The Samsung's submission to PolEval 2019 Machine Translation task

Marcin Chochowski, Paweł Przybysz

Samsung R&D Institute Poland

31 May 2019

Task plan

Initial plan:

- Train baseline model en-pl
- Train baseline model pl-en for backtranslation
- Filter out additional in-domain data
- Train full model with final finetunnig

Data inspection

- the domain is "lectures"
- sentences are splitted into several lines
- sentences are often misaligned
- significant overlap in train and dev sets

Polish	English
One sa z grubsza tej wielkości, ale różnia sie miedzy soba wymiarem, nie sa stworzone do tego,	Die sind in diesem Größenbereich, sehen aber alle unterschiedlich groß aus, sind natürlich auch nicht dafür gemacht
$2 * 16 \prod cm^2$	and now we have to figure out the surface area of this thing that goes around
Pawda? 1+2+3+...+k+k+1 jest suma wszystkich liczb z właczeniem k+1	Well we are assuming that we know what this already is.

Parallel-corpora filtering framework

Each sentence pair is scored with

- sentence-level language recognition tools
- sentences length
- fast-align score

These scores are input features for sentence pair classifier

- trained on sample of manually evaluated pairs
- output score resembles human judgment

Used data

Corpus	Sentence Pairs (raw)	Sentence Pairs (filtered)	filtering rate
Bible	0.03M	0.02M	79.81%
Books	0.002M	0.001M	40.09%
DGT	3M	2.8M	92.57%
ECB	0.07M	0.04M	66.74%
EMEA	0.9M	0.7M	78.36%
Eubookshop	0.5M	0.3M	65.68%
Euconst	0.008M	0.007M	91.40%
Europarl.7	0.6M	0.6M	98.09%
GlobalVoices	0.04M	0.03M	87.38%
GNOME	0.006M	0.005M	81.83%
IATE	0.1M	0.05M	44.81%
JRC-Acquis	1.3M	1.2M	93.08%
KDE4	0.1M	0.1M	63.62%
OpenSubtitles.2018	41.3M	34.7M	83.95%
Paracrawl.v1	1.3M	0.9M	75.76%
PHP	0.03M	0.02M	75.76%
Tanzil	0.1M	0.1M	88.35%
Tatoeba	0.002M	0.002M	98.20%
TED	0.2M	0.2M	93.54%
Ubuntu	0.006M	0.003M	46.16%
Wikipedia	0.1M	0.1M	68.50%
wikititles	0.7M	0.1M	18.79%
PolEval 2019 in-domain	10x 0.1M	10x 0.1M	81.35%
Synthetic in-domain	5x 0.1M	5x 0.1M	100%
Total	52.3M	43.7M	81.69%

Training

Preprocessing:

- sentencepiece tokenization and word segmentation (32k)

Network Configuration:

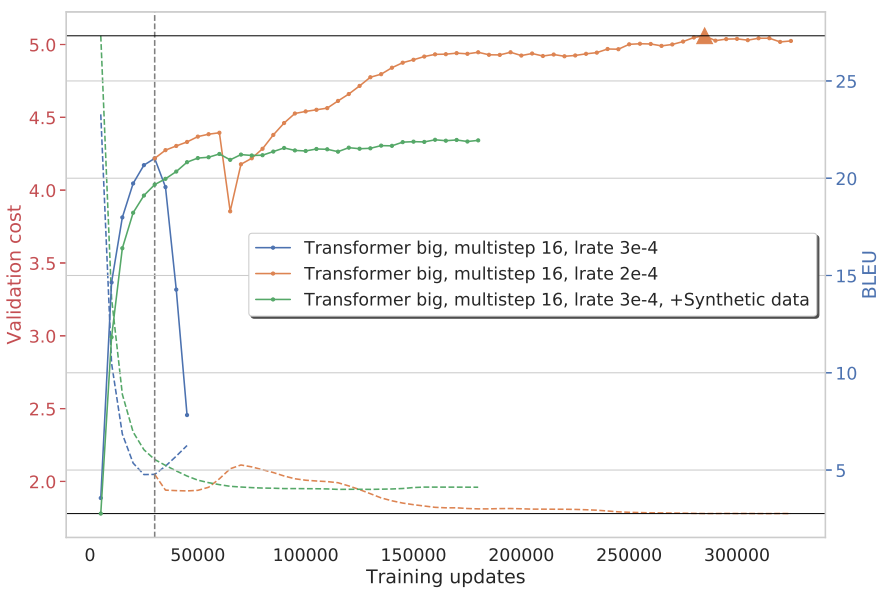
- *transformer_big*
- Hidden layer size 1024, filter size 4096
- 16-head attention

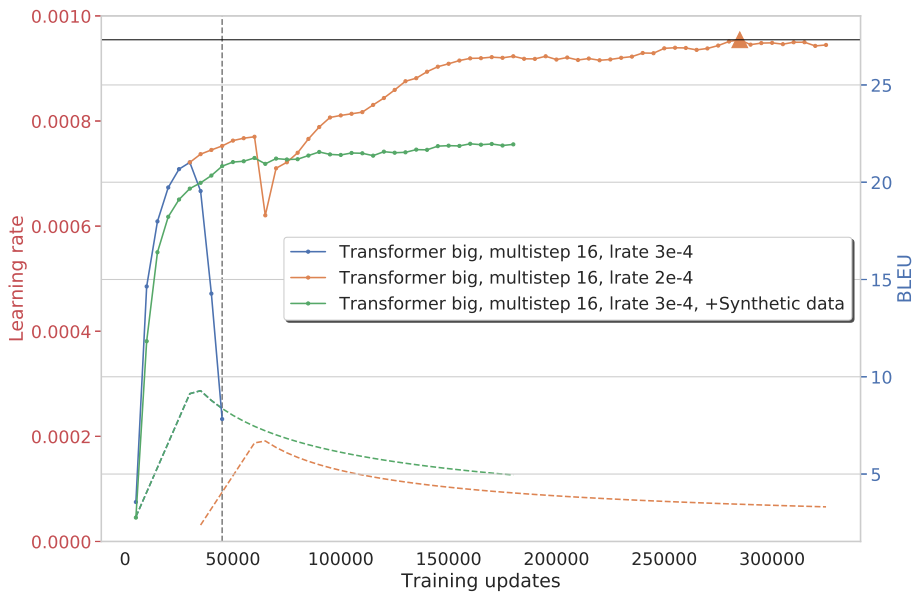
Training:

- Warmup (32k steps)
- 8x v100 GPUs
- increasing effective batch size x16

Results

- unstable trainings
- sensitive to hyperparameters
- synthetic data did not help
- final score





Results

Why synthetic data did not help ?

- high performing backtranslation system (33.51 BLEU)

Polish	English
do rzeczywistego kurczenia sie mieśni? Do tego, co obserwujemy codziennie jako energie mechaniczna?	contracting things, to actually doing what we see in our everyday world as mechanical energy?

Thank you!