The BEA-2019 Shared Task on Grammatical Error Correction

Christopher Bryant, Mariano Felice,
Øistein E. Andersen and Ted Briscoe

ALTA Institute
Computer Laboratory
UNIVERSITY OF CAMBRIDGE

The Task

- Correct all error types in tokenised sentences:

- Input:
  - Travel by bus is exspensive , bored and annoying .

- Output:
  - Travelling by bus is expensive , boring and annoying .

Goals

- Re-evaluate the field 5 years after CoNLL-2014
  - A lot has changed; e.g. approaches/datasets
- Introduce more diverse training/test data
  - Too much reliance on CoNLL-2014
- Carry out more detailed evaluation
  - Correction, Detection, Error Types, …

Data

- New corpus: Write & Improve + LOCNESS
  - 45k sentences, 800k tokens
  - Non-native and native student essays
  - Balanced across levels (A,B,C,N) by sentences
  - 5 references in test set
  - No native training data

- Other corpora: FCE, Lang-8, NUCLE
  - Standardised with ERRANT

Evaluation

- ERRANT scorer
  - Automatically extract edits from system output
  - Compare system edits and reference edits
  - Precision, Recall, F0.5
  - Blind evaluation on Codalab competition platform

Original	`I often look at TV`	
Reference	`[2, 4, watch]`	
Hypothesis 1	`[2, 4, watch]`	Match
Hypothesis 2	`[2, 4, see]`	No match
Hypothesis 3	`[2, 3, watch]`	No match

Tracks

1. Restricted
   - FCE, Lang-8, NUCLE, W&I+L learner data only
2. Unrestricted
   - Private and commercial resources allowed
3. Low Resource
   - W&I+L Dev set learner data only

- Public/free/artificial resources allowed in all tracks

Participants

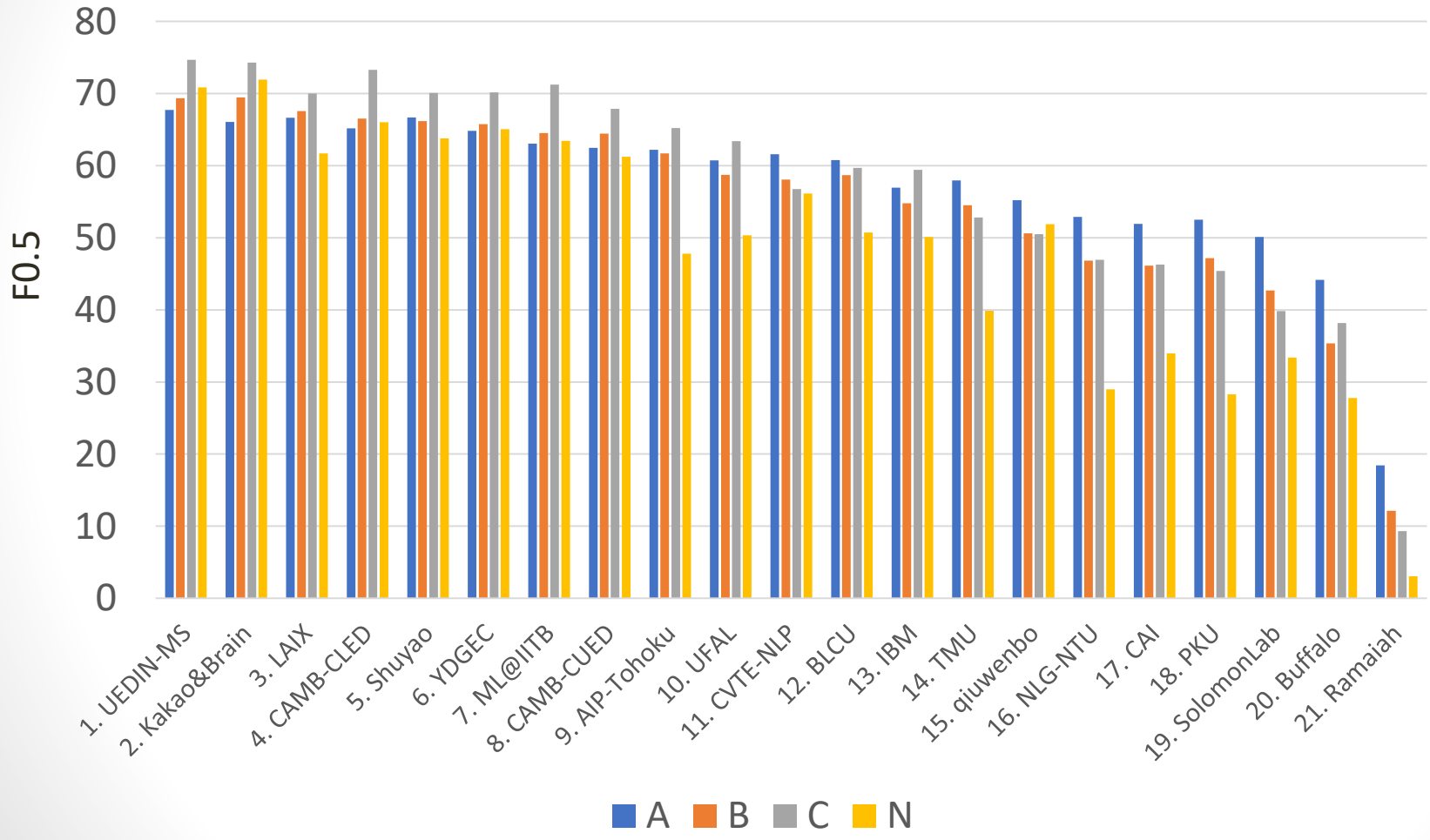
- 24 unique teams across all tracks
  - 21 Restricted, 7 Unrestricted, 9 Low Resource
  - 14 system papers, 4 email descriptions

