Defence of PhD thesis:

Machine Translation Using Syntactic Analysis

Martin Popel
popel@ufal.mff.cuni.cz

ÚFAL (Institute of Formal and Applied Linguistics)
2018-09-19

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

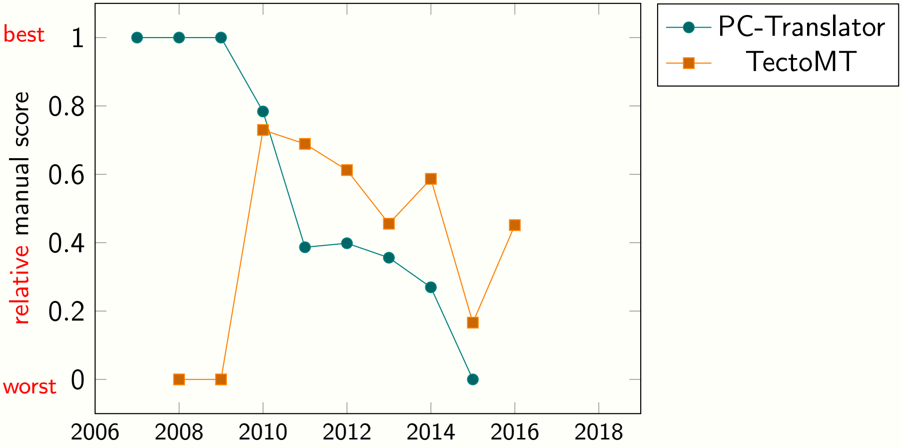
source	Chytá tlouště na višni.
Yandex	Catch a chub on a cherry.
Bing	Catching chub on Višni.
Google	Catch fat on cherry.
TectoMT	It catches chub to cherry.
Transformer	He catches the fat on the cherry.

Overview

- TectoMT – deep-syntactic MT
- Transformer – neural MT
- Evaluation
- Conclusion

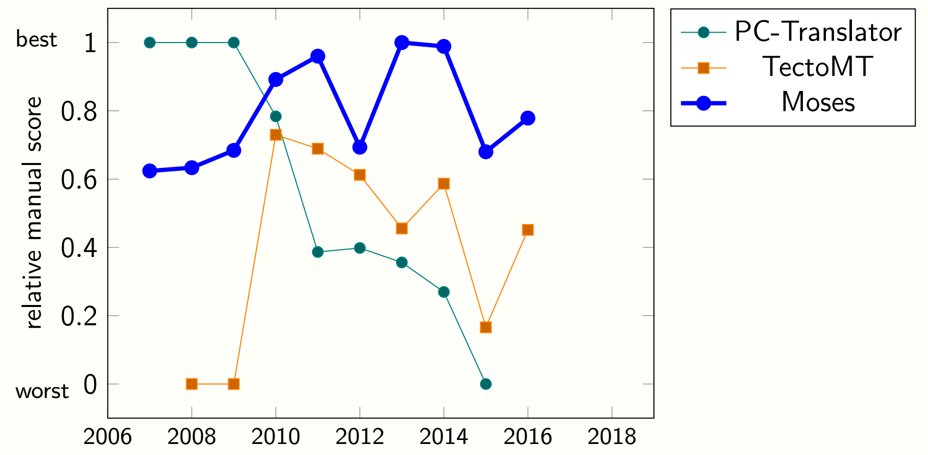
source	Great talkers are little doers.
Yandex	Velké talkers jsou trochu činitelé.
Bing	Velcí vysílačky jsou malí činitelé.
Google	Velcí mluvčí jsou malí lidé.
TectoMT	Velcí řečníci jsou malí vrazi.
Transformer	Velcí mluvkové jsou malí dříči.

