4. Controllable Generation

Presenter: Yue

Errors from Hallucinations

Hallucination: *generate[d] text that is nonsensical, or inconsistent with the provided input*

- Growing body of literature -- Here: taxonomy from *Ji et al., 2022* (pdf)
- **Factuality:** Quality of a statement being true or based in a fact
- Variants of hallucinations:

generated text contradicts source text

vs.

generated text is not grounded in the source text

Errors from Hallucinations

Speaker	**Utterance**
	Why did Federer withdraw from the tournament?
	He injured his back in yesterday's match.
	Did he have any other injuries?
---	Did **Roger** Federer have any other injuries besides his **leg**?

Adapted from: Jin et al., *Hierarchical Context Tagging for Utterance Rewriting*, AAAI 2022 (pdf).

Causes of Hallucinations

1. **Divergence of source texts and references** in training data
2. **Memorized (factual) knowledge** in models with a really high parameter count (e.g., T5 11B)
3. In general, **model quality** issues

(from Ji et al., 2022 (pdf))

4-1. Mitigating hallucinations with restricted vocabularies

Advantages of Text Editing over Generation

Natural protections against hallucination

A. Partial reuse of input tokens

B. Insertion from a restricted + hotfixable vocabulary

C. Supplemental edit operations for critical cases

A) Partial Reuse of Input Tokens

- Any reused token is one token not hallucinated
- Holds for text-editing models with unrestricted vocabulary or a seq2seq+copy model
- Statistic from a model for Utterance Rewriting:
  - In 75%+ of cases, the last user utterance is rewritten w/o adding new terms.
  - This is a great metric to monitor and set alerts on, e.g. to monitor for negative impact of the natural query distribution shift over time.

b) Insertion from a Restricted + Hotfixable Vocabulary

Error type	LASERTAGGER	SEQ2SEQBERT	Example
Imaginary words	not affected	affected	In: … Zenica (Cyrillic: "Зеница") is … Out: … Zenica (Cyrillic: "**gratulation**еница") is …
Repeated phrases	not affected	affected	In: I'm your employee, to serve on your company. Out: I'm your **company**, to serve on your **company**.
Premature end-of-sentence	less affected	affected	In: By the way, my favorite football team is Manchester United, they … Out: By the way, my favorite football team is.
Hallucinations	less affected	affected	In: Tobacco smokers may also experience … Out: **anthropology** smokers may also experience …
Coreference issues	affected	affected	In: She is the daughter of Alistair Crane … who secretly built … Out: She is the daughter of Alistair Crane … ⟨::::⟩ **She** secretly built …
Misleading rephrasing	affected	affected	In: … postal service was in no way responsible … Out: … postal service **was responsible** …
Lazy sentence splitting	affected	not affected	In: Home world of the Marglotta located in the Sagittarius Arm. Out: Home world of the Marglotta . ⟨::::⟩ Located in the Sagittarius Arm.

Table 7: Main error patterns observed in the output of the tagging and seq2seq models on their test sets (all tasks).

Malmi et al. *Encode, Tag, Realize: High-Precision Text Editing*. EMNLP 2019 (pdf)

b) Insertion from a Restricted + Hotfixable Vocabulary

- Some Text Editing models have restricted vocabularies

  → Easy to remove vocabulary elements in the case of observed losses.

- Made-up loss example: Spurious correlations in training data. Easy to hotfix by modifying the inference-time vocabulary.

[how old is **the President**] [does **he** have a partner] → [Does **Barack Obama** have a partner]

[how old is **the President of France**] [does **he** have a partner] → [Does **Barack Obama** have a partner]

[who is the richest person in the world] [how did **he** get rich] → [How did **Barack Obama** get rich?]

