MONASH University

By the Community & For the Community:
A Deep Learning Approach to Assist Collaborative Editing in Q&A Sites

Chunyang Chen, Zhenchang Xing, Yang Liu

MONASH University

Australian National University

NANYANG TECHNOLOGICAL UNIVERSITY

Chen, Chunyang, Zhenchang Xing, and Yang Liu. "By the community & for the community: a deep learning approach to assist collaborative editing in q&a sites." *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW (2017): 32.

Background

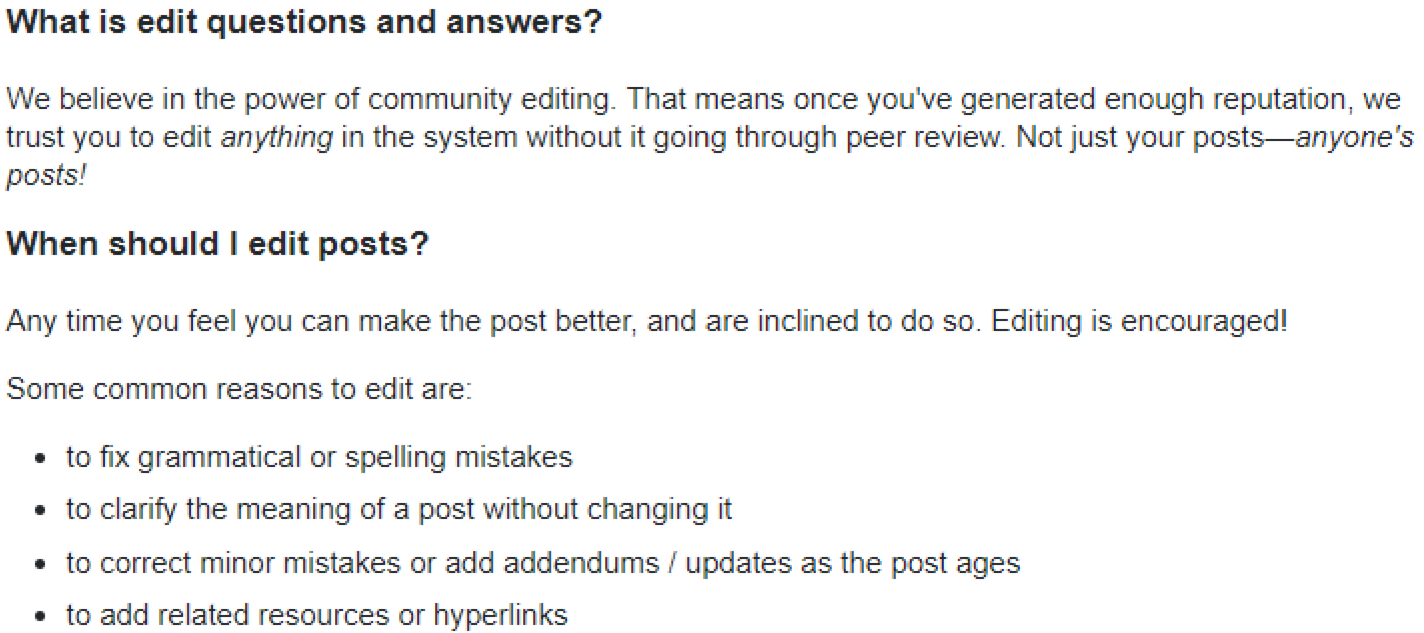
Collaborative edits

- **Collaborative editing** is the editing of groups producing works together through individual contributions. Effective choices in group awareness, participation, and coordination are critical to successful collaborative writing outcomes
- Widely used for crowd-sourcing websites or collaborative platforms:
  - Wikipedia
  - Google Doc, GitHub

Background

Collaborative edits in Stack Overflow

- Stack Overflow
  - 17M questions, 26M answers, 9.6M users
  - 7K new questions/day, many new users
- Edits:
  - https://stackoverflow.com/help/privileges/edit



What is edit questions and answers?

We believe in the power of community editing. That means once you've generated enough reputation, we trust you to edit *anything* in the system without it going through peer review. Not just your posts—*anyone's posts!*

When should I edit posts?

