CUNI System for the BEA Grammatical Error Correction

Jakub Náplava, Milan Straka
{naplava,straka}@ufal.mff.cuni.cz

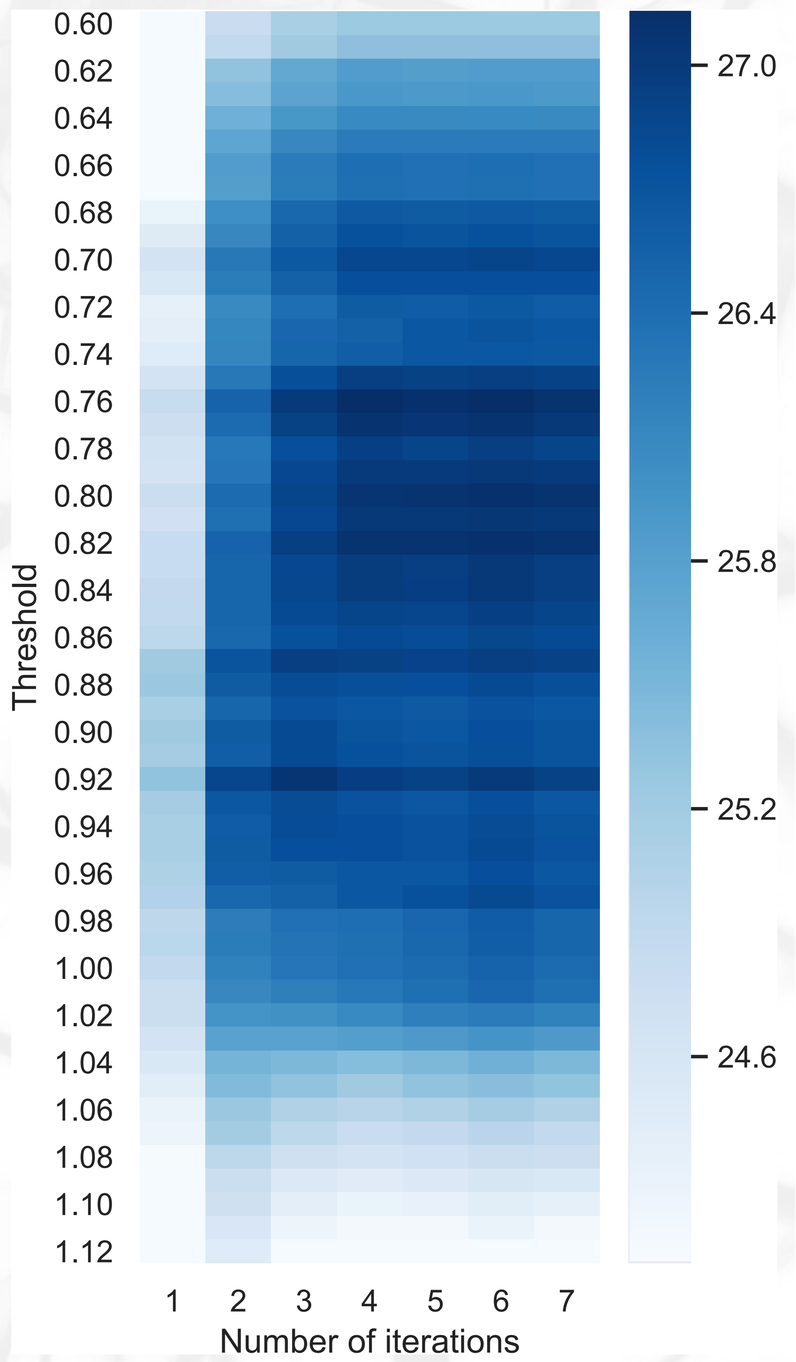
- we participate in all 3 Tracks
  - Restricted Track - 10th place
  - Unrestricted Track - 3rd place
  - Low-Resource Track - 5th place

Track	P	R	$F_{0.5}$	Best	Rank
Restricted	67.33	40.37	59.39	69.47	10 / 21
Unrestricted	68.17	53.25	64.55	66.78	3 / 7
Low Resource	50.47	29.38	44.13	64.24	5 / 9

Restricted Track

- we use T2T Transformer model with following enhancements:
  - dataset oversampling final training set = 10xW&I + 5xNUCLE + 5xFCE + 1xLang8
  - source and target word dropout zero entire source/target (sub)word embedding
  - edited MLE non-matching target words have bigger weight in MLE
  - checkpoint averaging average Transformer weights across several last checkpoints
  - iterative decoding iterate decoding process as long as new correction is more probable than current input multiplied by predefined threshold

System	A	B	C	N	Combined
Transformer-base architecture	39.98	32.68	23.97	14.49	32.47
Transformer-big architecture	39.70	35.13	26.22	20.20	34.20
+ 0.2 src drop, 0.1 tgt drop, 3 MLE	42.06	38.25	28.72	23.80	38.15
+ Extended dataset	45.99	41.79	32.52	27.89	40.86
+ Averaging 8 checkpoints	47.90	44.13	36.19	29.05	43.29
+ Iterative decoding	48.75	45.46	37.09	30.19	44.27



- The final combination is used in both Low-Resource Track (w/o dataset oversampling) and Restricted Track

Low-Resource Track

- we use Wikipedia data processed by T2T WikiRevision problem

- Transformer clean big seemed to perform slightly better than big configuration

Unrestricted Track

- we finetune best model from Low-Resource Track on data from Restricted Track

Acknowledgements: The work described herein has been supported by OP VVV VI LINDAT/CLARIN project (CZ.02.1.01/0.0/0.0/16 013/0001781) and it has been supported and has been using language resources developed by the LINDAT/CLARIN project (LM2015071) of the Ministry of Education, Youth and Sports of the Czech Republic. This research was also partially supported by SVV project number 260 453 and GAUK 578218 of the Charles University.