Phrase-based Machine Translation is State-of-the-art for Automatic Grammatical Error Correction

Marcin Junczys-Dowmunt
junczys@amu.edu.pl

Roman Grundkiewicz
romang@amu.edu.pl

Adam Mickiewicz University in Poznań

We study parameter tuning towards the M^2 metric, the standard metric for automatic grammatical error correction (GEC) tasks.

We find that a bare-bones phrase-based SMT setup with task-specific parameter-tuning outperforms all previously published results for the CoNLL-2014 test set by a large margin using the same publicly available data:

Corpus	Sentences	Tokens
NUCLE	57.15 K	1.15 M
CoNLL-2013 Test Set	1.38 K	29.07 K
CoNLL-2014 Test Set	1.31 K	30.11 K
Lang-8	2.23 M	30.03 M
Wikipedia	213.08 M	3.37 G
CommonCrawl (u)	59.13 G	975.63 G
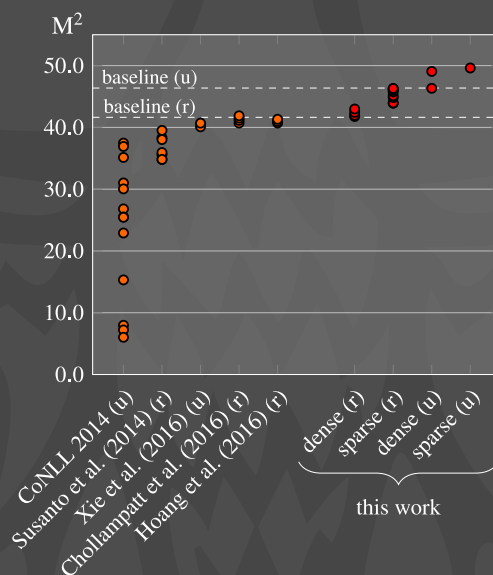
Scripts and example models required to reproduce the presented results are available from:

https://github.com/grammatical/baselines-emnlp2016
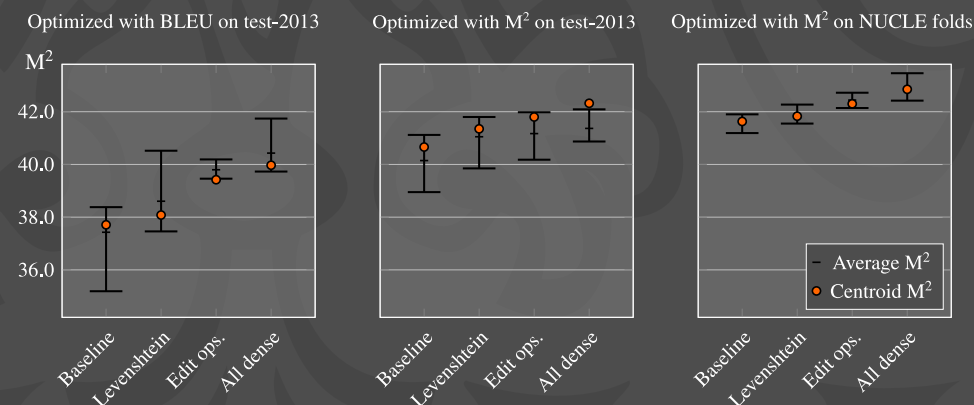
Dense features

Err: Then a new problem comes out .
Cor: Hence , a new problem surfaces .

Edit ops.: levenshtein=4, del=1, ins=1, subst=2

OSM: prob(TRAN_new_TO_new|TRAN_a_TO_a INS_, TRAN_hence_TO_then)

WCLM: prob(47|<s> 130 90 162 78 101)

Parameter tuning with dense features

Optimized with BLEU on test-2013 · Optimized with M^2 on test-2013 · Optimized with M^2 on NUCLE folds
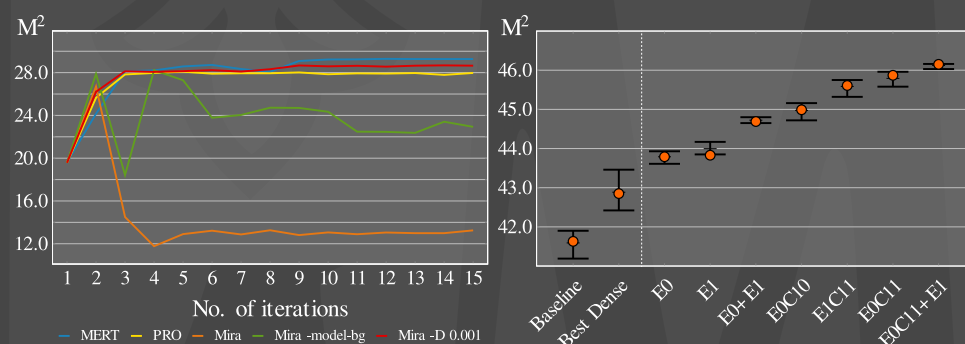
Sparse features

E0: subst(then,hence)=1 insert(,)=1
    subst(comes,surfaces)=1
    del(out)=1

E1: subst(130,130)=1 insert(90)=1
    subst(93,101)=1 del(165)=1

E0C11: <s>_subst(then,hence)_90=1
    130_ins(,)_90=1

Parameter tuning with sparse features

No. of iterations

MERT · PRO · Mira · Mira -model-bg · Mira -D 0.001

More corrected data

System	Prec.	Recall	M^2
R&R (np)	60.17	25.64	47.40
Best dense (np)	53.56	29.59	46.09
+CCLM	61.74	30.51	51.25
Best sparse (np)	58.57	27.11	47.54
+CCLM	63.52	30.49	52.21

Web-scale domain-selected language model

System	2014			2014-10		
	Prec.	Recall	M^2	Prec.	Recall	M^2
Baseline	48.97	26.03	41.63	69.29	31.35	55.78
+CCLM	58.91	25.05	46.37	77.17	29.38	58.23
Best dense	50.94	26.21	42.85	71.21	31.70	57.00
+CCLM	59.98	28.17	48.93	79.98	32.76	62.08
Best sparse	57.99	25.11	45.95	76.61	29.74	58.25
+CCLM	61.27	27.98	49.49	80.93	32.47	62.33