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

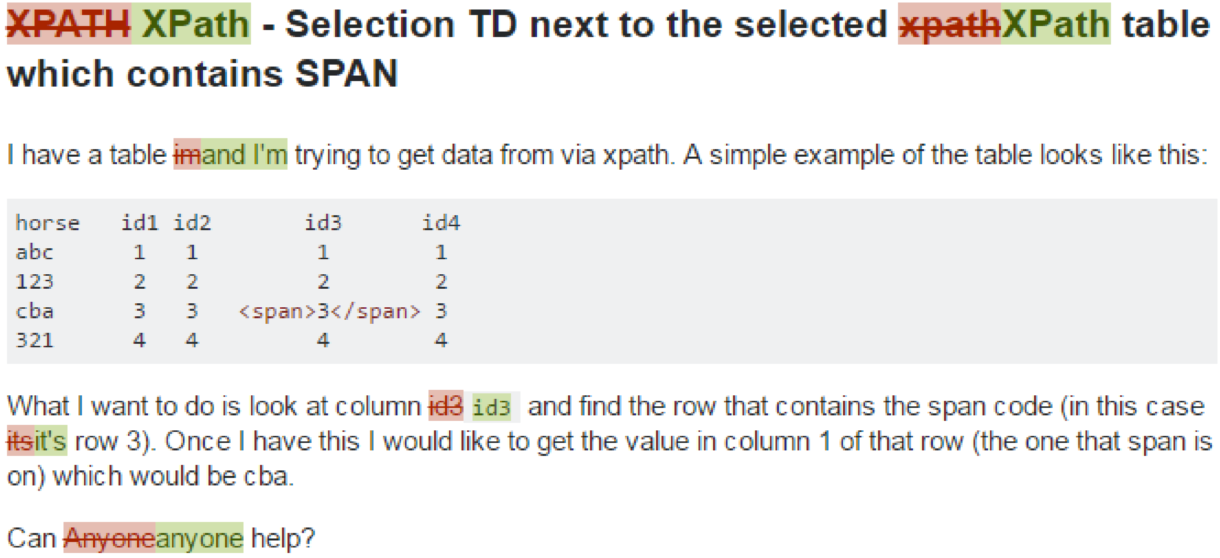
Some common reasons to edit are:

- 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

Observation

Community collaborative edits

- 21,759,565 edits before Dec 2017
  - 1,857,568 (9%) question-title edits
  - 2,622,955 (12%) question-tag edits
  - 17,279,042 (79%) post-body edits



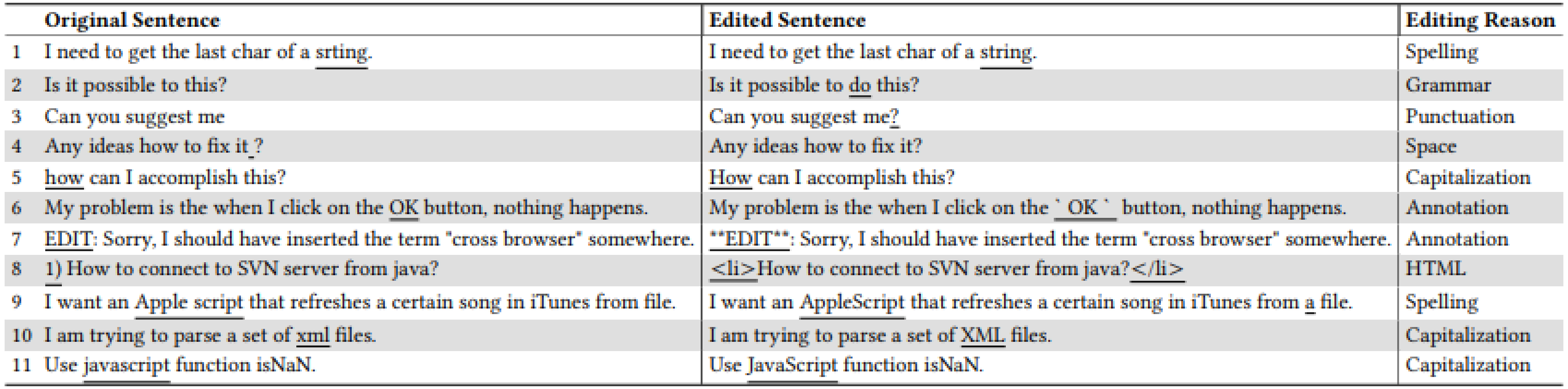
~~XPATH~~ XPath - Selection TD next to the selected ~~xpath~~XPath table which contains SPAN

