Diamonds in the Rough: Generating Fluent Sentences from Early-stage Drafts for Academic Writing Assistance

Takumi Ito[1,2], Tatsuki Kuribayashi[1,2], Hayato Kobayashi[3,4], Ana Brassard[4,1], Masato Hagiwara[5], Jun Suzuki[1,4] and Kentaro Inui[1,4]

1: Tohoku University, 2: Langsmith Inc., 3: Yahoo Japan Corporation, 4: RIKEN, 5: Octanove Labs LLC

The writing process

FIRST DRAFT: *"Model have good results."*

Revising

"Our model show good result in this task."

"Our model shows a excellent perfomance in this task."

Editing

"Our model shows good results in this task."

"Our model shows ~~a~~ excellent perfomance in this task."

Proofreading

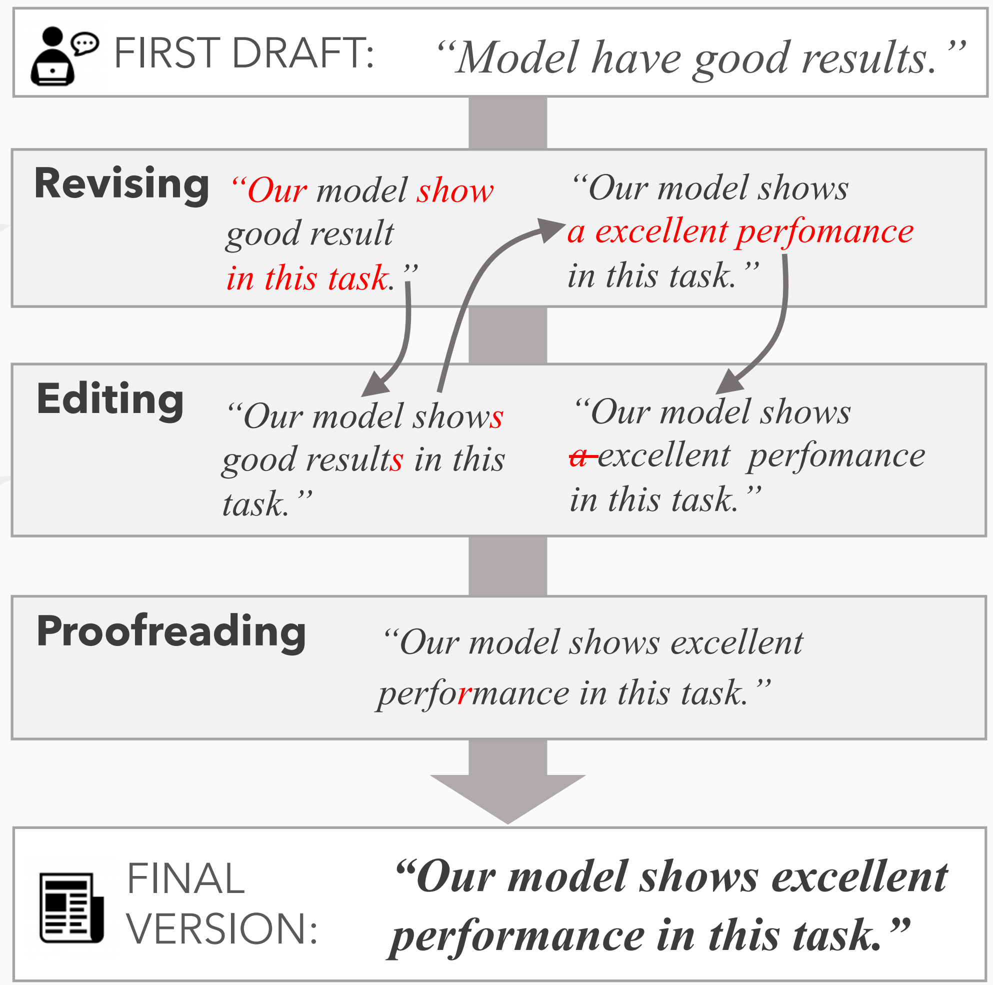
"Our model shows excellent performance in this task."