- Approaches
  - Two-thirds of all teams used transformer NMT
  - Most of the remainder used CNNs
  - Differences in terms of: artificial data, oversampled data, ensembling, re-ranking, custom components

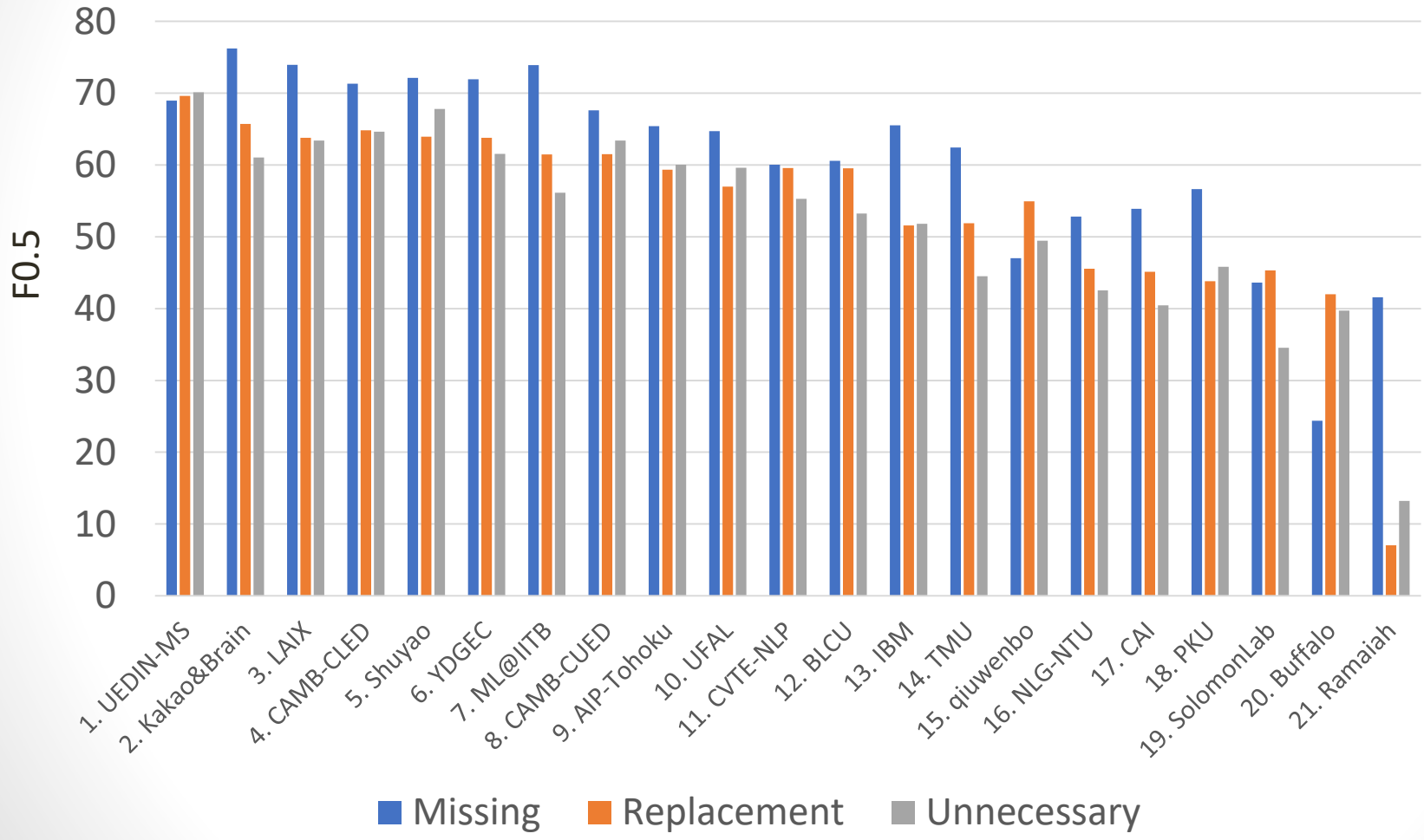
Restricted - Overall

Group	Rank	Teams	TP	FP	FN	P	R	F0.5
1	1	UEDIN-MS	3127	1199	**2074**	72.28	60.12	**69.47**
	2	Kakao&Brain	2709	894	2510	**75.19**	51.91	69.00
2	3	LAIX	2618	960	2671	73.17	49.50	66.78
	4	CAMB-CLED	2924	1224	2386	70.49	55.07	66.75
	5	Shuyao	2926	1244	2357	70.17	55.39	66.61
	6	YDGEC	2815	1205	2487	70.02	53.09	65.83
3	7	ML@IITB	**3678**	1920	2340	65.70	**61.12**	64.73
	8	CAMB-CUED	2929	1459	2502	66.75	53.93	63.72
4	9	AIP-Tohoku	1972	902	2705	68.62	42.16	60.97
	10	UFAL	1941	942	2867	67.33	40.37	59.39
	11	CVTE-NLP	1739	811	2744	68.20	38.79	59.22
5	12	BLCU	2554	1646	2432	60.81	51.22	58.62
6	13	IBM	1819	1044	3047	63.53	37.38	55.74
7	14	TMU	2720	2325	2546	53.91	51.65	53.45
	15	qiuwenbo	1428	854	2968	62.58	32.48	52.80
8	16	NLG-NTU	1833	1873	2939	49.46	38.41	46.77
	17	CAI	2002	2168	2759	48.01	42.05	46.69
	18	PKU	1401	1265	2955	52.55	32.16	46.64
9	19	SolomonLab	1760	2161	2678	44.89	39.66	43.73
10	20	Buffalo	604	**350**	3311	63.31	15.43	39.06
11	21	Ramaiah	829	7656	3516	9.77	19.08	10.83

