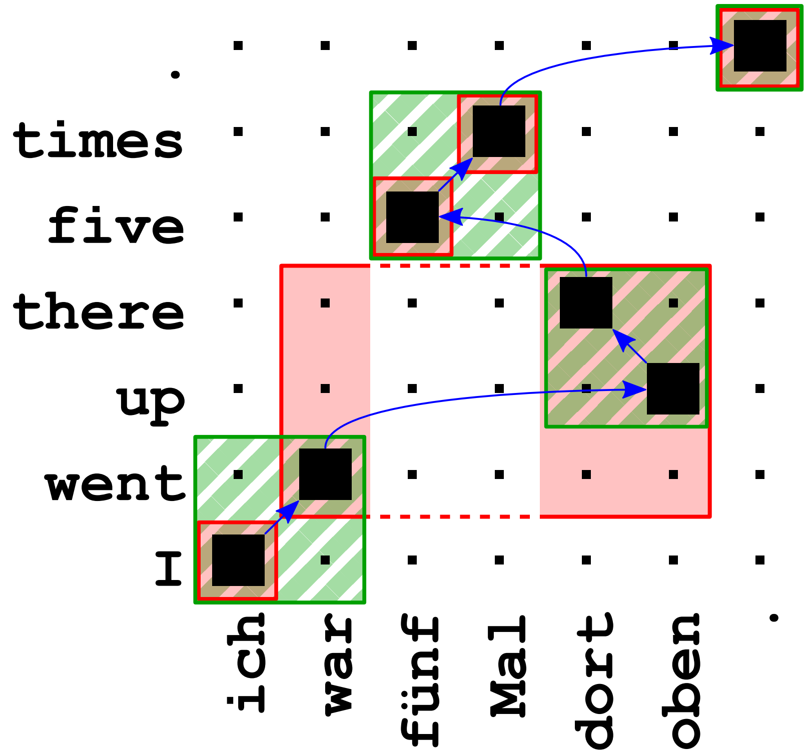
- **JTR models introduced by [Guta & Alkhouli^{+} 15]**
- **JTR sequence $(\tilde{f}, \tilde{e})_1^{\tilde{I}}$ is obtained from**
  - **Bilingual sentence pair f_1^J, e_1^I**
  - **GIZA++ word alignments b_1^I**

$$p(f_1^J, e_1^I, b_1^I) = \prod_{i=1}^{\tilde{I}} p((\tilde{f}, \tilde{e})_i | \underbrace{(\tilde{f}, \tilde{e})_{i-n+1}^{i-1}}_{h_i})$$

- ***Joint* model $p((\tilde{f}, \tilde{e})_i | h_i)$**
  - **Modified Kneser-Ney smoothing [Chen & Goodman 98]**
  - **KenLM toolkit [Heafield & Pouzyrevsky^{+} 13]**
- **Extension by *conditional* models $p(\tilde{f}_i | \tilde{e}_i, h_i)$ and $p(\tilde{e}_i | \tilde{f}_i, h_i)$**

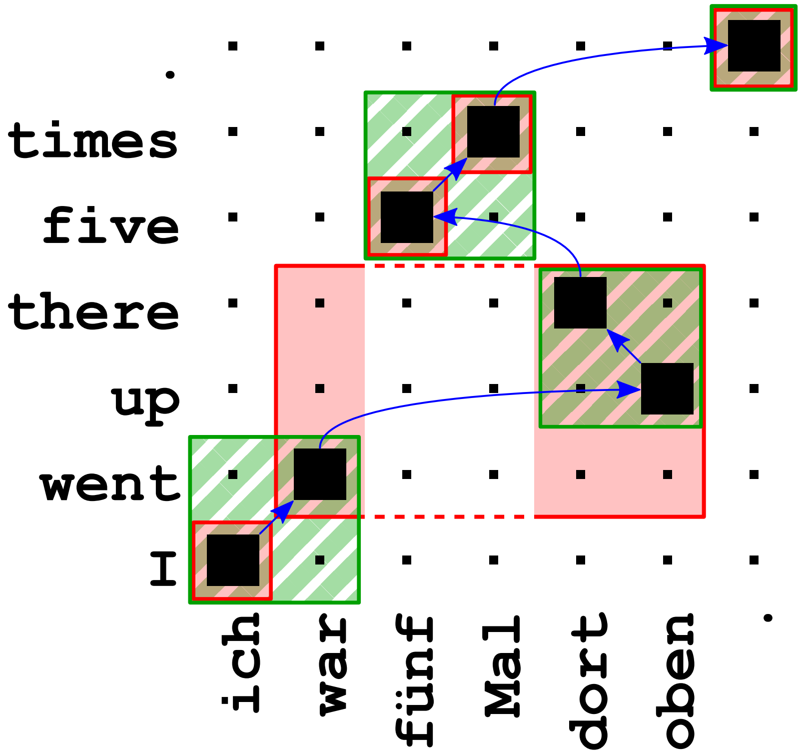
Word-level Models in Phrasal Translation

