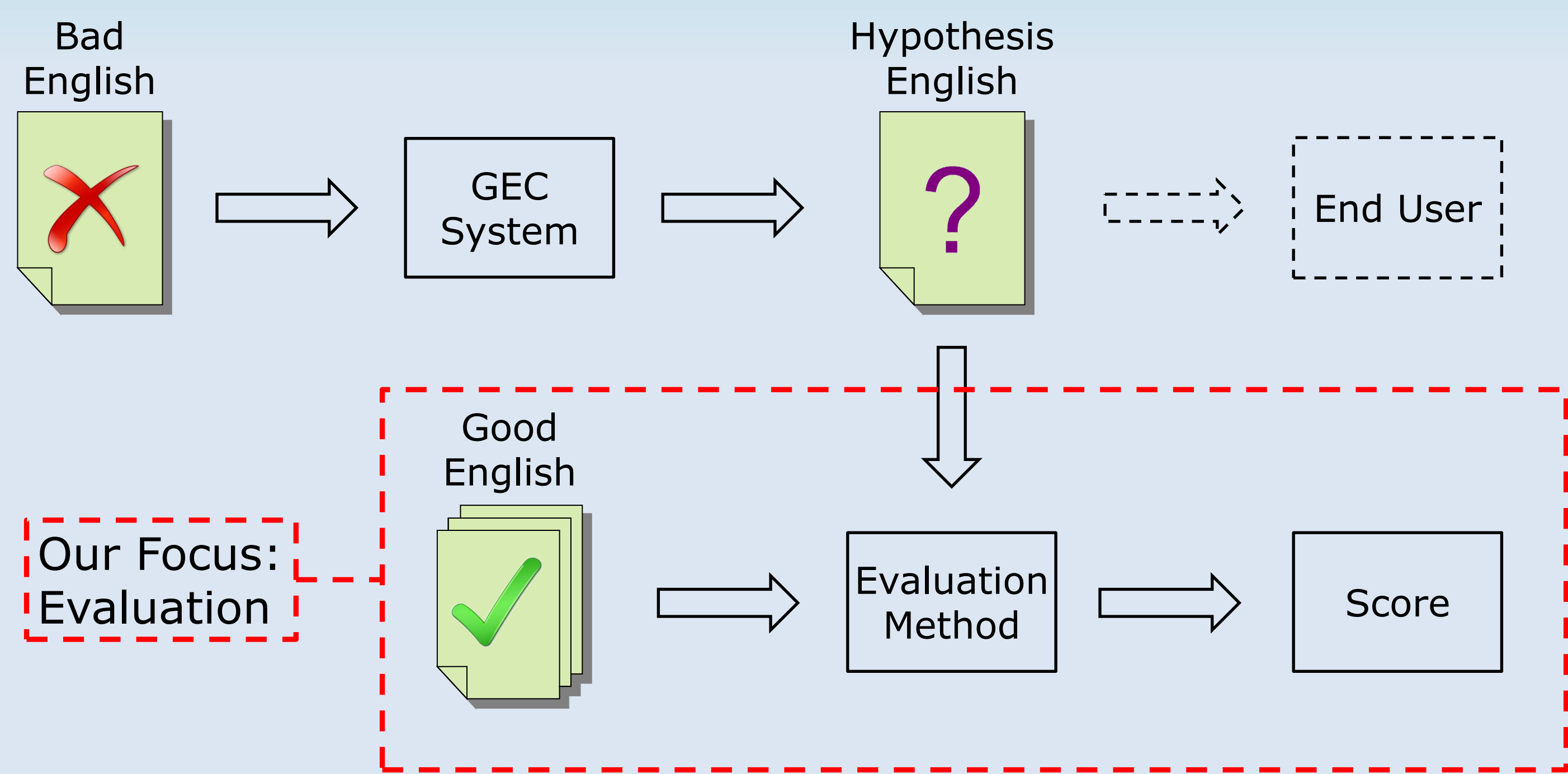


Christopher Bryant and Hwee Tou Ng

{bryant,nght}@comp.nus.edu.sg

Context



Problem:
What is "Good English"? There is often more than one way to make a correction...

Human Variation

Original	Social media has been playing a vital important role in our lives today .
A1	Social media plays an important role in our lives today .
A2	Social media plays a vital role in our lives today .
A3	Social media play a vitally important role in our lives today .
A4	Social media plays a vital role in our lives today .
A5	Social media plays a vital and important role in our lives today .
A6	Social media plays a vitally important role in our lives today .
A7	Social media has been playing a vital important role in our lives today .
A8	Social media plays a vital , important role in our lives today .
A9	Social media is playing a vital important role in our lives today .
A10	Social media has been playing a vital role in our lives today .