WMT 2007–2018 English→Czech manual evaluation



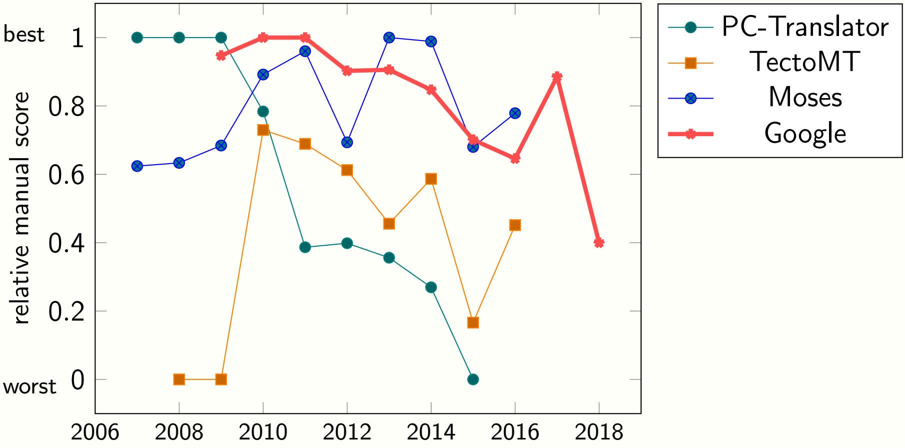
linearly scaled scores each year (best system = 1, worst system = 0)

WMT 2007–2018 English→Czech manual evaluation



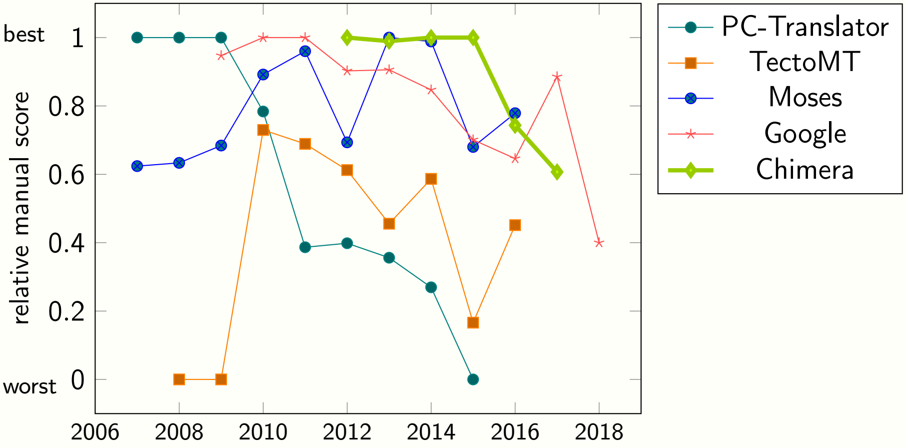
linearly scaled scores each year (best system = 1, worst system = 0)

WMT 2007–2018 English→Czech manual evaluation



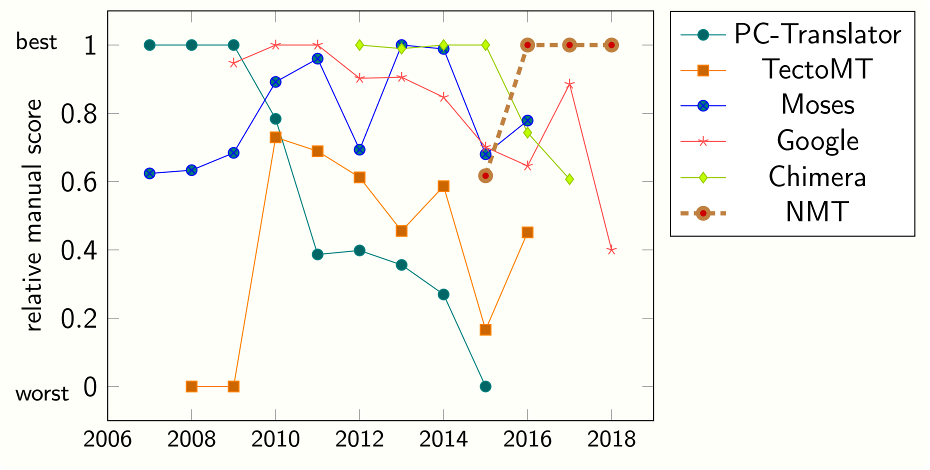
linearly scaled scores each year (best system = 1, worst system = 0)

