Towards a Consensus Taxonomy for Annotating Errors in Automatically Generated Text

Rudali Huidrom and **Anya Belz**
ADAPT Research Centre
Dublin City University
Ireland
{rudali.huidrom,anya.belz}@adaptcentre.ie

Abstract

Error analysis aims to provide insights into system errors at different levels of granularity. NLP as a field has a long-standing tradition of analysing and reporting errors which is generally considered good practice. There are existing error taxonomies tailored for different types of NLP task. In this paper, we report our work reviewing existing research on meaning/content error types in generated text, attempt to identify emerging consensus among existing meaning/content error taxonomies, and propose a standardised error taxonomy on this basis. We find that there is virtually complete agreement at the highest taxonomic level where errors of meaning/content divide into (1) Content Omission, (2) Content Addition, and (3) Content Substitution. Consensus in the lower levels is less pronounced, but a compact standardised consensus taxonomy can nevertheless be derived that works across generation tasks and application domains.

1 Introduction

Error analysis and error type annotation are widely considered important for diverse natural language processing (NLP) tasks (Popović and Burchardt, 2011; Costa et al., 2015). NLP has a long-standing track record in error analysis and error type annotation (Macklovitch, 1991; Costa et al., 2015; Rivera-Trigueros, 2021), not only for directly improving system performance but also for providing guidance in improving evaluation methods.

Errors of content (as opposed to errors of form such as grammatical or lexical-choice errors) are becoming more common in current language generation outputs, given the growing dominance of neural methods which are more prone to such errors than previous rule-based and statistical systems. Documenting and analysing what types of errors different systems make can help improve the semantic correctness (known as Adequacy in MT) of generated text. However, a large variety of different annotation schemes have been created (Huidrom and Belz, 2022), often task and/or domain-specific, which makes comparison between output annotations and thus incremental progress difficult. A standardised, task-agnostic error annotation taxonomy would not only help in comparing different NLP system outputs for performance analysis, but it would also aid in developing automatic or semi-automatic error metrics for various NLP tasks (van Miltenburg et al., 2021).

In this paper, we explore to what extent a standard has evolved in current error annotation schemes, and whether or not enough consensus is present to turn into a standardised consensus taxonomy for errors of content/meaning. Our exploration has resulted in the following contributions:

1. A systematic survey of error annotation schemes comprising content/meaning error types (Section 3 and Table 1);
2. A collated list of all content/meaning error type definitions found in the papers in the survey (see Appendix);
3. The minimally merged taxonomy comprising all non-task and non-domain-specific error types from the above list (Section 5.1 and Figure 1);
4. A standardised and generalised taxonomy of content/meaning error types derived directly from the minimally merged taxonomy (Section 5.2 and Figure 2), which is applicable across different input-controlled language generation tasks[1] and application domains.

The paper is organised as follows. Section 2 describes the paper selection and filtering pro-

[1]Tasks where the output content is wholly or largely determined by the input, in contrast to free text generation tasks, where the output is guided (but not determined) by a prompt.

Proceedings of Recent Advances in Natural Language Processing, pages 527–540
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_058

cess, Section 3 provides summaries of the selected papers, Section 4 presents the general meaning/content error concepts and definitions we use, Section 5 presents the minimally merged error taxonomy, and the maximally merged standardised version of the latter (i.e. our proposed consensus error taxonomy), Section 6 discusses our findings, and Section 7 concludes with a summary and future directions.

2 Paper Selection and Filtering

Our aim was to identify a set of papers reporting content error annotation schemes of any size and depth as a basis for deriving a consensus taxonomy. We followed the following selection/filtering process. First, we selected all papers from an existing survey on error types in machine and human-generated text (Huidrom and Belz, 2022) that described error taxonomies or error annotation schemes comprising errors of content/meaning. This gave us seven papers.

Second, to further expand the selection of papers, we searched the ACL Anthology[2] for papers that contained the terms "accuracy error" and "taxonomy" which yielded 15 results. We manually examined and selected five papers reporting work on content/meaning errors for generated text. Three of these papers used the same taxonomy, namely SCATE (Tezcan et al., 2017); we therefore included only the main paper on the SCATE taxonomy (Tezcan et al., 2017). In total, we obtained three further papers from this second step.

Third, we added one paper (Specia et al., 2021a) from the related work cited by Al Sharou and Specia (2022), and four relevant papers we were already aware of (Thomson and Reiter, 2020; Tang et al., 2022; Kasner and Dusek, 2022; Popović, 2020), the last of these as a (rare) example of work using the top-level content/meaning error type (Adequacy, Accuracy, see Section) in annotation.

Table 1 presents an overview of the final set of 15 papers, ordered by year of publication, and providing information about authors, language generation task,[3] number of error types, number of leaf nodes and depth of the taxonomy. The **number of error types** is the number of nodes in the tree including the root. For example, the (complete) error annotation scheme used by Popović (2020) is (error → (comprehensibility → (major, minor)), → (adequacy (major, minor))), and we count that as 7 different error types.

