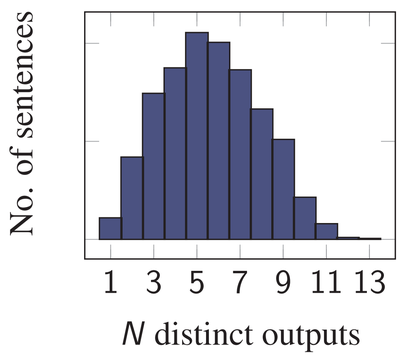
ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Human Evaluation of Grammatical Error Correction Systems

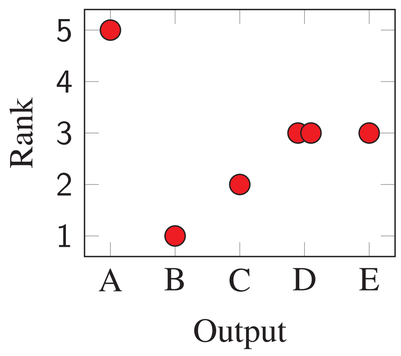
Roman Grundkiewicz, Marcin Junczys-Dowmunt, Edward Gillian
{romang,junczys}@amu.edu.pl, egillian@pwsz.pl

Judging system outputs from the CoNLL-2014 shared task

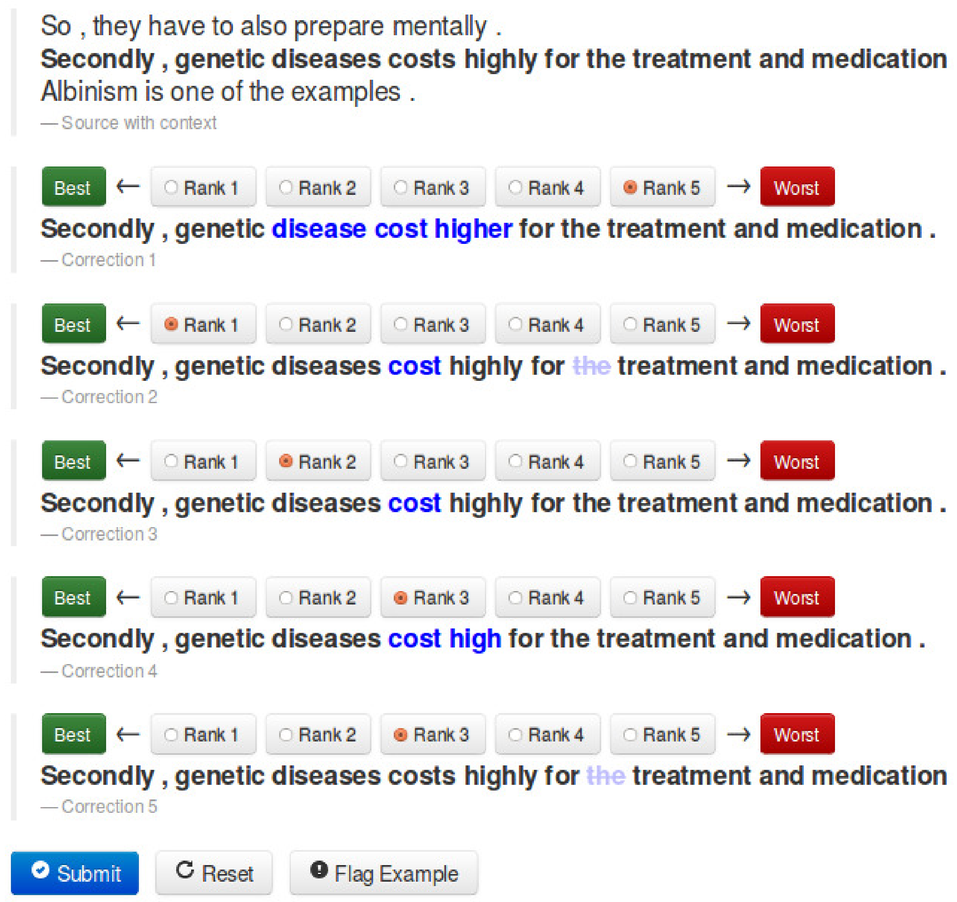
- We evaluate 13 system outputs from the CoNLL-2014 shared task on automatic grammatical error correction for English as a second language.
- The official test set contains 50 error-annotated essays (1,312 sentences).
- Our methods are based on the Workshop on Machine Translation (WMT) evaluation campaigns.
- Differences between GEC and MT (monolingual task, large overlap between system outputs, small number of changes w.r.t the input, etc.) require a number of adaptations.



Distribution of distinct outputs per test set sentence for 13 systems.



An example of overlapping rankings (Output D covers 2 systems).



Collected pairwise judgements

- The system outputs were ranked by 8 human judges, 2,319 collected rankings were expanded to 109,098 pairwise judgements of the form `A>B`, `A=B`, `A<B`
- Inter-annotator agreement: 0.29 (weak)
- Intra-annotator agreement: 0.46 (moderate)

Judge	Ranks	Unexpanded	Expanded
Total	2319	20516 (5694)	109098 (59117)

	1	2	3	4	5	6	7	8
1	.42	.26	.30	.37	.34	.26	.31	.24
2	–	.30	.25	.28	.23	.20	.10	.20
3	–	–	.50	.35	.44	.34	.46	.26
4	–	–	–	.34	.34	.30	.20	.26
5	–	–	–	–	.60	.36	.34	.32
6	–	–	–	–	–	.44	.35	.25
7	–	–	–	–	–	–	*	*
8	–	–	–	–	–	–	–	.48

Pairwise inter-annotator and intra-annotator agreement (Cohen's κ) per judge.

