A Typology of Errors for User Utterances in Chatbots

*Esmé Manandise[1], *Anu Singh[2]
[1]Parallel Cities,[2] Intuit AI Research
[1]Brussels, Belgium,[2] Mountain View, CA, USA
esmeman@comcast.net, anu_singh@intuit.com

Abstract

This paper discusses the challenges non-prescriptive language uses in chatbot communication create for Semantic Parsing (SP). To help SP developers improve their systems, we propose a flexible error typology based on an analysis of a sample of non-prescriptive language uses mined from a domain-specific chatbot logs. This typology is not tied to any specific language model. We also present a framework for automatically mapping errors to the typology. Finally, we show how our framework can help evaluate SP systems from a linguistic robustness perspective. Our framework can be expanded to include new error classes across different domains and user demographics.
Keywords: chatbot utterances, language prescription, semantic parsing, typology, errors

1. Introduction

Semantic Parsing (SP), which converts utterances to symbolic forms to represent meaning, fails to analyze utterances which are not modeled by underlying language models or grammars (Huang et al., 2021; Manandise and de Peuter, 2020; Zhang et al., 2020). Ideally, robust parsing would produce analyses even when utterances are not well-formed. In practice, as there are no 100% parsing guarantees, a language framework in which SP is integrated might rely on additional processes such as Grammar Error Correction (GEC) to rectify input prior to parsing. For example, the *English Resource Grammar* (Copestake and Flickinger, 2000; Oepen and Flickinger, 2019) and its parser *Answer Constraint Engine* are highly sensitive to prescriptive English. Consider the utterances in Table 1. While the prescriptive utterance (1a) parses accurately, (1b) without a determiner in front of the singular noun '*daughter*' fails to parse. Correcting '*daughter*' into its plural form in (1c) or substituting it by a mass noun like '*bread*' in (1d) ensure parsing. English grammar sanctions determiner-less nominal heads for plural or mass nouns.

(a) Can I claim my daughter as a dependent?
(b) Can I claim daughter as a dependent?
(c) Can I claim daughters as a dependent?
(d) Can I claim bread as a dependent?

Table 1: Some Non-/Prescriptive Utterances

Many utterances produced on-the-fly by users while engaging with a chatbot[1] stray from prescriptive English. Consider the utterances in Table 2. Though interpretable by human agents, these utterances *in the wild* are difficult for SP to parse (Jurcıcek et al.; Huang et al., 2021; Manandise and Srivastava, 2022). Even after GEC, utterances can remain non-prescriptive. For instance, a T5-based GEC[2] for spelling and grammar corrects some tokens in 2(d), respectively, '*eligibne*' to '*eligible*', '*vertion*' to '*version*', and '*wwbsitw*' to '*website*'. However, word-level misspellings still remain, and the syntactic ungrammaticalities in 2(a-d) persist *as is*.

(a) ho*w d*elete the state return?
(b) to eliminate*s* fee down grade
(c) tax return not free why
(d) I am *eligibne* for free *vertion byt* the *wwbsitw* keep telling me *I to* sign up for deluxe to *dile* why *I* being charged for *Turbm* Tax

Table 2: Examples of Real-Time User Utterances

In Computational Linguistics, typologies of errors are established artifacts. Typically, these are tied to specific end tasks and defined to measure the quality of automatically-generated output. For instance, in machine translation (MT) [3] (Lommel, 2018), a typology of errors helps set the features relevant to measure the accuracy of the machine output given a source text. Our typology of errors does not classify inaccurate parses output by SP, but rather the language-specific features in the source that contribute to SP failures.

Our Main Contributions:

- We present a typology that categorizes errors

*Both authors contributed equally to this work.

[1]The customer-service chatbot in our study is specific to the tax domain. See section 2.

[2]A T5-based GEC trained on C4_200M dataset

[3]http://qt21.eu/mqm-definition/

LREC-COLING 2024, pages 1789–1794

and non-prescriptive language uses (NPLUs) discovered during our experiments with four symbolic SP systems. Table 3 displays our expandable typology of errors and NPLUs.

- Our proposed typology has multiple advantages as it is model-independent, serves as a template for error analysis in SP and GEC systems. It can guide SP development prioritization based on linguistic robustness, be adapted to new domains and languages, and can aid in language model selection through error distribution analysis.
- We develop an automated error classification framework using a Large Language Model (LLM) to identify errors in NPLUs and categorize them according to our proposed typology. Furthermore, we present an analysis of the parse failures of the SP systems used in our study using this framework.

In the following sections, we present our analysis of SP failures, which laid the foundation to create an extendable error typology and a system for automated error analysis.

