Regularization Techniques for Fine-tuning in Neural Machine Translation

Antonio Valerio Miceli Barone • Barry Haddow • Ulrich Germann • Rico Sennrich

ModernMT

TraMOOC
Translation for Massive Open Online Courses

Overview

- Good translation quality requires lots of parallel training data
- Only small datasets may be available in some domains
- Fine tuning
  - Train on a large out-of-domain dataset first
  - Continue training on a small in-domain dataset
  - How do we avoid overfitting to the in-domain dataset?

Regularization

- Overfitting can be prevented with early stopping
  - Effective, but requires a separate in-domain validation set
- We empirically investigate explicit regularization techniques
- Variational dropout (Gal and Ghahramani, 2016)
  - Randomly drop activations to zero the same way for each time step
    $v = W \cdot \frac{1}{p}\mathrm{diag}(\mathrm{Bernoulli}^{\otimes n}(p)) \cdot h$
  - Not a specific domain adaptation method
- MAP-L2 penalization (Chelba and Acero, 2006)
  - Penalize the L2-distance between the weights of the in-domain and out-of-domain models
    $L_W = \lambda \cdot \|W - W_{\text{out-of-domain}}\|_2^2$
  - We are the first to apply it to the domain adaptation of neural networks
- Tuneout
  - For each layer, randomly drop activations towards those computed with the weights of the out-of-domain model
    $v = (W_{\text{out-of-domain}} + \Delta W \cdot \frac{1}{p}\mathrm{diag}(\mathrm{Bernoulli}^{\otimes n}(p))) \cdot h$

Experimental setup

- Language pairs: English-to-German and English-to-Russian
- Out-of-domain data: WMT16 parallel + backtranslated monolingual data
- In-domain data: IWSLT 2015 (En→De) / 2014 (En→Ru)
- Model: GRU sequence-to-sequence with attention
- System: Nematus toolkit with BPE subword segmentation

Results

Table: Translation BLEU scores

System	En→De validation	En→De test (avg.)	En→Ru validation	En→Ru test (avg.)
Out-of-domain only	27.19	27.76	15.74	16.81
Early-stopping baseline	30.53	31.20	17.47	18.67
Early-stopping + dropout	30.63	31.33	17.68	18.80
Early-stopping + MAP-L2	30.81	31.25	**17.77**	18.91†
Early-stopping + tuneout	30.49	30.78†	17.51	18.78
Early-stopping + dropout + MAP-L2	30.80	**31.48**†	17.74	**19.10**†

†: different from the fine-tuning baseline at 5% significance.

Training curves

Figure: English→German validation BLEU over training mini-batches.

Effects of data size

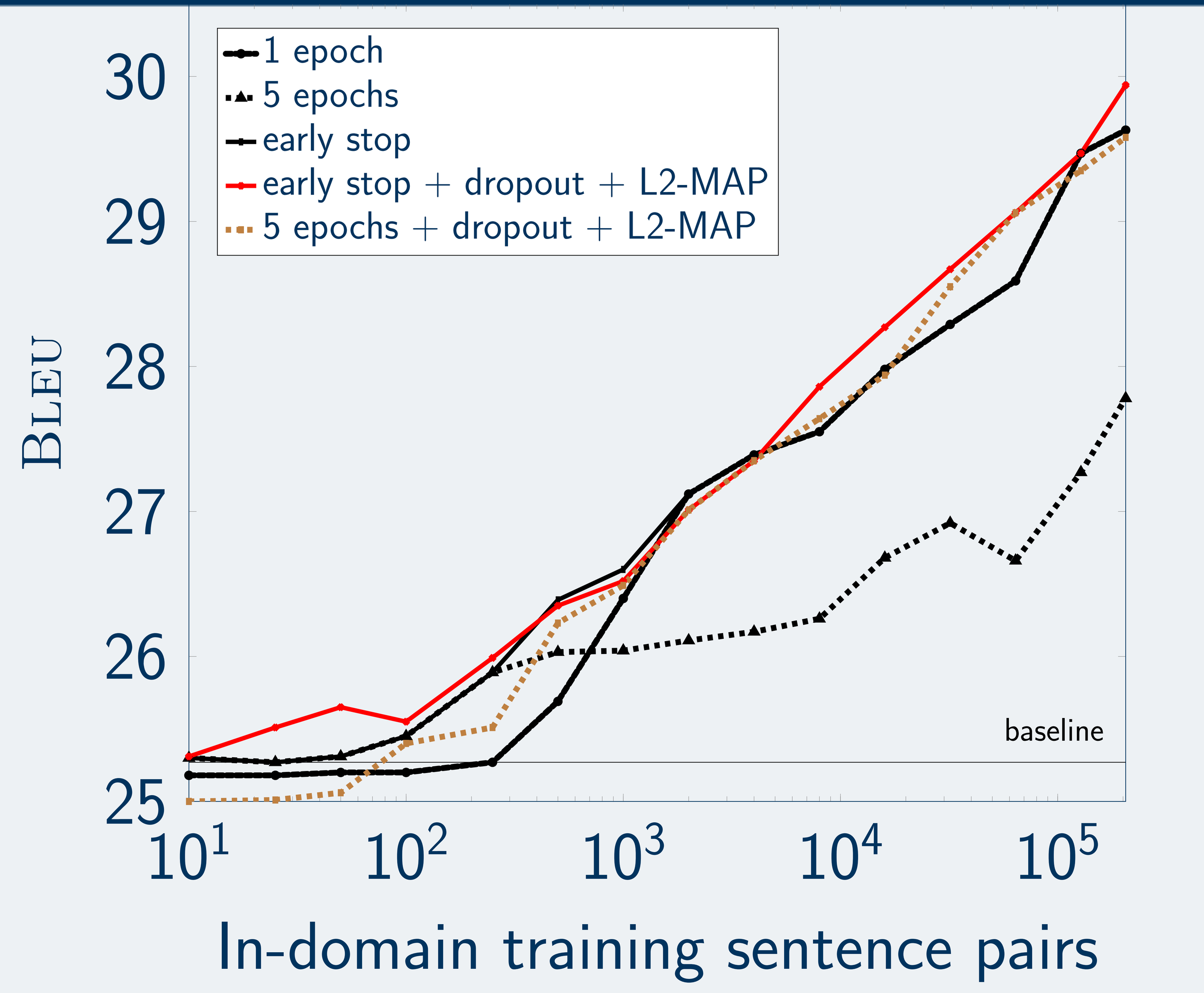

Figure: English→German test BLEU with fine-tuning on different in-domain data set size.

Findings

- On full-sized IWSLT training data
  - Dropout and MAP-L2 stabilize training, preventing overfitting
  - Dropout + MAP-L2 significantly improve over Early-stop alone
  - Tuneout did not yield improvements
- We evaluate Dropout + MAP-L2 over different indomain data sizes (10-206,000)
  - Logarithmic relation between data size and BLEU
  - Even for fixed number of epochs perform equally or better than Early-stop
  - Don't require held-out validation set
  - Fine-tuning without Early-stop or reguluarizers underfits or overfits
- We recommend using Dropout + MAP-L2 for fine-tuning, especially for very small amounts of in-domain data

Links

Nematus (includes Dropoout and MAP-L2) https://github.com/EdinburghNLP/nematus

Nematus (Tuneout branch) https://git.io/v7jSZ

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 644333 (TraMOOC) and 645487 (ModernMT).
We also thank Booking.com for their support.