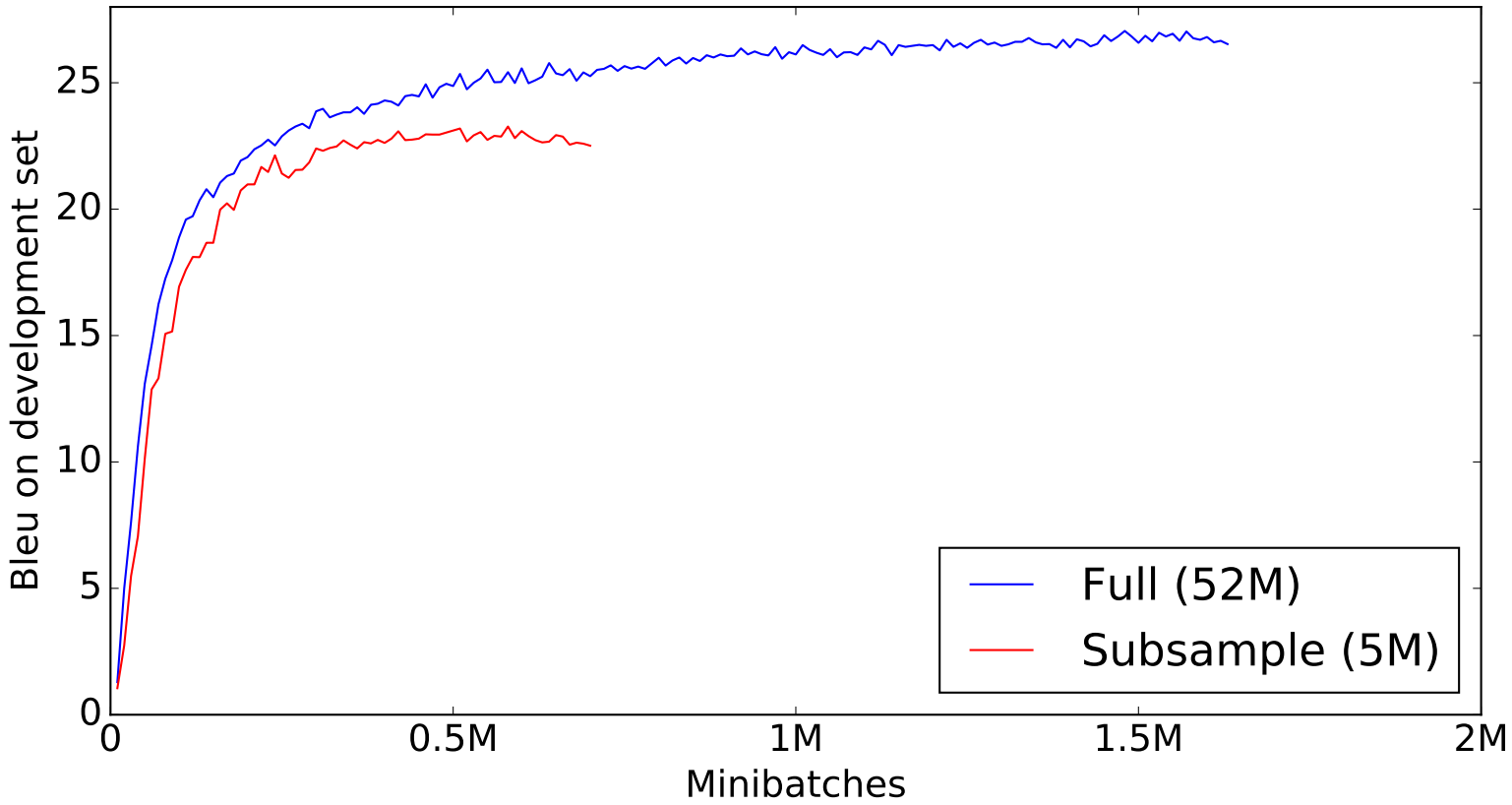
Training with Nematus on Large Data

Barry Haddow

University of Edinburgh

DGT, Luxembourg
March 15th, 2017

Nematus Learning Curves



- Czech→English, Full WMT data vs. Subsampled to 5M
- 10× increase in data leads to 3–4× increase in iterations to converge
- NB: subsampled has mini-batch 60, full has mini-batch 80

Nematus Resource Usage

Factors Influencing Speed/Memory Usage

- Number of model parameters, especially vocabulary size
- Size of training instance (max. length $\times$ batch size)
- Hardware and library versions