FINAL VERSION: ***"Our model shows excellent performance in this task."***

Automatic writing assistance

- insufficient fluidity
- awkward style
- collocation errors
- missing words

- grammatical errors
- spelling errors



Automatic writing assistance

- ✗ insufficient fluidity
- ✗ awkward style
- ✗ collocation errors
- ✗ missing words

- ✓ grammatical errors
- ✓ spelling errors

Grammatical error correction (GEC)

FIRST DRAFT: *"Model have good results."*

Revising *"Our model show good result in this task."* *"Our model shows a excellent perfomance in this task."*

EXISTING STUDIES

Editing *"Our model shows good results in this task."* *"Our model shows ~~a~~ excellent perfomance in this task."*

Proofreading *"Our model shows excellent performance in this task."*

FINAL VERSION: ***"Our model shows excellent performance in this task."***

Automatic writing assistance

- ✓ insufficient fluidity
- ✓ awkward style
- ✓ collocation errors
- ✓ missing words

- ✓ grammatical errors
- ✓ spelling errors

Grammatical error correction (GEC)

Sentence-level revision (SentRev)

FIRST DRAFT: *“Model have good results.”*

OUR FOCUS

Revising *“Our model show good result in this task.”* *“Our model shows a excellent perfomance in this task.”*

Editing *“Our model shows good results in this task.”* *“Our model shows ~~a~~ excellent perfomance in this task.”*

Proofreading *“Our model shows excellent performance in this task.”*

FINAL VERSION: ***“Our model shows excellent performance in this task.”***

Proposed Task: Sentence-level Revision

Our aproach idea is <> at read patern of normal human.*

draft

revising, editing, proofreading

The idea of our approach derives from the normal human reading pattern.

final version

- input: early-stage draft sentence
  - has errors (e.g., collocation errors)
  - has Information gaps (denoted by **<*>**)
- output: final version sentence
  - error-free
  - correctly filled-in sentence

Proposed Task: Sentence-level Revision

Our aproach idea is <> at read patern of normal human.*

draft

revising, editing, proofreading

The idea of our approach derives from the normal human reading pattern.

final version

- input: early-stage draft sentence
  - has errors (e.g., collocation errors)
  - has Information gaps (denoted by **<*>**)
- output: final version sentence
  - error-free
  - correctly filled-in sentence
- issue: lack of evaluation resource

Our contributions

Our aproach idea is <> at read patern of normal human.*

draft

revising, editing, proofreading

The idea of our approach derives from the normal human reading pattern.

final version

- Created an evaluation dataset for SentRev
  - Set of Modified Incomplete TecHnical paper sentences (SMITH)
- Analyzed the characteristics of the dataset
- Established baseline scores for SentRev

Evaluation Dataset Creation

Goal: collect pairs of draft sentence and final version

Our model <> results*	*Our model **shows competitive** results*
draft	final

Evaluation Dataset Creation

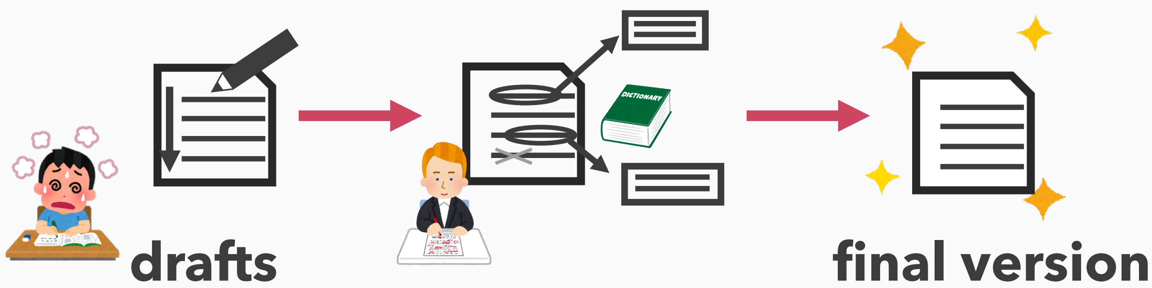
Goal: collect pairs of draft sentence and final version

