# The Accuracy Evaluation Shared Task as a Retrospective Reproduction Study

Craig Thomson & Ehud Reiter
Department of Computing Science
University of Aberdeen, UK

## Data-to-text evaluation by annotation

- *Text is ~300 basketball game summary*
- *Three human annotators mark errors in text*
- *Gold Standard Mistake List (GSML)*
- *Proof of concept extended to shared task*
- *Goal:  Representative errors*

**Error Categories**

**NAME$^{N}$** – Named entities (people, etc)
**NUMBER$^{U}$** – Ordinal, cardinal, percent
**WORD$^{W}$** – Incorrect word or phrase
**CONTEXT$^{C}$** – Contextual error
**NOT_CHECABLE$^{X}$** – Too time consuming
**OTHER$^{O}$** – Last resort for nonsensical

## Example annotated text (detokenized)

The Atlanta Hawks defeated the New York Knicks, 142-139, at Phillips Arena on **Saturday** night.  The Hawks (28-20) learned earlier in the day that they can't disappoint this season, and the Hawks (28-**28**) were able to escape with a win.  Carmelo Anthony led the way for the **Hawks** (21-**20**) with a 45-point, **9**-rebound, four-assist **triple-double** that also included **four** steals, a **pair** of blocks, a **pair** of blocks and a block.  **Dwight Howard** posted a 19-point, 12-rebound double-double that also included an assist and a block.

## Dataset (90 texts) collected in 3 stages

- *Same annotators*
- *Same platform (MTurk)*
- *Same interface (MS Word => WebAnno)*
- *Same three systems (as ensemble)*

- *Different sample sizes:*
  - *21 initial texts (July 2020)*
  - *39 to complete train set (Jan 2021)*
  - *30 for shaed task test set (March 2021)*
- *Different sample from test-gens*

## Results

Table 1: Mean Error Count (MEC) for Ensemble

| experiment MEC | | | |
|---|---|---|---|
| **A** | **B** | **C** | **CV*** |
| 19.62 | 20.56 | 20.73 | 3.61 |

Table 2: Mean Error Count (MEC) for each type within the Ensemble

| | experiment MEC | | | |
|---|---|---|---|---|
| **error type** | **A** | **B** | **C** | **CV*** |
| NAME | 5.33 | 5.26 | 7.07 | 21.26 |
| NUMBER | 8.86 | 7.38 | 7.47 | 12.80 |
| WORD | 4.43 | 6.18 | 4.67 | 22.80 |
| CONTEXT | 0.76 | 0.90 | 0.27 | 63.22 |
| N-CHECK | 0.19 | 0.85 | 1.27 | 86.35 |
| OTHER | 0.05 | 0.00 | 0.00 | 211.73 |

Table 3: Mean Error Count (MEC) for each system

| | experiment MEC | | | |
|---|---|---|---|---|
| **system** | **A** | **B** | **C** | **CV*** |
| cond-copy | 21.57 | 25.54 | 26.60 | 13.19 |
| doc-plan | 21.86 | 17.77 | 18.90 | 13.23 |
| h-encoder | 15.43 | 18.38 | 16.70 | 10.77 |

## Complex annotations – exact reproduction of errors may not be the goal

*There can be multiple ways to annotation the same underlying problem.*

**Annotator A:**  The **only other** **Raptor** to reach double figures in points was **Dwayne** Dragic, who **came off the bench** for 22 points (**9**-**17** FG, 3-7 3Pt, 3-3 FT), **six** rebounds and five assists.

**Annotator B:**  The **only other** Raptor to reach double figures in points was **Dwayne Dragic**, who came off the bench for **22** points (9-**17** FG, **3**-7 3Pt, **3**-**3** FT), **six** rebounds and **five** assists.

*Annotator A thought the sentence should be about Goran Dragic, whilst B made annotations based on another player, Lou Williams.  Having exhaustive rules for annotation could improve agreement, but would take longer and stop the above problem showing up in error analysis.*