Machine Translation

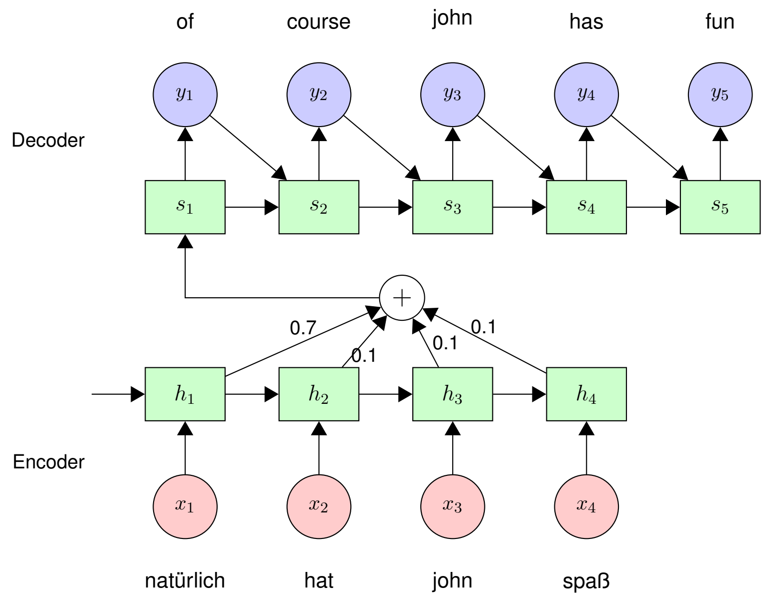
06: Attention Models
Analysis and Variants