Our model <> results*

*Our model **shows competitive** results*

Straight-forward approach：

Experts modify collected drafts to final version



limitation:

early-stage draft sentences are not usually publicly available

Note:

We can access plenty of final version sentences

Evaluation Dataset Creation

Goal: collect pairs of draft sentence and final version

Our model <> results*

*Our model **shows competitive** results*

Straight-forward approach：

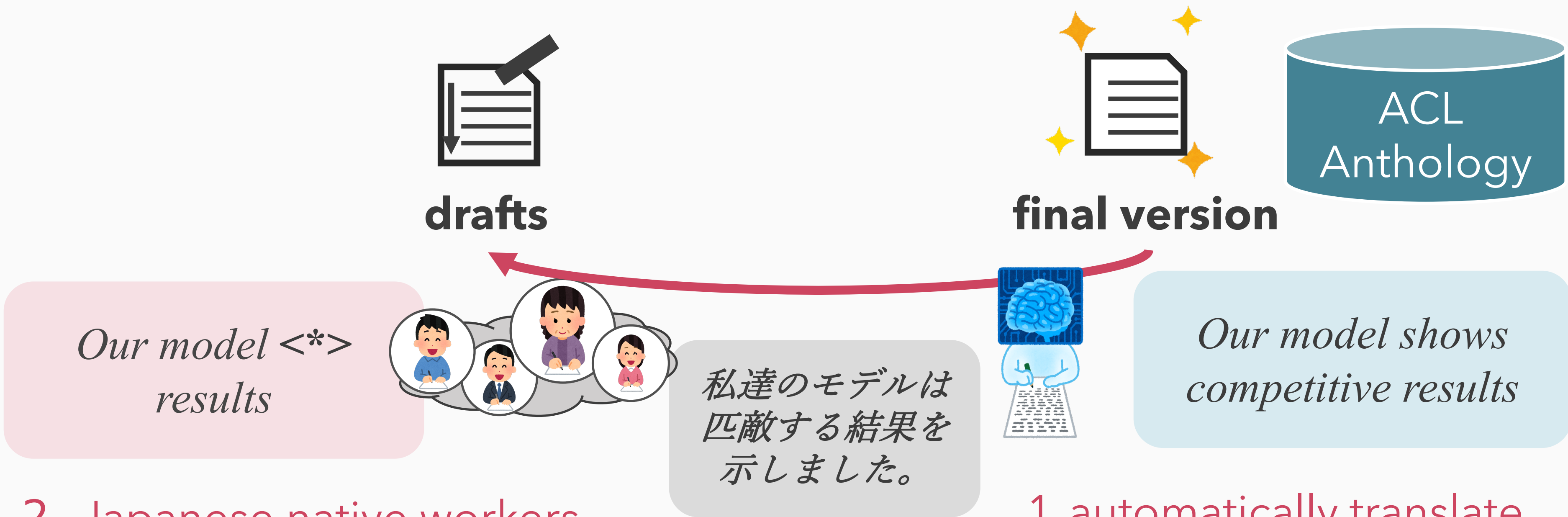
Experts modify collected drafts to final version

Our approach:

create draft sentences from final version sentences

Crowdsourcing Protocol for Creating an Evaluation Dataset

Our approach:
create draft sentences from final version sentences



2. Japanese native workers translate into English

1. automatically translate the final sentence into Japanese

Crowdsourcing Protocol for Creating an Evaluation Dataset

Our approach:
create draft sentences from final version sentences

Statistics

Dataset	size	w/<*>	w/change	Levenshtein distance
Lang-8	2.1M	-	42%	3.5
AESW	1.2M	-	39%	4.8
JFLEG	1.5K	-	86%	12.4
SMITH	10K	33%	99%	47.0

w/<*>: percentage of source sentences with <*>

w/change: percentage where the source and target sentences differ

- collected 10,804 pairs
- SMITH simulates significant editing
- Larger Levenshtein distance ⇨ more drastic editing

Examples of SMITH

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping .*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Examples of SMITH

(1) Wording problems

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping .*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Examples of SMITH

(1) Wording problems

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping.*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Examples of SMITH

