The RWTH Aachen University German→English and English→German Machine Translation System for IWSLT 2017

Parnia Bahar, Jan Rosendahl, Nick Rossenbach and Hermann Ney
Human Language Technology and Pattern Recognition, RWTH Aachen University

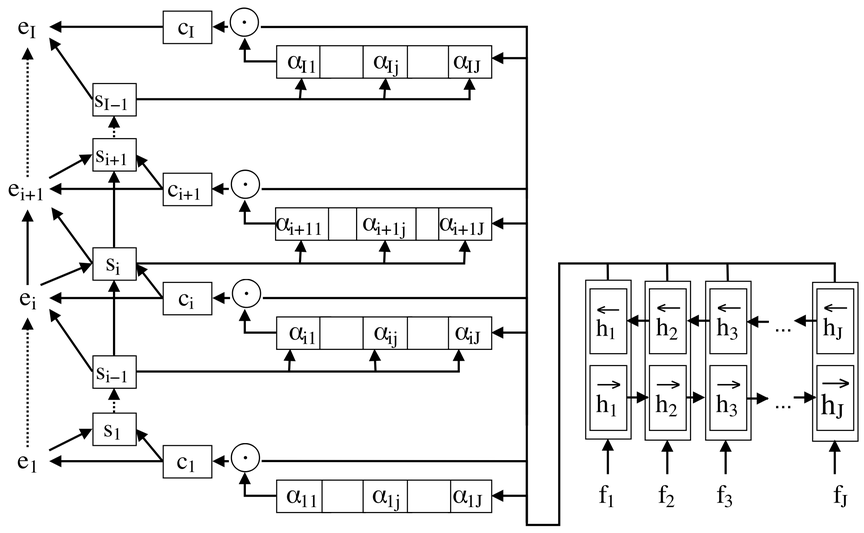
Pre-Processing

- Simplified preprocessing
  - Only tokenization, frequent casing, and simple categories (numbers are treated as characters)
- Byte pair encoding (BPE) with 90k merging operations trained jointly [Sennrich et al., 2016]
- Data filtering based on the ratio: $\frac{\text{source sentence length}}{\text{target sentence length}} > 0.7$

English→German	
# total sentences	20.9 M
# removed sentences	1.7 M
% removed sentences	8 %

Neural Machine Translation (NMT) System

- Baseline
  - Attention-based NMT system [Bahdanau et al., 2015]
  - Bidirectional encoder and unidirectional decoder composed of LSTM nodes
  - Two maxout-layers followed by a softmax operation as an output layer
- Multilayer encoder-decoder
  - Two stacked LSTM layers in both encoder and decoder
  - We connect all internal states of the first LSTM layer to the second



Overview over the attention-based encoder-decoder approach.

Extensions of the Attention Mechanism

- Drawbacks of baseline
  - No explicit alignment or coverage information
- Fertility feedback
  - Solution by [Tu et al., 2016]: Feed back the sum of past alignments to the computation of the attention energies
  - Fertility parameter $2\sigma(\upsilon_\phi^\top \cdot \boldsymbol{h_j})$ determines how many target words should be generated by a single source word

$$\beta_{i,j} = \frac{1}{2\sigma(\upsilon_\phi^\top \cdot \boldsymbol{h_j})} \sum_{k=1}^{i-1} \alpha_{k,j}$$

$$e_{i,j} = \upsilon^\top \tanh\left(W\boldsymbol{s_{i-1}} + U\boldsymbol{h_j} + V\beta_{i,j}\right)$$

- Convolutional feedback
  - Convolutional operation G_n over the last attention weights proposed by [Feng et al., 2016]

$$\gamma_{i,j} = \sum_{l=j-k}^{j+k} G_{n,j-l} \cdot \alpha_{i-1,l} \qquad \text{for all } n = 1, \cdots, N$$

$$e_{i,j} = \upsilon^\top \tanh\left(W\boldsymbol{s_{i-1}} + U\boldsymbol{h_j} + V\gamma_{i,j}\right)$$

- We use $N = 5$ filters with a window of size $k = 5$

Experimental Setup

- Structure
  - 620-dimensional word embedding, LSTM layers with 1k nodes
  - Dropout of 20%
  - Adam optimization algorithm
  - Domain adaptation by weighting data: TED corpus (11x) and QED corpus (6x)
  - Continued training using an Adam Annealing scheme with a factor of 0.75 or averaging of best snapshots (avg4)
  - Ensemble of up to 4 networks
- Implementation based on Blocks [Merriënboer er al., 2015] and Theano [Bergstra et al., 2010, Bastien et al., 2012]

German→English Results

#	System	TED.tst2014 BLEU	TED.tst2014 TER	TED.tst2014 cTER	TED.tst2015 BLEU	TED.tst2015 TER	TED.tst2015 cTER
1	multilayer enc-dec baseline	31.3	49.0	51.6	31.1	48.2	50.6
2	+ annealing Adam	33.8	46.2	49.3	34.0	46.0	48.3
3	+ fertility feedback	33.9	46.2	49.0	34.6	45.4	48.4
4	+ fine tuned	33.9	46.6	48.3	34.5	45.7	48.6
5	+ convolutional feedback (avg4)	33.1	46.4	49.2	33.1	45.8	49.0
6	ensemble 2, 3, 5	35.5	44.9	47.8	35.5	44.5	47.6

English→German Results

#	System	TED.tst2014 BLEU	TED.tst2014 TER	TED.tst2014 cTER	TED.tst2015 BLEU	TED.tst2015 TER	TED.tst2015 cTER
1	baseline	24.6	57.1	53.7	27.2	55.2	51.1
2	+ dropout	25.3	56.4	53.9	27.4	55.1	51.3
3	+ annealing Adam	27.9	53.9	50.7	30.2	53.0	48.4
4	+ fertility feedback	25.3	56.7	53.8	27.0	55.5	50.9
5	+ avg4	27.2	54.3	50.5	29.9	52.4	47.9
6	+ multilayer enc-dec	25.2	56.2	53.5	27.3	54.5	50.7
7	+ annealing Adam	27.6	54.0	50.0	29.9	52.6	48.3
8	+ fine tuned	27.5	54.3	49.8	29.9	52.7	47.4
9	+ dropout	27.6	54.1	50.3	30.5	52.3	46.8
10	ensemble 3, 5, 8, 9	29.2	52.8	48.8	31.5	51.1	45.9

Observations

- Additional fine tuning does not improve system
- Only small gains from larger BPE (20k→90k)
- Fertility feedback is more helpful than convolutional feedback
- Adam benefits from annealing scheme or snapshot averaging

Evaluation Results

- Results are obtained using ensemble of the best networks

MT Task	TED.tst2016 BLEU	TED.tst2016 TER	TED.tst2016 NIST	TED.tst2017 BLEU	TED.tst2017 TER	TED.tst2017 NIST
De→En	35.38	44.48	7.8947	30.22	49.44	7.1608
En→De	28.09	55.23	6.5995	25.12	59.09	6.1239

Acknowledgement

The work reported in this paper has been funded by three projects, SEQCLAS, QT21 and DFG-Core-Tec. SEQCLAS has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement n° 694537. QT21 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 645452. It was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under Contract No NE572/8-1. The work reflects only the authors' views and neither the European Commission nor the European Research Council Executive Agency nor the DFG are responsible for any use that may be made of the information it contains.