Rico Sennrich

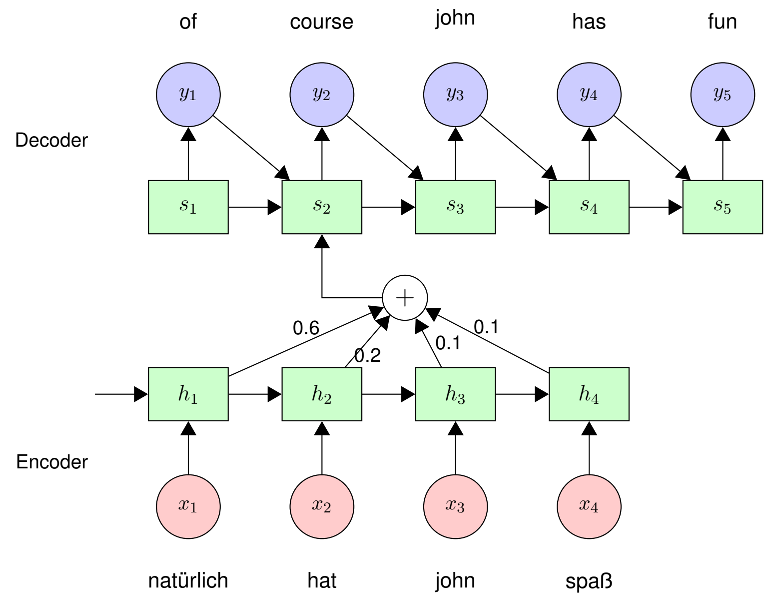
University of Edinburgh

MT – 2018 – 06

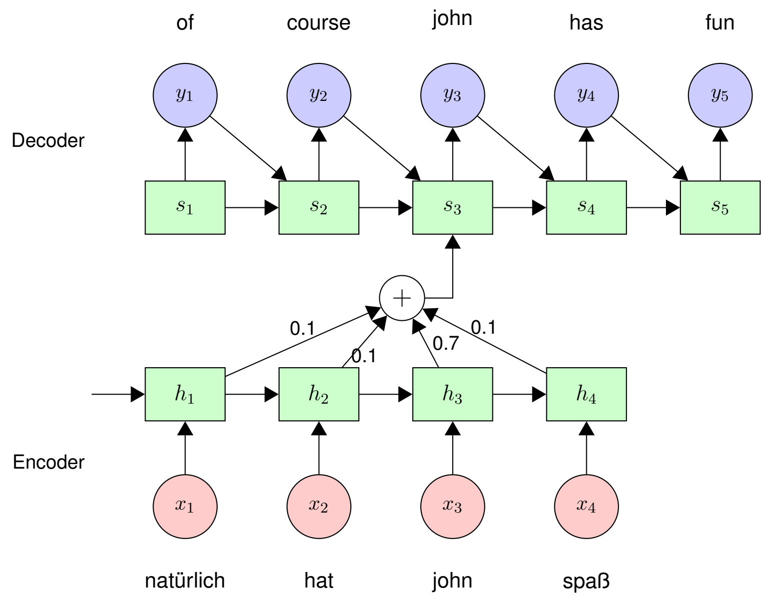
Encoder-Decoder with Attention



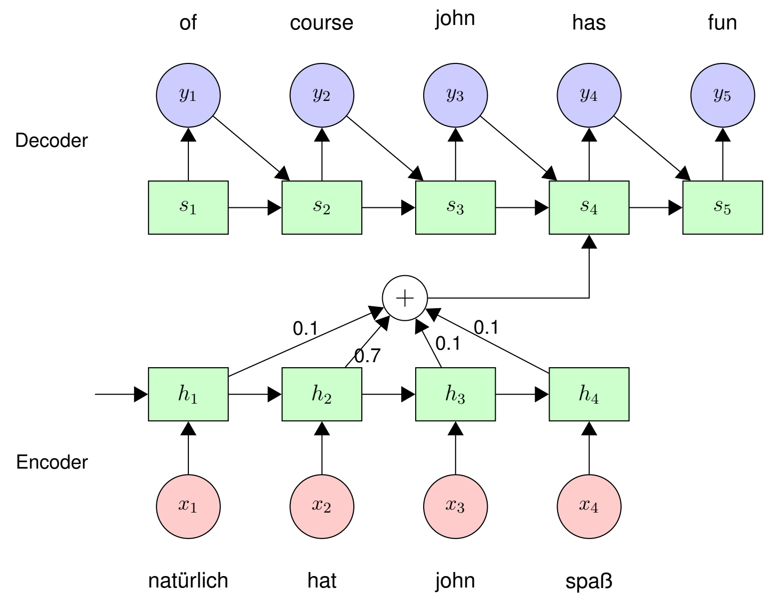
Encoder-Decoder with Attention



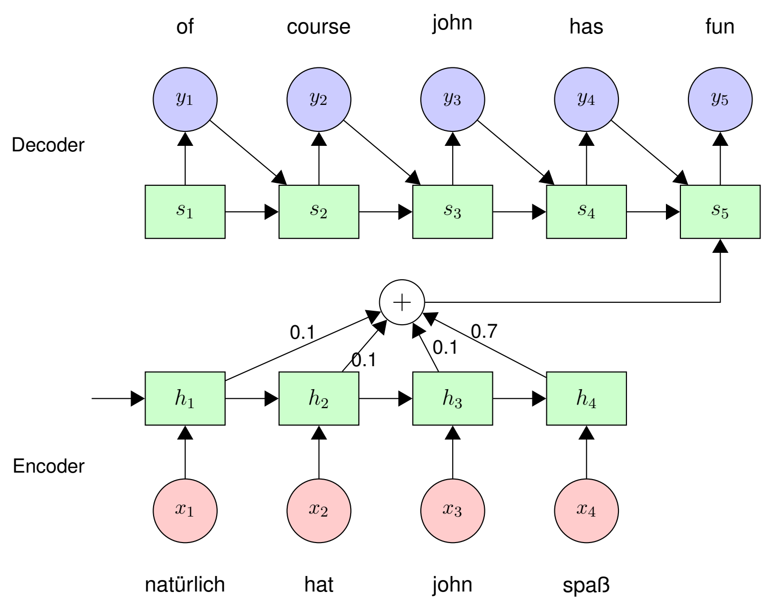
Encoder-Decoder with Attention



Encoder-Decoder with Attention



Encoder-Decoder with Attention



Attentional encoder-decoder: Maths

(one type of) attention model

$$e_{ij} = v_a^\top \tanh(W_a s_{i-1} + U_a h_j)$$

$$\alpha_{ij} = \mathrm{softmax}(e_{ij})$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Attention model

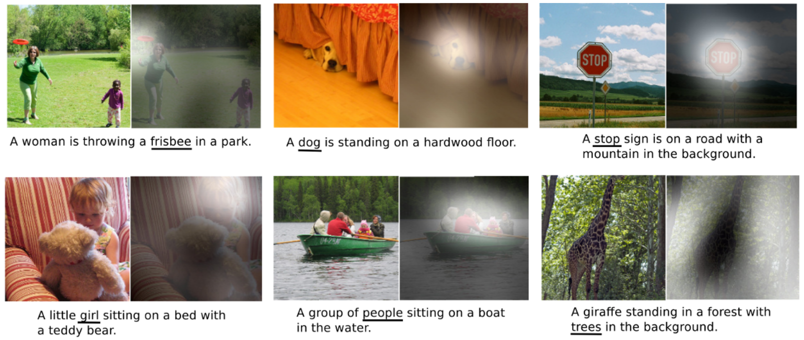
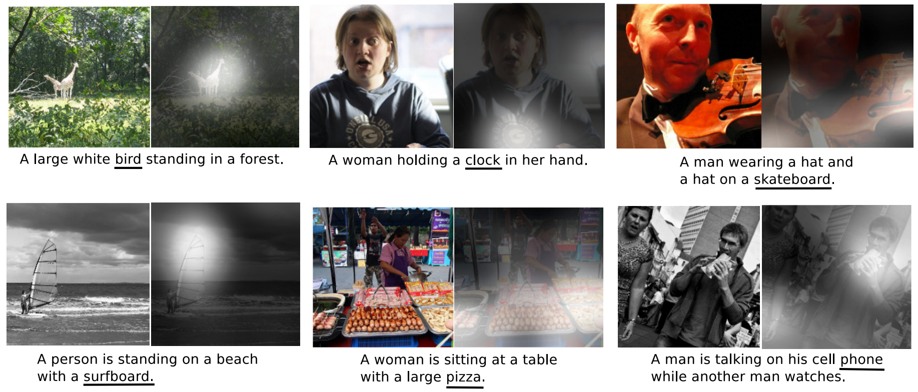


Fig. 5. Examples of the attention-based model attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word) [22]

[Xu et al., 2015]

Attention model

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



[Xu et al., 2015]

Alignment

- word-alignment between source and target words is used for various applications
- translate rare/unknown words with back-off dictionary:

source	The **indoor temperature** is very pleasant.
reference	Das **Raumklima** ist sehr angenehm.
[Bahdanau et al., 2015]	Die **UNK** ist sehr angenehm
[Jean et al., 2015]	Die **Temperatur** ist sehr angenehm.

(more on open-vocabulary MT in future lecture)
- attention has been used to obtain alignments. **However, ...**

MT – 2018 – 06

Attention is not alignment

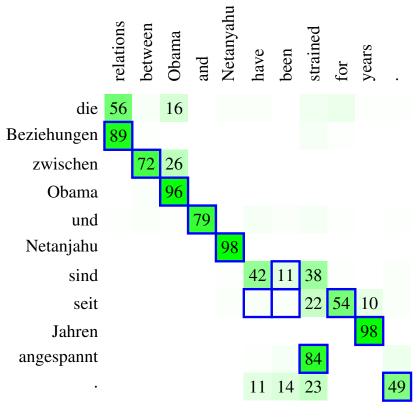


Figure 8: Word alignment for English–German: comparing the attention model states (green boxes with probability in percent if over 10) with alignments obtained from fast-align (blue outlines).

[Koehn and Knowles, 2017]

Attention is not alignment

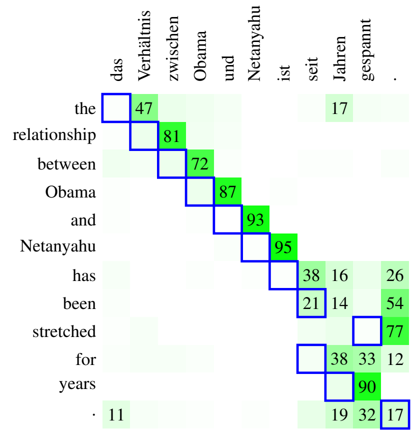


Figure 9: Mismatch between attention states and desired word alignments (German–English).

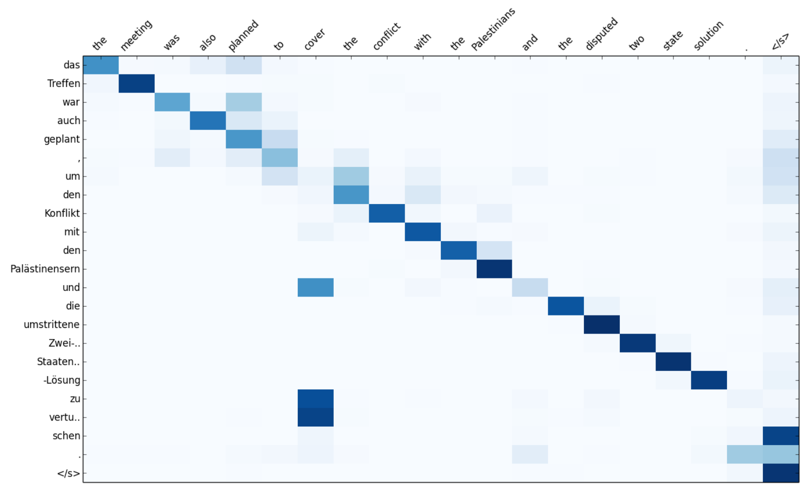
[Koehn and Knowles, 2017]

Attention is not alignment

discuss in pairs

how can NMT model translate text, even if attention is off?

Attention is not alignment



MT – 2018 – 06

Obtaining Attention Scores

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \textit{concat} \end{cases}$$

attention variants from [Luong et al., 2015]

- many ways to score encoder states:
- *concat*: attention as introduced by [Bahdanau et al., 2015]
- *dot*: more attention on similar vectors

Conditioning Attention on Past Decisions

attention in dl4mt-tutorial (and Nematus):

$$\begin{aligned} s'_i &= GRU_1(s_{i-1}, y_{i-1}) \\ c_i &= ATT(C, s'_i) \\ s_i &= GRU_2(c_i, s'_i) \end{aligned}$$

motivation

- (simple) attention model from lecture 4 is only conditioned on s_{i-1}...
  ...but it also matters which word we predicted last (y_{i-1})
- more transitions per timestep $\rightarrow$ more depth
  [Miceli Barone et al., 2017])

Guided Alignment Training [Chen et al., 2016]

core idea

1. compute alignment with external tool
   (IBM models; discussed in later lecture)
2. if multiple source words align to same target words,
   normalize so that $\sum_j A_{ij} = 1$
3. modify objective function of NMT training:
   - minimize target sentence cross-entropy (as before)
   - minimize divergence between model attention α and external alignment A:

$$H(A, \alpha) = -\frac{1}{T_y} \sum_{i=1}^{T_y} \sum_{j=1}^{T_x} A_{ij} \log \alpha_{ij}$$

Incorporating Structural Alignment Biases

core idea [Cohn et al., 2016]

we know that alignment has some biases, which are exploited in statistical word alignment algorithms [Brown et al., 1990, Koehn et al., 2003]:

- position bias: relative position is highly informative for alignment
- fertility/coverage: some words produce multiple words in target language
  all source words should be covered (respecting fertility)
- bilingual symmetry: $\alpha^{s \leftarrow t}$ and $\alpha^{s \rightarrow t}$ are symmetrical

Incorporating Structural Alignment Biases

position bias

- provide attention model with positional information
- found to be especially helpful with non-recurrent architectures
- different choices for positional encoding:
  - [Cohn et al., 2016]: $\log(1 + i)$
  - [Gehring et al., 2017]: positional embedding: $E(i)$
  - [Vaswani et al., 2017]: sine/cosine function