The **number of leaf nodes** is simply the number of terminal nodes in a taxonomy, 4 in the above example. Note that in some cases, both internal and leaf nodes are used in annotation, in other cases just leaf nodes. The **depth of the tree** is the longest path from the root to a leaf. In the above example, the depth is 2. If there is no underlying hierarchical structure, then depth=1 (as we always assume a default top-level root error category, even if an explicit one is not included).

3 Summaries of Papers

This section presents high-level summaries of the papers that directly fed into our consensus error taxonomy, focusing on content/meaning aspects.

Costa et al. (2012) provide a corpus of 6,000 questions that have been manually translated into Portuguese. Error annotation addresses two types of errors that arose during the manual translation: semantic-level errors and structure-level errors.

Federico et al. (2014) propose a statistical framework to analyse the impact of different error types, employing linear-mixed models. The experiments are designed for English as the source language and languages that are distant from English as the target language. The paper uses a set of four error classes which partially overlap with those used by Vilar et al. (2006): reordering errors, lexicon errors, missing words, morphological errors.

Costa et al. (2015) introduce a linguistically motivated taxonomy of errors in machine-translated text. The taxonomy has five high-level error categories: Orthography, Lexis, Grammar, Semantic, and Discourse.

Specia et al. (2017) present a large-scale machine translation (MT) dataset that combines various degrees of human annotation with automatically recorded productivity features. Errors are annotated using the Multidimensional Quality Metrics (MQM) error annotation framework (Lommel et al., 2014). The errors are broadly categorised into three main categories: Accuracy, Fluency and Terminology. Additionally, these errors are populated with detailed error categories from MQM.

Tezcan et al. (2017) introduce the SCATE (Smart Computer-aided Translation Environment) MT error taxonomy, which is hierarchical and categorises errors into Accuracy errors (detected by examining both source and target sentences), and Fluency er-

[2] https://aclanthology.org

[3] Note that our taxonomy is task-agnostic.

Paper and Taxonomy Name (where named)	Language Generation Task	# Error types	# Leaf nodes	Depth
Costa et al. (2012)	Machine Translation [MT]	11	9	2
Federico et al. (2014)	Machine Translation [MT]	5	4	1
Costa et al. (2015)	Machine Translation [MT]	36	25	4
Specia et al. (2017)	Machine Translation [MT]	21	15	4
Tezcan et al. (2017), SCATE	Machine Translation [MT]	45	33	4
Caseli and Inácio (2020)	Machine Translation [MT]	17	12	2
Popović (2020)	Machine Translation [MT]	7	4	2
Huang et al. (2020), PolyTope	Text Summarisation [TS]	11	8	2
Thomson and Reiter (2020)	Data-to-Text Generation [D2T]	7	6	1
Specia et al. (2021a)	Machine Translation [MT]	19	15	2
Mahmud et al. (2021a)	Textual Summarisation of source code [TS(SC)]	39	31	2
Zou et al. (2022)	Machine Translation [MT]	5	4	1
Al Sharou and Specia (2022)	Machine Translation [MT]	25	21	2
Tang et al. (2022)	Text Summarisation [TS]	19	8	5
Kasner and Dusek (2022)	Data-to-Text Generation [D2T]	6	5	1
Minimally merged error taxonomy	Task-agnostic	40	30	4
Maximally merged consensus error taxonomy	Task-agnostic	15	11	3

Table 1: Overview of properties of the error annotation schemes that form the basis of the merged taxonomies presented in this paper (last two rows).

rors (relating to the wellformedness of the target sentence, regardless of content or meaning).

Caseli and Inácio (2020) address error analysis of neural MT (NMT) system outputs for Brazilian Portuguese, comparing the errors made by the NMT system with those made by a phrase-based machine translation (PBSMT) system. The error analysis adopted by the paper extends the taxonomy put forward by Martins and Caseli (2015), which consists of four broad error categories: syntactic errors, lexical errors, n-gram, reordering errors.

Popović (2020) introduce a manual evaluation method for MT outputs which marks up errors in the translated text. The proposed method uses two quality criteria: Comprehensibility and Adequacy. Comprehensibility refers to the degree to which a translated text can be understood (as distinct from fluency). Adequacy refers to the degree to which the translation conveys the meaning of the original source text. These error types each subdivide into Major and Minor.

Huang et al. (2020) introduce PolyTope, a set of eight metrics for Accuracy and Fluency error types, designed to quantify primary errors for 10 representative models for text summarisation. Accuracy-type errors occur when a target summarisation does not match or accurately reflect the source text, while Fluency-type errors relate to linguistic properties of the text that are independent of how source and target relate. These categories subdivide into three levels of severity: Critical, Minor and Major.