I have a table ~~im~~and I'm trying to get data from via xpath. A simple example of the table looks like this:

```
horse    id1  id2        id3      id4
abc        1    1          1        1
123        2    2          2        2
cba        3    3   <span>3</span>  3
321        4    4          4        4
```

What I want to do is look at column ~~id3~~ `id3` and find the row that contains the span code (in this case ~~its~~it's row 3). Once I have this I would like to get the value in column 1 of that row (the one that span is on) which would be cba.

Can ~~Anyone~~anyone help?



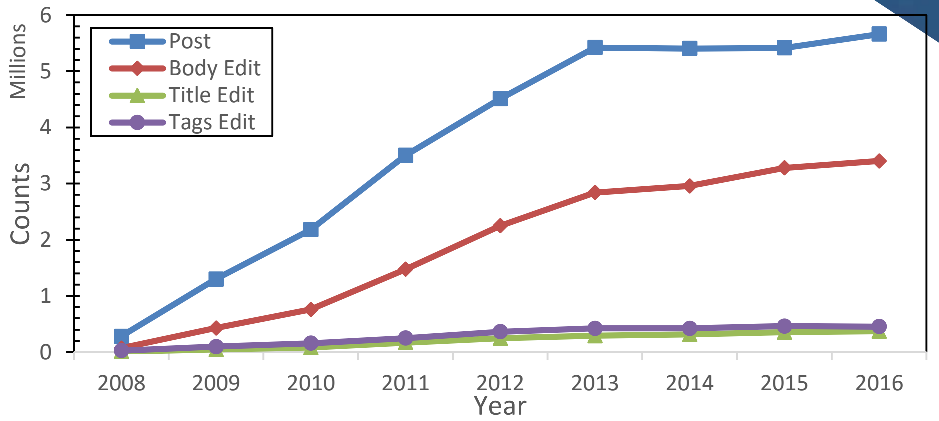
	Original Sentence	Edited Sentence	Editing Reason
1	I need to get the last char of a srting.	I need to get the last char of a string.	Spelling
2	Is it possible to this?	Is it possible to do this?	Grammar
3	Can you suggest me	Can you suggest me?	Punctuation
4	Any ideas how to fix it ?	Any ideas how to fix it?	Space
5	how can I accomplish this?	How can I accomplish this?	Capitalization
6	My problem is the when I click on the OK button, nothing happens.	My problem is the when I click on the ` OK ` button, nothing happens.	Annotation
7	EDIT: Sorry, I should have inserted the term "cross browser" somewhere.	**EDIT**: Sorry, I should have inserted the term "cross browser" somewhere.	Annotation
8	1) How to connect to SVN server from java?	<li>How to connect to SVN server from java?</li>	HTML
9	I want an Apple script that refreshes a certain song in iTunes from file.	I want an AppleScript that refreshes a certain song in iTunes from a file.	Spelling
10	I am trying to parse a set of xml files.	I am trying to parse a set of XML files.	Capitalization
11	Use javascript function isNaN.	Use JavaScript function isNaN.	Capitalization

