SHELL: Scoring Human-like Errors in Generated Language

Bryan Wilkinson, Ashwinkumar Ganesan & Tim Oates

Dept. Of Computer Science & Electrical Eng.,
University Of Maryland Baltimore (UMBC)

UMBC
AN HONORS UNIVERSITY IN MARYLAND

cognition | robotics | learning
CORAL

Assessing Machine Generated Text

Content Summarization

Grass pollen levels for Friday have increased from the moderate to high levels of yesterday with values of around 6 to 7 across most parts of the country. However, in Northern areas, pollen levels will be moderate with values of 4.[1]

↓ ↓

Pollen counts are expected to remain high at level 6 over most of Scotland, and even level 7 in the south east. The only relief is in the Northern Isles and far northeast of mainland Scotland with medium levels of pollen count.

Machine Translation

Bei der Begegnung soll es aber auch um den Konflikt mit den Palästinensern und die diskutierte Zwei-Staaten-Lösung gehen.

→

At the meeting, however, it is also a question of the conflict with the Palestinians and the two-state solution that is being discussed.

The meeting was also planned to cover the conflict with the Palestinians and the disputed two state solution.

Image Captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Quantitative Assessment Metrics

	BLEU	METEOR	RR	DA	HUME
Human Involvement	No	No	Yes	Yes	Yes
Alignment Based	Yes	Yes	No	No	No
Additional Comments	Widely Used	Widely Used	Primary metric for WMT 2016	Evaluate translation fluency & adequacy	Checks segments are semantically correct

What are Human-Like Errors (HLE)?

- Error made by native speakers, errors by non-native people and errors made by children that are learning a language for the first time.

```
Mrs. Moss said that it was the biggest number of dogs
she had ever come across.
```

biggest is a HLE, the correct word being ***largest***

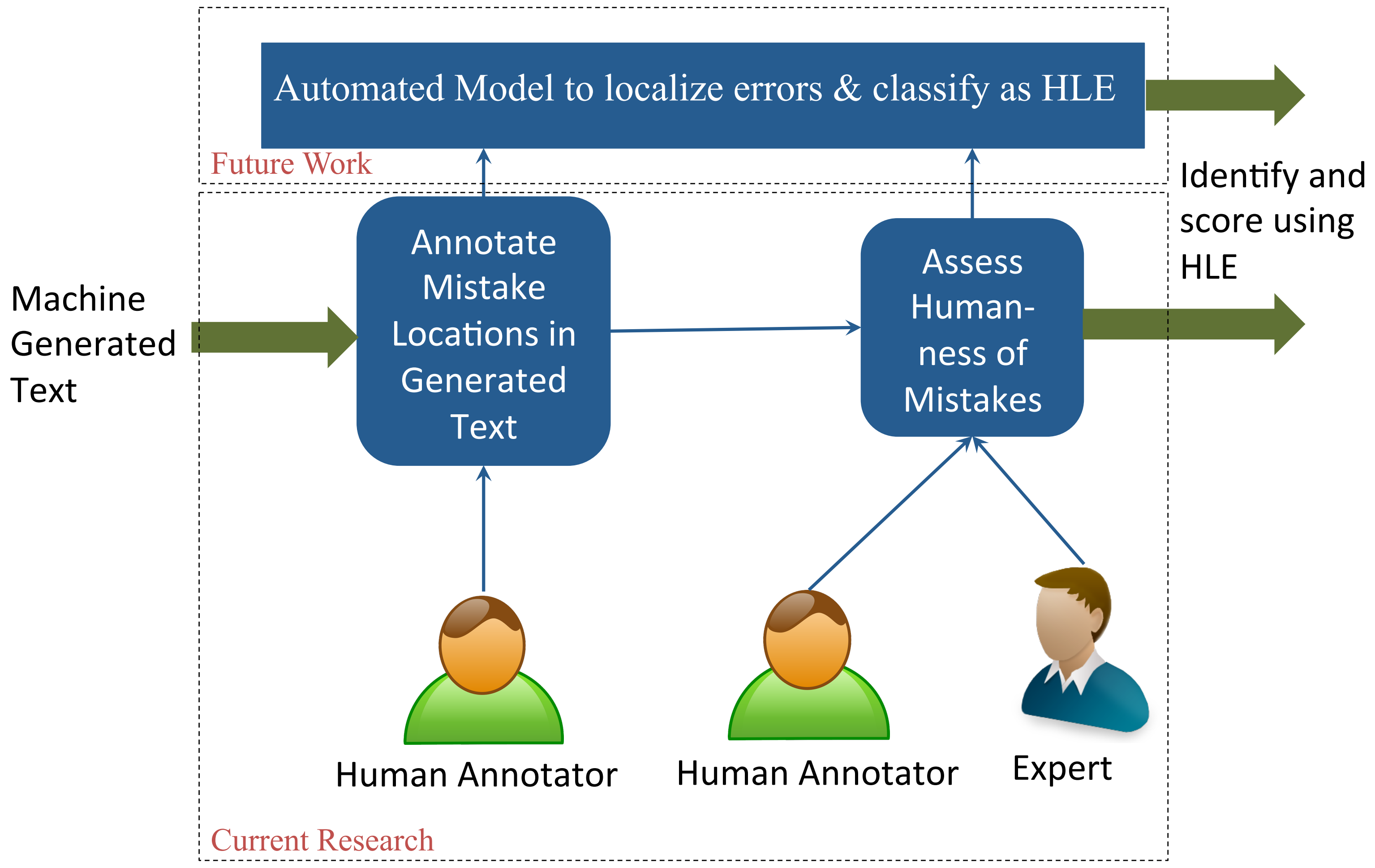
- We assume that HLEs are errors in generated text that are similar to ones made by non-native speakers.
- Misused forms, incorrect omissions, unnecessary words, misplaced words & confused words
- **Hypothesis: Human's can correct for HLEs as compared to *gibberish* words**

