The University of Cambridge's Machine Translation Systems for WMT18

Felix Stahlberg, Adria de Gispert, Bill Byrne

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Introduction

- Large batch training (delayed SGD)
- Comparison and combination of
  - SMT (phrase-based)
  - NMT
    - RNN
    - SliceNet (convolutional)
    - Transformer (`transformer_big` and `transformer_relative_big`)

MBR-Based System Combination

- Generalisation of (Stahlberg et al., 2017)

Transformer + Rel. Transformer ensemble

$$S(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{T} \Big(\sum_{i=1}^{p} \lambda_i \log P(y_t|y_1^{t-1}, \mathbf{x}, \mathcal{M}_i) + \underbrace{\sum_{j=p+1}^{q} \lambda_j \sum_{n=1}^{4} P(y_{t-n}^{t}|\mathbf{x}, \mathcal{M}_j)}_{\text{MBR-based } n\text{-gram scores}} \Big)$$

PBMT, LSTM, SliceNet, R2L Transformer

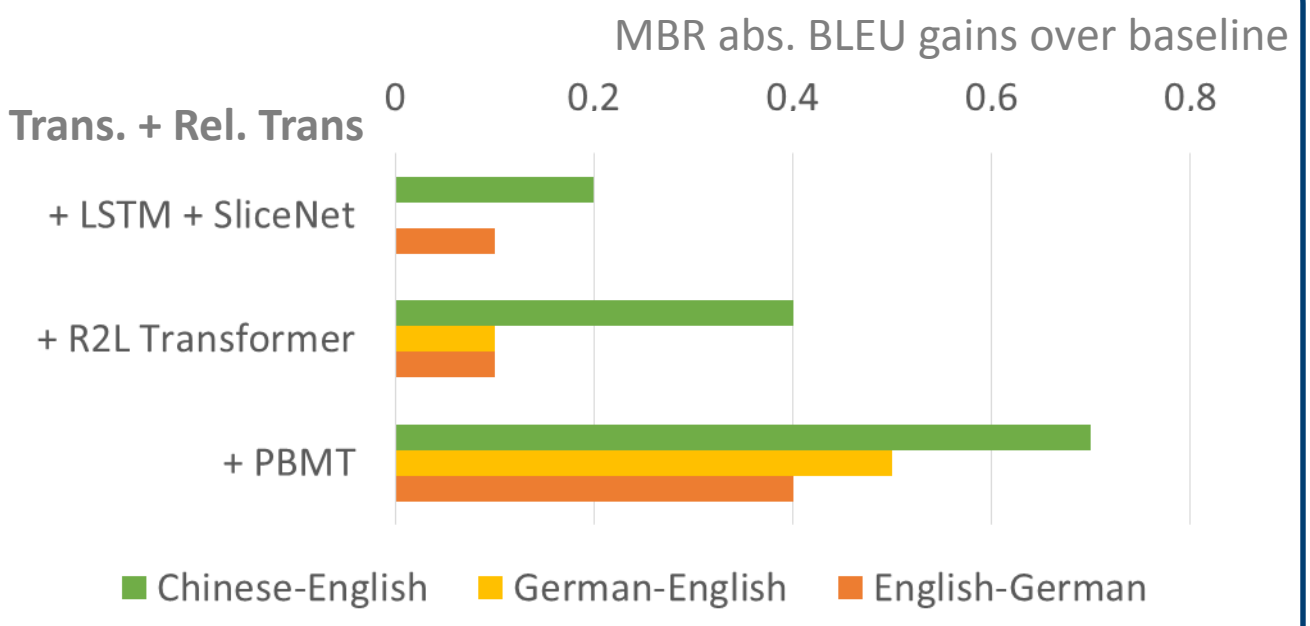
Delayed SGD

- Optimise batch size under GPU memory constraints by accumulating gradients over n batches

#Physical GPUs (g)	Delay factor (d)	#Effective GPUs (g'=gd)	Effective batch size (b'=bg')	BLEU
1	1	1	2,048	28.2
4	1	4	8,192	29.5
4	4	16	32,768	30.3
4	16	64	131,072	29.8

Baselines

System	en-de	de-en	zh-en
PBMT	20.0	28.2	15.8
LSTM	28.5	35.3	23.6
SliceNet	28.3	34.3	23.4
Transformer	30.5	37.9	25.6
Rel. Trans	31.1	38.1	25.8
Trans. + Rel. Trans	31.3	38.2	26.4



English-German

System	BLEU	Av. %	Av. z
MS-Marian	48.3	81.9	0.551
UCAM	46.6	82.3	0.537
NTT	46.5	80.2	0.491
KIT	46.3	79.3	0.454
…			

German-English

System	BLEU	Av. %	Av. z
RWTH	48.4	79.9	0.413
UCAM	48.0	79.4	0.395
NTT	46.8	78.2	0.359
MLLP-UPV	45.1	77.4	0.321
…			

Chinese-English

System	BLEU	Av. %	Av. z
NiuTrans	28.7	78.8	0.140
UCAM	27.7	77.9	0.109
Unisound-A	28.4	78.0	0.108
Tencent-Ens.	29.3	77.5	0.099
…			

- English-German: Best accuracy (on par with NTT) for paradigm contrast features (Burlot et al., 2018)
- German-English: Single best accuracy on a variety of linguistic phenomena (Macketanz et al., 2018)