Computing ranks

The official CoNLL-2014 ranking is based on the MaxMatch (M^2) metric. A $F_{0.5}$ score computed from system outputs and gold standard annotations.

Our pairwise judgements were compiled into a single human-created ranking with the ExpectedWins-method: Scores reflect the probability that a given system will be ranked better than another randomly chosen system. The rank clusters group systems with overlapping rank values at $p \leq 0.05$.

#	System	P	$M^2_{0.5}$
1	CAMB	0.397	0.373
2	CUUI	0.417	0.367
3	AMU	0.416	0.350
4	POST	0.345	0.308
5	NTHU	0.350	0.299
6	RAC	0.331	0.266
7	UMC	0.312	0.253
8	PKU	0.322	0.253
9	SJTU	0.301	0.151
10	UFC	0.700	0.078
11	IPN	0.112	0.071
12	IITB	0.307	0.059
13	INPUT	0.000	0.000

Official CoNLL-2014 ranking.

#	Score	Range	System
1	0.628	1	AMU
2	0.566	2-3	RAC
	0.561	2-4	CAMB
	0.550	3-5	CUUI
	0.539	4-5	POST
3	0.513	6-8	UFC
	0.506	6-8	PKU
	0.495	7-9	UMC
	0.485	7-10	IITB
	0.463	10-11	SJTU
	0.456	9-12	INPUT
	0.437	11-12	NTHU
4	0.300	13	IPN

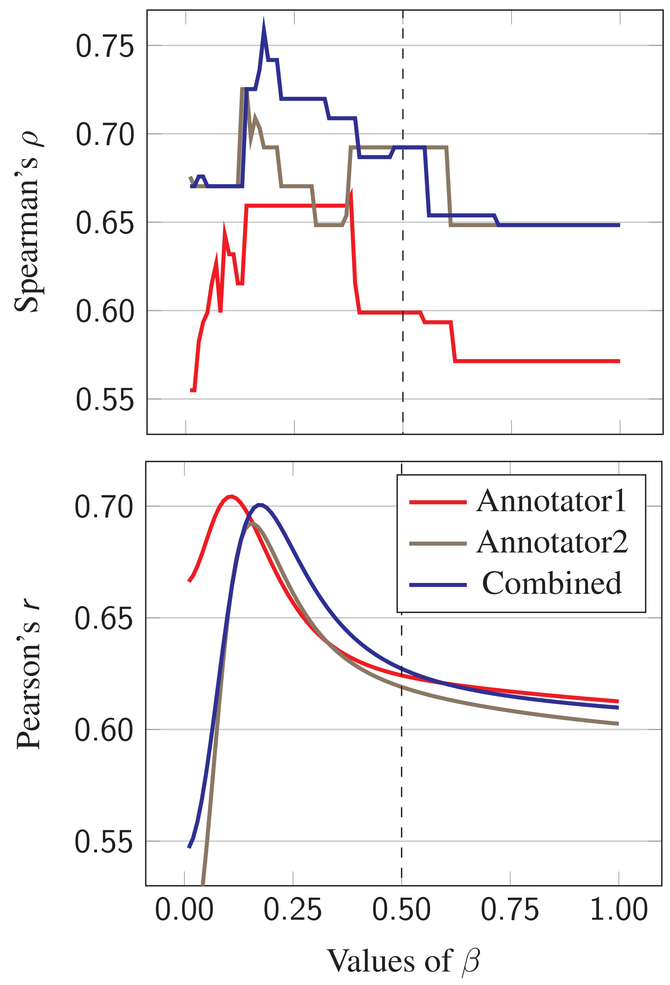
Human ExpectedWins ranking.

Correlation with GEC metrics

Based on the human ranking we could assess the quality of automatic evaluation metrics for GEC in terms of correlation with the collected human judgments.

Metric	Spearman's ρ	Pearson's r
M^2 $F_{1.0}$	0.648	0.610
M^2 $F_{0.5}$*	0.692	0.627
M^2 $F_{0.25}$	0.720	0.680
M^2 $F_{0.18}$	**0.758**	**0.701**
M^2 $F_{0.1}$	0.670	0.652
I-WAcc	-0.154	-0.098
BLEU	-0.346	-0.240
METEOR	-0.374	-0.241

Correlation results for various metrics and human ranking. Star (∗) marks official metric.



Spearman's ρ and Pearson's r correlation of M^2 with human judgment w.r.t. β. Dashed line marks official CoNLL-2014 choice $\beta = 0.5$.

	1	2	3	4	5	6	7	8	ρ	$\bar{\rho}$
1	–	.70	.31	.76	.74	.19	.62	.48	.70	
2	.72	–	.77	.84	.90	.57	.59	.64	.93	
3	.53	.89	–	.66	.70	.58	.42	.64	.63	
4	.82	.79	.69	–	.91	.42	.67	.54	.91	.72
5	.65	.85	.82	.87	–	.63	.63	.51	.93	
6	.32	.71	.67	.56	.86	–	.63	.39	.42	
7	.72	.74	.57	.76	.72	.63	–	.63	.76	
8	.64	.85	.86	.69	.72	.57	.75	–	.60	
r	.67	.93	.82	.87	.92	.66	.80	.82	–	
$\bar{r}$					.80					

Inter-annotator ranking correlation (Spearman's ρ above the diagonal, Pearson's r below).

Obtaining the data

The presented data and tools are available from:
`https://github.com/grammatical/evaluation`

ul. H. Wieniawskiego 1, Collegium Minus, 61-712 Poznań, Poland
tel. +48 61 829 00 00, fax +48 61 829 00 00
rectorof@amu.edu.pl