Data-Driven Proactive Policy Assurance of Post Quality in Community Q&A Sites

Chunyang Chen, Xi Chen, Jiamou Sun, Zhenchang Xing, Guoqiang Li

Background

Q&A sites are popular for sharing knowledge

• Social Q&A sites
• Technical Q&A sites
Motivation

The quality of Q&A sites are decaying

- Stack Overflow
  - 17M questions, 26M answers, 9.6M users
  - 7K new questions/day, many new users

- Complains:
  - Why do so many good programmers waste their time on Stack Overflow?
  - Farewell Stack Exchange
  - The decline of Stack Overflow
Motivation

To keep the quality of content

1. Publish community norms
   - [https://stackoverflow.com/help/how-to.ask](https://stackoverflow.com/help/how-to.ask)
   - [https://stackoverflow.com/help/how-to-answer](https://stackoverflow.com/help/how-to-answer)

Problem: Users do not read or understand the instructions.

Motivation

To keep the quality of content

2. Peer review

- https://stackoverflow.com/help/privileges/edit
- 2M question-title edits (17.6%)
- 3M question-tag edits (12.9%)
- 21M post-body edits (36.2%)

Problem:
- Require significant community efforts;
- Some edits are difficult to locate;
- The policy violation has hurt readers before edits

When should I edit posts?

Any time you feel you can make the post better, and are inclined to do so. Editing is encouraged!

Some common reasons to edit are:
- to fix grammatical or spelling mistakes
- to clarify the meaning of a post without changing it
- to correct minor mistakes or add addendums / updates as the post ages
- to add related resources or hyperlinks
Goal

To keep the quality of content

• We need a way to help policy assurance of post quality
  • **Proactive**: remind users before they publish the posts
  • **Data-driven**: learn from real existing edits
Observation

Observe the existing edits

Four different kinds of middle-level edits

- Code format edit
- Text format edit
- Link modification
- Image revision
Observation

Observe the existing edits

Each edit including
- Insert
- Replace
- Delete
Data Collection

Collecting the dataset of <original-post, post-body-edit-type>

• Regular expression and text differencing
• Data for different edits
  • Adding code format: 1,567,272
  • Adding text format: 52,945
  • Adding hyperlinks: 1,126,252
  • Adding images: 219,215
CNN model for edit prediction

• Word embedding
  • Convert the word into vector representation

• Convolutional Layer
  • Kernel filter sliding within the input matrix

• Maxpooling
  • Preserve the salient information

• Fully-connected layer
  • Final prediction
Approach

Locating the Key Phrases in Posts to Explain the Edit Prediction

• Tracing back through the model to locating the filtered phrases in the input layer
• Predicting the contribution score of the phrases’ corresponding features in the fully connected layer to the prediction class
Evaluation

Performance comparison between our model and baselines

• Evaluation metrics
  • Precision, recall, F1-score

• Baseline
  • Logistic regression, SVM, FastText, Attention-based LSTM
Evaluation

Understanding of edit predictions

• Locate key phrase to help understand the prediction
  • Add code format

• Add images