(2) Information gaps

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping .*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Examples of SMITH

(2) Information gaps

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping.*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Examples of SMITH

(3) Spelling and grammatical errors

draft: *I research the rate of workable SQL <*> at the generated result.*

final: *We study the percentage of executable SQL queries in the generated results.*

draft: *For <*>, we used Adam using weight decay and gradient clipping.*

final: *We used Adam with a weight decay and gradient clipping for optimization.*

draft: *In the model aechitecture, as shown in Figure 1 , it is based an AE and GAN.*

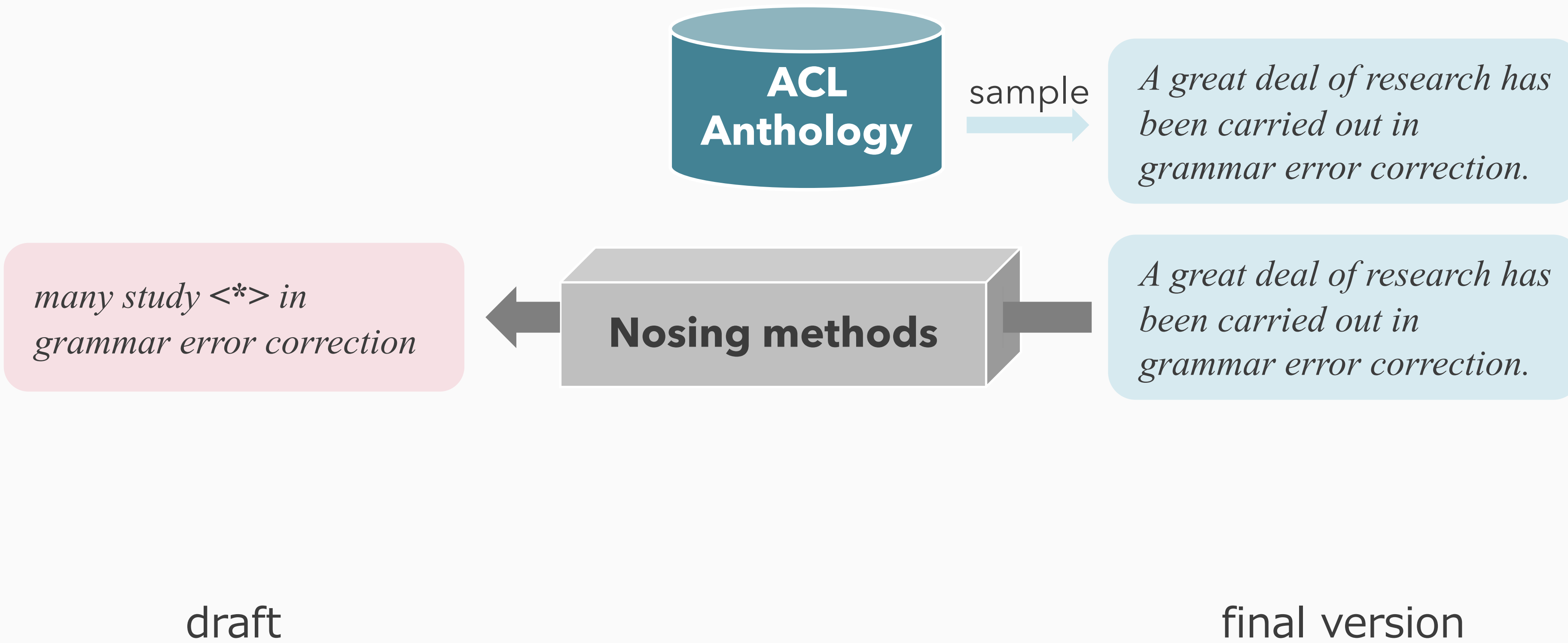
final: *The model architecture, as illustrated in figure 1 , is based on the AE and GAN.*

Experiments

- built baseline revision models (draft ⇨ final version)
  - training data: generated synthetic data with noising methods
- evaluated the performance on SMITH
  - using various reference and reference-less evaluation metrics

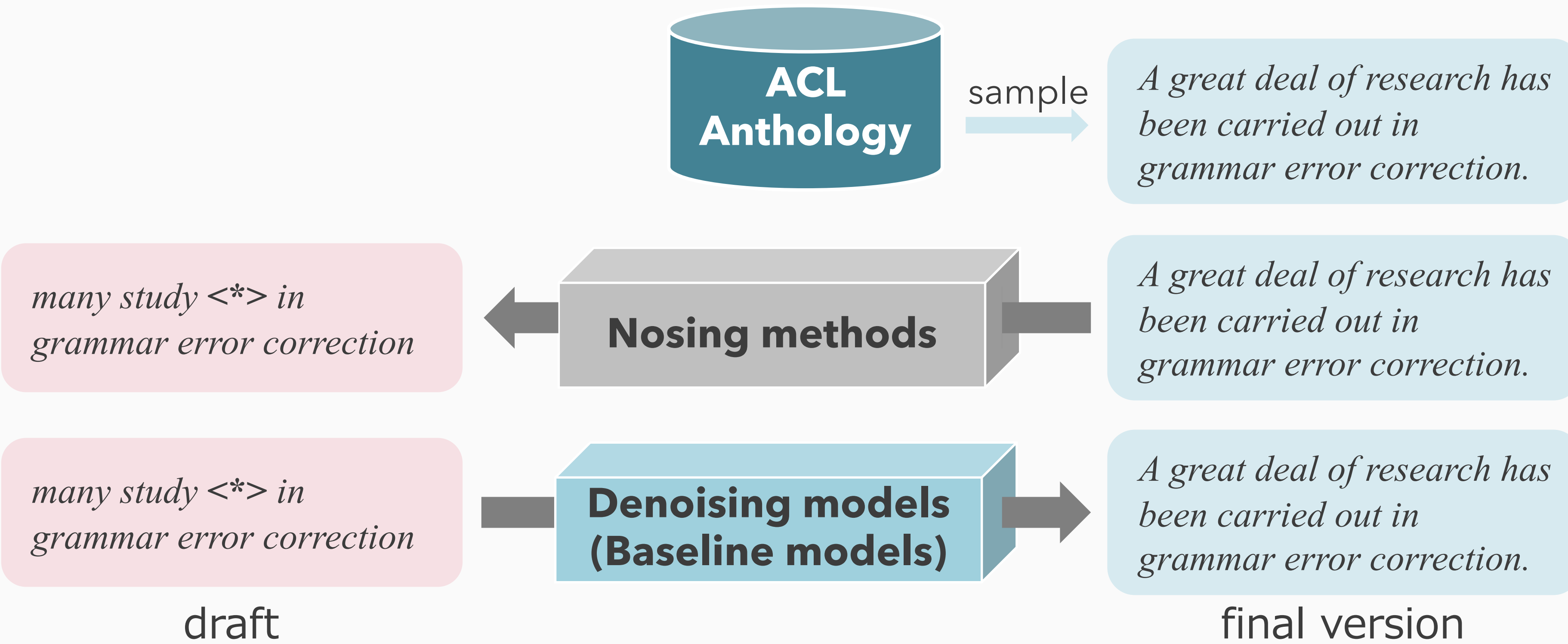
Noising and Denoising

Noising: automatically generate drafts from the final versions



Noising and Denoising

Denoising: generate final versions from the drafts



Noising methods

drafts	Noising methods	final versions
it is not surprisingly that the random policy have the worst performing.	Grammatical error generation	*it is not surprising that the random policy has the worst performance.*
we see the same on larger data.	Style removal	*we observe a similar trend on larger datasets.*
Figure 2 illustrates effectiveness	Entailed sentence generation	*Figure 2 illustrates the effectiveness of different features.*
perplexity indicates a <> model.*	Heuristic	*lower perplexity indicates a better model.*

Noising methods

drafts	Noising methods	final versions
it is not surprisingly that the random policy have the worst performing.	Grammatical error generation	*it is not surprising that the random policy has the worst performance.*
perplexity indicates a <> model.*	Heuristic	*lower perplexity indicates a better model.*

train Enc-Dec noising model (clean ⇨ erroneous)
using Lang8[Mizumoto+ 11], AESW[Daudaravicius+ 15],
and JFLEG[Napoles+ 17]

