Smaller, faster, deeper: University of Edinburgh MT submittion to WMT 2017

Rico Sennrich, Alexandra Birch, Anna Currey,
Ulrich Germann, Barry Haddow, Kenneth Heafield,
Antonio Valerio Miceli Barone, Philip Williams

University of Edinburgh

July 19 2017

Main collaborators

Rico Sennrich

Barry Haddow

Linear models in MT

Phrase-based machine translation

- log-linear model: $p(t) = exp \sum_{i=1}^{n} \lambda_i h_i(x)$

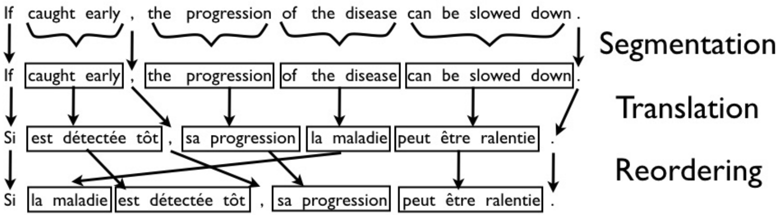
Weighted Model

- number of feature funtions n
- random variables $x = (e, f, start, end)$
- feature functions

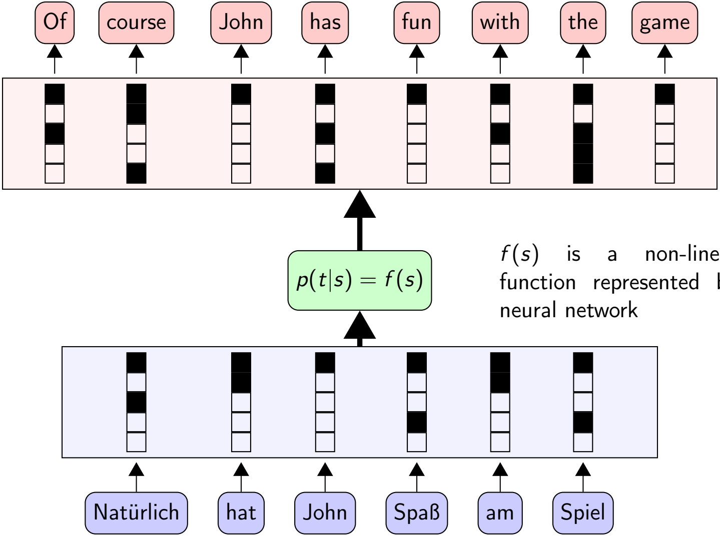
$$h_1 = p(e|f) \text{ translation probability}$$
$$h_2 = d(start_e - start_f) \text{ distortion}$$
$$h_3 = d(p_{LM}) \text{ language model}$$

- weights λ

SMT framework



Neural MT



$f(s)$ is a non-linear function represented by neural network

Neural versus Phrase-base MT

Phrase-based SMT

Learn segment-segment correspondances from bitext

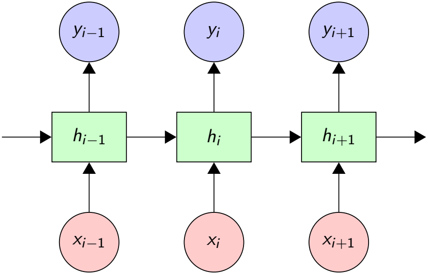
- Training is multistage pipeline of heuristics
- Fixed weights for features
- Limited ability to encode history
- Strong independence assumptions

Neural MT

Learn mathematical function on vectors from bitext

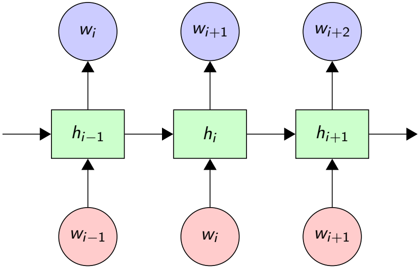
- End-to-end trained model
- Output conditioned on full source text and target history
- Non-linear dependence on information sources

