Appendix to ChatGPT Incorrectness Detection in Software Reviews

Anonymous Author(s)

ACM Reference Format:
Anonymous Author(s). 2024. Appendix to ChatGPT Incorrectness Detection in Software Reviews. In *Proceedings of 46th International Conference on Software Engineering (ICSE 2024)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICSE 2024, April 2024, Lisbon, Portugal

© 2024 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Library Aspect	Collected Post Count
Feature	33
Stability	16
Performance	14
Active Maintenance	11
Documentation	10
Ease of Use	8
Security	8
Total	100

Table 1: Number of Stack Overflow posts collected for each of the seven library-related aspects

1 DATA COLLECTION PROCESS

For collecting the evaluation posts from Stack Overflow, we applied following selection criteria:

1. The post has an accepted answer
    - Reason: we could evaluate the ChatGPT response against the accepted answer when we manually label the dataset.
2. The post has to have a tag of any of three popular text processing libraries spaCy, NLTK, or GSON
    - Reason: since these libraries are popular, we have a fair chance of collecting diversified posts. Also, having a common theme of libraries (text processing libraries) allowed us to select posts based on feature sets of the libraries (that comes as the following criteria)
3. The post must discuss about any of 6 aspects (Active Maintenance, Documentation, Ease of use, Performance, Security, Stability) or 9 features (dependency parsing, Entity linking, Lemmatization, NER, POS, Rule-based matching, Sentence segmentation, Text classification, tokenization) that we listed.
    - Reason: these aspects are commonly used by developers when comparing multiple libraries [40]
4. The question and the accepted answer should not be excessively large (not more than 2500 characters each)
    - Reason: Since we had to manually collect each post and verify that the question and accepted answer actually discussed about our target aspect or feature, we tried to avoid excessively long posts (questions or answers).

With these selection criteria, we used the search interface of Stack Overflow as shown in Figure 1.

We sorted the posts by score so that we get popular and meaningful discussions from where we can find our keyword specific posts. Then we visited each of the posts to manually evaluate the selection criteria. For each of the 6 aspects and 9 features, we manually analyzed up to 30 posts (nearly 500 posts in total) from the search results and selected only the relevant post. Since, the manual search was a labour-intensive process and we ensured diversity of the posts based on the aspects and features, we stopped searching for more posts after reaching 100 posts. The number of posts under each library aspect is shown in Table 1.

2 ENQUIRER PROMPT AND RESPONSE EXAMPLE

The sample provided between lines 908 and 921 in the original paper shows a prompt asked by the Enquirer component and the response from ChatGPT. Due to the lack of space in the paper, we omitted the original Stack Overflow question and answer in the paper by simply referring to them as <SO Question> and <SO Answer>. Similarly, we also referred to the initial base question asked by the Enquirer as <Base Question> and the response of ChatGPT to that base question as <Base Answer>. The actual prompt generated by the Enquirer component is shown in Figure 2. This prompt is represented by the Enquirer prompt shown in the original paper from lines 910 to 915.