2. Background

We analyze the chat logs of the TurboTax Virtual Assistant (TTA), a chat application embedded in the TurboTax[4] Online product. TurboTax customers engage in writing in English via free-form input texts with TTA to get help on various product-, tax- and customer-specific questions relevant to Canada and United States tax filing.

The current TTA is **intent-based**. TTA does not always pair queries with accurate pre-existing intents or know what to do with some queries, labeling them as '*fallback*'. Other venues for interpreting TTA '*fallback*' queries would be to perform deeper semantic analysis with SP. However, SP relies on prescriptive language uses.

Our experiments using a corpus of 34K non-prescriptive utterances show that, after GEC, 43% of the utterances remain irreparable. For chatbots that rely on SP for query interpretation this class of utterances contribute to parse and, consequently, conversation failures.

With the goal of assessing the robustness of SP against the '*fallback*' queries, we experimented with four symbolic SP frameworks. For our study, we chose symbolic rather than neural SP as NPLUs are out-of-distribution examples, which are more problematic for neural SP. As discussed in (Huang et al., 2021) neural SP is also challenged by input quality (such as example 1(b) in Table 1); the robustness of SOTA neural SP drops by 10%-15% in the presence of meaning preserving perturbed utterances. Contrary to neural SP, symbolic SP allows for insights into the underlying meaning-composition process of the utterances.

(Qorib and Ng, 2022) presents robustness analysis of GEC systems. While some SOTA GEC systems such as gT5 (Rothe et al., 2021) and GECToR (Omelianchuk et al., 2020) outperform humans by a wide margin on the CoNLL-2014 benchmark test set, there remains classes of errors that GEC systems fail to correct. For instance, with double negation '*I didnt enter nothing in box c*' in Table 6, the T5-based GEC model corrects the utterance into '*I didnt enter anything in box c.*'. However, without additional user sociolinguistic markers, GEC ignores the possibility of negative concord where two negation elements for a user of non-standard English is to be interpreted as a single negation '*I entered something in box c.*'.

Class/Level	Word	Phrase	Utterance
Orthography	Spelling, Capitalization, Insertion, Omission, Abbreviation, Acronym		Capitalization, Erroneous Punctuation, Omitted Punctuation
Lexicon	Idiom, Negation, Out-of-Vocabulary	Word Order, Idiom, MWE, Out-of-Vocabulary, Nominal Compound	Coordination, Subordination
Morphology	Affix, Comparative, Superlative	Case, Deverbal Noun, Agreement, Tense, Negation	Agreement, Tense, Negation
Grammar	Part of Speech, Omission	Case, Omission, Comparative, Superlative, Subject Omission	Coordination, Prepositional Attachment, Word Order, Ellipsis, Anaphora, Apposition, Topicalization
Semantics	Sense, Fuzzy Choice, Negation, Temporal Reference, Spatial Reference, Pronoun	Double Negation, Quantification, Temporal Reference	Double Negation, Negative Concord

Table 3: Our Proposed Error and NPLU Typology

[4] https://turbotax.intuit.com/

Orthography
• **Spelling:** i have Medicare Health insursnce • **Abbreviations, Acronyms:** Sch B Prt 2 • **Character Insertion:** can5t access return • **Space Insertion:** re load w2 • **Spelling Variants:** Trying to add a W 2 from my 2nd job
Grammar: Part of Speech
• in 0000 I work for an employer and self-employed whats the best option
Semantics: Fuzzy Choice
• I have other job and I thought I could send in my w2s separately

Table 4: Examples of NPLUs at Word-Level

3. Linguistic Classification of Errors and NPLUs

In order to discover problematic utterances, we conducted SP experiments over the TTA *fallback* dataset to surface which—not how many, fragmented utterances and NPLUs parse successfully, break parsing, and necessitate correction to prevent parse failures.

We used four off-the-shelf symbolic SP frameworks, namely *(i) Abstract Meaning Representation (AMR) (Banarescu et al., 2013; Damonte et al., 2017), (ii) Combinatory Categorical Grammar (CCG) (Martínez-Gómez et al., 2016), (iii) Minimal Recursion Semantics (MRS) (Copestake et al., 2005), (iv) Frame Semantics (Chanin, 2022)*; and one GEC framework. These semantic parsers each have their strengths and weaknesses. For instance, while an MRS-based parser fails to parse '*Can I claim daughter as a dependent?*', the AMR-based parser succeeds.

3.1. Formulating a Typology of Errors and NPLUs

For the typology generation (see Table 3), we randomly selected 10K utterances from the 34K *fallback* set that the four parsers all fail to parsec[5]. The 10K utterances were manually annotated by 4 human annotators to expose specific linguistic problems. The descriptions were consolidated to select linguistic labels. After several iterations of human-annotated linguistic analyses of the utterances, 3 error levels and 5 top classes were identified. Each class subsumes any number of sub-

[5]Given the small size of our working corpus (10K utterances with NPLUs that the 4 symbolic SP systems in our study **all** fail to parse), we consider NPLUs with low frequency of occurrence. For our typology, NPLUs are independent of their quantitative footprint.

Lexicon
• **Word Order:** identification employer num • **Idiom:** bank account wrong routining number
Morphology
• **Deverbal Noun:** didn't receive unemploy compensation or family leave • **Tense:** my daughter college living expenses apartment food car insurance can be add • **Comparative:** continue when I get my data Is that a gooder plan
Grammar
• **Comparative:** I've estimated more closer to 0000 Can I get support with where such difference is emerging • **Omission + WO:** tax return not free why • **Subject Omission:** claim daughter as a dependent
Semantics: Temporal Reference
• I just update my TurboTax 0000 program and it wants the serial number

Table 5: Examples of NPLUs at Phrase-Level

Orthography
• **Omitted Punctuation:** Please help me retreive my W2 from last year I've downloaded 2x and now I'm getting frustrated • **Erroneous Punctuation:** I am trying to enter my state tax information , for box 00 of my W0 , there is no texas in the drop down menu need to enter texas
Morphology: Agreement
• I updates my TurboTax 0000 program and it wants the serial number Why
Grammar
• **Apposition:** i uploaded file, my return • **Coordination:** my daughter college living expenses apartment food car insurance can be add • **Word Order:** taxes paid get proof how • **Prepositional Attachment:** to account update didnt apply • **Ellipsis:** in 0000 I work for an employer and self-employed whats the best option for me to file
Semantics: Double Negation
• I didnt enter nothing in box c.

Table 6: Examples of NPLUs at Utterance-Level

Level	Class	Subclass
92.88%	90.47%	75.78%

Table 7: Accuracy of the LLM-Based Classifier

Class/Level	Word	Phrase	Utterance
Orthography	18.18%	0.37%	0.21%
Lexicon	1.89%	3.54%	0.04%
Morphology	0.66%	1.66%	0.26%
Grammar	17.99%	18.89%	8.99%
Semantics	5.7%	15.98%	2.28%

Table 8: Error Distribution in NPLUs (10K) Using Our Automated Analysis

classes that relate to linguistic problems encountered in the data; subclasses are expandable and not fixed, to allow for new data discovery and additonal coverage.

An error in an utterance can exist at three levels, namely *word, phrase, and utterance*. A single utterance can have errors and NPLUs at all 3 levels and in more than 1 class and subclass. ***Word level*** refers to errors within single-token words; for instance, wrong characters within a word or when a word is erroneously split (see Table 4). ***Phrase level*** refers to expressions that have spaces like multi-word expressions (MWE) as well as to syntactic phrases (see Table 5). ***Utterance level*** refers to errors and NPLUs that affect the grammar of an utterance, namely, errors and NPLUs that are distributed throughout the utterances (see Table 6). The errors in the examples in Tables 4, 5, and 6 are highlighted in red.

This typology allows for the training of models and designing of processes to address domain-specific errors. Although it may not cover all errors, it can identify which ones disrupt SP. Moreover, this error classification can be customized not only for a specific domain but also for non-standard or non-mainstream language varieties.

3.2. From Typology to Automated Error Analysis of Semantic Parsing Failures

We developed an automated approach to classify errors and NPLUs per our proposed typology using a Large Language Model (LLM), specifically we used GPT-3.5 Turbo[6].

To perform error classification, we instruct the LLM via a prompt which consists of:

- Typology of errors
- Few-shot examples (6 in our case) for mapping of utterances to errors
- Utterance

[6] https://platform.openai.com/docs/models/gpt-3-5

Subclass	Error %
Part of Speech	19.77
Omission	14.9
Subject Omission	2.29
Coordination	4.01
Prepositional Attachment	0.29
Word Order	16.91
Ellipsis	4.58
Anaphora	0.86
Apposition	1.15

Table 9: Subclass Error Distribution for Errors in Grammar Class at All Levels

Class/Parser	AMR	CCG	MRS	Frame
Orthography	19.91%	18.03%	18.31%	12.16%
Lexicon	6.21%	5.99%	5.63%	4.05%
Morphology	2.22%	2.24%	4.22%	1.35%
Grammar	44.47%	44.67%	35.21%	47.29%
Semantics	23.79%	25.36%	36.62%	31.08%

Table 10: Error Distribution of Failures for 4 Symbolic SPs Using Our Automated Error Analysis

We report the accuracy of our error classifier (see Table 7) for Level, Class, and Subclass, using the ground truth we curated (10K samples) in the typology creation process.