Objectives

1. How many annotators do we need in the gold standard?
2. How well do human corrected texts score against each other?
3. Do annotators agree on some error types more than others?

Data

- 50 essays: 25 non-native speakers (2 essay topics) ~600 words each
- 10 annotators: 2 from CoNLL-2014
1 is the first author
7 recruited via online recruitment agency Elance
- All edits classified into one of the 28 error categories used in CoNLL-2014.
- *Freely available:* http://www.comp.nus.edu.sg/~nlp/sw/10gec_annotations.zip

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Total
ArtOrDet	879	639	443	503	665	620	331	358	390	624	5452
Mec	227	376	493	325	411	336	228	733	598	780	4507
Prep	755	488	390	421	502	556	211	276	362	459	4420
Wci	623	476	479	446	456	595	340	250	212	346	4223
Nn	404	290	228	264	360	300	215	254	277	365	2957
...
Total	5560	3982	3317	3528	4016	3959	2391	3286	3397	4231	37667

The top 5 most common error categories and their counts, along with the total number of edits for each annotator.

Methodology

- Use the M2 scorer (Dahlmeier and Ng, 2012) as the evaluation method.
- Use the system output from the top 3 teams in CoNLL-2014 as hypotheses.
- Hypotheses are scored on a sentence-by-sentence basis.
- Output score is $F_{0.5}$, which prioritises Precision over Recall.
- We calculated the average score for:
 - any human vs. a *specific* combination of gold standard annotators (Eq. 1).
 - any human vs. *any* combination of i gold standard annotators (Eq. 2).
 - a system vs. *any* combination of i gold standard annotators (Eq. 3).

$$\text{Equation 1} \quad g(X) = \frac{1}{|A| - |X|} \sum_{a \in A \setminus X} f(a, X)$$

$$\text{Equation 2} \quad h_i = \frac{1}{\binom{|A|}{|X|}} \sum_{X: |X|=i} g(X)$$

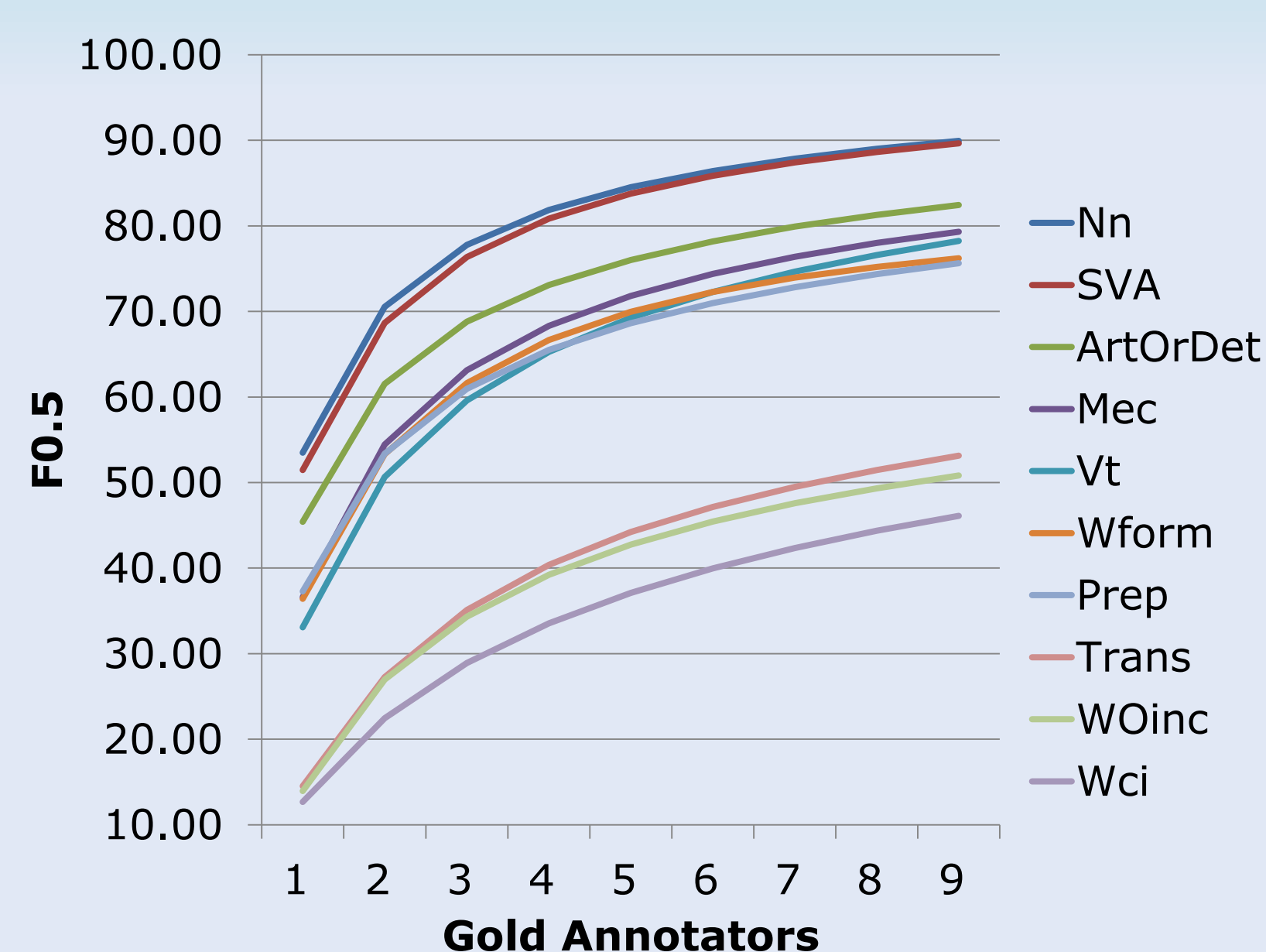
$$\text{Equation 3} \quad s_i = \frac{1}{\binom{|A|}{|X|}} \sum_{X: |X|=i} f(s, X)$$