WMT 2007–2018 English→Czech manual evaluation



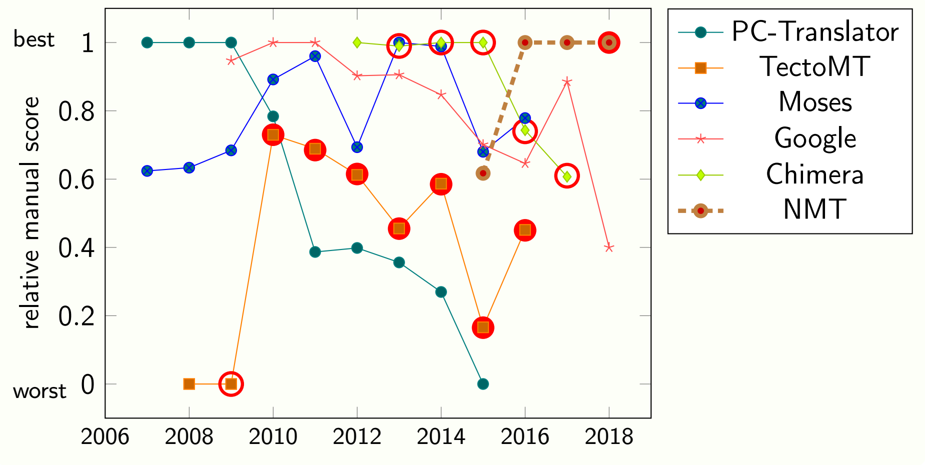
linearly scaled scores each year (best system = 1, worst system = 0)

WMT 2007–2018 English→Czech manual evaluation



linearly scaled scores each year (best system = 1, worst system = 0)

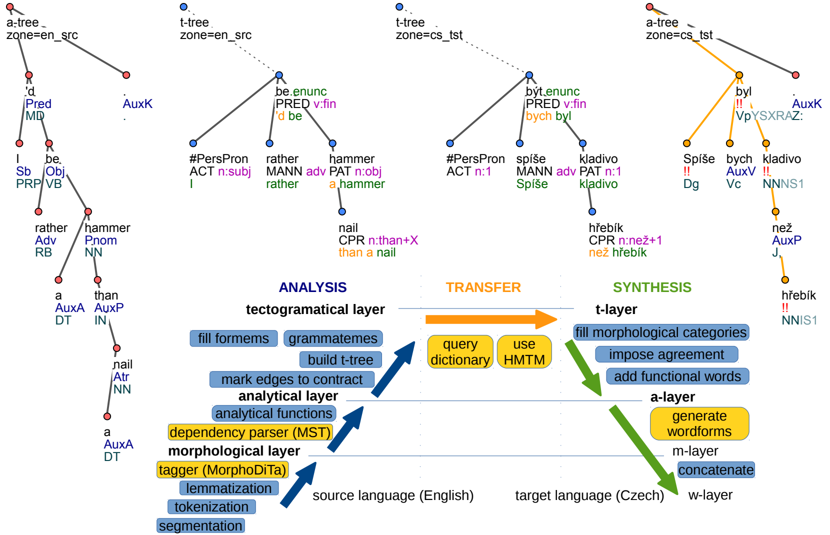
WMT 2007–2018 English→Czech manual evaluation



linearly scaled scores each year (best system = 1, worst system = 0)

my contribution (partial)

TectoMT: analysis, transfer, synthesis



I'd rather be a hammer than a nail. Spíše bych byl kladivo než hřebík/nehet.

Discriminative TM features

output_label=hřebík#N	
feature	λ
child_formeme_n:in+X=1	1.64
is_member_of_coord=1	1.30
child_formeme_v:fin=1	1.04
next_lemma=down	0.84
is_capitalized=1	0.79
+precedes_parent=0	**0.75**
tense_g=post	0.74
+voice_g=active	**0.66**
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+prev_lemma=hammer	**0.59**
child_lemma_few=1	0.55
child_lemma_remove=1	0.54
sempos=n.denot	0.50
next_lemma=and	0.50
formeme_g=v:until+fin	0.49
child_lemma_rusty=1	0.47
...	

MaxEnt (logistic regression)

$$Z(\mathbf{x}) = \sum_y exp \sum_i \lambda_i f_i(\mathbf{x}, y)$$

Discriminative TM features

output_label=hřebík#N

feature	λ
child_formeme_n:in+X=1	1.64
is_member_of_coord=1	1.30
child_formeme_v:fin=1	1.04
next_lemma=down	0.84
is_capitalized=1	0.79
+precedes_parent=0	**0.75**
tense_g=post	0.74
+voice_g=active	**0.66**
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+prev_lemma=hammer	**0.59**
child_lemma_few=1	0.55
child_lemma_remove=1	0.54
sempos=n.denot	0.50
next_lemma=and	0.50
formeme_g=v:until+fin	0.49
child_lemma_rusty=1	0.47
...	