MONASH University

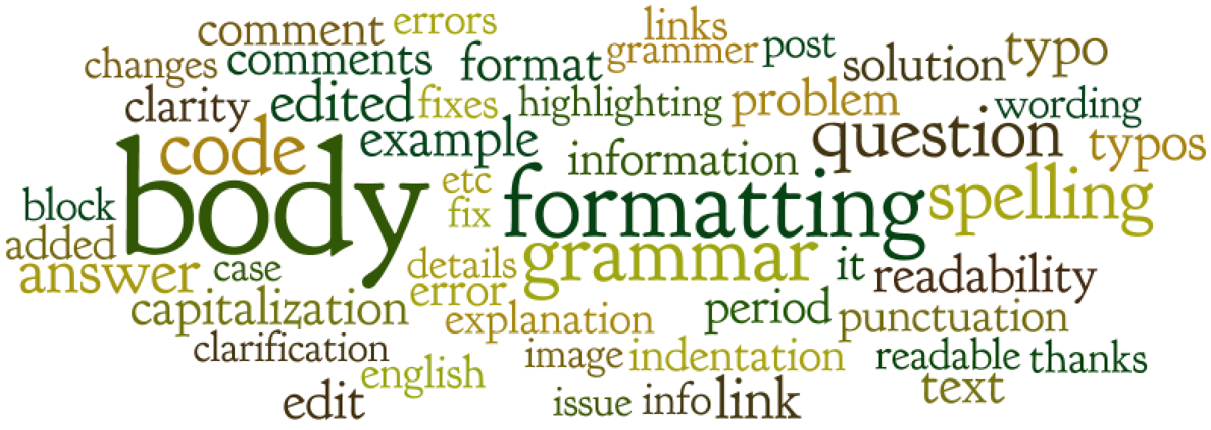
Observation

What has been edited

- Edit type
  - Title, tag and post body



- Edit content
  - https://stackoverflow.com/help/privileges/edit
  - Formatting, spelling, grammar, readability



comment errors links
changes comments format grammer post solution typo
clarity edited fixes highlighting problem wording
code example question typos
information
block body etc fix formatting spelling
added
answer case details grammar it readability
capitalization error explanation period punctuation
clarification image indentation readable thanks
english
edit issue info link text

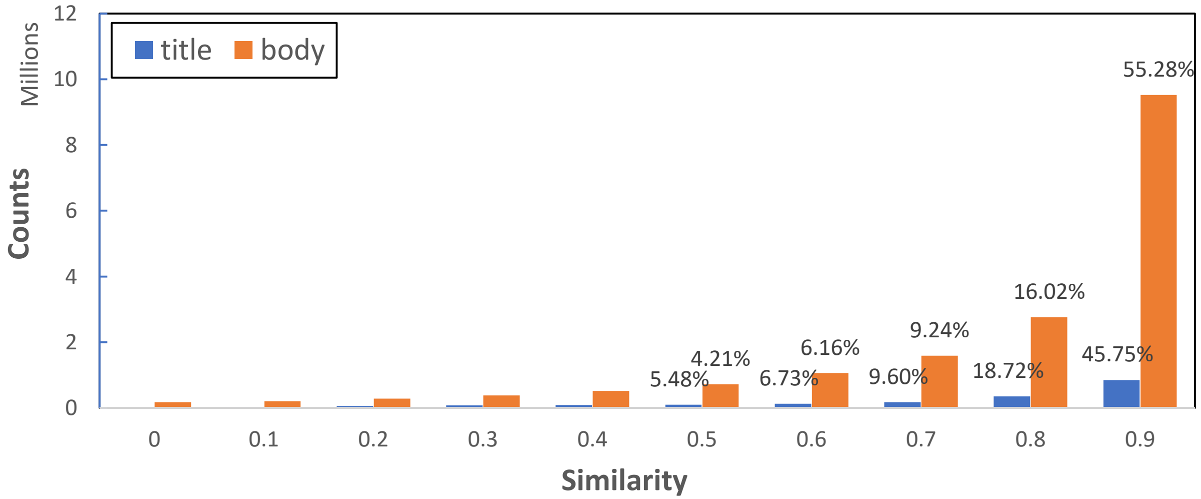
Observation

What is the scale of changes that post edits involve

- Calculate the change scale
  - character-level Longest Common Subsequence (LCS)

$$similarity(original, edited) = \frac{2 * N_{match}}{N_{total}}$$

- Similarity
  - 71.29% post body edits are with similarity score between 0.8 and 1
  - 64.47% post body edits are with similarity score between 0.8 and 1