- Flexibility of word-level models
  - JTR joint and conditional models
  - 3 language models
  - 2 lexical models for smoothing
  - RNN Model with LSTM [Sundermeyer & Alkhouli^{+} 14]
- Heuristic features: Reordering, gap, phrasal frequency, word penalty, ...
- Log-linear model combination optimized on BLEU using MERT

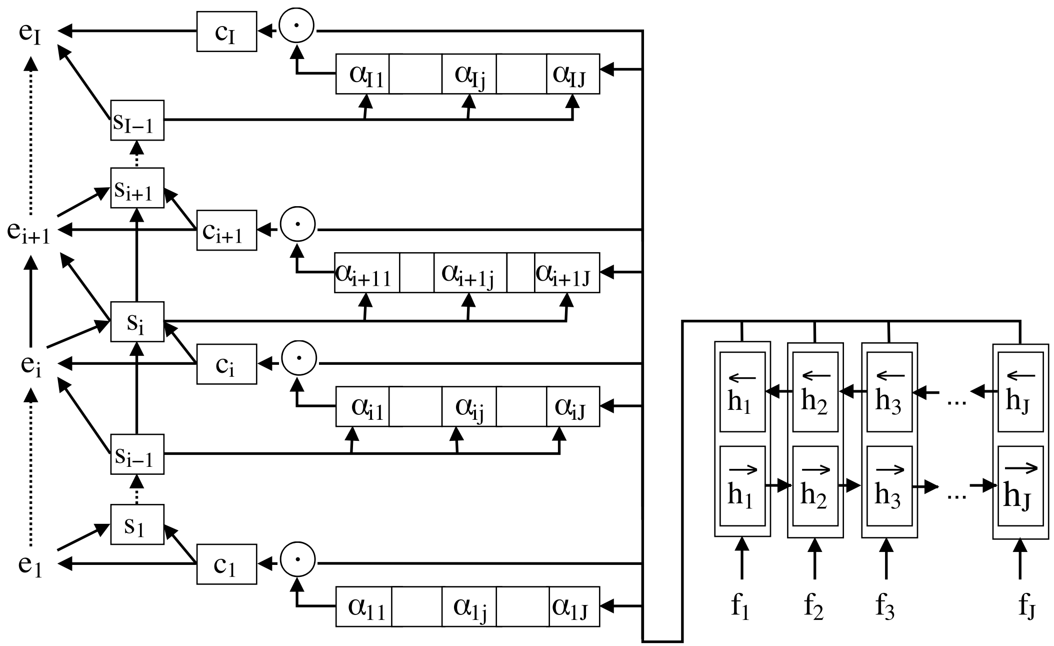


Word-level Models in Phrasal Translation

- SCSS decoding
  - GIZA++ word alignment annotations
  - Discontinuous source side
- Search states:
  - JTR and language model histories
  - Last aligned source position
  - Source coverage
- Recombination of equal states
- Issue: Phrasal segmentations can result in equal word alignments
  - Hash (full) JTR history
  - Delete states that are JTR hash duplicates



Neural Machine Translation System (NMT)



- **LSTM bidirectional encoder, unidirectional decoder**
- **Attention layer**
- **Forked from blocks-examples by Montreal**

Configuration

- Use byte-pair encoding generated on joint data
  - Using 20000 merge operations
  - Resulting vocabulary of $\sim$22000 on source and target side
- Word embedding with 620 dimension
- LSTM encoder and decoder with 1000 hidden units
- Maxout layer with 500 nodes before Softmax
  - Dropout is applied after this layer (if used)