c) Supplemental Edit Operations for Critical Cases

Bias in NLG is an Active Research Area

Demo. Dim.	NLG Task	Works
Gender	Autocomplete	Bordia and Bowman (2019); Qian et al. (2019); Solaiman et al. (2019); Sheng et al. (2019, 2020); Vig et al. (2020); Yeo and Chen (2020); Brown et al. (2020); Dhamala et al. (2021); Schick et al. (2021); Nozza et al. (2021); Kirk et al. (2021)
	Dialogue	Henderson et al. (2018); Dinan et al. (2020a); Liu et al. (2020a,b); Cercas Curry et al. (2020); Sheng et al. (2021a,b)
	MT	Vanmassenhove et al. (2018); Elaraby et al. (2018); Prates et al. (2019); Stanovsky et al. (2019); Escudé Font and Costa-jussà (2019); Cho et al. (2019); Moryossef et al. (2019); Saunders and Byrne (2020); Saunders et al. (2020); Kocmi et al. (2020); Costa-jussà and de Jorge (2020); Costa-jussà et al. (2020); Basta et al. (2020); Farkas and Németh (2020); Stafanovičs et al. (2020); Gonen and Webster (2020); Hovy et al. (2020); Roberts et al. (2020); Cho et al. (2021); Savoldi et al. (2021); Renduchintala and Williams (2021); Choubey et al. (2021); Saunders et al. (2021); Tomalin et al. (2021)
	Re-writing	Habash et al. (2019); Zmigrod et al. (2019); Alhafni et al. (2020); Sun et al. (2021)
Profession	Autocomplete	Huang et al. (2020); Dhamala et al. (2021)
Race	Autocomplete	Solaiman et al. (2019); Sheng et al. (2019, 2020); Groenwold et al. (2020); Brown et al. (2020); Dhamala et al. (2021); Schick et al. (2021); Kirk et al. (2021)
	Dialogue	Sheng et al. (2021a,b)
Religion	Autocomplete	Solaiman et al. (2019); Brown et al. (2020); Dhamala et al. (2021); Kirk et al. (2021); Abid et al. (2021)
Sexuality	Autocomplete	Sheng et al. (2019, 2020); Kirk et al. (2021)
	Dialogue	Sheng et al. (2021a)
Other	Autocomplete	Shwartz et al. (2020); Peng et al. (2020); Huang et al. (2020); Dhamala et al. (2021); Kirk et al. (2021)
	Dialogue	Sheng et al. (2021a)
	Re-writing	Pryzant et al. (2020); Ma et al. (2020)

Table 1: Existing bias studies on different demographic dimensions in various NLG tasks: autocomplete generation, dialogue generation, machine translation (MT), and text re-writing.

Sheng et al. *Societal Biases in Language Generation: Progress and Challenges*. ACL 2021 (pdf)

c) Supplemental Edit Operations for Critical Cases

Bias in Pronominalization

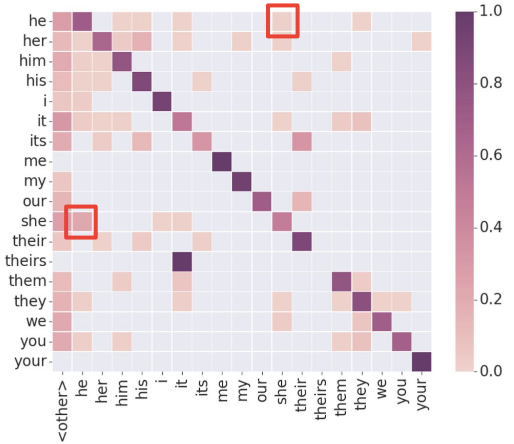
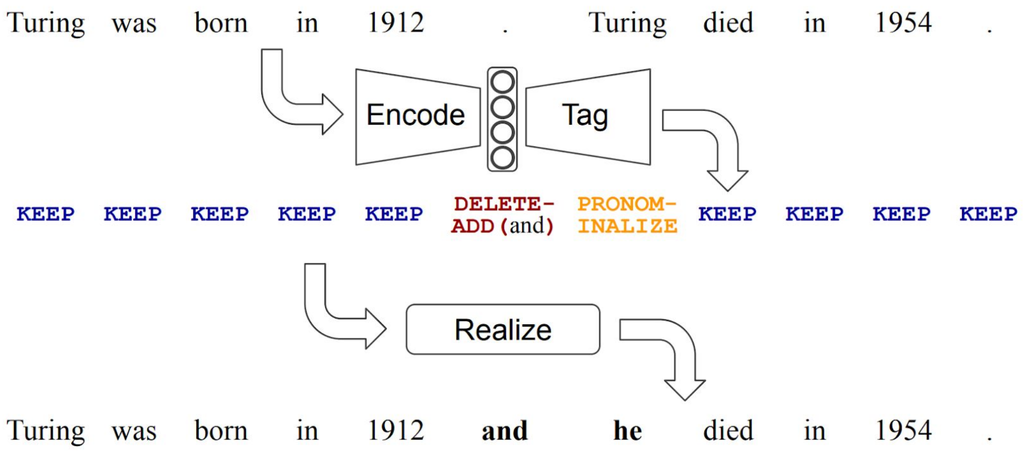


