LLM of Babel: An analysis of the behavior of large language models when performing Java code summarization in Dutch

SUPERVISORS

Prof. Dr. Arie van Deursen

Assistant Prof. Dr. Maliheh Izadi

Ir. Jonathan Katzy

AUTHOR

Gopal-Raj Panchu
gpanchu@student.tudelft.nl

AFFILIATIONS

Delft University of Technology

01. Introduction

How well do large language models (LLMs) infer text in a non-English context when performing code summarization? The goal of this paper was to understand the mistakes made by LLMs when performing code summarization in Dutch. We categorized the mistakes made by CodeQwen1.5-7b when inferring Java code comments in the Dutch language through an open coding methodology to create a taxonomy of errors by which to categorize these mistakes

02. Research Questions

RQ1: What are the most common mistakes made by LLMs in the context of Java code summarization in Dutch?
RQ2: Are BLEU and ROUGE-L metrics reliable indicators for the accuracy of code summarization inferences?
RQ3: To what extent does the multilingual code summarization error taxonomy overlap with existing error taxonomies in machine translation and code summarization?

03. Methodology

03.01 Processing

A dataset[1] was generated by scraping github for code files containing Dutch comments

Exclusion Criteria

- **Autogenerated comments**
- **Comments with < 3 words**
- **Files with > 8192 tokens**
- **Files with no comments**

We used DelftBlue [2] to run a huggingface inference pipeline:

1. **Identify comments with regex**
2. **Span-mask the comments**
3. **Fill-in-the-middle inference using CodeQwen1.5**

03.02 Qualitative analysis

- BLEU-1[3]
- ROUGEL[4]

03.03 Quantitative analysis

1. **Classify 200-300 code comments with open coding**
2. **Identify outliers, inclusion criteria and categories**
3. **Determine inference accuracy, inaccurate(0), partially accurate(1), accurate(2)**
4. **Improve the taxonomy with the new findings**
5. **Repeat**
6. **Final taxonomy 600 comments**

03.04 Taxonomy comparison

Leaf nodes were compared with:

- **Mahmud et al[5].**
- **Sharou and Specia[6].**
- **Huidrom and Belz[7].**

04. Results

Accuracy occurrences

- **Fully accurate: 131**
- **Partially accurate: 287**
- **Inaccurate: 182**

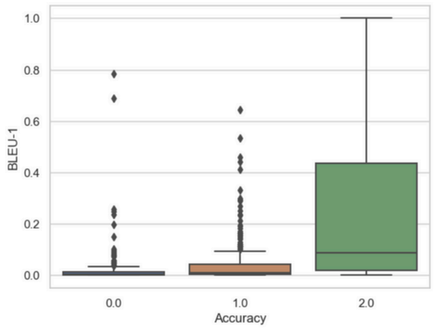

figure 1: BLEU-1 per accuracy group

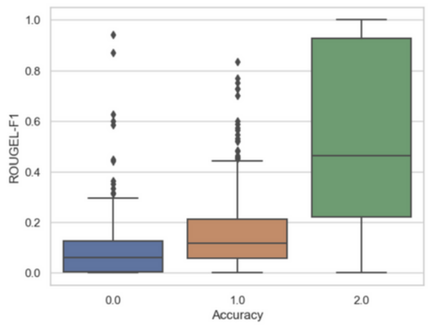

figure 2: ROUGEL-F1 against accuracy group

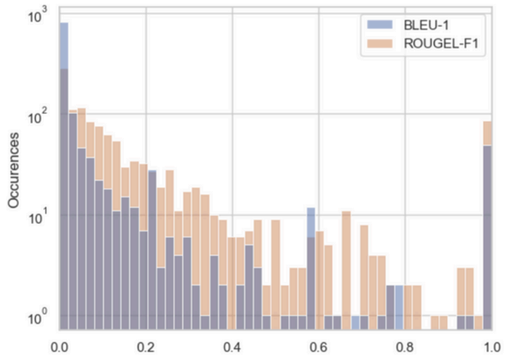

figure 3: BLEU-1 and ROUGEL-F1 occurrence counts

Table 1: Taxonomy of errors including occurrence count and number of overlapping leaf categories in compared taxonomies