Recurrent neural network



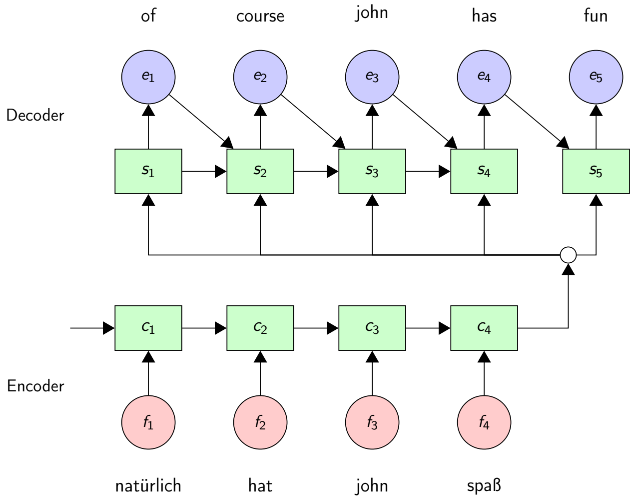
$$h_i = f(x_i, h_{i-1})$$
$$y_i = g(h_i)$$

RNN for Languagel Modelling

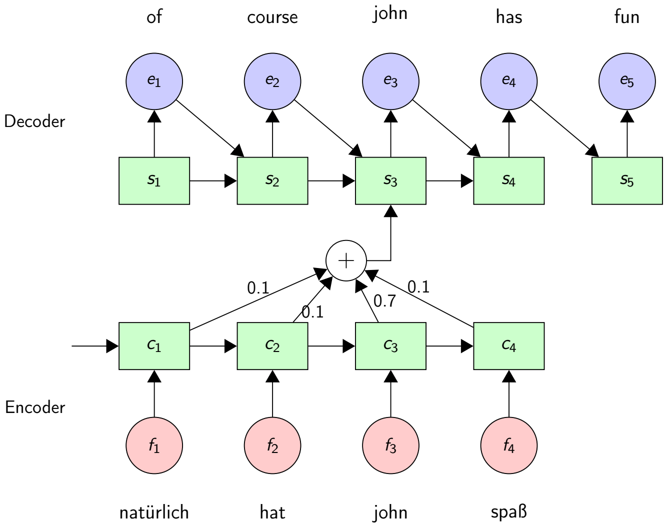


- Predict w_i conditioned on $w_1 \ldots w_{i-1}$
- Allows unlimited history
- Outperforms traditional count-based n-gram models

Encoder-Decoder



Encoder-Decoder with Attention



Limitations of Neural MT

Limitations

- Limited memory on GPUs
- Slow training times
- Not really deep deep learning models
- Training is not very stable

Improvements to Subword Segmentation

Byte-pair encoding [Sennrich et al., 2016]

- iterative, frequency-based merging of subword units into larger units
- "joint BPE" on parallel corpus for more consistent segmentation

Problems

- subword unit can be part of (frequent) larger unit, but rare on its own

Allergikerzimmer	330
Allergiker:	10

- subword unit can be frequent in one language, but rare in the other

nationalities	541 (EN)	1 (DE)

Consequences

- model is unlikely to learn good representation for rare subwords
- BPE may even produce subword that is unknown at test time
- having rare subwords in vocabulary is wasteful

Improvements to Subword Segmentation

Solution

- require that subword has been observed in source training corpus
- optionally require minimum frequency
- even in joint BPE, condition is checked for each side individually
- split up (reverse merge of) subwords that don't meet this requirement

Vocabulary size (EN→DE with joint BPE; 90k merge operations)

BPE	EN	DE
no filter	83227	91921
threshold 50	52652	73297

- only minimal change in sequence length and BLEU
- advantage: smaller models; no UNK
- disadvantage: additional hyperparameter: frequency threshold

Parameter Tying

Embedding and Output Layer

- in Nematus, last hidden layer has same size as target-size embedding
  $\rightarrow$ output matrix: vocabulary size $\times$ embedding layer size
  $\rightarrow$ target embedding matrix: embedding layer size $\times$ vocabulary size
- [Press and Wolf, 2017] propose tying weights of embedding matrix and transpose of output matrix.
- little effect on quality, but smaller models.

Optimizers

Optimizer

- adaptive learning rates tend to speed up training
- this year we used adam [Kingma and Ba, 2015] instead of adadelta [Zeiler, 2012]

Learning Rate Annealing

- our WMT systems use Adam without annealing
- SGD with annealing is popular [Sutskever et al., 2014, Wu et al., 2016]
- Adam with annealing recommended by [Denkowski and Neubig, 2017]
  → "--anneal_restarts 2 --patience 3" in Nematus

Layer Normalization

- if input distribution to NN layer changes, parameters need to adapt to this **covariate shift**.
- normalization of layers reduces shift, and improves training stability.
- for layer $\mathbf{a}$ with H units, re-center and re-scale layer.
- normalization changes representation power:
  two bias parameters, $\mathbf{g}$ and $\mathbf{b}$, restore original representation power

$$\mu = \frac{1}{H}\sum_{i=1}^{H} a_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i - \mu)^2} \tag{2}$$

$$\mathbf{h} = \left[\frac{\mathbf{g}}{\sigma} \odot (\mathbf{a} - \mu) + \mathbf{b}\right] \tag{3}$$

Stacked RNNs

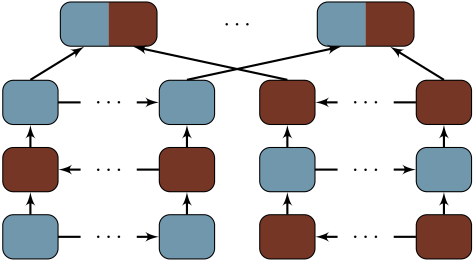


Figure: Alternating stacked encoder [Zhou et al., 2016].

Deep Transition Networks

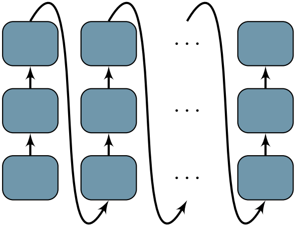


Figure: Deep transition network.

Deep Architectures: Discussion

- we use depth of 4 for submission systems
- all models were trained on single GPU
- in post-submission experiments [Miceli Barone et al., 2017], BiDeep architecture (combination of deep transition and stack) performed best

Monolingual Data

Back-translations

- all systems in the news shared task use target-side news data, automatically translated into source language

Copying

- the EN$\leftrightarrow$TR systems use monolingual data that is paired with a copy on the source

Biomedical task

- pseudo in-domain monolingual data is extracted from commoncrawl as follows:
  - automatically translate in-domain source corpus
  - perform Moore-Lewis data selection in target-language commoncrawl corpus

Ensembling and Reranking: Research Questions

Ensembling

- last year, we used checkpoint ensemble (of last 4 checkpoints).
- this year, we contrast this with ensemble of independent models.

Reranking with right-to-left models

- last year, reranking with right-to-left models gave significant improvements for three translation directions.
- this year, we evaluate strategy on stronger baseline and more systems.

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
WMT-16 single system	25.9	31.1	—	29.6	—	—
baseline	27.5	32.0	16.4	31.3	19.7	21.7
+layer normalization	28.2	32.1	17.0	32.3	18.8	22.5
+deep model	28.9	33.5	16.6	32.7	20.6	22.9
+checkpoint ensemble	29.4	33.8	17.7	33.3	21.0	23.6
+independent ensemble	30.3	34.4	18.5	33.6	21.6	25.1
+right-to-left reranking	31.1	35.1	19.0	34.6	22.3	25.7
WMT-17 submission	30.9	35.1	19.0	30.8	20.1	25.7

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
WMT-16 single system	25.9	31.1	—	29.6	—	—
baseline	27.5	32.0	16.4	31.3	19.7	21.7

- we start from stronger baselines (more data, adam, new BPE)

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
baseline	27.5	32.0	16.4	31.3	19.7	21.7
+layer normalization	28.2	32.1	17.0	32.3	18.8	22.5
+deep model	28.9	33.5	16.6	32.7	20.6	22.9

- we start from stronger baselines (more data, adam, new BPE)
- layer normalization and deep models generally help

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
+deep model	28.9	33.5	16.6	32.7	20.6	22.9
+checkpoint ensemble	29.4	33.8	17.7	33.3	21.0	23.6
+independent ensemble	30.3	34.4	18.5	33.6	21.6	25.1

- we start from stronger baselines (more data, adam, new BPE)
- layer normalization and deep models generally help
- checkpoint ensembles help, but independent ensembles are better

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
+independent ensemble	30.3	34.4	18.5	33.6	21.6	25.1
+right-to-left reranking	31.1	35.1	19.0	34.6	22.3	25.7

- we start from stronger baselines (more data, adam, new BPE)
- layer normalization and deep models generally help
- checkpoint ensembles help, but independent ensembles are better
- reranking helps

News Task: Into English

system	CS→EN 2017	DE→EN 2017	LV→EN 2017	RU→EN 2017	TR→EN 2017	ZH→EN 2017
WMT-16 single system	25.9	31.1	—	29.6	—	—
baseline	27.5	32.0	16.4	31.3	19.7	21.7
+layer normalization	28.2	32.1	17.0	32.3	18.8	22.5
+deep model	28.9	33.5	16.6	32.7	20.6	22.9
+checkpoint ensemble	29.4	33.8	17.7	33.3	21.0	23.6
+independent ensemble	30.3	34.4	18.5	33.6	21.6	25.1
+right-to-left reranking	31.1	35.1	19.0	34.6	22.3	25.7
WMT-17 submission	30.9	35.1	19.0	30.8	20.1	25.7

- we start from stronger baselines (more data, adam, new BPE)
- layer normalization and deep models generally help
- checkpoint ensembles help, but independent ensembles are better
- reranking helps
- large improvements over baseline

News Task: Out of English

system	EN→CS 2017	EN→DE 2017	EN→LV 2017	EN→RU 2017	EN→TR 2017	EN→ZH 2017
WMT16 single system	19.7	24.9	—	26.7	—	—
baseline	20.5	26.1	14.6	28.0	15.6	31.3
+layer normalization	20.5	26.1	14.9	28.7	15.7	32.3
+deep model	21.1	26.6	15.1	29.9	16.2	33.4
+checkpoint ensemble	22.0	27.5	16.1	31.0	16.7	33.5
+independent ensemble	22.8	28.3	16.7	31.6	17.6	35.8
+right-to-left reranking	22.8	28.3	16.9	–	18.1	36.3
WMT-17 submission	22.8	28.3	16.9	29.8	16.5	36.3

Biomedical Task

	EN→PL		EN→RO	
system	**Coch**	**NHS24**	**Coch**	**NHS24**
baseline	26.2	18.2	36.8	23.0
+layer normalization	25.5	20.2	35.6	24.7
+deep model	25.9	20.2	37.8	27.3
+checkpoint ensemble	28.4	21.3	39.1	27.0
+independent ensemble	28.1	21.6	40.5	28.3
+right-to-left reranking	28.6	22.5	40.8	29.0
WMT17 submission	29.0	23.2	41.2	29.3

Copied Monolingual Data

Table: BLEU scores for EN$\leftrightarrow$TR when adding copied monolingual data.

	TR$\rightarrow$EN		**EN$\rightarrow$TR**	
system	**2016**	**2017**	**2016**	**2017**
baseline	20.0	19.7	13.2	14.7
+copied	20.2	19.7	13.8	15.6

Biomedical Task: Domain Adaptation

	EN→PL		EN→RO	
system	**Coch**	**NHS24**	**Coch**	**NHS24**
generic (single)	22.8	16.6	37.6	26.5
generic (ensemble 4)	23.6	19.9	39.2	27.9
fine-tuned (single)	27.2	19.5	38.6	27.0
fine-tuned (ensemble 4)	27.4	20.9	39.9	26.0

Thank you!

Bibliography I

Denkowski, M. and Neubig, G. (2017).
Stronger Baselines for Trustable Results in Neural Machine Translation.
ArXiv e-prints.

Kingma, D. P. and Ba, J. (2015).
Adam: A Method for Stochastic Optimization.
In The International Conference on Learning Representations, San Diego, California, USA.

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.

Press, O. and Wolf, L. (2017).
Using the Output Embedding to Improve Language Models.
In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Neural Machine Translation of Rare Words with Subword Units.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to Sequence Learning with Neural Networks.
In
Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, pages 3104–3112, Montreal, Quebec, Canada.

Bibliography II

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016).
Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.
ArXiv e-prints.

Zeiler, M. D. (2012).
ADADELTA: An Adaptive Learning Rate Method.
CoRR, abs/1212.5701.

Zhou, J., Cao, Y., Wang, X., Li, P., and Xu, W. (2016).
Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation.
Transactions of the Association of Computational Linguistics – Volume 4, Issue 1, pages 371–383.