output_label=nehet#N

feature	λ
child_formeme_n:poss=1	1.32
child_lemma_finger=1	1.07
child_formeme_n:of+X=1	0.98
precedes_parent=1	0.88
prev_lemma=black	0.77
child_lemma_broken=1	0.76
child_formeme_v:attr=1	0.70
formeme=n:at+X	0.67
formeme_g=n:attr	0.67
child_lemma_long=1	0.67
next_lemma=file	0.60
child_lemma_false=1	0.58
prev_lemma=false	0.58
+number=sg	**0.56**
formeme=n:obj	0.53
formeme=n:by+X	0.52
...	

Discriminative TMs evaluation

TectoMT version	BLEU No LM	BLEU TreeLM
Baseline TM	10.70	12.15
MaxEnt TM	12.78	13.57
VowpalWabbit TM	13.07	13.77

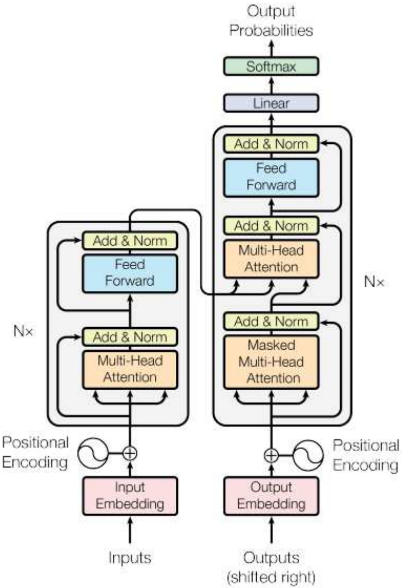
Advantages of VowpalWabbit over MaxEnt:

- Training more than 1000 times faster.
- One model for all source lemmas, multi-task learning.
- Advanced features: quadratic, novel “label-dependent”.

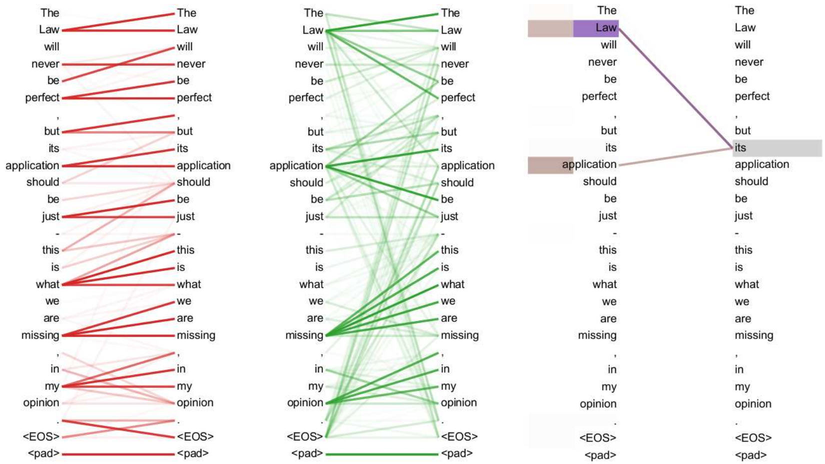
TectoMT Achievements

- Complementary strengths to SMT. Chimera won in 2013–2015.
- Adapted for English $\leftrightarrow$ **Czech, Spanish, Portuguese**, Dutch and Basque.
- Domain adaptation for an IT helpdesk application.
- For 5 language pairs TectoMT outperformed SMT.
- TectoMT is linguistically interpretable (cf. clitic reordering).

Transformer architecture (Vaswani et al., 2017)



Transformer: self-attention visualization



Adapted from Vaswani et al. (2017).

Experiments with Transformer Training

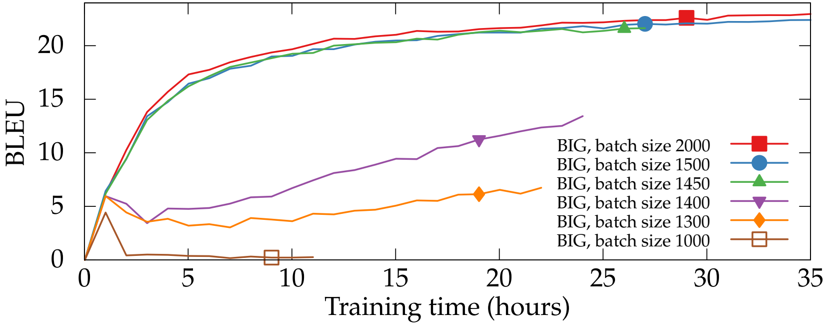
learning curves (dev-set BLEU vs. training time) for

- #GPUs, model size, learning rate, warmup steps,
- maximum sentence length, checkpoint averaging, ...