Training

- Optimized with AdaDelta on 500k or 700k iterations
- Mini-batches of size 50
- Evaluation on dev set each 10k iterations

Optional Settings:

- Finetuning on various indomain-corpora for 1000 iterations
  - Evaluated each 100
- Dropout of 20% on maxout layer
- Using alignment-feedback / linguistic coverage [Cohn & Hoang$^+$ 16]
- Using guided-alignment [Chen & Matusov$^+$ 16]

Used Setups

Using two different setups:

- **IWSLT 2013**
  - **Used since the models where already trained**
  - **Same preprocessing as RWTH 2013 system**
  - **All data from IWSLT 2013**

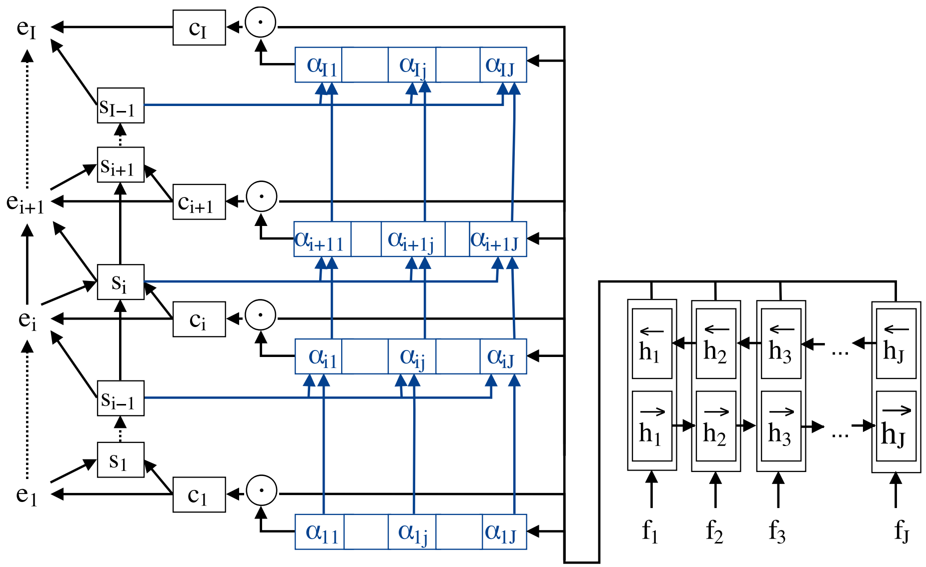
- **IWSLT 2016**
  - **Same preprocessing as RWTH 2015 system**
  - **All data from IWSLT 2016**
  - **Main difference: More date (OpenSubtitles2016, QED)**

Attention Features

- **Add sum over previous alignment (β) to energy computation**
- **Apply additional transformation W_β**

$$\alpha_{i,j} = v_\alpha^\top \tanh\left(W_\alpha s_{i-1} + U_\alpha h_j + W_\beta \beta_{i,j}\right)$$

$$\beta_{i,j} = \frac{1}{\Phi_j} \cdot \sum_{k=1}^{i-1} \tilde{\alpha}_{k,j}$$



- **Similar approaches as [Cohn & Hoang$^+$ 16] and [Tu & Lu$^+$ 16]**

Alignment Feedback w. Fertility

- Fertility Φ for each source words $\rightarrow$ depending on encoder state

$$\Phi_j = 2 * \text{sigmoid}\left(v_\Phi^\top \cdot h_j\right)$$

- Context fertility $\rightarrow$ add dependency on first and last encoder state

$$\Phi_j = 2 * \text{sigmoid}\left(v_\Phi^\top \cdot [h_j h_0 h_J]\right)$$

System	MSLT 2016 BLEU	TER	cTER	length
Baseline	35.1	46.4	42.5	100.9
+ Fertility	35.6	43.6	38.9	98.7
+ Context-Fertility	35.8	43.4	38.7	99.6

Guided Alignment Training [Chen & Matusov$^+$ 16]

- Utilize GIZA++ alignment
- Introducing alignment A as additional objective function
- Cross-Entropy cost $\mathcal{L}_{\text{align}}$ between the attention weights α and alignment A

