Automatic Metric Validation for Grammatical Error Correction

Leshem Choshen and Omri Abend

Hebrew University Jerusalem Israel

17 July 2018

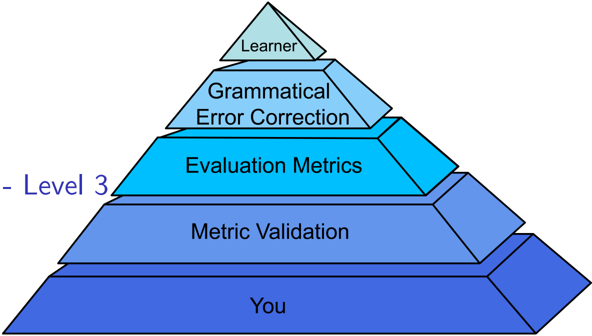
Meta view

The task - Level 1

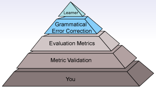
Evaluation - Level 2

Evaluation of evaluation - Level 3

Peers - Level 4



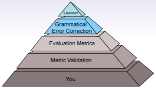
the task



- Input: a text which is perhaps ungramatical
- Output: a grammatical text saying the same meaning/content.

Example: However , there are both sides of stories

The task



- Input: a text which is perhaps ~~ungramatical~~ ungrammatical
- Output: a grammatical text ~~saying~~ conveying the same meaning/content.

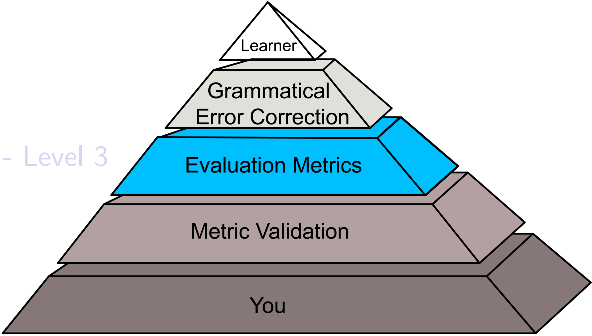
Example: However , there are ~~both sides of stories~~ →
However , there are two sides to the story.

The task - Level 1

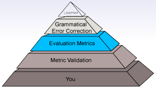
Evaluation - Level 2

Evaluation of evaluation - Level 3

Peers - Level 4



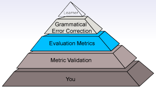
Test Set



- Learner sentences (perhaps ungrammatical)
- References - word edits and the error type corrected by them

Since ancient times , human interact with others face by face . $\rightarrow$
Since ancient times , ~~human~~ humans (Noun number) interact with others face ~~by~~ to (Wrong Preposition) face .

Metrics



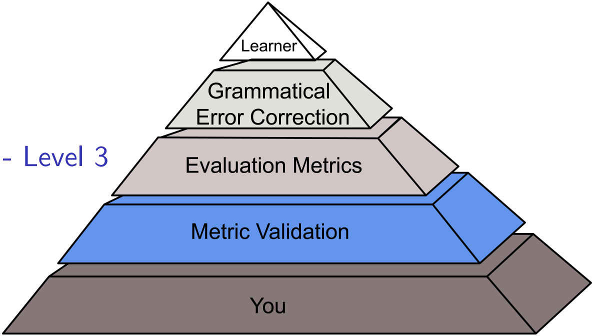
There are many suggestions for evaluation metrics:
M^2, GLEU, I-measure, LT, etc.
More on that in the paper.

The task - Level 1

Evaluation - Level 2

Evaluation of evaluation - Level 3

Peers - Level 4

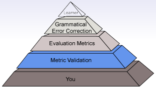


Human Rankings

	Sentence
1	You have become powerful, I sense the dark side in you.
2	**Powerful** you have become, I sense the dark side in you.
2	You have become powerful, the dark side **I sense** in you.
3	**Powerful** you have become, the dark side **I sense** in you.

Existing Metric Validation

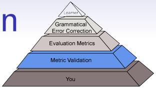
Human Rankings



- Annotation – Humans rank system corrections
  - Two benchmarks – GJG15 (Grundkiewicz et al. 2015), and NSPT15 (Napoles et al. 2015).
- Score – correlation between metric and human rankings
  - Rank each system by the metric scores of its outputs
  - Rank each system by the human ranks of its outputs
    - Methodologically troublesome
  - Correlate the two

Human Rankings - not a perfect solution

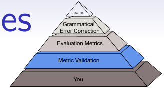
What Machine Translation has already found



- Costly
- Low agreement
  - Ranking is hard (correcting is easy)
  - Some sentences are uncomparable
- Not detailed
- ...

