One Document, Many Revisions: A Dataset for Classification and Description of Edit Intents

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Jauhar, Diyi Yang, Eduard Hovy
dheeraj@cs.cmu.edu

Language Technologies Institute, Carnegie Mellon University
https://tinyurl.com/editsumm

Edit Description

- **Goal:** When an author edits a document, we aim to understand the motivation behind the edit.
- We use wikipedia as our dataset where each page is edited multiple times over time
- We present a publicly available dataset that contains first 100 edits of about 147 wikipedia pages with annotated intent
- We also present the task of Edit Comment Generation, which uses the intention to generate free-form version of the edit description

Dataset

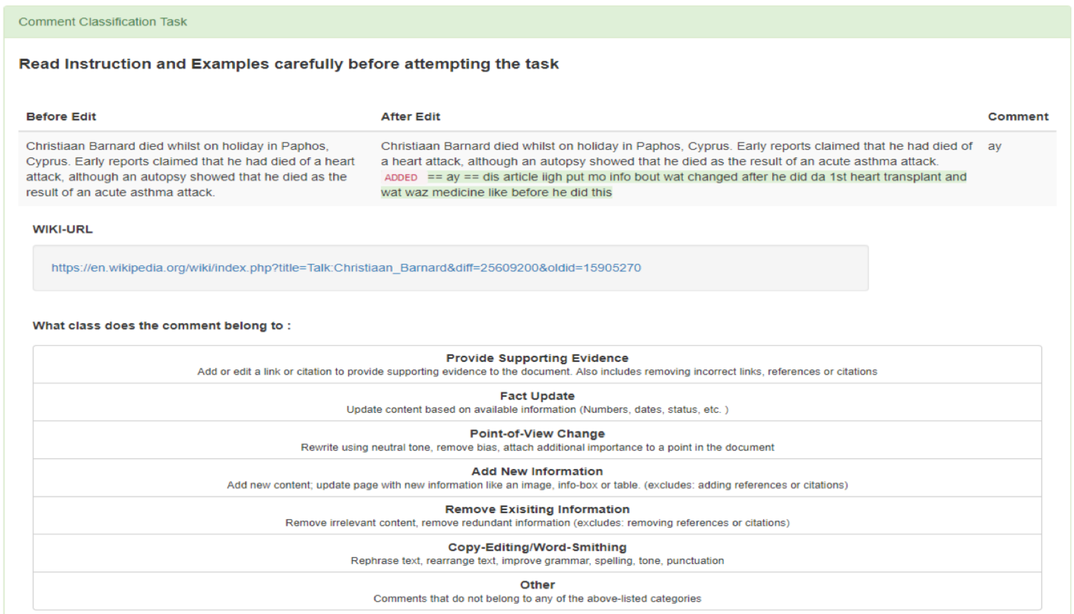
Class	Percentage(%)
Provide Supporting Evidence	45.9
Word-Smithing	19.0
Add New Information	23.2
Fact Update	3.7
Point-of-view Change	0.7
Remove Existing Information	4.7
Other	2.6

Table: Dataset statistics for **Edit Intention Classification**

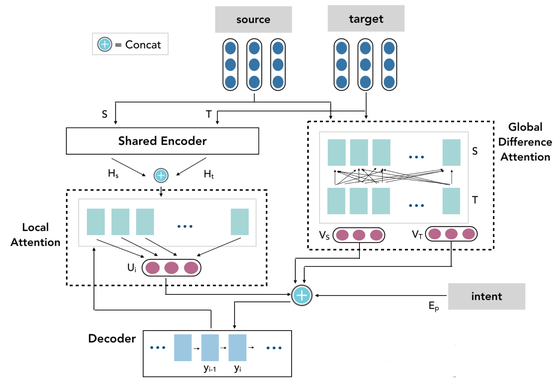
Field	Description
Title	Title of the page
Bef_Rev	Content of page before revision
Aft_Rev	Content of page after revision
Comment	Author's comment
Edit Intent	**Intent of the Edit (crowd-sourced)**
Time Stamp	Time information of revision
Author	Author ID
URL	URL corresponding to the revision
Mod_Bef	Edited block before revision
Mod_Aft	Edited block after revision
Minor	Major/Minor revision

Table: Dataset fields for each revision

Data Collection Interface



Model Architecture



Results

Classifier	Accuracy
Random Baseline	0.14
Majority Class	0.46
KNN*	0.58 ± 0.32
SVM (linear)*	0.58 ± 0.07
Logistic Regression*	**0.89 ± 0.14**

Table: Edit intent classification Results, with their corresponding 95% confidence interval estimates with 10-fold cross validation. * - denotes that these classifiers use BoW features

Model	R-1	R-2	R-L
S2S + LA	13.0	5.6	12.5
S2S + LA + PI	14.5	6.6	14.1
S2S + LA + PI + GA	**28.0**	**17.0**	**27.9**

Table: Results for **Edit Summary Generation Task**

Conclusion

- We present a dataset for studying edits in documents
- Pretrained language models still struggle at this task
- Future work can explore document evolution over time

References

[1] Yang et al., "Identifying semantic edit intentions from revisions in wikipedia.", EMNLP 2017