Failure category plus label ID	Count	Mahmud et al.[5]	Sharou and Specia[6]	Huidrom and Belz[7]
MS Model-oriented Errors	**398**	-	-	-
MS-IG Incoherent Generation	3	-	-	1
MS-CC Copy context	58	-	-	2
MS-ME Memorization	13	-	-	-
MS-ME1 PII	9	-	-	-
MS-ME2 URL	3	-	-	-
MS-ME3 Training Memorization	1	-	-	-
MS-ET Early Termination	50	8	-	-
MS-LT Late Termination	107	-	-	1
MS-NG No Generation	0	-	-	-
MS-RE Repetition	167	-	-	-
MS-RE1 Pattern Repetition	57	2	-	1
MS-RE2 Verbatim Repetition	110	2	-	1
LG Linguistic Error	**66**	-	-	-
LG-GR Grammar	33	2	-	-
LG-GR1 Plurality	1	-	-	-
LG-GR2 Conjugation	2	-	-	-
LG-GR3 Gender	6	-	-	-
LG-GR4 Language Syntax	7	-	-	-
LG-GR5 Capitalization	1	-	-	-
LG-GR6 Cohesion	16	-	1	3
LG-IS Incorrect synonym	0	-	-	2
LG-WL Wrong language	33	-	-	-
LG-WL1 Undesired translations	5	-	1	1
LG-WL2 Incorrect language	28	-	1	1
SE Semantic error	**520**	-	-	1
SE-MD Missing Details	38	15	1	1
SE-TS Too specific	16	3	-	-
SE-HA Hallucination	290	-	-	-
SE-HA1 Misplaced Facts	33	1	3	1
SE-HA2 Out of Context	21	1	2	1
SE-HA3 In context	236	3	3	1
SE-CS Completion includes code	185	-	-	-
SE-CS1 Code commented out	30	-	-	1
SE-CS2 Code intended to run	155	-	-	1
ST Syntax	**8**	-	-	-
ST-IF Incorrect comment format	8	-	-	-
ST-IF1 Comment Syntax	1	-	-	-
ST-IF2 Omitted Identifier	7	-	-	-
+ Fully accurate	**131**			

Responsible Research:

- Dataset was open-source
- Inclusion and exclusion criteria
- All used code is openly available

05. Conclusion

- Four main categories of errors: **Linguistic, Semantic, Syntactic, and Model-behaviour**
- Most common sub-categories were related to **Repetition, Code snippets, and Hallucinations.**
- The created taxonomy had **significant overlap with similar taxonomies** but **left semantic gaps regarding "Toxicity" and "User safety"** in addition to in-depth **linguistic analysis.**
- **BLEU-1 and ROUGEL-F1 scores** are **generally unreliable** for accuracy evaluation in this context

07. References

1. Jonathan Katzy. Llm-of-babel-nl2. https://huggingface.co/datasets/AISE-TUDelft/LLM-of-Babel-NL2. [Accessed 2024-05-20].
2. Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase2). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2, 2024.
3. Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002.
4. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.
5. Junayed Mahmud, Fahim Faisal, Raihan Islam Arnob, Antonios Anastasopoulos, and Kevin Moran. 2021. Code to Comment Translation: A Comparative Study on Model Effectiveness & Errors. In Proceedings of the 1st Workshop on Natural Language Processing for Programming (NLP4Prog 2021).
6. Khetam Al Sharou and Lucia Specia. 2022. A Taxonomy and Study of Critical Errors in Machine Translation. In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation.
7. Rudali Huidrom and Anya Belz. 2022. A Survey of Recent Error Annotation Schemes for Automatically Generated Text. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM).