L21

Learning to Generate Textual Data

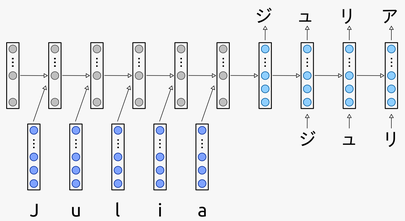
Guillaume Bouchard*
University College London
Bloomsbury AI
g.bouchard@cs.ucl.ac.uk

Pontus Stenetorp*
University College London
p.stenetorp@cs.ucl.ac.uk

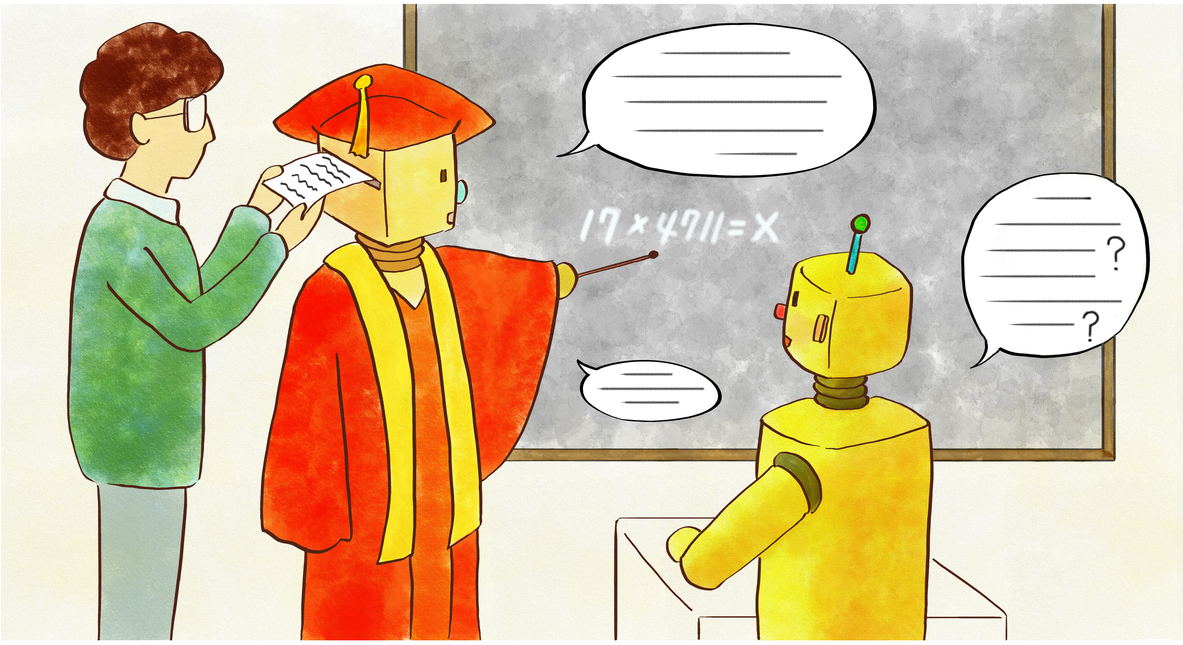
Sebastian Riedel
University College London
s.riedel@cs.ucl.ac.uk

Vision

- ~~Big~~ Small data, the next big thing?
- Regularise
  - Model *shrinking* (underfit)
  - Model *bias* (less general)
- Virtual samples (Niyogi et al., 1998)
- Generation, but *what* and *how much*?
- "Learning to Teach"
  - *Generic* model
  - Bias through *learned generation*



Encoder-decoder (Sutskever et al. 2014)



Results

	AI2	IL	CC	Avg.
RR2015	82.4	75.4	55.5	71.1
100% Data	72.5	53.7	95.0	73.7
100% Gen	60.3	51.2	92.0	67.8
85%Gen + 15%Data	74.0	55.4	97.5	75.6
GeneRe	77.9	56.7	98.5	77.7

- Model details
  - 1-layer GRU with 256-dimensional hidden state
  - Projected 300 dimensional pre-trained vectors
- Using only non-generated or generated is insufficient
- Mixing generated and non-generated helps
- Using GeneRe improves the results even further

Generative Regularisation (GeneRe)

Require: $\hat{P}$: real data sampler
Require: P_γ: parametric data generator
Require: λ: generative regularization strength
Require: η: learning rate
Require: α: baseline smoothing coefficient

1: Initialize parameters θ, sampling coefficients γ and baseline μ
2: **for** $t = 1, 2, \cdots$ **do**
3: $\quad x, y \sim \frac{1}{1+\lambda}\hat{P} + \frac{\lambda}{1+\lambda}P_\gamma$
4: $\quad g_\theta \leftarrow \nabla_\theta \log p_\theta(y|x)$
5: $\quad g_\gamma \leftarrow (\log p_\theta(y|x) - \mu)\, \nabla_\gamma \log p_\gamma(x, y)$
6: $\quad (\theta, \gamma) \leftarrow (\theta, \gamma) - \eta(g_\theta, g_\gamma)$
7: $\quad \mu \leftarrow \alpha\mu + (1-\alpha) \log p_\theta(y|x)$
8: **end for**

- Variant of `Reinforce` (Williams, 1988)

Math Word Problems

Example:

"A pet store had 81 puppies. In one day they sold 41 of them and put the rest into cages with 8 in each cage. How many cages did they use?"

Template:

"S1 V1 Q1 O1. C1 S1(pronoun) V2 Q2 of O1(pronoun) and V2 the rest into O3(plural) with Q3 in each O3. How many O3(plural) V3?"

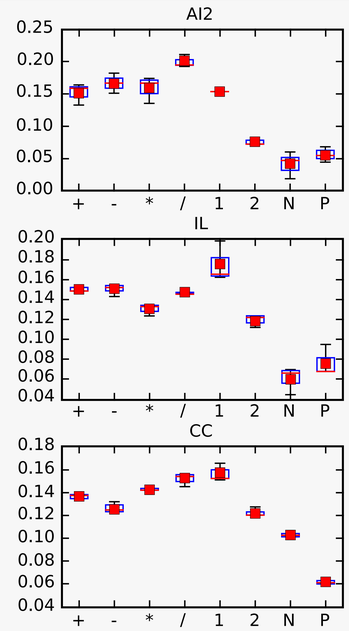
	AI2	IL	CC
Train	198	214	300
Dev	66	108	100
Test	131	240	200
Total	395	562	600

AI2	IL	CC
X + Y	X + Y	X + Y − Z
X + Y + Z	X − Y	X * (Y + Z)
X − Y	X * Y	X * (Y − Z)
	X/Y	(X + Y)/Z
		(X − Y)/Z

	The problem...
+	contains at least one addition
-	contains at least one subtraction
*	contains at least one multiplication
/	contains at least one division
1	has a single mathematical operation
2	has a couple of mathematical operations
N	is a training sample with words removed
P	is based on word position frequencies

- Current state of the art: RR2015 (Roy and Roth, 2015)

Optimal γ proportions



*Contributed equally to this work.
Illustration by Mika Riedel (http://mikariedel.github.io/).

The 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), November 3rd 2016, Austin, Texas, USA