Annotation Method

- The annotation will be done in two phases
  - Annotators will identify which regions of the translation they believe are erroneous
  - Annotators will label aggregated error regions in a sentence as human or not human.
- Both tasks will be carried out on output from 4 MT systems

	German (WMT16)	Hindi (WMT14)
High BLEU	uedin-nmt-ensemble[4]	CMUHIEN[6]
Low BLEU	NeuralMT-BPE-IF[5]	dcu-hien-stem[7]

- As a control, annotators will also be asked to annotate sentences from NUCLE, corpus of error annotated English learner text.

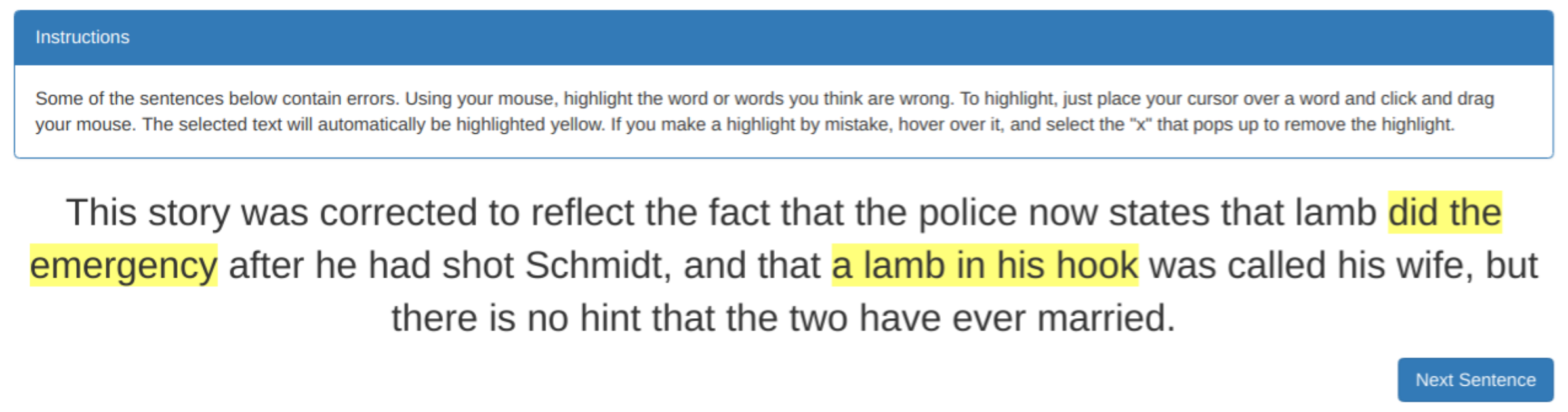


Automated Model to localize errors & classify as HLE
Future Work
Machine Generated Text
Annotate Mistake Locations in Generated Text
Assess Human-ness of Mistakes
Identify and score using HLE
Human Annotator
Human Annotator
Expert
Current Research

References

- [1] https://en.wikipedia.org/wiki/Natural_language_generation#Example
- [2] Callison-Burch, et al. 2012. Findings of the 2012 Workshop on Statistical Machine Translation
- [3] Styme & Ahrenberg. 2012 On the practice of error analysis for machine translation evaluation.
- [4] The Edinburgh/LMU Hierarchial Machine Translation System for WMT 2016
- [5] WMT 2016 - The University Of Melbourne, Australia
- [6] WMT 2014 – The CMU machine translation systems at wmt2014
- [7] WMT 2014 – The IIIT Hyderabad machine translation systems at wmt 2014