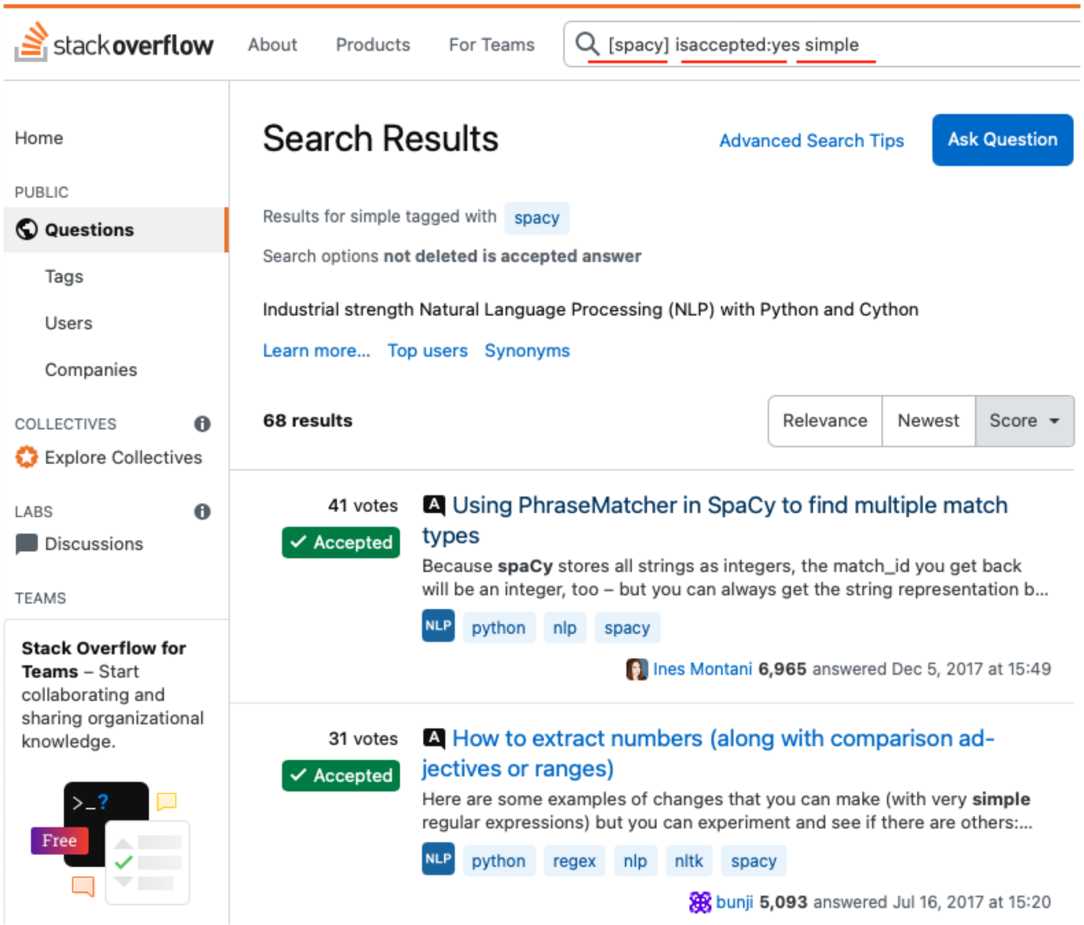
And following was the response from ChatGPT:

```
[
{
"title": "Bug in v2.3.0", "explanation": "The conversation
highlights a specific bug that affects Spacy 2.3.0 and its tagger
training functionality."
},
"title": "Active open-source community", "explanation": "The
recommendation to install the fix from the current master
branch highlights the active community of developers who are
contributing to Spacy's stability and reliability."
{
"title": "Alternative solutions provided", "explanation": "Solutions
such as using the previous version or alternative methods to
train taggers demonstrate a responsive development team and
their attention to user feedback."
},
{
"title": "Upcoming v2.3.1 release", "explanation": "The Spacy
development team is actively working on a fix for the bug,
which indicates regular maintenance and updates of the library."
},
]
```

In the shortened example provided in the original paper (from lines 918 to 921), we mentioned only the first three explanations. The first two explanations were labeled correct (highlighted as tick mark in the paper and in green color here) and the third explanation was labeled incorrect (highlighted as cross mark in the paper and red color here).

3 PROFILE OF PILOT SURVEY PARTICIPANTS

The survey questionnaire was designed based on a pilot survey. In the pilot survey, we provided the questionnaire to 24 industry professionals from big software companies across the world. The recruitment of these professionals was conducted by mostly convenient sampling from the professional network of the authors. Before distributing the survey, we provided them with a short briefing of our intention. Finally, 20 professionals participated in the survey. Based on their survey responses and post-survey feedback, we re-adjusted our final survey. The profile of the participants is shown in Table 2.