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

IWSLT 2013 Setup	MSLT 2016			
System	BLEU	TER	cTER	length
Baseline + DEV12 finetune	32.1	49.7	44.0	104.0
+ Guided Alignment + DEV12 finetune	32.7	47.9	44.9	100.2

Fine-tuning

Continued training on fully trained network only on indomain data

- Used QED or TED Corpora
- Additional 1000 iterations
- Evaluated each 100 iterations

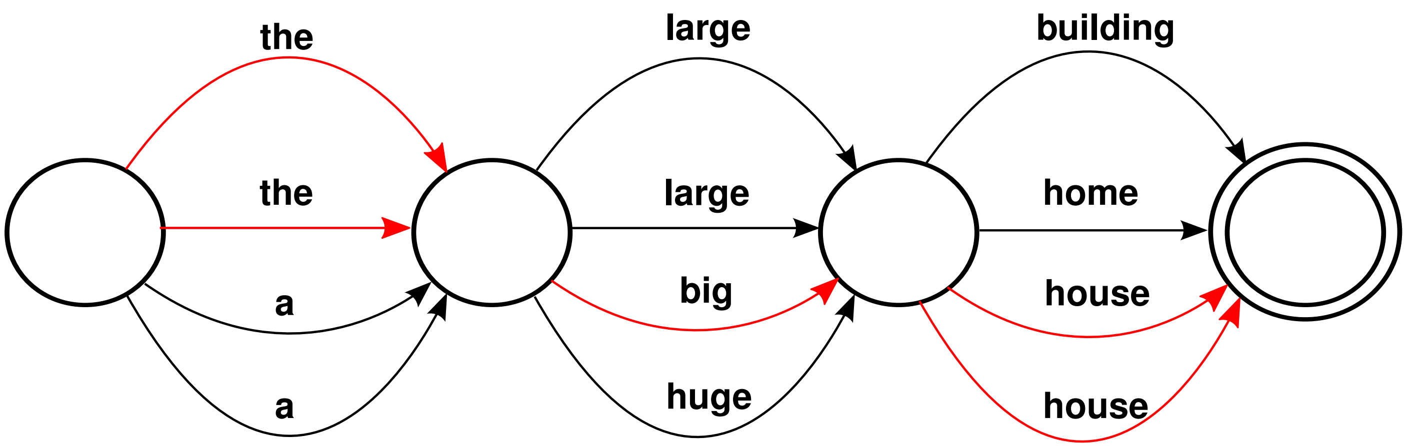
	MSLT 2016			
System	BLEU	TER	cTER	length
Baseline	35.1	46.4	42.5	100.9
+ QED-finetune	36.5	44.6	40.1	100.7
+ TED-finetune	36.9	43.3	37.7	99.7

Ensemble

- Combined best performing networks for ensemble
  - 5-best for IWSLT 2013, only 5 systems trained
  - 8-best for IWSLT 2016, selected out of 24

System	MSLT 2016			
	BLEU	TER	cTER	length
Baseline	35.1	46.4	42.5	100.9
+ QED-finetune	36.5	44.6	40.1	100.7
+ TED-finetune	36.9	43.3	37.7	99.7
Fertility	35.6	43.6	38.9	98.7
+ TED-finetune	36.6	43.5	38.1	100.0
+ QED-finetune	36.3	44.9	40.3	101.1
Context-Fertility	35.8	43.4	38.7	99.6
+ QED-finetune	37.0	43.7	39.6	100.0
Context-Fertility w/o Dropout	34.6	46.1	41.2	100.4
+ TED-finetune	35.8	44.5	39.3	99.7
NMT 2016 8best	40.8	39.3	34.8	99.1

System Combination



- Confusion network generation using n translation hypotheses
- Compute alignment using METEOR [Banerjee and Lavie, 2005]
- System combination features: word penalty, 3-gram LM, binary primary system, and binary voting feature
- Used best working combination out of 25 different combinations