Experiments with Transformer Training

learning curves (dev-set BLEU vs. training time) for

- #GPUs, model size, learning rate, warmup steps,
- maximum sentence length, checkpoint averaging, . . .
- batch size



Backtranslation (Sennrich et al., 2016)

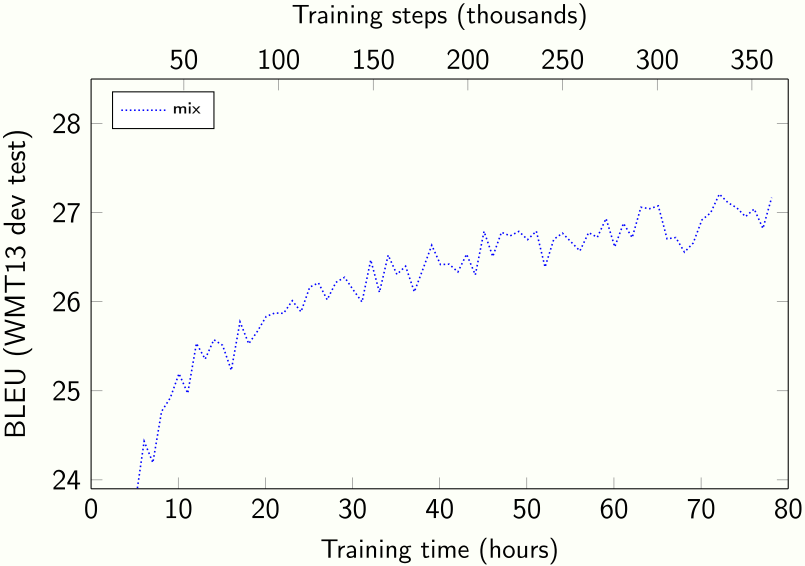
- For EN→CS translation, we can exploit monolingual CS data.
- Translate the data back to English (with any CS→EN MT).
- Prepare synthetic parallel data (orig-CS, synth-EN).
- Train on both authentic and synthetic
    - **fine-tune**: first auth then auth+synth
    - **mixed**: shuffle auth and synth sentences 1:1

previous SotA

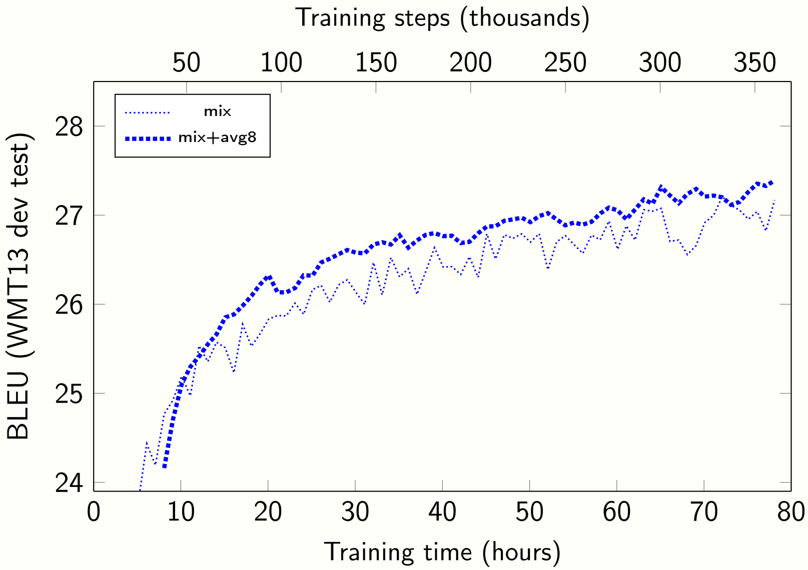
Backtranslation (Sennrich et al., 2016)