Goal

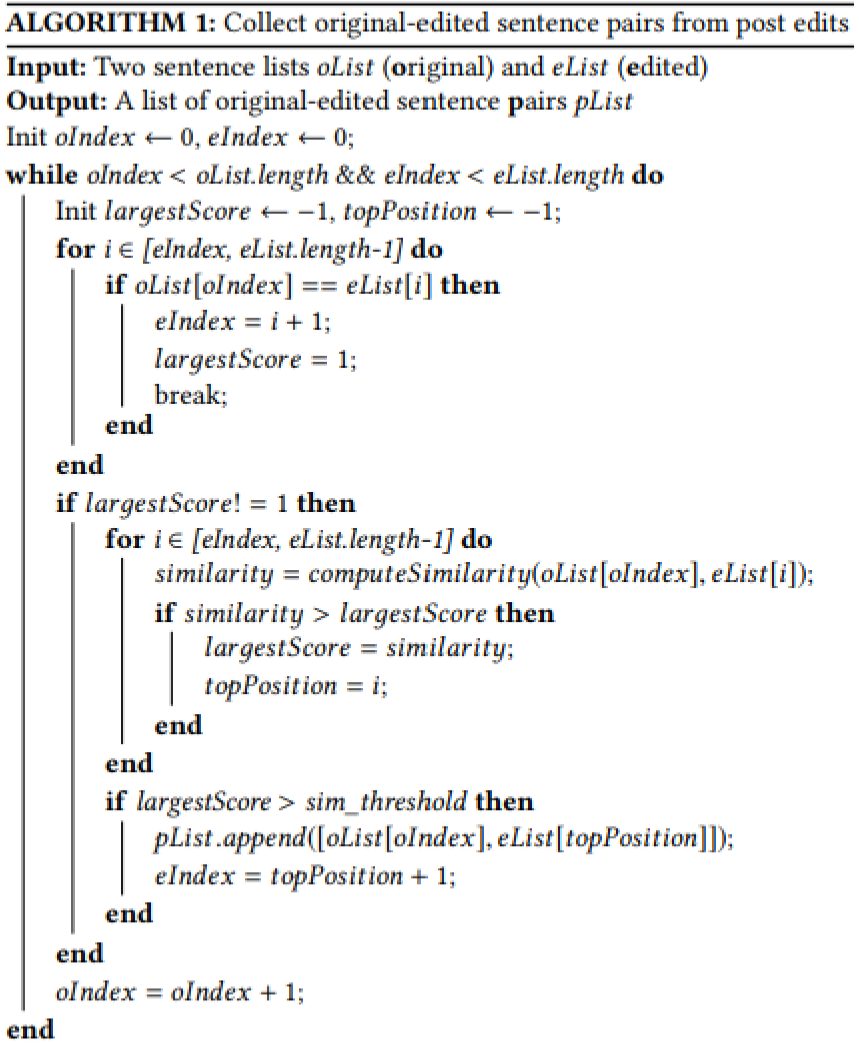
Assist or even automate collaborative edits

- Develop a tool for reminding users about the potential edits
  - **Minor revision:** many posts edits are about spelling, formatting, grammar …
  - **Sentence-level:** most minor revision happen in single sentences.

Data Collection

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

- Regular expression and text differer
- 13,806,188 sentence pairs



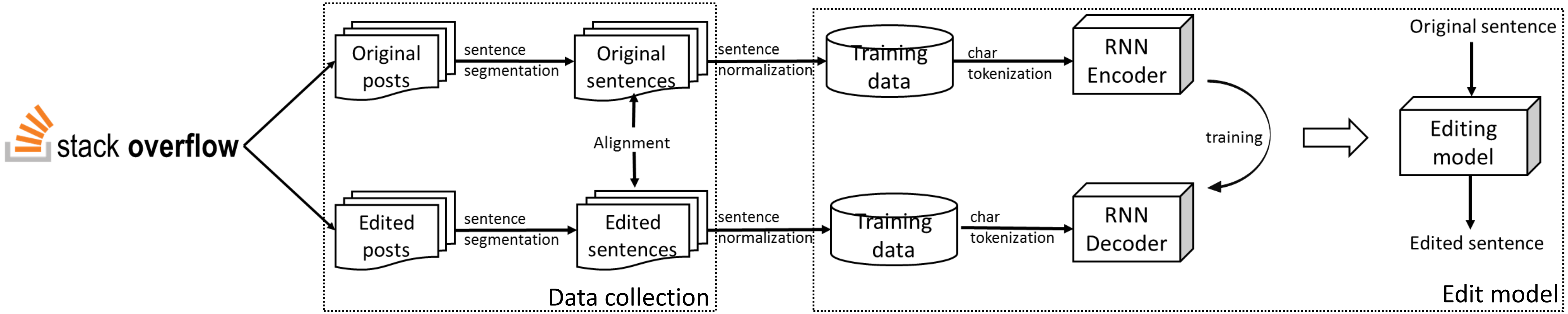
ALGORITHM 1: Collect original-edited sentence pairs from post edits

```
Input: Two sentence lists oList (original) and eList (edited)
Output: A list of original-edited sentence pairs pList
Init oIndex ← 0, eIndex ← 0;
while oIndex < oList.length && eIndex < eList.length do
    Init largestScore ← −1, topPosition ← −1;
    for i ∈ [eIndex, eList.length-1] do
        if oList[oIndex] == eList[i] then
            eIndex = i + 1;
            largestScore = 1;
            break;
        end
    end
    if largestScore! = 1 then
        for i ∈ [eIndex, eList.length-1] do
            similarity = computeSimilarity(oList[oIndex], eList[i]);
            if similarity > largestScore then
                largestScore = similarity;
                topPosition = i;
            end
        end
        if largestScore > sim_threshold then
            pList.append([oList[oIndex], eList[topPosition]]);
            eIndex = topPosition + 1;
        end
    end
    oIndex = oIndex + 1;
end
```