Annotation Interface

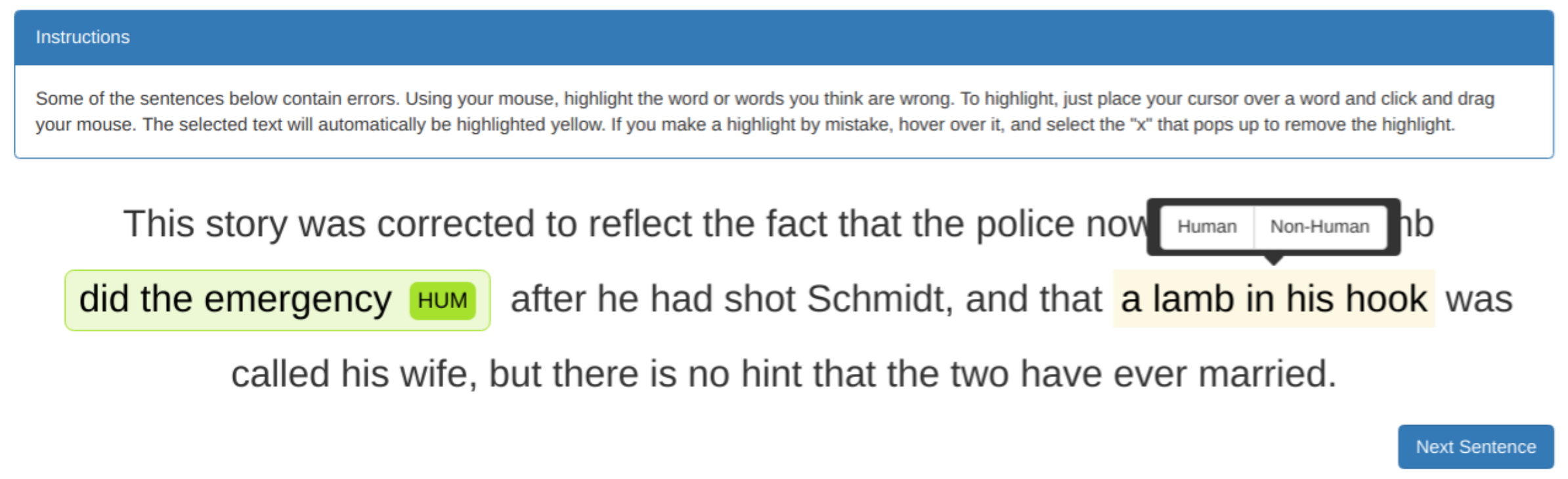


Instructions

Some of the sentences below contain errors. Using your mouse, highlight the word or words you think are wrong. To highlight, just place your cursor over a word and click and drag your mouse. The selected text will automatically be highlighted yellow. If you make a highlight by mistake, hover over it, and select the "x" that pops up to remove the highlight.

This story was corrected to reflect the fact that the police now states that lamb did the emergency after he had shot Schmidt, and that a lamb in his hook was called his wife, but there is no hint that the two have ever married.

Task A => Highlight parts of the sentence that are erroneous. This can be the *entire* statement too.



Instructions

Some of the sentences below contain errors. Using your mouse, highlight the word or words you think are wrong. To highlight, just place your cursor over a word and click and drag your mouse. The selected text will automatically be highlighted yellow. If you make a highlight by mistake, hover over it, and select the "x" that pops up to remove the highlight.

This story was corrected to reflect the fact that the police now nb did the emergency HUM after he had shot Schmidt, and that a lamb in his hook was called his wife, but there is no hint that the two have ever married.

Task B => For the highlighted parts of the sentence check if the error is likely an error a human might make while writing

Shell Score

$$\text{SHELL} = \frac{N_h}{N_h + N_g + \epsilon}$$

This story was corrected to reflect that the police now say that Lamb made the emergency call after he shot Schmidt, and that lamb in his emergency call HUM Prentiss was called HUM his wife, but there was no indication that the two were ever married.

SHELL ≈ 1.0

This story was corrected to reflect the fact that the police now states that lamb did the emergency HUM after he had shot Schmidt, and that a lamb in his hook NON was called his wife, but there is no hint that the two have ever married.

SHELL ≈ 0.5

Comparisons & Future Work

- **Quality Estimation**[2]

This story was corrected to reflect the fact that the police now states that lamb did the emergency after he had shot Schmidt, and that a lamb in his hook was called his wife, but there is no hint that the two have ever married.

Quality Estmation Score = 4

- **Error Detection**[3]

This story was corrected to reflect the fact that the police now states that lamb did RV the emergency MN after he had shot Schmidt, and that a lamb in his hook R was called his wife, but there is no hint that the two have ever married.

- Analyze if humans can correct for mistakes made while reading descriptions or during conversations
- Automatic detection of human like errors