Noising methods

drafts	Noising methods	final versions
it is not surprisingly that the random policy have the worst performing.	Grammatical error generation	*it is not surprising that the random policy has the worst performance.*
we see the same on larger data.	Style removal	*we observe a similar trend on larger datasets.*
Figure 2 illustrates ef…	…led sentence	*Figure 2 illustrates the effectiveness of different…*
pe… model.	Heuristic	*a better model.*

train Enc-Dec noising model (academic ⇨ non-academic)
using the ParaNMT-50M dataset [Wieting+18]

Noising methods

drafts | Noising methods | final versions

it is not surprisingly that the random policy have the worst performing. ← Grammatical error generation ← *it is not surprising that the random policy has the worst performance.*

train Enc-Dec noising model (⇨ entailed sentence)
using SNLI [Bowman+ 15], MultiNLI [Williams+ 18]

we see th... data.

Figure 2 illustrates effectiveness ← Entailed sentence generation ← *Figure 2 illustrates the effectiveness of different features.*

perplexity indicates a <> model.* ← Heuristic ← *lower perplexity indicates a better model.*

Noising methods

drafts | Noising methods | final versions

it is not surprisingly that the random policy have the worst performing. ← Grammatical error generation ← *it is not surprising that the random policy has the worst performance.*

we see the same on larger data. ← Style removal ← *we observe a similar trend on larger datasets.*

generation

features.

heuristic noising rules:
randomly deleting, replacing with <*> or common terms, and swapping

perplexity indicates a <> model.* ← Heuristic ← *lower perplexity indicates a better model.*

Baseline models

- Noising and Denoising models
  - Heuristic noising and denoising model (H-ND)
    - Rule-based Heuristic noising (e.g., random token replacing)
  - Enc-Dec noising and denoising model (ED-ND)
    - Rule-based Heuristic noising
      + trained error generation models (e.g., grammatical error generation)
- SOTA GEC model [Zhao+ 19]

Experiment settings

- Noising and Denoising Model architecture
  - Transformer [Vaswani+ 17]
  - Optimizer: Adam with $\alpha = 0.0005$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10e^{-8}$
- Evaluation metrics
  - BLEU
  - ROUGE-L
  - F0.5
  - BERTscore [Zhang+ 19]
  - Grammaticality score [Napoles+ 16]: 1 – (#errors in sent /#tokens in sent)
  - Perplexity (PPL): 5-gram LM trained on ACL Anthology papers

Results

Model	BLEU	ROUGE-L	BERT-P	BERT-R	BERT-F	P	R	$F_{0.5}$	Gramm.	PPL
Draft X	9.8	46.8	75.9	78.2	77.0	-	-	-	92.9	1454
H-ND	8.2	45.0	77.0	76.1	76.5	5.4	2.9	4.6	94.1	406
ED-ND	**15.4**	**51.1**	**80.9**	**80.0**	**80.4**	21.8	**12.8**	**19.2**	96.3	**236**
GEC	11.9	49.0	80.8	79.1	79.9	**22.2**	6.2	14.6	**96.7**	414
Reference Y	-	-	-	-	-	-	-	-	96.5	147