Figure 3: DFWIKI outputs versus the gold pronouns. Rows refer to gold pronouns and columns refer to aligned model outputs at the gold pronoun position.



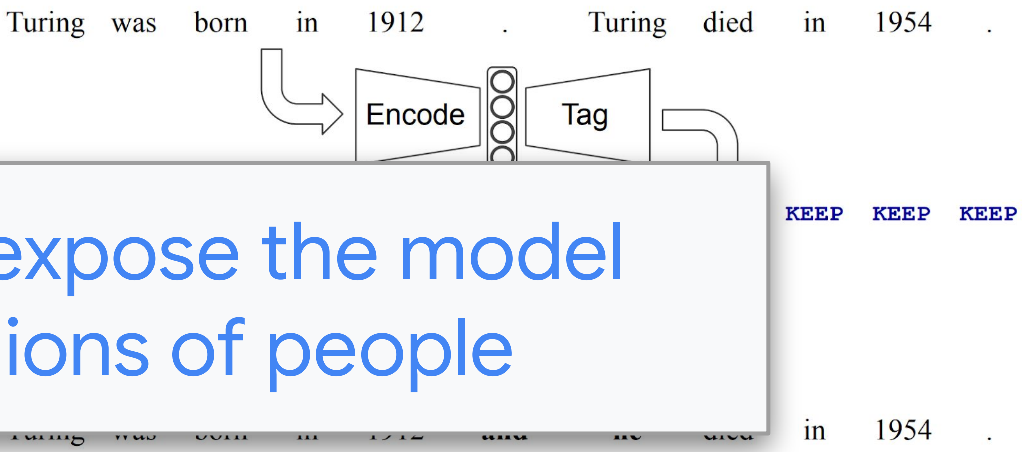
Geva et al. *DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion.* NAACL 2019 (pdf)

Fig: Leveraging external knowledge to select the appropriate pronoun with LaserTagger.

c) Supplemental Edit Operations for Critical Cases

Bias in Pronominalization

Figure 3: DFWIKI
Rows refer to gold pronouns and columns refer to aligned model outputs at the gold pronoun position.



Safer to expose the model to millions of people

Geva et al. *DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion.* NAACL 2019 (pdf)

Fig: Leveraging external knowledge to select the appropriate pronoun with LaserTagger.

4-2. Biasing the edit types

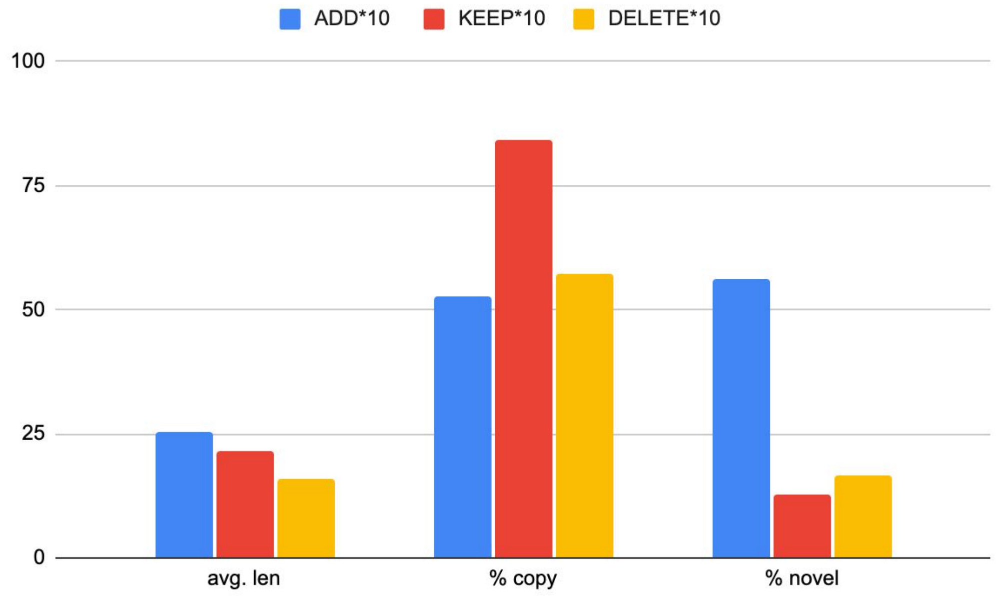
Edit Type Bias

Controlled Generation

Assigning bias/weights for each edit type results in different model behavior

- ***Confidence bias*** for **KEEP** (Omelianchuk et al., 2020)
  - Added to the probability of **KEEP** tag for not changing the source token
- ***Threshold values*** and relative weights (Kumar et al., 2020)
  - Added to control when to perform edit
- ***Edit label ratio*** (Dong et al., 2019)
  - Added to control the ratio for each edit operation

Edit Type Bias



- [Dong et al., 2019]

Reward **ADD**:

- Long output
- More novel words

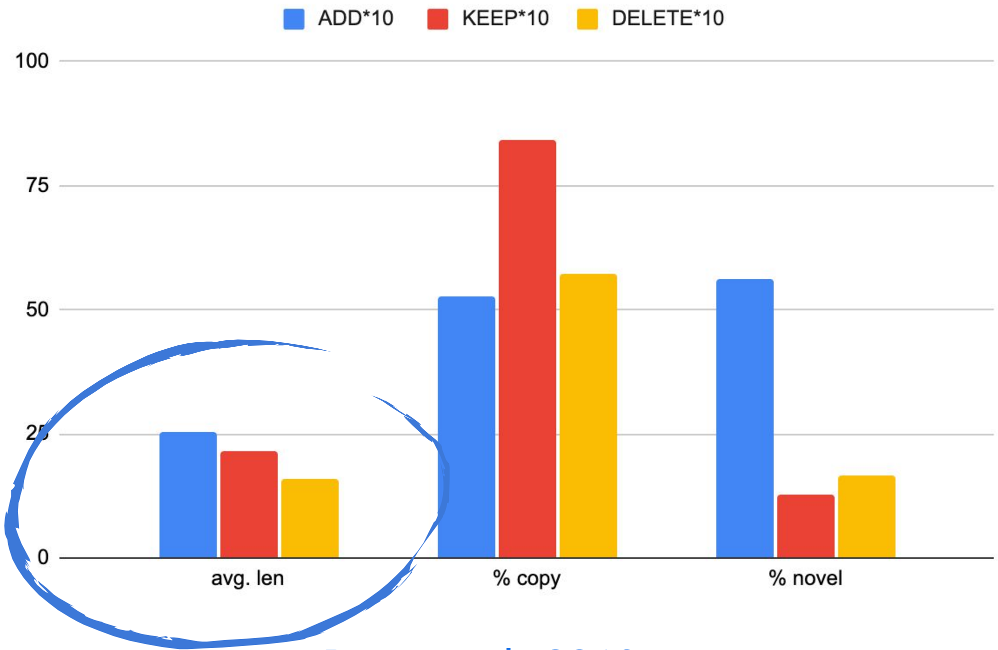
Reward **KEEP**:

- More copy

Reward **DELETE**:

- Short output

Edit Type Bias



○ Dong et al., 2019

Reward **ADD**:

- Long output
- More novel words

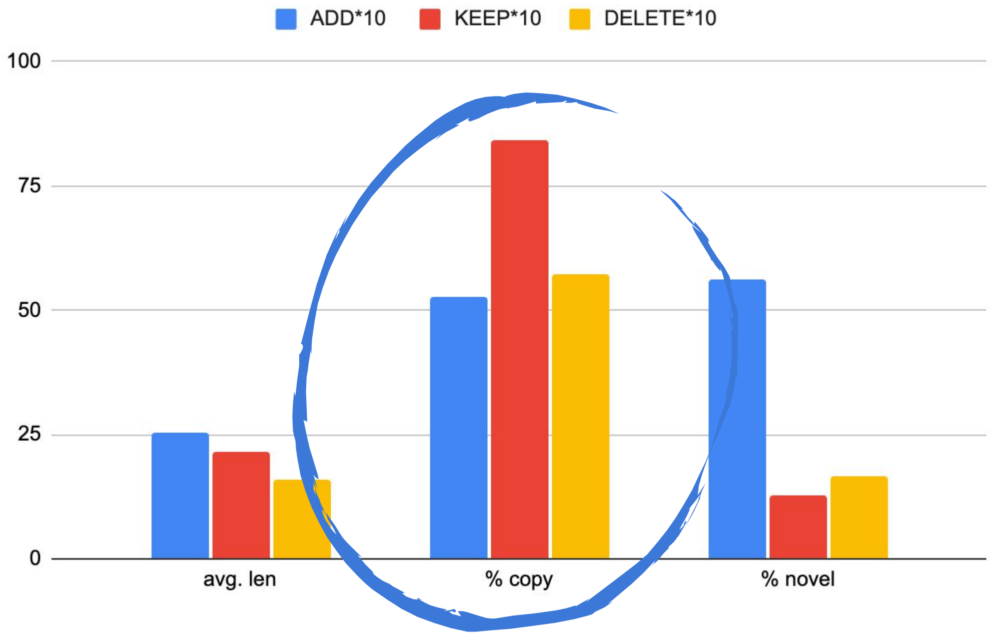
Reward **KEEP**:

- More copy

Reward **DELETE**:

- Short output

Edit Type Bias



- [Dong et al., 2019](#)

Reward **ADD**:

- Long output
- More novel words

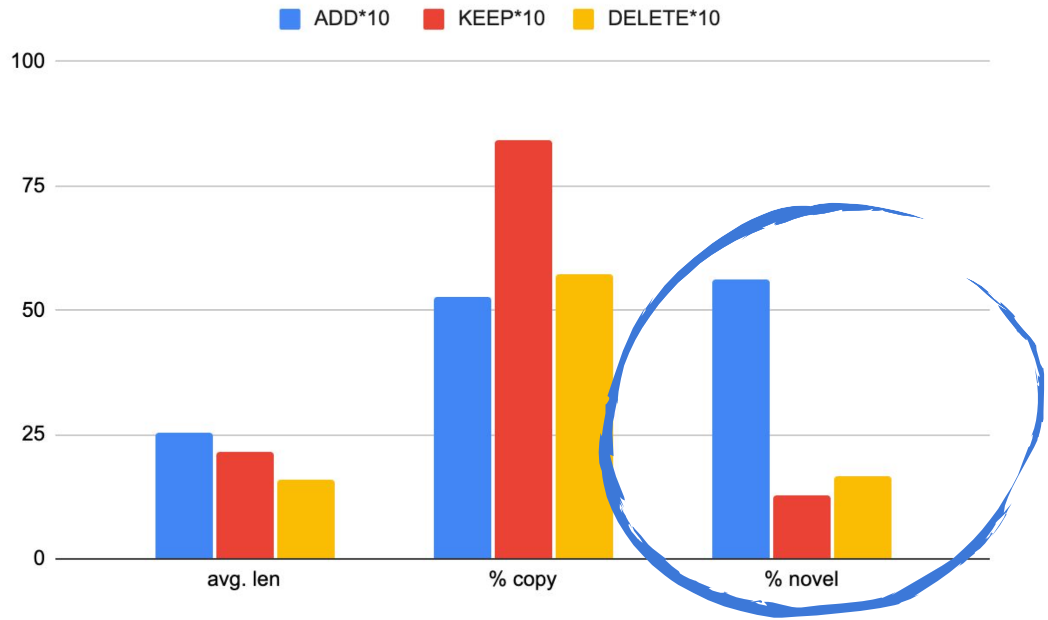
Reward **KEEP**:

- More copy

Reward **DELETE**:

- Short output

Edit Type Bias



Reward **ADD**:
- Long output
- More novel words

Reward **KEEP**:
- More copy

Reward **DELETE**:
- Short output

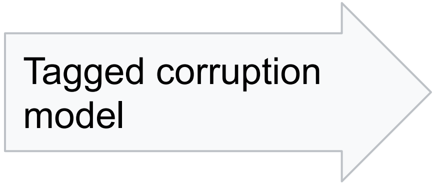
- [Dong et al., 2019]

4-3. Controllable dataset generation

Tagged corruption models for synthetic GEC training data generation

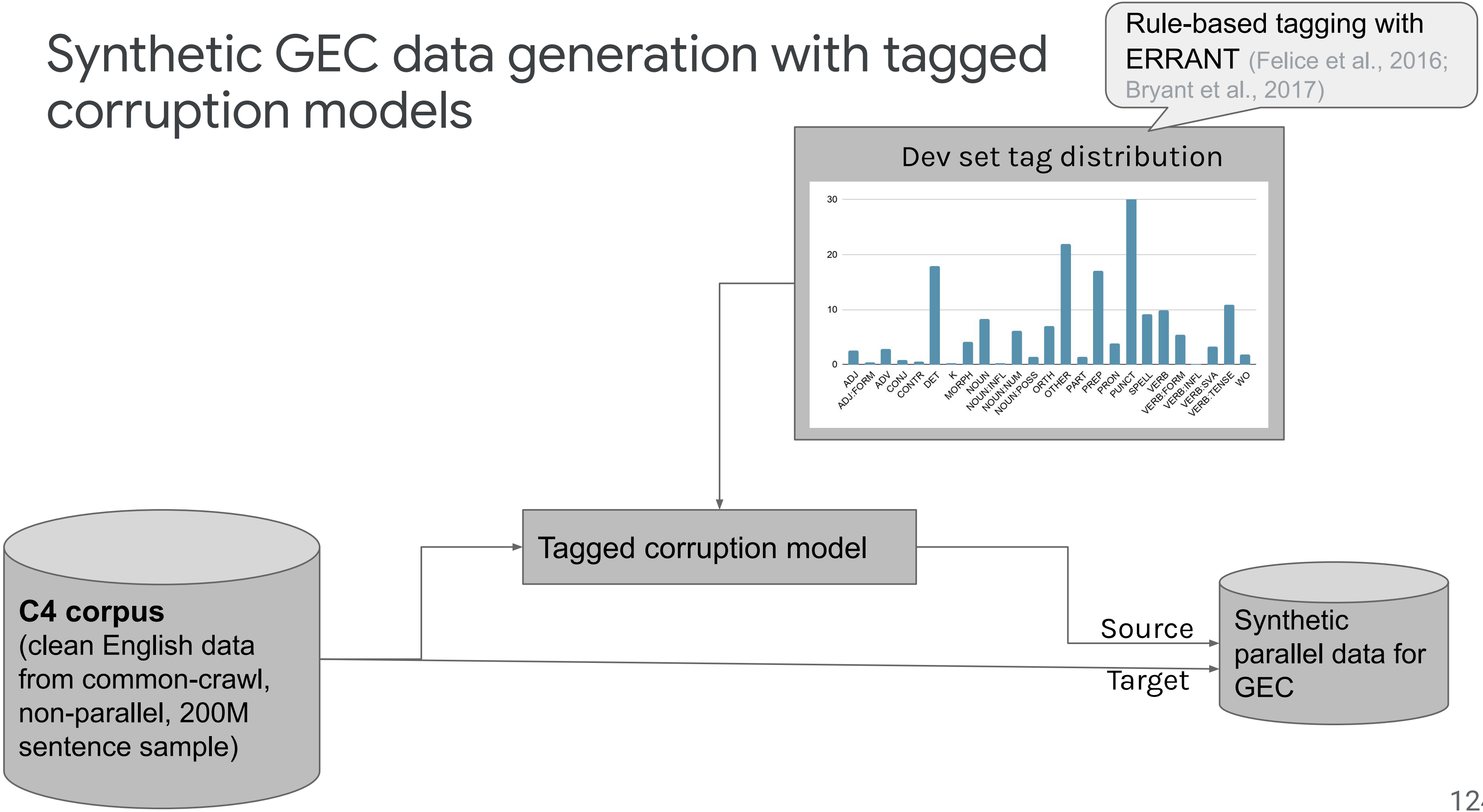
- Applying back-translation to grammatical error correction does not always generate realistic data
  - Not enough diversity
  - Tendency to synthesize only trivial errors
- Can we use error type tags (Bryant et al., 2017) to generate more diverse and more realistic grammatical errors? (Stahlberg and Kumar, 2021)