Nematus Training Benchmark (`test_train.sh`)

hardware	CPU	GPU	+CuDNN	+Theano 0.9b	+Theano cuda
sentences/s	2.5	83	138	172	227

(On Titan X Pascal – 50% faster than previous generation)

Estimating Training Time

$$\text{Days per 100k iterations} \approx 1.16 \times \frac{\text{batch size}}{\text{sentences per second}}$$

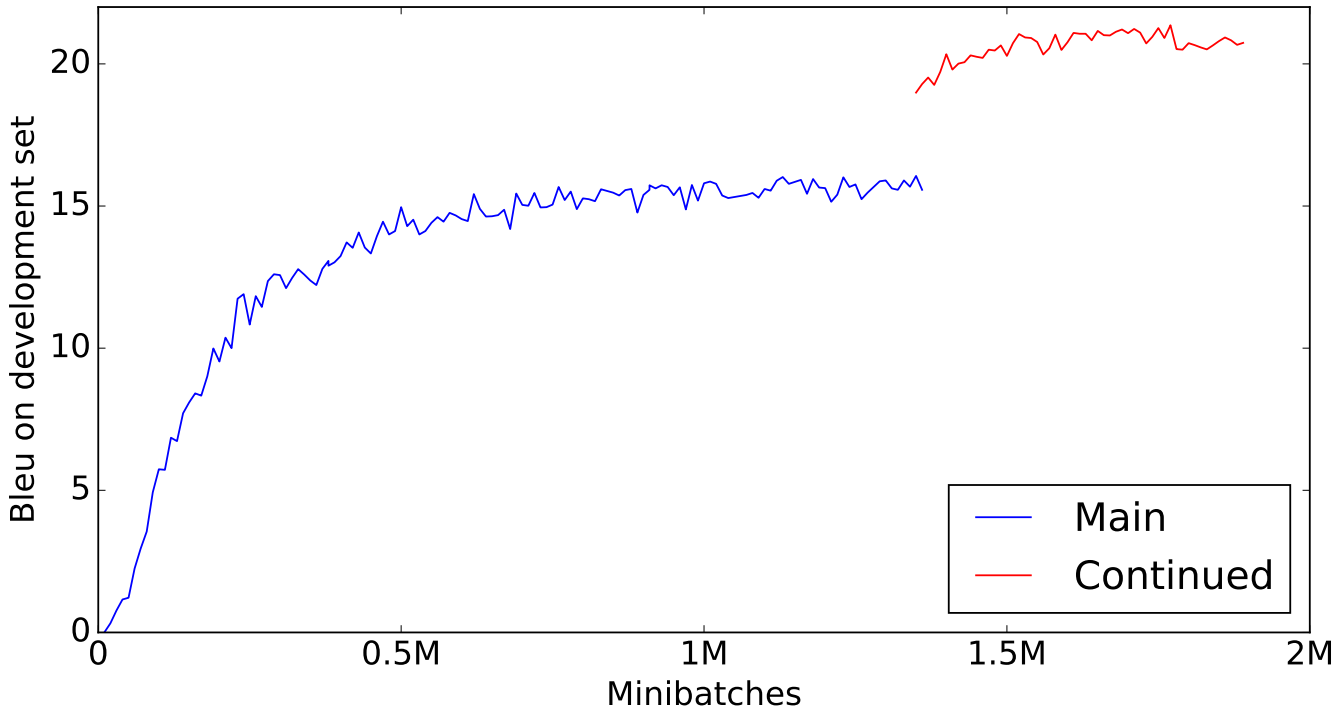
Minibatch 80 @ 200 sentences per sec $\rightarrow$ 11 hours per 100k

Continued Training

- New data does not require a full re-train
- Using BPE means open vocabulary $\rightarrow$ no need to retrain.
- e.g. domain specialisation can be achieved by continue training

Continued Training

- New data does not require a full re-train
- Using BPE means open vocabulary → no need to retrain.
- e.g. domain specialisation can be achieved by continue training



- General purpose system – finetuned with 50/50 mix of generic/in-domain synthetic

Bibliography


Britz, D., Goldie, A., Luong, T., and Le, Q. (2017).
Massive Exploration of Neural Machine Translation Architectures.
ArXiv e-prints.


Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Valerio Miceli Barone, A., Mokry, J., and Nădejde, M. (2017).
Nematus: a Toolkit for Neural Machine Translation.
In *Proceedings of EACL (demo session)*.