[spacy] isaccepted:yes simple

Search Results

Results for simple tagged with spacy

Search options **not deleted is accepted answer**

Industrial strength Natural Language Processing (NLP) with Python and Cython

68 results

41 votes ✓ Accepted

Using PhraseMatcher in SpaCy to find multiple match types

Because **spaCy** stores all strings as integers, the match_id you get back will be an integer, too – but you can always get the string representation b...

NLP python nlp spacy

Ines Montani **6,965** answered Dec 5, 2017 at 15:49

31 votes ✓ Accepted

How to extract numbers (along with comparison adjectives or ranges)

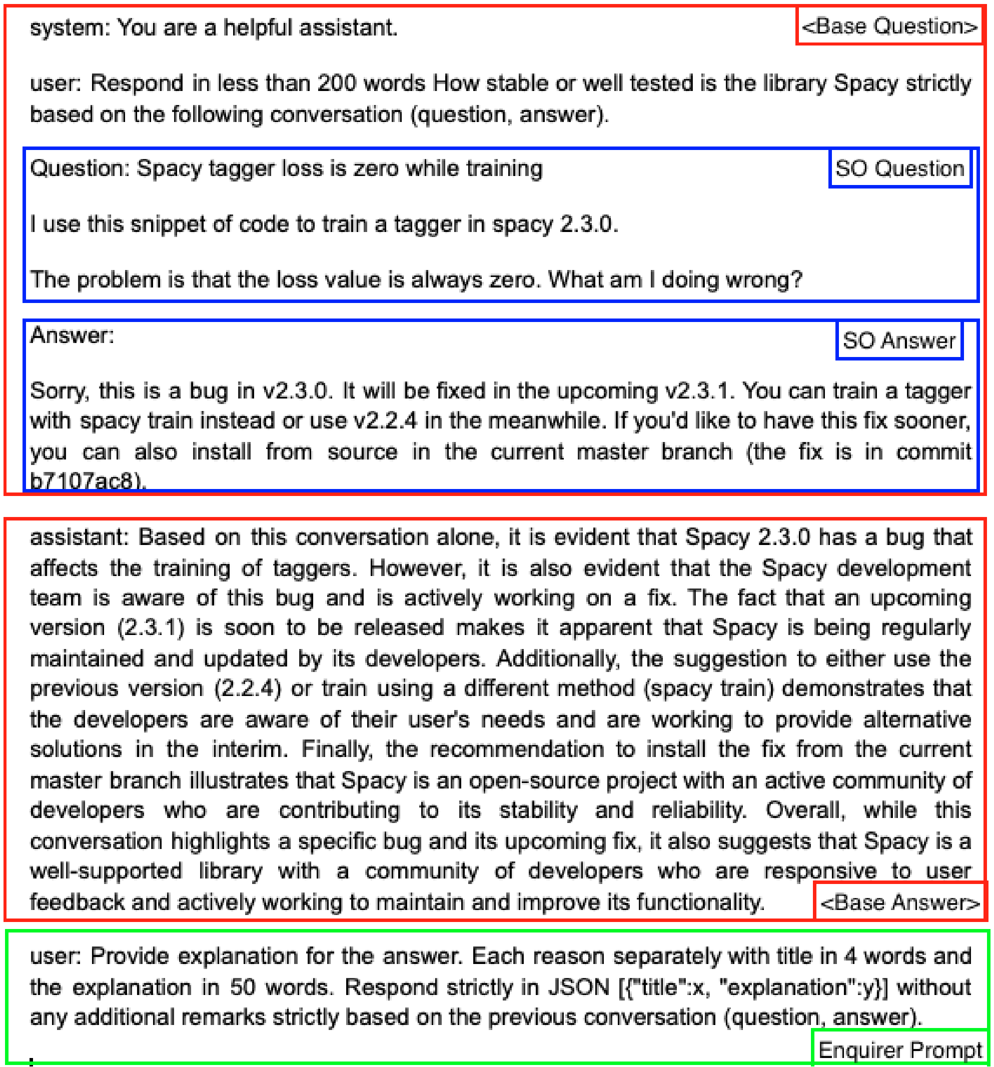
Here are some examples of changes that you can make (with very **simple** regular expressions) but you can experiment and see if there are others:...