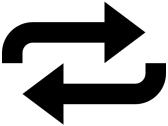
	Combined		GJG15		NSPT15	
	ρ	P-val	ρ	Rank	ρ	Rank
GLEU	0.771	0.001	0.512	1	0.758	1
LT	0.692	0.006	0.358	4	0.615	3
M^2	0.626	0.017	0.398	3	0.703	2
BLEU	0.143	0.626	0.455	2	-0.126	6

Human Rankings (CHR) - inherent biases



The vicious loop

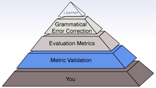
1. Metrics are favored if they discern high-performing and low-performing **existing** systems
2. Systems are fitted against metrics



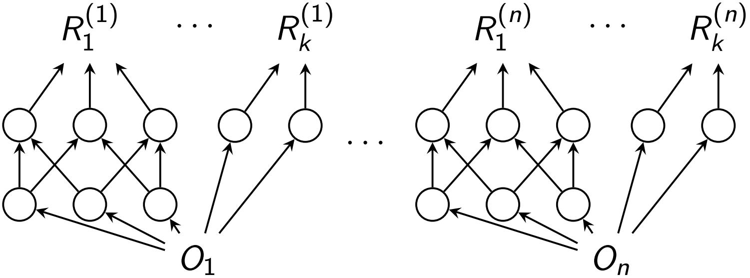
- Problematic:
  - Systems have similar biases – under-correct & favor correcting specific error types (Choshen & Abend 2018)
  - Metrics are evaluated based on distribution of errors in outputs, rather than true distribution

MAEGE

Methodology for Automatic Evaluation of GEC Evaluation



- Annotation – Humans correct errors in sentences
  - Widely available – regular GEC corpora
- Lattice – graded quality
  - Original sentences O_i
  - Partial corrections, apply some edits
  - Reference sentences $R_i^{(j)}$

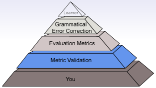


Human Rankings

Since ancient times , ~~human~~ humans (Noun number) interact with others face ~~by~~ to (Wrong Preposition) face .

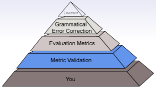
Corrections	Sentence
2	Since ancient times , humans interact with others face to face .
1	Since ancient times , human interact with others face to face .
0	Since ancient times , human interact with others face by face .

Corpus Level



- Models – Set of randomly chosen corrections
- Model's score
  - MAEGE score – the expected number of applied edits
  - We sample models from the lattices with different distributions
- Score – correlation between the two rankings
- Interesting results
  - Positive low correlation with CHR
  - The best metric is LT (number of detected errors)
  - With precision-oriented models MAEGE is similar to CHR
    - Indication that CHR is biased due to precision-oriented models

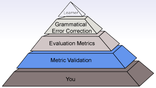
Types



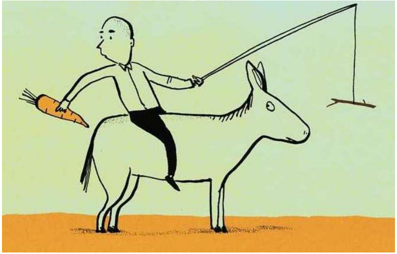
1. Pick sentence pairs with one correction difference
2. Find Δ: the change in metric score
3. Compute average Δ per type

Types - sensitivity analysis

Surprising results



1. All metrics penalize for validly correcting certain error types
2. Some error types (close class) are more commonly penalized than others (open class)

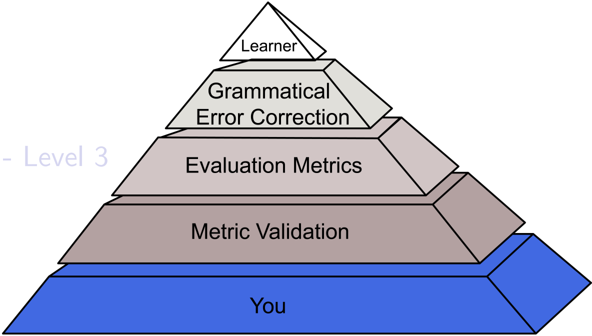


The task - Level 1

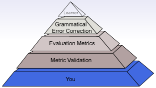
Evaluation - Level 2

Evaluation of evaluation - Level 3

Peers - Level 4

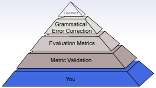


Take-home message



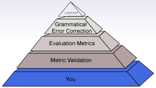
- Metrics emphasize some aspects of the task over others.
  - Metric validation should tell you which
  - If validation is opaque, metrics and systems may tune towards one another (vicious loop)
- MAEGE breaks the loop by not relying on system outputs
- Instead compile naturally ranked corpus

Take-home message



- Metrics emphasize some aspects of the task over others.
- MAEGE breaks the loop by not relying on system outputs
- Instead compile naturally ranked corpus
- Use MAEGE

Take-home message



- Metrics emphasize some aspects of the task over others.
- MAEGE breaks the loop by not relying on system outputs
- Instead compile naturally ranked corpus
- Use MAEGE

UCCA Semantic Parsing shared task
SemEval 2019