MONASH University

Method

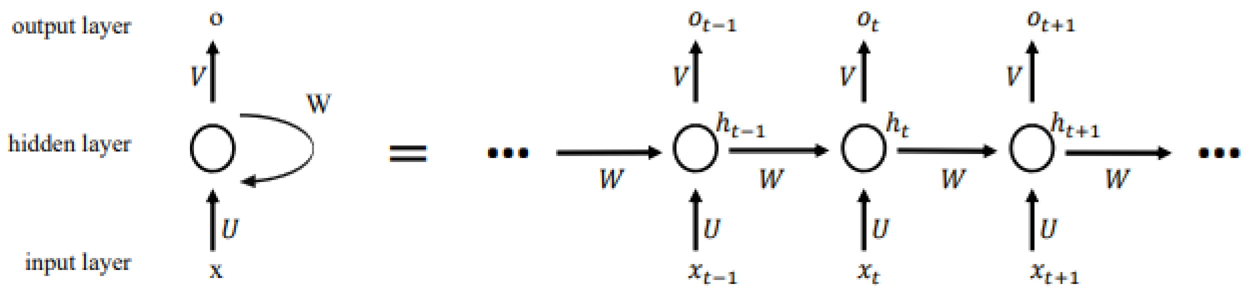
- Data collection
  - 7,545,979 original-edited sentence pairs
- Edit model
  - Char-level seq2seq
  - Copy domain-specific word



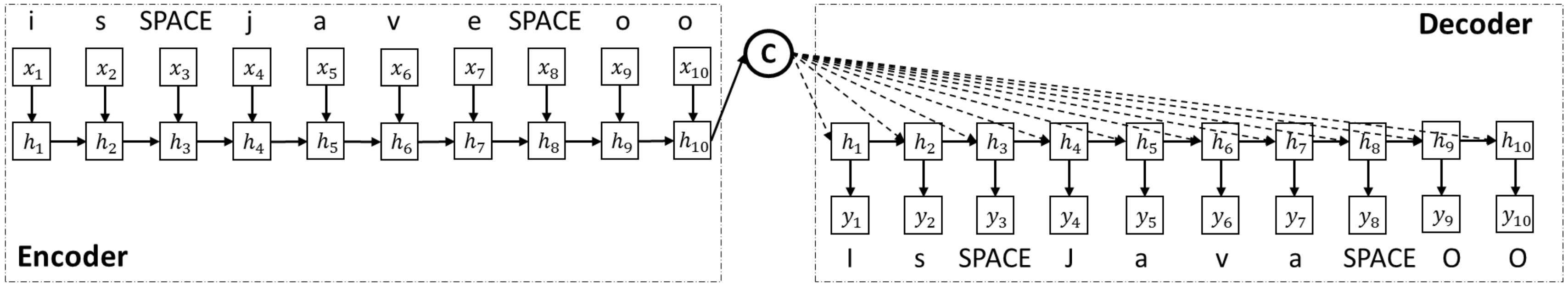
Method

Character-Level RNN Encoder-Decoder Model

- Basic RNN model



- Char-based seq2seq



i s SPACE j a v e SPACE o o
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10}
Encoder
C
Decoder
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10}
y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10}
I s SPACE J a v a SPACE O O

Method

Character-Level RNN Encoder-Decoder Model

- Copy domain-specific words
  - URL, API calls, variable names

Original: pls check https://docs.python.org/2/library/itertools.html#itertools.groupby for itertools.groupby() ...