NLP python regex nlp nltk spacy

bunji **5,093** answered Jul 16, 2017 at 15:20

Stack Overflow for Teams – Start collaborating and sharing organizational knowledge.

Figure 1: Stack Overflow search interface for collecting posts with given criteria



system: You are a helpful assistant.

user: Respond in less than 200 words How stable or well tested is the library Spacy strictly based on the following conversation (question, answer).

Question: Spacy tagger loss is zero while training

I use this snippet of code to train a tagger in spacy 2.3.0.

The problem is that the loss value is always zero. What am I doing wrong?

Answer:

Sorry, this is a bug in v2.3.0. It will be fixed in the upcoming v2.3.1. You can train a tagger with spacy train instead or use v2.2.4 in the meanwhile. If you'd like to have this fix sooner, you can also install from source in the current master branch (the fix is in commit b7107ac8).

assistant: Based on this conversation alone, it is evident that Spacy 2.3.0 has a bug that affects the training of taggers. However, it is also evident that the Spacy development team is aware of this bug and is actively working on a fix. The fact that an upcoming version (2.3.1) is soon to be released makes it apparent that Spacy is being regularly maintained and updated by its developers. Additionally, the suggestion to either use the previous version (2.2.4) or train using a different method (spacy train) demonstrates that the developers are aware of their user's needs and are working to provide alternative solutions in the interim. Finally, the recommendation to install the fix from the current master branch illustrates that Spacy is an open-source project with an active community of developers who are contributing to its stability and reliability. Overall, while this conversation highlights a specific bug and its upcoming fix, it also suggests that Spacy is a well-supported library with a community of developers who are responsive to user feedback and actively working to maintain and improve its functionality.

user: Provide explanation for the answer. Each reason separately with title in 4 words and the explanation in 50 words. Respond strictly in JSON [{"title":x, "explanation":y}] without any additional remarks strictly based on the previous conversation (question, answer).

Figure 2: Original prompt generated by Enquirer component that was submitted to ChatGPT API.

Table 2: Interview participants by reference ID, years in role, primary technology, geographic location (AS-Asia, AU-Australia, EU-Europe, NA-North America, SA-South America), industry, and company size.

P#	Years	Role	Tech	Location	Industry	Size
P01	12	Architect	Java	EU	Automotive	500
P02	6	SDE (Software Dev Engineer)	Python	NA	Cloud Service	80,000
P03	12	SDE	A/IOS	NA	Automotive	600
P04	20	CEO (Chief Executive Office)	.NET	AS	Broadcast Media	54
P05	16	EM (Engineering Manager)	.NET	AU	Financial	12
P06	17	SDE	Perl	EU	Tech	20
P07	9	CTO (Chief Technical Office)	Javascript	NA	Data Analytics	6
P08	9	EM	Any	NA	Cloud Service	10,000
P09	13	Architect	Python	NA	Web	100
P10	15	EM	Javascript	EU	Energy	300
P11	7	MLE (Machine Learning Engineer)	Python	NA	Data Analytics	30
P12	22	Consultant	Perl	AS	Tech	1,000
P13	15	Architect	Java	NA	Retail	200,000
P14	6	SDE	Mobile (Android/IOS)	AS	Financial	100
P15	22	CTO	.NET	AS	Enterprise	300
P16	9	SDE	Java	AS	Cyber Security	300
P17	15	CTO	Ruby on Rails	EU	Custom Software	6
P18	27	CEO	C++	NA	Financial	40
P19	15	EM	Ruby on Rails	NA	Cloud Service	75,000
P20	10	SDE	Mobile	NA	Food Service	10
P21	13	SDE	Ruby on Rails	NA	CI/CD	900
P22	30	Architect	Java	NA	Operating Sys.	9,000
P23	7	MLE	Python	SA	Custom Software	750
P24	6	MLE	Python	NA	Medical	80