- For EN→CS translation, we can exploit monolingual CS data.
- Translate the data back to English (with any CS→EN MT).
- Prepare synthetic parallel data (orig-CS, synth-EN).
- Train on both authentic and synthetic
  - **fine-tune**: first auth then auth+synth
  - **mixed**: shuffle auth and synth sentences 1:1
  - **concat**: no shuffle, just concatenate auth and synth blocks

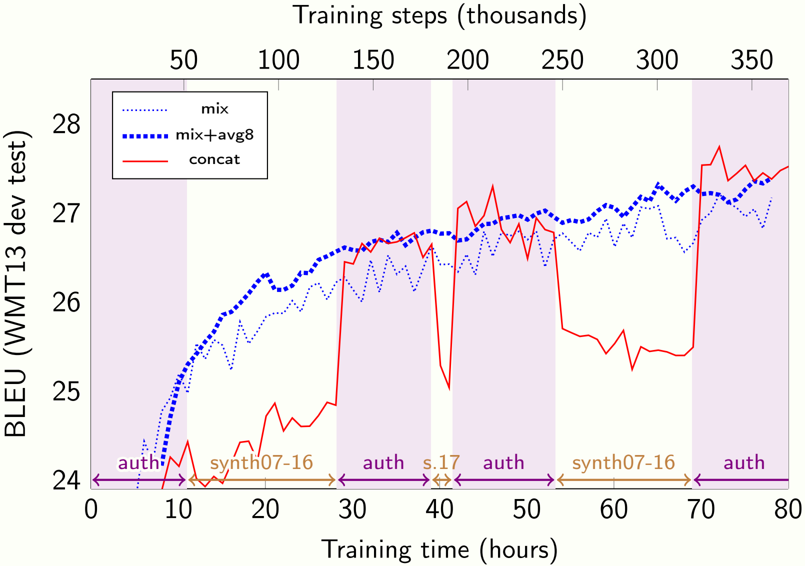
Concat Backtranslation



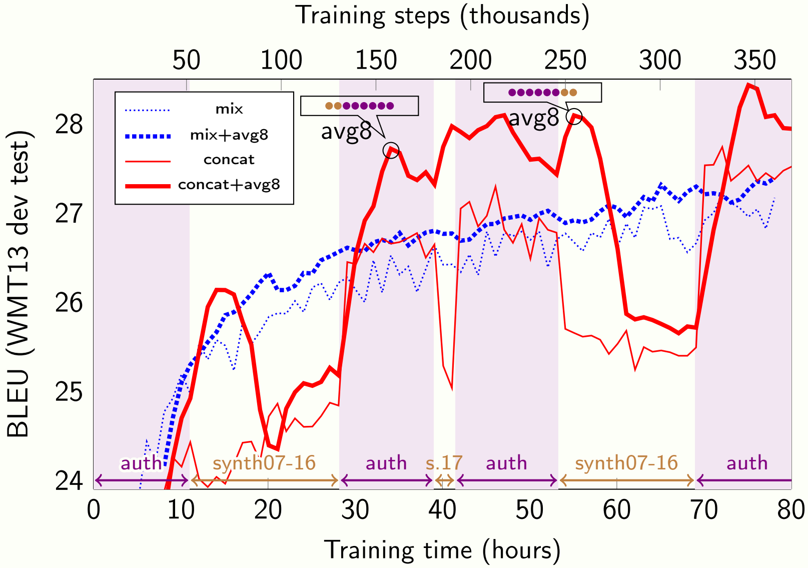
Concat Backtranslation



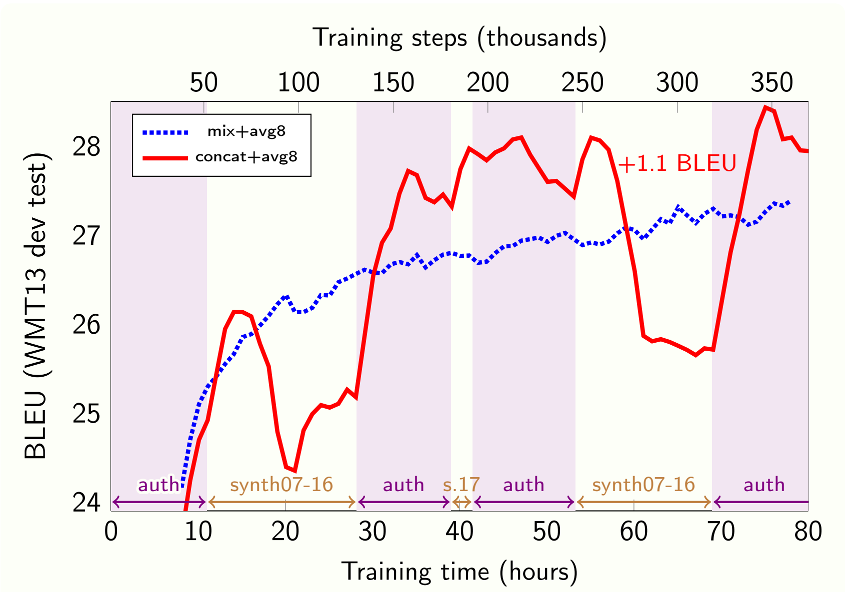
Concat Backtranslation



Concat Backtranslation



Concat Backtranslation



WMT2018 Evaluation: EN→CS BLEU

system	BLEU uncased	BLEU cased	chrF2 cased
CUNI-Transformer	**26.82**	**26.01**	**0.5372**
UEdin NMT	24.30	23.42	0.5166
Chimera	21.43	19.81	0.4838
Online-B	20.16	19.45	0.4854
Moses	17.88	16.36	0.4594
Online-A	16.84	15.74	0.4584
Online-G	16.33	15.11	0.4560
TectoMT	13.09	12.43	0.4332

WMT2018 Evaluation: EN→CS Manual (SrcDA)

	Ave. %	Ave. z	System
1	**84.4**	**0.667**	CUNI-Transformer
2	79.8	0.521	uedin
	78.6	0.483	newstest2018-ref
4	68.1	0.128	online-B
5	59.4	−0.178	online-A
6	54.1	−0.354	online-G

Significantly different ($p < 0.05$, Wilcoxon rank-sum test) systems are separated by a line.

Conclusion: Achievements

- Improved English↔Czech MT.
  Transformer significantly better than all other MT systems.
  online demo at LINDAT, WMT18 outputs at
  `http://wmt.ufal.cz`
- Explored the impact of syntactic structures in TectoMT.
- Explored domain-adaptation techniques (IT, translationese).
- Contributed to training data (CzEng) preparation,