Preprocess: pls check $\underline{UNK\_URL_1}$ for $\underline{UNK\_API_1}$...

Edited: Please check $\underline{UNK\_URL_1}$ for $\underline{UNK\_API_1}$...

Post-process: Please check https://docs.python.org/2/library/itertools.html#itertools.groupby for itertools.groupby() ...

MONASH University

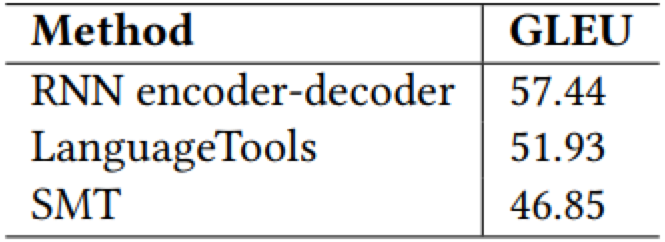
Results

Performance comparison between our model and baselines

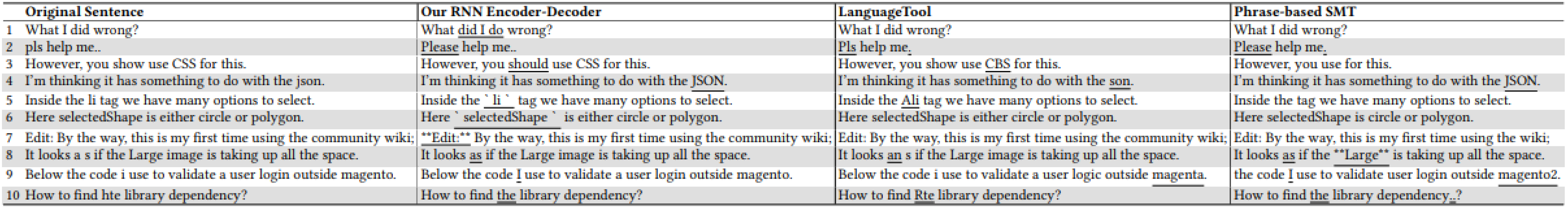
- Evaluation metrics
  - GLEU:

$$GLEU(S,R,C) = BP \cdot \exp(\sum_{n=1}^{N} w_n \log p_n)$$

- Baseline
  - LanguageTools, Statistical Machine Translation (SMT)



