[Automatically Neutralizing Subjective Bias in Text]

Focus on: inappropriate subjectivity (= subjective bias)

WPOV policy: - avoiding stating opinions as facts
- preferring nonjudgemental language

Aim: To debias text by suggesting edits that would make it more neutral.
(= extends the detection/classification problems into a generation task w/ otherwise similar meaning)

Bias
- ① Framing Bias
- ② Epistemological Bias
- ③ Demographic Bias

Dataset: WNC
- properties
  - text's topic ∝ realization of bias
  - the complexity of neutralizing text is typically reserved for more senior editors

Methods
- MODULAR : human-based
- CONCURRENT : simple

<Method - Modular> : better at reducing bias & has higher accuracy

① **Detection Module** The detection module is a neural sequence tagger that estimates p_i, the probability that each input word w_i^s is subjectively biased (Figure 2).

② **Editing Module** The editing module takes a subjective source sentence $\mathbf{s}$ and is trained to edit it into a more neutral compliment $\mathbf{t}$.

③ **Final System** Once the detection and editing modules have been pre-trained, we join them and fine-tune together as an end to end system for translating $\mathbf{s}$ into $\mathbf{t}$.

<Method- Concurrent> : fluent, preserves meaning better, higher BLEU

3.2 CONCURRENT (동시 발생의)

Our second algorithm takes the problematic source **s** and directly generates a neutralized $\hat{\mathbf{t}}$. While this renders the system easier to train and operate, it limits interpretability and controllability.

Distribution of Model Errors:

Error Type	Proportion (%)	Valid (%)
No change	38	0
Bad change	42	80
Disfluency	12	0
Noise	8	87

Most errors are due to the subtlety and complexity of language understanding required for bias neutralization, rather than the generation of fluent text.

Limitations

- single-word edits
- when a presupposition is in fact true and hence not subjective