  organizing shared tasks (WMT16 IT-task, CoNLL 17–18),
  evaluation tools (MT-ComparEval),. . .

Thanks for ~~self~~ your attention

source	As good be an addled egg as an idle bird.
Bing	Jako dobrý být popletený vejce jako nečinný pták.
Google	Jako dobrá být včleněná vejce.
T2009	Dobré je feťácké vejce jako činný pták.
T2018	Dobří buďte plete vejce jako nečinný pták.
Trans.	Stejně dobré je být pomateným vejcem jako zahálejícím ptákem.

Birds of a feather flock together.	source	A miss by an inch is a miss by a mile.
Ptáci peří stáda dohromady.	Bing	Miss o palec je Miss o míli.
Vrána k vráně sedá.	Yandex	Slečna tím, že palec je vedle o míli.
Vrána k vráně sedá.	Google	Chybějící palcem je míle vzdálená míle.
Ptáci v bederním hejnu spolu.	T2009	Slečna palec je slečna miliónu.
Ptáci péřového hejna spolu.	T2018	Slečna palce je slečna míle.
Vrána k vráně sedá.	Trans.	Minutí o centimetr je o kilometr.

Comparison with RNMT

Transformer (Vaswani et al., 2017) introduced several novelties:

A self-attention instead of RNN

B multihead-attention, layer normalization, label smoothing, linear warmup of learning rate schedule, synchronous training, variable batch size

Chen et al. (2018) designed RNMT+ by enhancing RNMT with the techniques (B). According to their BLEU evaluation:

- RNMT+ is competitive with Transformer.
- Best result achieved using Transformer encoder with RNMT+ decoder.

Improvements introduced in my thesis (concat backtranslation + checkpoint avg, iterated backtranslation,. . .) can be applied to RNMT (or RNMT+ or RNN/self-attention hybrids).