Table 8 shows the distribution of errors in NPLUs (10K datasize) identified using our error classification framework. Table 9 displays the subclass errors at the '*Grammar*' class at all levels. We used our automated framework to gather statistics (see Table 10) on the error classification of semantic parsers[7]. For example, 23.79% of the AMR-based parse failures are due to semantic errors. Frame Semantics-based parser failures are mostly due to grammatical errors (47.29%).

These results can provide useful insights into the weaknesses of a SP system, and be suggestive of areas for improvement and extensions.

4. Conclusion

The typology can be extended by analyzing errors in more chat data sources. Automated tooling can be developed to integrate robustness metrics in development of SP systems based on a typology. Aided by a typology of errors and NPLUs, a venue could be to model utterance productions as variants of a *standard* language. While some NPLUs are adhoc and hard to predict no matter the size of a corpus, some NPLUs might fall within some parameters of a linguistic class or subclass.

[7] The numbers reported in Table 10 are for 5K of previously unseen utterances from the *fallback* dataset.

Limitations

In this paper, we focused on data selected from Turbotax chatbot logs. In our prior experiments, we also analyzed chatbot logs from the Quickbooks accounting app. Annotation efforts for both Turbotax and Quickbooks logs showed similarities in the non-prescriptive language uses for English. However, as there are many more TurboTax than Quickbooks customers, we discovered error classes that were not present in the Quickbooks chat logs. With more seed error classes from TurboTax, we were able to expand a typology of errors to experiment with.

While Turbotax chatbot has a narrow focus of helping users file their taxes, the users often embark on exchanges with the chatbot where they share personal stories and information as part of the context to file taxes. These exchanges are broad enough to be tax-independent and representative of general conversations with chatbots. The range of exchanges vary from tax-specific content to raging about their lives and their tax status.

It will be useful to consider more chat data sources to extend the error typology proposed in this paper.

Ethics Statement

This work has been conducted keeping in mind best ethical practices. The selection of the data used in our experiments (34k chats) was purely random. The only fix points for the retrieval of chats were dates (peak usage) and tax years. The chatbot allows free text input with no editing of the input. For annotation, full-time specialists were not used. The annotators consisted of computer scientists (engineers and computational linguists) working on the project. The annotators were either native speakers of English or highly proficient in English as a second language (if not bilingual English/other language). We did not include purposely instances specific to any vernacular use of English. The data is completely anonymized, and there has been no bias against a segment of people using a certain dialect or variation of the English language. The socioeconomic demographics of the users of TurboTax can range from highly educated users to users not having completed high school. Furthermore, the users are distributed geographically across all states, counties and cities in the US, and they engage with the chatbot using the language of their geolocation. The age of TurboTax users ranges from 18 and up.

Acknowledgements

We thank our organizations for supporting this work, as well as our colleagues Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das for their comments. Furthermore, we extend our appreciation to the anonymous reviewers of LREC-COLING 2024 for their valuable feedback.

5. References

L. Banarescu, Claire Bonial, Shu Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, Martha Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

David Chanin. 2022. Frame semantic transformer. https://github.com/chanind/frame-semantic-transformer.

Ann Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ann A. Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3:281–332.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for abstract meaning representation. In *Proceedings of EACL*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342, Online. Association for Computational Linguistics.

F Jurcıcek, F Mairesse, M Gašic, S Keizer, B Thomson, K Yu, and S Young. Error corrective learning for semantic parsing.

Arle Lommel. 2018. *Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies*, pages 109–127. Springer International Publishing.

Esme Manandise and Conrad de Peuter. 2020. Mitigating silence in compliance terminology during parsing of utterances. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 204–212, Barcelona, Spain (Online). COLING.

Esme Manandise and Raj Srivastava. 2022. Assessing the natural language understanding dilemna of chatbots: Seeing or not seeing the forest for the trees. In *CONVERSATIONS 2022, 6th International Workshop on Chatbot Research, Applications and Design*. University of Amsterdam, Netherlands, Online.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90, Berlin, Germany. Association for Computational Linguistics.

Stephan Oepen and Dan Flickinger. 2019. The ERG at MRP 2019: Radically compositional semantic dependencies. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 40–44, Hong Kong. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Muhammad Reza Qorib and Hwee Tou Ng. 2022. Grammatical error correction: Are we there yet? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2794–2800, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3).