Restricted - CEFR



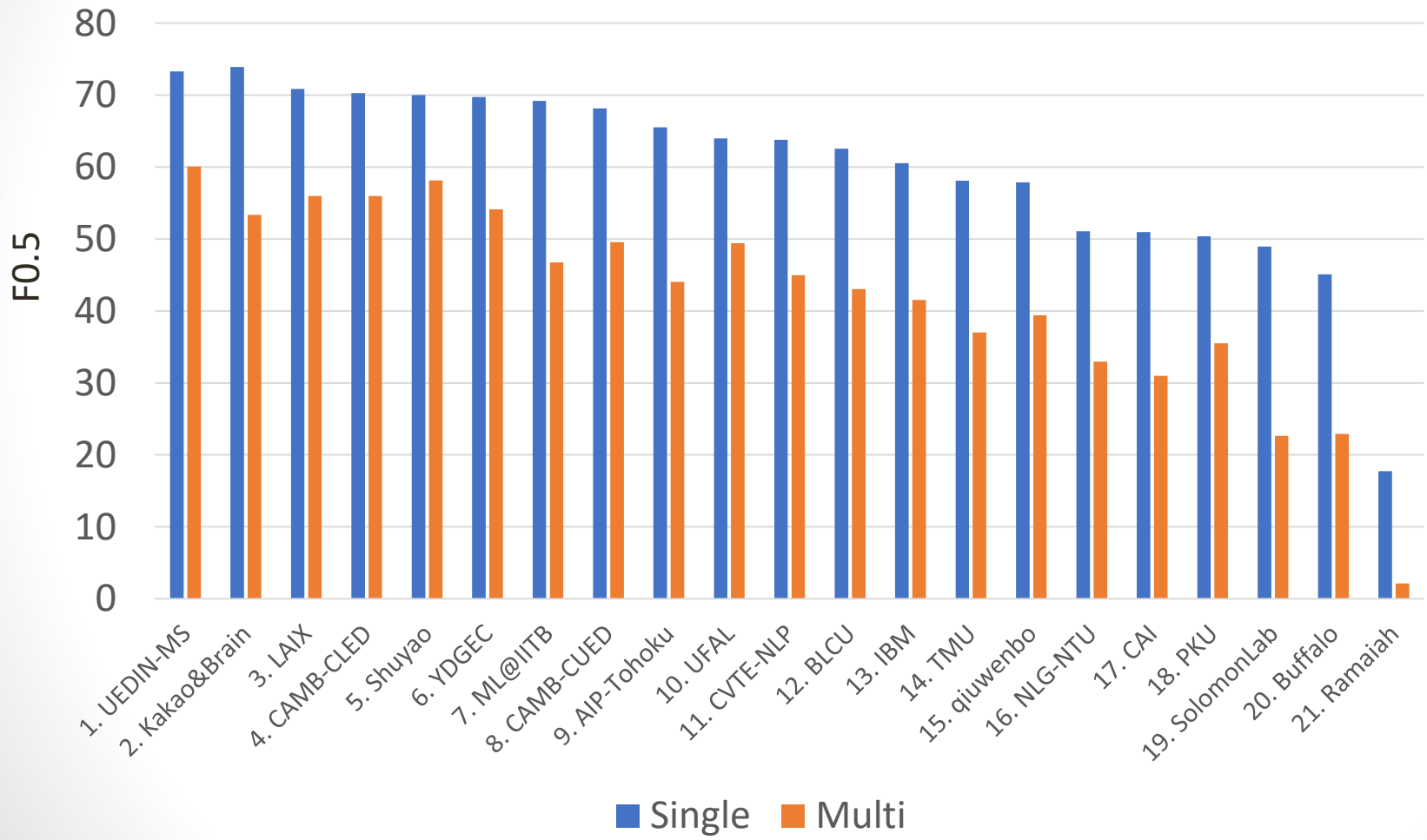
Restricted - Edit Operation



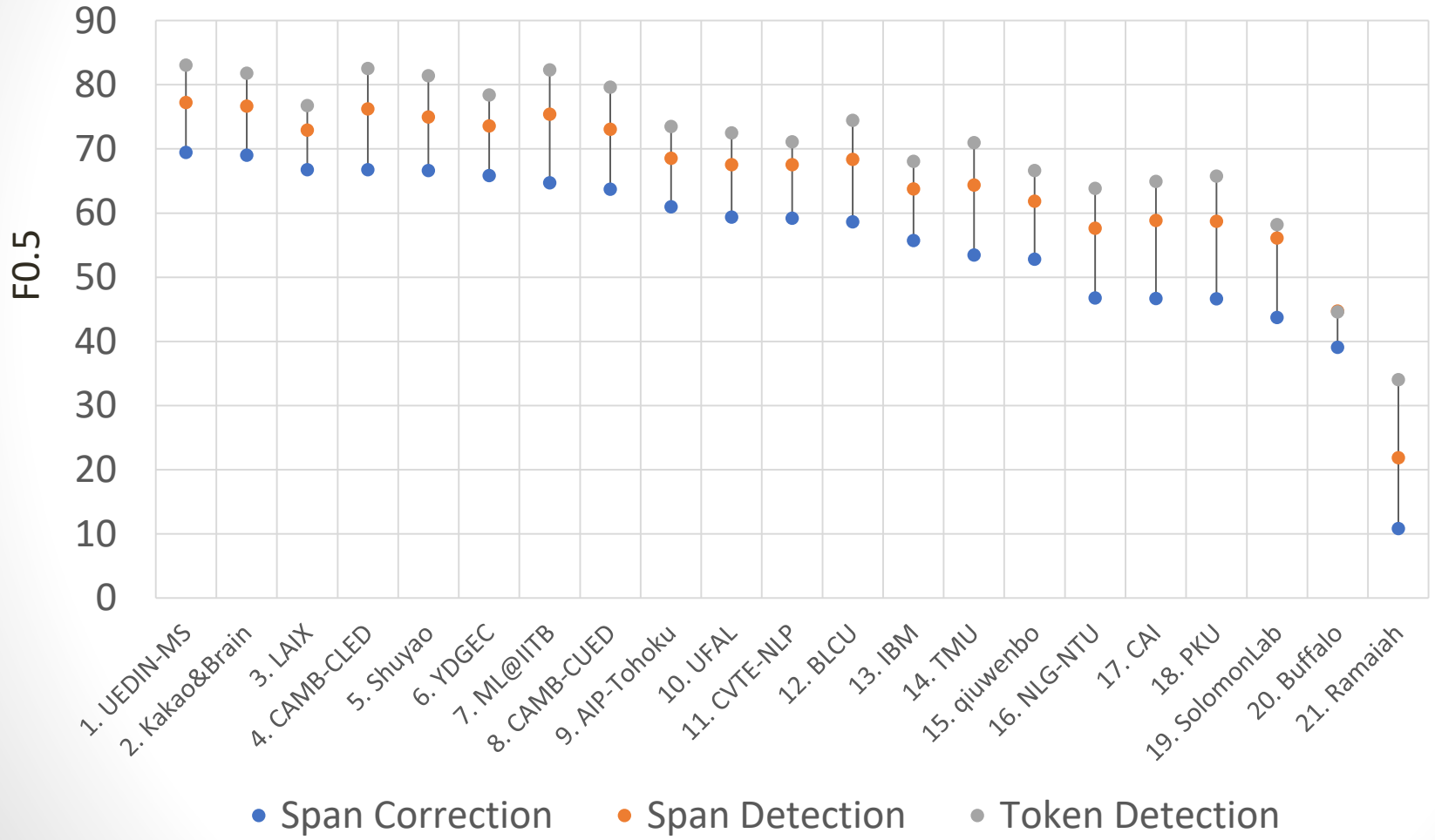
Restricted - Main Error Types

- Almost all teams attempted to correct all error types
- Best system scored highest in 15/24 types
- ADJ, ADV, CONJ, NOUN, OTHER, VERB most difficult
  - Top systems scored <50 F0.5
  - Content words constitute >10% of all edits
  - These types often require semantic understanding

Restricted - Single Vs. Multi



Restricted - Det. Vs. Corr.