Overview Results

#	System	Opt.	TED 2014 BLEU	TED 2014 TER	MSLT 2016 BLEU	MSLT 2016 TER
1	NMT 2013 5best	TED	32.3	48.4	36.9	43.9
2	NMT 2016 8best	TED	33.7	47.4	39.0	41.9
3	NMT 2013 5best	TEDX	32.3	47.9	37.9	42.4
4	NMT 2016 8best	TEDX	32.6	47.1	40.8	39.3
5	PBT	TEDX	29.4	51.6	38.6	39.9
6	+ JTR	TEDX	30.4	50.1	39.8	38.5
7	+ LSTM LM + NMT	TEDX	30.8	49.6	41.6	36.4
8	JTR	TEDX	30.6	49.7	38.9	38.7
9	PBT + JTR + NMT	TED	32.1	49.6	39.9	40.2
10	JTR + LSTM BTM	TED	30.8	50.3	37.6	40.7
11	NMT syscomb$_{1-4}$	-	33.4	47.1	40.3	40.8
12	MSLT syscomb$_{1-4,5,7,8}$	-	33.8	46.7	43.0	37.6
13	TED syscomb$_{1-4,9,10}$	-	34.2	46.5	42.9	37.6

Comparison to last Year's System

- Last year's system was optimized for TEDX
- Improvement of 1.7 BLEU on TEDX test set
- Improvement of 3.1 BLEU on TED test set

	TED test 2010			TEDX test 2014		
System	BLEU	TER	CTER	BLEU	TER	CTER
2015-Submission	31.9	47.6	45.5	26.2	54.7	54.6
TED-system	35.0	44.1	42.7	27.6	53.1	55.6
MSLT-system	34.7	44.1	42.9	27.9	53.2	54.3

Conclusion

- **Ensembles give the largest improvement for NMT**
- **System Combination did not work using only NMT models**
- **Strongest single system for:**
  - **TED task: Ensemble of NMT Systems**
  - **MSLT task: PBT + LSTM LM + NMT**
- **Strong improvement to last years system**
  - **1.7 BLEU on TEDX**
  - **3.1 BLEU on TED**

Thank you for your attention

Jan-Thorsten Peter, Andreas Guta,
Nick Rossenbach, Miguel Graça,
and Hermann Ney

`<surname>@cs.rwth-aachen.de`

- **D. Bahdanau, K. Cho, Y. Bengio.**
  Neural machine translation by jointly learning to align and translate.
  In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

- **S. F. Chen, J. Goodman.**
  An Empirical Study of Smoothing Techniques for Language Modeling.
  Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, 63 pages, Aug. 1998.

- **W. Chen, E. Matusov, S. Khadivi, J. Peter.**
  Guided alignment training for topic-aware neural machine translation.
  CoRR, Vol. abs/1607.01628, 2016.

- **K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio.**
  On the properties of neural machine translation: Encoder-decoder approaches.
  In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103—-111, Doha, Qatar, October 2014.

- K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio.
  Learning phrase representations using rnn encoder–decoder for statistical machine translation.
  In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
- T. Cohn, C. D. V. Hoang, E. Vymolova, K. Yao, C. Dyer, G. Haffari.
  Incorporating structural alignment biases into an attentional neural translation model.
  CoRR, Vol. abs/1601.01085, 2016.
- J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, J. Makhoul.
  Fast and robust neural network joint models for statistical machine translation.
  In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1370–1380,

Baltimore, Maryland, June 2014. Association for Computational Linguistics
.

- A. Guta, T. Alkhouli, J.-T. Peter, J. Wuebker, H. Ney.
A Comparison between Count and Neural Network Models Based on Joint Translation and Reordering Sequences.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

- K. Heafield, I. Pouzyrevsky, J. H. Clark, P. Koehn.
Scalable modified Kneser-Ney language model estimation.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 690–696, Sofia, Bulgaria, August 2013.

- M. Luong, H. Pham, C. D. Manning.
Effective approaches to attention-based neural machine translation.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412—-1421, Lisbon, Portugal, 2015.

- **M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney.**
  Translation modeling with bidirectional recurrent neural networks.
  In *Conference on Empirical Methods in Natural Language Processing*, pp. 14–25, Doha, Qatar, Oct. 2014.
- **I. Sutskever, O. Vinyals, Q. V. V. Le.**
  Sequence to sequence learning with neural networks.
  In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., Monteal, Canada, 2014.
- **Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li.**
  Coverage-based neural machine translation.
  CoRR, Vol. abs/1601.04811, 2016.