Decoding Speed

transformer_big (800 MiB model) on a single 1080Ti GPU, times including 35 seconds for loading the model:

decoding params			decoding time per	
beam	alpha	batch	3000 sents (s)	1 sentence (ms)
1	0.6	32	135	45
4	0.6	32	276	92
4	0.6	16	328	109
4	1.0	32	398	132
1	1.0	1	1689	562
4	1.0	1	2870	956

Comparison with Ensembles

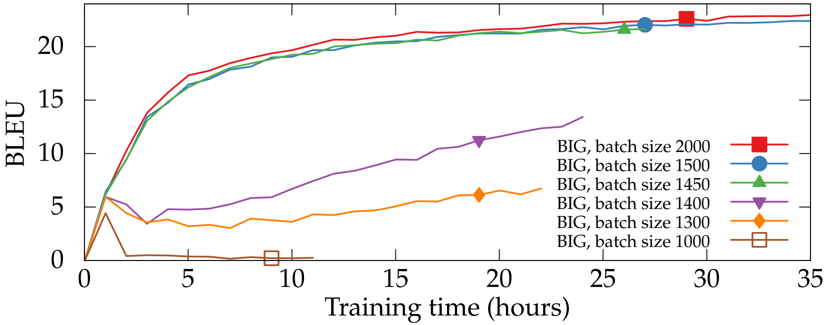
checkpoint averaging	checkpoint ensembles
semi-independent averaging	independent ensembles

- T2T does not currently support ensembles.
- Ensembles are not practical for deployment (cf. distillation).
- I tried semi-independent averaging of 3 models (+0.5 BLEU), checkpoint averaging (+0.5 BLEU) and combination of both (+0.7 BLEU).
- Junczys-Dowmunt et al. (2016, §6.3) report that averaging ten checkpoints "slightly outperforms the real four-model ensemble".

Contradiction: scaling batch size vs. number of GPUs

Considering BLEU after a given amount of training examples:

- Figure 4.6: batch_size > 1450 has no effect.



Contradiction: scaling batch size vs. number of GPUs

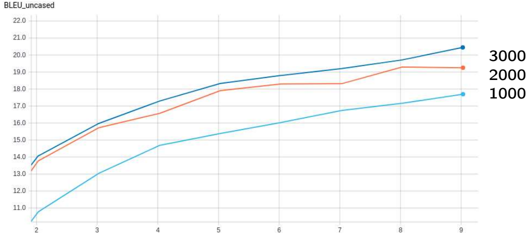
Considering BLEU after a given amount of training examples:

- Figure 4.6: batch_size > 1450 has no effect.
- Section 4.3.7: effective batch size > 1500 (using 2 or 6 GPUs) has a positive effect.
- "1 GPU & 4000 batch" = "2 GPUs & 2000 batch" etc.
- Is it a contradiction?
- I confirmed the reported experimental results: $1400<1450=1500=2000?3000$ vs. $1500<3000<9000=12000$

Contradiction: scaling batch size vs. number of GPUs

Considering BLEU after a given amount of training examples:

- Figure 4.6: batch_size > 1450 has no effect.
- Section 4.3.7: effective batch size > 1500 (using 2 or 6 GPUs) has a positive effect.
- "1 GPU & 4000 batch" = "2 GPUs & 2000 batch" etc.
- Is it a contradiction?
- I confirmed the reported experimental results: 1400<1450=1500=2000?3000 vs. 1500<3000<9000=12000
- but not when tried with Adafactor instead of Adam:



Reproducibility

- I replicated *some* experiments with different seeds: $<$ 0.2 BLEU variance ($\approx$ checkpoint-to-checkpoint variance).
- I could not afford replicating *all* experiments (or even try 10 runs), cf. over 4 years of total GPU time.
- My current English-French experiments show similar effects (as Chapter 5: concat backtranslation).

Data Block Size Effect in Concat Backtranslation

- The best ratio: 6 (one-hour-)checkpoints auth and 2 synth.
- Concat+avg advantage: tries all ratios, finds the best on dev.
- By default: auth block takes 11 hours.
- When auth longer: (the same) best result still after 6 hours.
- When auth shorter than 6 hours: optimal ratio not achieved.
- With shorter blocks, we would need more frequent checkpoints or less checkpoints in the average.
- Future work: mixed backtranslation, but varying ratio of synth:auth during training.

Extra slides

source	Loví tlouště na višni.
Yandex	Fishing for chub on a cherry.
Bing	They hunt chub on višni.
Google	He's hitting fat on sour cherries.
TectoMT	It hunts chub to cherry.
Transformer	He hunts fat on cherry.