Method	GLEU
RNN encoder-decoder	57.44
LanguageTools	51.93
SMT	46.85

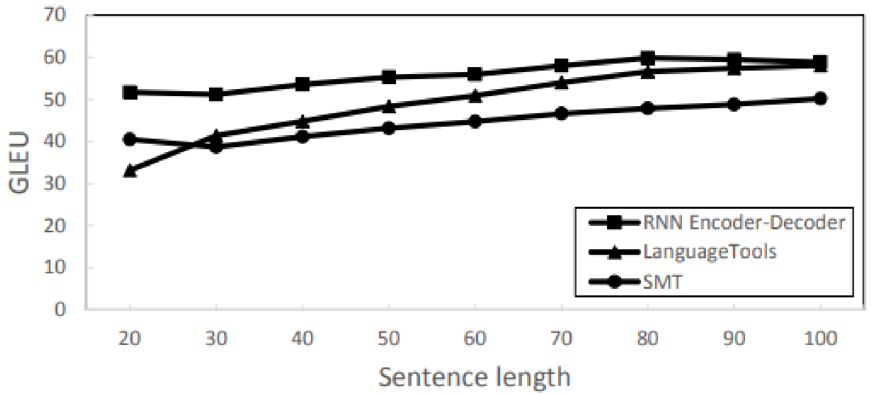


	Original Sentence	Our RNN Encoder-Decoder	LanguageTool	Phrase-based SMT
1	What I did wrong?	What did I do wrong?	What I did wrong?	What I did wrong?
2	pls help me..	Please help me..	Pls help me.	Please help me.
3	However, you show use CSS for this.	However, you should use CSS for this.	However, you show use CBS for this.	However, you use for this.
4	I'm thinking it has something to do with the json.	I'm thinking it has something to do with the JSON.	I'm thinking it has something to do with the son.	I'm thinking it has something to do with the JSON.
5	Inside the li tag we have many options to select.	Inside the ` li ` tag we have many options to select.	Inside the Ali tag we have many options to select.	Inside the tag we have many options to select.
6	Here selectedShape is either circle or polygon.	Here ` selectedShape ` is either circle or polygon.	Here selectedShape is either circle or polygon.	Here selectedShape is circle or polygon.
7	Edit: By the way, this is my first time using the community wiki;	**Edit:** By the way, this is my first time using the community wiki;	Edit: By the way, this is my first time using the community wiki;	Edit: By the way, this is my first time using the wiki;
8	It looks a s if the Large image is taking up all the space.	It looks as if the Large image is taking up all the space.	It looks an s if the Large image is taking up all the space.	It looks as if the **Large** is taking up all the space.
9	Below the code i use to validate a user login outside magento.	Below the code I use to validate a user login outside magento.	Below the code i use to validate a user logic outside magenta.	the code I use to validate user login outside magento2.
10	How to find hte library dependency?	How to find the library dependency?	How to find Rte library dependency?	How to find the library dependency..?

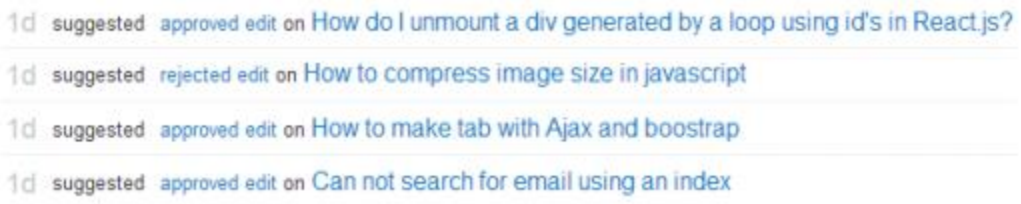
Results

Further evaluation

- Influence by length
  - With the length increase, the baselines are comparable to our model



- Assist real edit
  - Within randomly selected 50 latest posts, our model recommend edits to 39 of them.
  - 36 edits (92.3%) are accepted, while 3 (7.7) rejected



1d suggested approved edit on How do I unmount a div generated by a loop using id's in React.js?

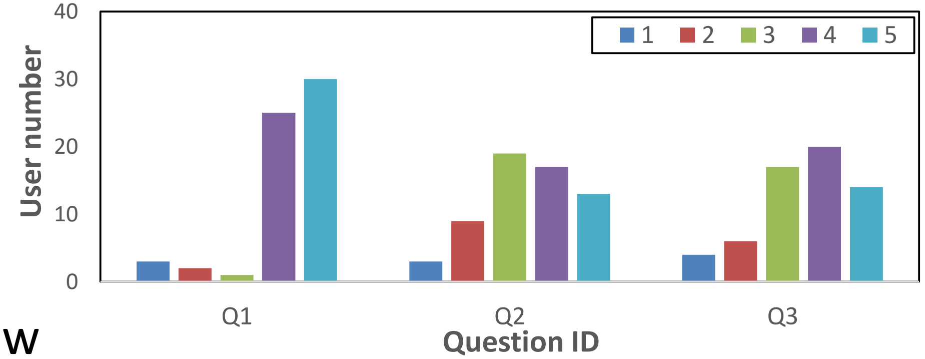
1d suggested rejected edit on How to compress image size in javascript

1d suggested approved edit on How to make tab with Ajax and boostrap

1d suggested approved edit on Can not search for email using an index

Feedback from Real Users

- Survey
  - Q1: How much do yo care about spelling, grammar, formatting edits?
  - Q2: What percentage of your edits are spelling, grammar, formatting edits?
  - Q3: How much could our tool help with such edits?



- Feedback
  - Send email to 410 users who ranked top 2000 in Stack Overflow
  - 61 valid replies
  - Other suggestions:
    - *“SO needs this tool and I hope to see it in action soon. I believe the resulting tool might be useful outside the context of SO websites.”*
    - *“How will it get integrated with the SO site?”*
    - *“Not interested. Same reason I abhor spell and grammar checkers. Generally way too many false positives”*