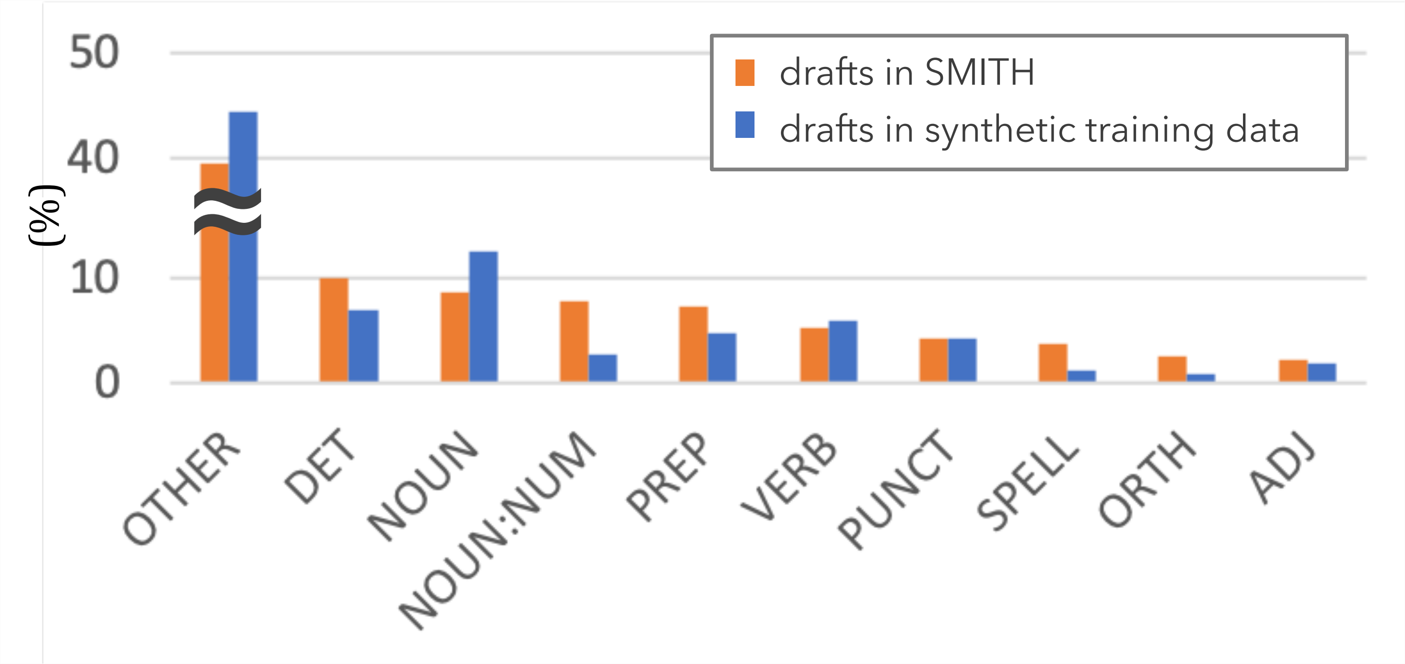
- ED-ND model outperforms the other models
  - the HD-ND noising methods induced noise closer to real-world drafts
- SOTA GEC model showed higher precision but low recall
  - the GEC model is conservative

Examples of the baseline models' output

Draft	*The input and output <*> are one - hot encoding of the center word and the context word , <*> .*
H-ND	*The input and output are one - hot encoding of the center word and the context word , respectively .*
ED-ND	*The input and output layers are one - hot encoding of the center word and the context word , respectively .*
GEC	*The input and output are one - hot encoding of the center word and the context word , .*
Reference	*The input and output layers are center word and context word one - hot encodings , respectively .*

ED-ND models replaced the **<*>** token with plausible words

Analysis:
error types of drafts in SMITH & training data



Similar error type distribution

Conclusions

- proposed the SentRev task
  - Input: a incomplete, rough draft sentence
  - Output: a more fluent, complete sentence in the academic domain.
- created the SMITH dataset with crowdsourcing for development and evaluation of this task
  - available at https://github.com/taku-ito/INLG2019_SentRev
- established baseline performance with a synthetic training dataset
  - training dataset available at the same link as above