Incorporating Structural Alignment Biases

coverage without fertility

reminder:

$$\sum_{j}^{T_x} \alpha_{ij} = 1 \qquad \text{(softmax)}$$

idea: model should attend to each source word exactly once:

$$\sum_{i}^{T_y} \alpha_{ij} \approx 1 \qquad \text{(our goal)}$$

we can bias model towards this goal with regularisation term:

$$\sum_{j}^{T_x} (1 - \sum_{i}^{T_y} \alpha_{ij})^2 \qquad \text{(to be minimized)}$$

discuss in pairs

is this the right goal? why / why not?

Incorporating Structural Alignment Biases

coverage with fertility [Cohn et al., 2016, Tu et al., 2016]

idea: learn fertility of words with neural network:

$$f_j = N\sigma(W_j h_j)$$

coverage objective that takes fertility into account:

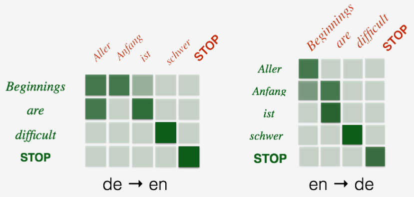
$$\sum_{j}^{T_x} (f_j - \sum_{i}^{T_y} \alpha_{ij})^2 \qquad \text{(to be minimized)}$$

Incorporating Structural Alignment Biases

bilingual symmetry

joint training objective with *trace bonus* B, which rewards symmetric attention:

$$B(\alpha^{s\leftarrow t}, \alpha^{s\rightarrow t}) = \sum_{i=1}^{T_y} \sum_{j=1}^{T_x} \alpha_{ij}^{s\rightarrow t} \alpha_{ji}^{s\leftarrow t}$$



Further Reading

- Philipp Koehn and Rebecca Knowles (2017). Six Challenges for Neural Machine Translation.

Coursework

Coursework

- available at the end of this week
- deadline: March 15, 3pm
- you are encouraged to work in pairs. More details to follow
- training models takes hours or days, so **start early**
- I will have no sympathy if you don't realize you can't do this coursework last minute

Lab Sessions

- two lab sessions will provide support getting started (installation of tools and virtual environment)
  - Tuesday, February 6, 15.10-16.00
    Room 4.12, Appleton Tower
  - Wednesday, February 7, 15.10-16.00
    Room 5.08, North Lab, Appleton Tower
- attendance **not** mandatory

Bibliography I

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural Machine Translation by Jointly Learning to Align and Translate.
In Proceedings of the International Conference on Learning Representations (ICLR).

Brown, P., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990).
A Statistical Approach to Machine Translation.
Computational Linguistics, 16(2):79–85.

Chen, W., Matusov, E., Khadivi, S., and Peter, J. (2016).
Guided Alignment Training for Topic-Aware Neural Machine Translation.
CoRR, abs/1607.01628.

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., and Haffari, G. (2016).
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model.
In
Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua
pages 876–885, San Diego, California.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017).
Convolutional Sequence to Sequence Learning.
CoRR, abs/1705.03122.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015).
On Using Very Large Target Vocabulary for Neural Machine Translation.
In
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference o
pages 1–10, Beijing, China. Association for Computational Linguistics.

Bibliography II

Koehn, P. and Knowles, R. (2017).
Six Challenges for Neural Machine Translation.
In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003).
Statistical Phrase-based Translation.
In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Lang
pages 48–54, Edmonton, Canada.

Luong, T., Pham, H., and Manning, C. D. (2015).
Effective Approaches to Attention-based Neural Machine Translation.
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Miceli Barone, A. V., Helcl, J., Sennrich, R., Haddow, B., and Birch, A. (2017).
Deep Architectures for Neural Machine Translation.
In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers, Copenhagen, Denmark. Association for Computational Linguistics.

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016).
Modeling Coverage for Neural Machine Translation.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 76–85. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention Is All You Need.
CoRR, abs/1706.03762.

Bibliography III

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015).
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.
In Bach, F. and Blei, D., editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2048–2057, Lille, France. PMLR.