Error type:	NOUN:INFL
Sentence:	There were a lot of sheep.



There were a lot of sheeps.

Synthetic GEC data generation with tagged corruption models

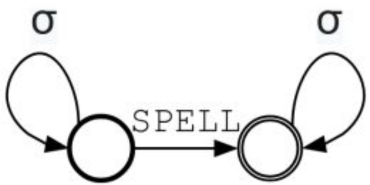


Tagged corruption models

Option 1: Train on tagged source sentences (full sequence and edit-based models)

NOUN:INFL There were a lot of sheep.	There were a lot of sheeps.
DET There were a lot of sheep.	There were lot of sheep.
PART There were a lot of sheep.	There were a lot off sheep.
…	

Option 2: Finite state transducer constraints (tagged edit-based models only)



Full sequence vs. edit-based corruption models for GEC

***Corruption* model type**	***Correction* F0.5 score**		
	Untagged	**Tagged (FST constraint)**	**Tagged (input)**
Full sequence	42.4	-	38.8
Seq2Edits	40.4	46.2	46.3

Tagged edit-based corruption models outperform tagged full sequence corruption models (Stahlberg and Kumar, 2021).

Tagged corruption models in fine-tuning

System	Test set (F0.5)			
	CEFR-A	**CEFR-B**	**CEFR-C**	**Native**
Real data	**<u>50.3</u>**	**<u>51.5</u>**	**<u>44.1</u>**	42.1
Tagged corruptions ~ CEFR-A	47.4	46.2	39.0	39.0
Tagged corruptions ~ CEFR-B	47.1	46.0	40.9	38.0
Tagged corruptions ~ CEFR-C	47.1	46.2	37.1	39.1
Tagged corruptions ~ Native	47.8	49.2	42.8	**<u>42.9</u>**

Matching the tag distribution improves GEC performance for native speakers.

Tagged corruption models in pre-training (C4_200M)

Tag distribution	BEA-dev	CoNLL-13	JFLEG-dev
P*()	**F0.5**	**F0.5**	**GLEU**
None (no tags)	51.4	47.9	57.1
BEA-dev	54.7	51.9	58.5
CoNLL-13	53.9	50.8	58.1
JFLEG-dev	53.8	50.9	58.4
Uniform	54.5	51.1	58.3

The BEA-dev distribution generalizes well to other test sets
The Uniform distribution is also a good choice

200M synthetic GEC training set (C4_200M) available here:
https://github.com/google-research-datasets/C4_200M-synthetic-dataset-for-grammatical-error-correction

Questions?