- A is the set of all gold standard annotators.
- X is a proper, non-empty subset of A .
- $f(a, X)$ is the calculation performed by the M2 Scorer to evaluate annotator a against gold standard combination X .
- $\binom{|A|}{|X|}$ is the binomial coefficient for $|A|$ choose $|X|$.
- $f(s, X)$ is the calculation performed by the M2 Scorer to evaluate system s against gold standard combination X .

Results

CAMB	P	R	$F_{0.5}$
A1	39.64	14.06	29.06
A2	35.73	17.35	29.48
A3	35.22	20.29	30.70
A4	32.69	17.88	28.04
A5	35.74	17.26	29.43
A6	35.76	17.73	29.72
A7	24.96	19.62	23.67
A8	29.17	16.92	25.48
A9	32.03	18.28	27.84
A10	35.52	16.26	28.72

The CAMB system vs. each individual annotator



Error category scores for select categories as the number of gold standard annotators increases

Gold Annotators (i)	Human	AMU		CAMB		CUUI	
	$F_{0.5}$	$F_{0.5}$	Ratio	$F_{0.5}$	Ratio	$F_{0.5}$	Ratio
1	45.91	24.20	52.71%	28.22	61.46%	26.76	58.29%
2	56.68	33.47	59.05%	37.77	66.64%	36.04	63.59%
3	61.83	38.35	62.03%	42.68	69.03%	40.76	65.92%
4	65.05	41.53	63.85%	45.87	70.51%	43.77	67.29%
5	67.33	43.84	65.11%	48.17	71.54%	45.94	68.23%
6	69.07	45.62	66.06%	49.93	72.29%	47.60	68.92%
7	70.45	47.06	66.80%	51.34	72.87%	48.94	69.46%
8	71.60	48.26	67.40%	52.50	73.32%	50.05	69.89%
9	72.58	49.28	67.90%	53.47	73.67%	50.99	70.25%

Human shows the average $F_{0.5}$ performance for any one human vs. increasing numbers of other humans. AMU, CAMB and CUUI show the same but for their respective systems. Ratio scores show system performance as a percentage fraction of equivalent human performance.

Discussion

- CAMB's system score varied by as much as 15% Precision, 6% Recall or 7% $F_{0.5}$.
- Error categories involving a more restricted type of edit (e.g., the addition or removal of an $-s$ suffix on a noun (Nn)) score much higher than error categories where there are many more possible corrections (e.g., word choice errors (Wci)).
- Scores increase diminishingly as the number of gold annotators also increases.
- The best system, CAMB, is able to perform 73.67% as well as a human; this information is not apparent if we just look at the 53.47% $F_{0.5}$ score.
- Additional experiments showed that similar results could be obtained from a smaller dataset of only 10 essays (~6000 words).

Conclusion

- The first large scale annotation of all error types by multiple annotators in GEC.
- If even humans versus other humans are unable to score 100% $F_{0.5}$, it is unreasonable to expect machines to be able to do the same.
- Ratio scoring is a more informative way of evaluating system performance as a function of human performance.
- Annotators agree more on error categories that have smaller confusion sets.

References

Daniel Dahlmeier and Hwee Tou Ng, 2012. Better evaluation for grammatical error correction. In HLT-NAACL, pages 568–572.