CASIMIR: A Corpus of Scientific Articles enhanced with multiple Author-Integrated Revisions

Léane Jourdan, Florian Boudin, Richard Dufour, Nicolas Hernandez

CASIMIR: A Corpus of Scientific Articles enhanced with multiple Author-Integrated Revisions

Léane Jourdan ♣, Florian Boudin ♦♣, Richard Dufour ♣, Nicolas Hernandez ♣
{firstname.lastname}@univ-nantes.fr
♣ Nantes Université, CNRS, LS2N, UMR 6004, Nantes, France ♦ Japanese French Laboratory for Informatics, CNRS, NII, Tokyo, Japan

Summary

- **15K+ scientific articles with revisions**, metadata and peer reviews
- 3.7 M **aligned sentences** and 5.2M edits
- Automatic **annotation of edits' intention**
- Evaluation of **writing assistance** models

Content

CASIMIR

Article pairs

- 15 646 different articles
- (3.5 versions per article on average)
- 36 733 pairs of versions

Metadata

- Authors
- Keywords
- Conference
- Dates...

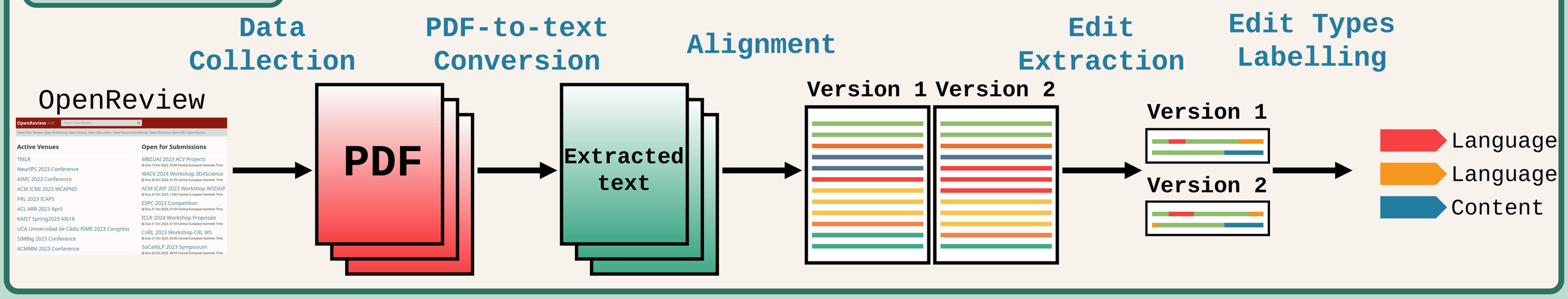
Reviews

- Acceptance decision
- Comments (can contain grades)
- Dates...

29 conferences

Domains : machine learning (ICLR, ICML, NeurIPS), robotics (RSS, CoRL), NLP (ACL) and computer vision (ECCV)

Methodology



Example

Label of edits:
Content | **Language** | **Improve-grammar-Typo**

Source text

Nevertheless, challenges exist for developing deep learning-based models to predict mutational effects on protein-protein **binding.**
The major challenge is the scarcity of experimental data — only a few **thousands of** protein **mutations** annotated with **the change** in binding **affinity** are publicly available (Geng et al., **2019b). This hinders** supervised learning **as** the **insufficiency of** training **data tends to cause over-fitting.**

Revised text

However, developing deep learning-based models to predict mutational effects on protein-protein **binding is challenging due to the scarcity of experimental data.**
Only a few **thousand** protein **mutations,** annotated with **changes** in binding **affinity,** are publicly available (Geng et al., **2019b), making** supervised learning **challenging due to** the **potential for overfitting with insufficient** training **data.**

Statistics

5.2M of individual edits distributed in 3.7M of edited sentences

Quantity of edits			
Min	1	Average	142.12
Max	4432	Median	74
Edits length			
Min	1	Average	34.88
Max	9316	Median	13

Table 1: Distribution of the quantity of **edits per article** and their length

Edit intention	Percentage
Content	41.97%
Improve-grammar-typo	22.73%
Format	20.38%
Language	14.92%

Table 2: Distribution of **edits' intention**

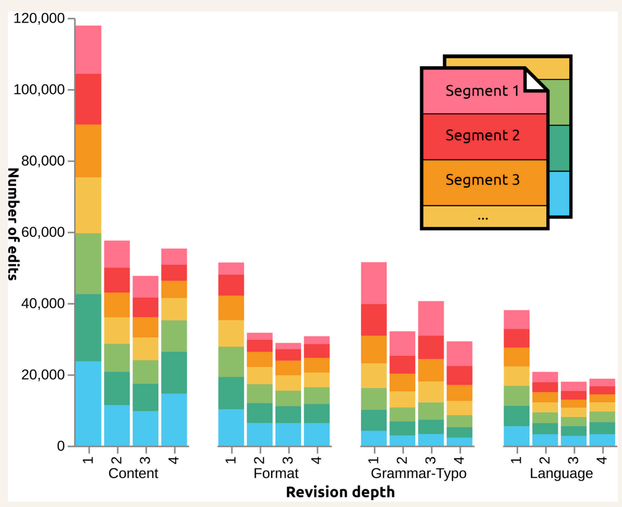


Figure : Evolution of the **position of edited** text per intention and revision depth

Experiments

Task: **Sentence revision**

Sentence to revise + Intention → Model → Generated revised sentence

Models

- IteraTeR-PEGASUS (Grammarly)
- CoEdIT (XL) (Grammarly)
- Llama2-7B (Meta)

Metrics

- SARI
- BLEU
- ROUGE-L
- Bert-score

Every metric measures the predicted and the gold sentences similarity.

Results

Model/Metric	BLEU	ROUGE	SARI	BERT
CopyInput	**66.31**	**74.19**	61.38	94.46
Iterater-Pegasus (best intention)	60.99	73.25	55.27	95.93
Iterater-Pegasus (all intentions)	58.68	72.36	53.77	93.29
CoEdIT (best intention)	58.88	70.89	53.94	**96.08**
CoEdIT (all intentions)	56.44	69.22	51.62	95.99
Llama2-7B (best intention)	61.91	73.02	**62.07**	92.84
Llama2-7B (all intentions)	57.46	68.18	58.39	92.37

- **Results are close**, even the control approach (CopyInput) performs best with two metrics
- The task is **hard to evaluate automatically** (a sentence can have several valid revisions)
- Promising directions:
  - **Aggregation of metrics** based on the improvement between predicted and gold sentences (grammaticality, readability, ...)
  - **Multiple ground truth revisions**, either produced manually or generated automatically

Article Corpus

LREC-COLING 2024

LS2N Nantes Université TALN taln.ls2n.fr