Restricted - Other Metrics

	ERRANT				MaxMatch					
Teams	P	R	F0.5	#	F0.5	#	I	#	GLEU	#
UEDIN-MS	77.87	62.29	**69.47**	1	**76.48**	1	**38.92**	1	**77.93**	1
Kakao&Brain	**80.18**	53.28	69.00	2	74.09	2	36.84	2	75.87	5
LAIX	77.03	50.19	66.78	3	70.78	7	28.20	7	74.33	9
CAMB-CLED	74.59	56.53	66.75	4	72.51	3	34.10	3	76.62	3
Shuyao	74.41	56.31	66.61	5	72.22	4	33.22	4	76.55	4
YDGEC	74.50	54.49	65.83	6	71.60	6	29.21	6	75.39	7
ML@IITB	69.69	**63.29**	64.73	7	71.97	5	30.75	5	77.89	2
CAMB-CUED	71.49	55.63	63.72	8	70.37	8	26.37	8	75.82	6
AIP-Tohoku	72.79	43.05	60.97	9	65.95	9	19.22	9	73.16	11
UFAL	71.56	41.21	59.39	10	65.70	10	17.46	10	72.79	12
CVTE-NLP	72.12	39.12	59.22	11	63.04	12	16.71	11	72.51	13
BLCU	65.11	52.54	58.62	12	64.82	11	13.04	12	74.33	8
IBM	66.19	37.45	55.74	13	59.47	14	8.84	14	71.48	15
TMU	57.69	53.15	53.45	14	61.44	13	-0.54	17	73.96	10
qiuwenbo	66.56	32.84	52.80	15	57.70	15	8.94	13	71.30	16
LG-NTU	52.54	39.20	46.77	16	53.38	17	-1.45	18	71.13	17
CAI	51.49	42.61	46.69	17	53.68	16	-1.49	19	71.68	14
PKU	54.84	32.17	46.64	18	52.84	18	-0.32	15	71.06	18
SolomonLab	47.05	39.69	43.73	19	50.00	19	-3.50	20	70.56	19
Buffalo	65.09	15.08	39.06	20	40.95	20	-0.32	15	68.32	20
Ramaiah	10.29	19.04	10.83	21	18.68	21	-21.78	21	56.31	21

Unrestricted - Overview

- Expected higher scores than Restricted track
  - Highest scoring Unrestricted team submitted same system to Restricted track

Group	Rank	Teams	TP	FP	FN	P	R	F0.5	Restricted F0.5
1	1	LAIX	2618	960	2671	73.17	49.50	66.78	66.78
	2	AIP-Tohoku	2589	1078	2484	70.60	51.03	65.57	60.97
2	3	UFAL	2812	1313	2469	68.17	53.25	64.55	59.39
3	4	BLCU	3051	2007	2357	60.32	56.42	59.50	58.62
…									

- Other teams increased scores by ~1-5 F0.5
- Least popular track

Low Resource - Overview

- Impressive scores despite increased difficulty
  - UEDIN-MS outperformed 14 Restricted Track systems

Group	Rank	Teams	TP	FP	FN	P	R	F0.5	Restricted F0.5
1	1	UEDIN-MS	2312	982	2506	70.19	47.99	64.24	69.47
2	2	Kakao&Brain	2412	1413	2797	63.06	46.30	58.80	69.00
3	3	LAIX	1443	884	3175	62.01	31.25	51.81	66.78
	4	CAMB-CUED	1814	1450	2956	55.58	38.03	50.88	63.72
…									

- Quality of artificial data is key
- GEC for low-resource languages a real possibility

Conclusions

- Significant progress has been made in 5 years
- Systems generalised fairly well to a wider range of texts
- Artificial data is likely a key component of a neural system
- Content word/semantic errors are hard

Final Remarks

- More details in the shared task paper + appendix
  - https://www.cl.cam.ac.uk/research/nl/bea2019st/
- Codalab open for further submissions
- Codalab outage
- Questions?