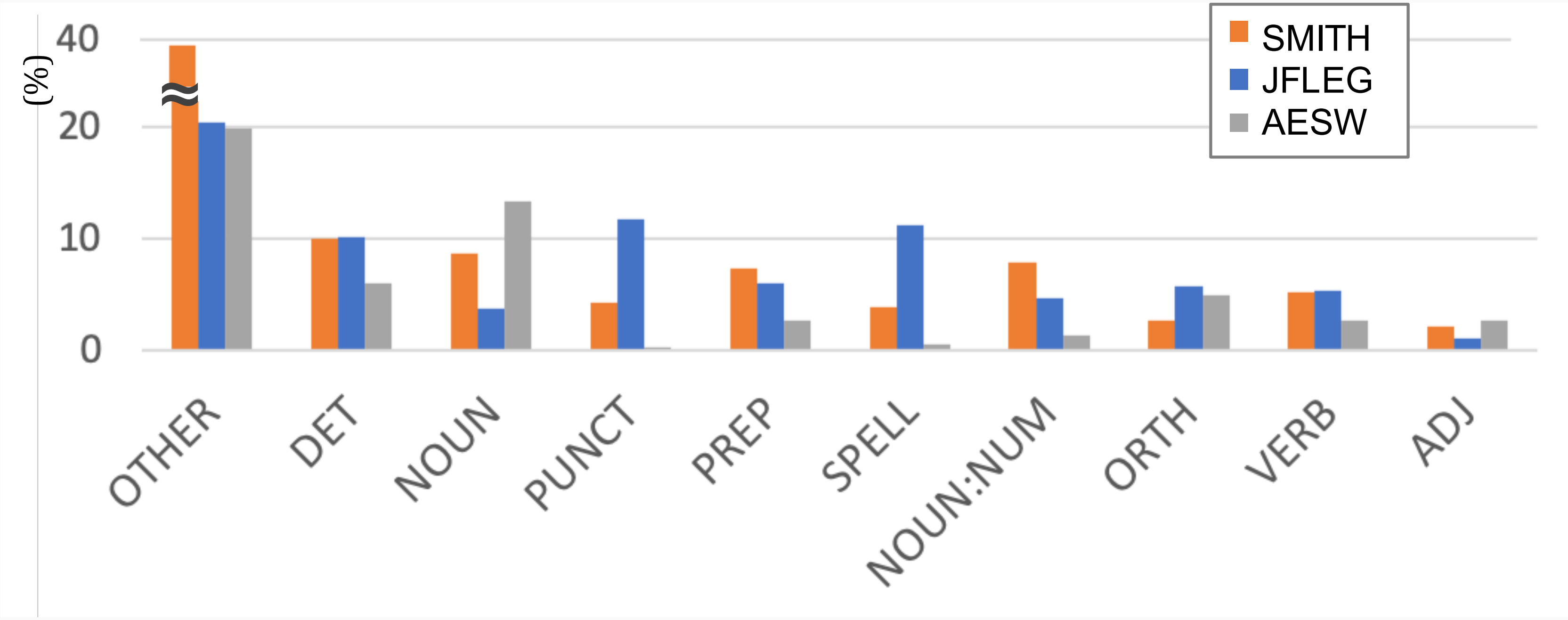
Appendix

Criteria for evaluating crowdworkers

Criteria	Judgment
Working time is too short (< 2 minutes)	Reject
All answers are too short (< 4 words)	Reject
No answer ends with "." or "?"	Reject
Contain identical answers	Reject
Some answers have Japanese words	Reject
No answer is recognized as English	Reject
Some answers are too short (< 4 words)	-2 points
Some answers use fewer than 4 kinds of words	-2 points
Too close to automatic translation (20 <= L.D. <= 30)	-0.5 points/ans
Too close to automatic translation (10 <= L.D. <= 20)	-1.5 points/ans
Too close to automatic translation (L.D. <= 10)	Reject
All answers end with "." or "?"	+1 points
Some answers have **`<*>`**	+1 points
All answers are recognized as English	+1 points

- filtered the crowdworkers' answers using the criteria
- accepted answers with score 0 or higher

Comparison of the top 10 frequent errors observed in the 3 datasets



SMITH included more “OTHER” than the other two datasets

Examples of "OTHER" in SMITH

Draft: *the best models are very effective on the* $\boxed{}$ *condition that they are far greater than human.*

OTHER

Reference: *The best models are very effective in the* $\boxed{\textit{local context}}$ *condition where they significantly outperform humans.*

Draft: *Results show MARM tend to generate* $\boxed{\text{<*>}}$ *and very short responces.*

OTHER

Reference: *The results indicate that MARM tends to generate* $\boxed{\textit{specific}}$ *but very short responses.*

SMITH emphasizes "completion-type" task setting for writing assistance.