Taking the Correction Difficulty into Account in Grammatical Error Correction Evaluation

Takumi Gotou[†], Ryo Nagata[†‡], Masato Mita[‡♠], Kazuaki Hanawa[‡♠]
[†]Konan University [‡]RIKEN AIP [♠]Tohoku University

1. Purpose

Original sentence: It has been more easy
Correct sentence: It was easier
(difficult: has been → was; easy: more easy → easier)

There're easy and difficult errors to correct.

I want to give incentive to tackling more difficult errors.

2. Basic Idea

Error correction difficulty based on correction success rate:

$$w = 1 - \text{correction success rate}$$

Example: With two different systems

Output of system1: It was easier

Correct sentence: It was easier
Error correction difficulty (w): It 0, was 1/2, easier 0

Output of system2: It had been easier

was: $1 - \underbrace{\frac{1}{2}}_{\text{correction success rate}} = \underbrace{\frac{1}{2}}_{w}$

easier: $1 - \underbrace{\frac{2}{2}}_{\text{correction success rate}} = \underbrace{0}_{w}$

3. Performance Measures

$$\text{Weighted Recall} = \frac{\sum_{i \in E} w_i l_i}{\sum_{i \in E} w_i} \qquad \text{Weighted Precision} = \frac{\sum_{i \in E} w_i l_i}{\sum_{i \in C} w_i}$$

$l = 1$ (succeed) or 0 (failed)
C = Set of indices of tokens aligned to erroneous tokens
E = Set of indices of tokens to which error correction is applied

4. Experiments and Discussion

Evaluation in Difficulty-Weighted $F_{0.5}$, and Conventional $F_{0.5}$ (M² scorer)

Difficulty-weighted $F_{0.5}$

CoNLL-2013		CoNLL-2014		KJ		ICNALE	
Transformer	18.68	Transformer	15.17	Transformer	18.33	Transformer	18.17
SMT	15.16	SMT	13.66	CNN	17.39	LSTM	15.16
CNN	12.32	LSTM	11.01	LSTM	16.88	CNN	14.56
LSTM	11.94	CNN	9.75	SMT	8.51	SMT	12.88

Conventional $F_{0.5}$

CoNLL-2013		CoNLL-2014		KJ		ICNALE	
Transformer	36.20	Transformer	48.62	LSTM	45.64	LSTM	43.02
LSTM	33.76	LSTM	48.57	CNN	45.40	CNN	40.78
CNN	33.67	SMT	46.80	Transformer	42.80	Transformer	37.72
SMT	32.30	CNN	46.16	SMT	32.04	SMT	32.91

Examples of Difficulty Heat Map

This had caused panic among the people who had flooded the local police department with ...

Personally I am study in oversea , busy study life keeps me away from contact my old friend.

(Both excerpted from CoNLL-2014.)

ERRANT's Error Types Sorted by Difficulty Weights

easy → difficult

Error type	Average	SD
ADJ	0.982	0.074
VERB	0.891	0.254
VERB:TENSE	0.876	0.213
DET	0.747	0.292
VERB:FORM	0.590	0.393
NOUN:NUM	0.539	0.340
SPELL	0.533	0.342

(CoNLL-2014, eight systems.)

5. Conclusions

- Performance measures that consider correction difficulty
- The measures give incentive to tackling more difficult errors
- Stable ranking in the cross-corpora evaluation

(Scorer and difficulty weight data are available on the web !)