Thomson and Reiter (2020) propose a methodology for gold-standard accuracy evaluations in texts generated by data-to-text systems. There are six main categories: Incorrect Number, Incorrect Named Entity, Incorrect Word, Context Error, Not Checkable and Other Error.

Specia et al. (2021a) report the WMT 2021 Shared Task on Quality Estimation, where the aim is to predict the quality of outputs of neural machine translation (MT) systems at the word and sentence levels. Three main categories of meaning deviation are involved: Mistranslation, Omission and Hallucination. For each meaning deviation category, there are five critical errors. Annotators are instructed to ignore minor grammatical or typographical errors.

Mahmud et al. (2021a) report a qualitative and quantitative comparative analysis of recently proposed source code summarisation models. A taxonomy of different error types across various models is used, with seven top-level categories: Missing Context, Missing Information, Incorrect Semantic Information, Incorrect Construction, Consistent with Ground Truth, Extraneous/Unnecessary, and Over-generalisation.

Zou et al. (2022) explore the effect of translation briefs and search conditions on the quality of post-editing performed by participants with varying levels of translation expertise, using the error

categorisation scheme adopted by the ATA.[4] Mistranslations and addition/omission errors fall under as single Accuracy error type, while usage, grammar and others fall under Fluency. Each category has two levels of severity: Accuracy_Critical, Accuracy_Minor, Fluency_Critical and Fluency_Minor. Note that errors of omission and addition are (unusually) treated as the same error type, rather than two different types, in this study.

Al Sharou and Specia (2022) adds two new categories of critical errors to that defined by Specia et al. (2021a): deviation in instructions (INS) and other critical meaning deviation (OTH).

Tang et al. (2022) investigate factual errors in summarisation system outputs, in the context of which they unify nine existing factual error annotation schemes into a single, non-hierarchical typology. The latter distinguishes errors on a number of different dimensions, of which however just two are used in the reported work: intrinsic (misrepresented words from the source text) vs. extrinsic (added words not in the source text) errors, involving a noun phrase vs. a predicate.

Kasner and Dusek (2022) present a zero-shot alternative for data-to-text generation using ordering, aggregation, and paragraph compression. A manual error analysis is performed using five error types: Hallucination, Incorrect Fact Merging, Omissions, Redundancy, Grammar Error, and Disfluency.

4 General Error Concepts

The consensus error taxonomy we propose is intended for input-controlled text generation, rather than free text generation (see also footnote 1). In the case of the former, only content/meaning from the input must be present in the output, and all content in the input must be present in the output, except in contexts where only task-relevant parts of the input are required (e.g. in Summarisation and arguably also in Paraphrasing). What constitutes an error is therefore relatively clear in input-controlled text generation. If we think of the output as rendering the input, errors in input-controlled text generation are mismatches between input and output, where the input (1) is missing something (often referred to as an error of *Omission*), adds something it shouldn't (error of *Addition*), or renders something from the input wrongly (error of *Substitution*). Definitions of these and other error types are provided in the following section.

It is much less clear what constitutes an error in free text generation. Factual incorrectness and faulty common-sense reasoning are at the clearer end of the spectrum, but deviation from an intended reference continuation and relevance to the prompt are less clear to judge or measure. The term 'hallucination' is often used as something of a coverall term for anything that is undesirable in the output in free text generation.

In contrast, in input-controlled text generation, factual incorrectness or common-sense faults have no relevance; what matters is whether what is in the output can be justified by (a) overlap between input and output content, and (b) whether the given NLP task requires all content in the input to be rendered, or just part of it.

In other fields such as psychology, the term 'hallucination' is defined e.g. as "a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world" Blom (2010). Because of its association with mental health conditions, using the term for errors made by a computational system is controversial, and we prefer to use the more sober 'addition error' or just addition.

Omission errors are also a recognised phenomenon in neuroscience, defined (Perri et al., 2017) e.g. as "infrequent errors consisting in missing responses to the target stimuli," which is fairly close to how the concept is used in NLP error assessment.

5 Towards a Consensus Taxonomy

Our overall goal in the work presented here is a consensus taxonomy of errors of meaning and content for use in error annotation and analysis that is based on a representative sample of existing taxonomies and is agnostic with respect to NLP task and domain. We proceed towards this goal in two steps: (1) directly deriving a single hierarchy of error types from our sample of existing taxonomies, minimally merging only those categories that are identical in scope (even if a different category name is used); (2) merging further error categories that are very similar (but not necessarily identical) in scope, yielding what we call a maximally merged taxonomy which standardises over, and encodes the consensus among, the original error type schemes.

Section 5.1 describes the first of these steps, Sec-

[4] https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/

tion 5.2 the second. Section 5.3 outlines how the final consensus taxonomy is used in practice.

The error taxonomies that form the starting point for our process of consensus identification often address errors of content/meaning and errors of form both. We only use the former, although the orthogonal error types below (Section 5.2) can in principle apply to errors of either form or content/meaning.

We draw the line between the two as follows. Errors of content/meaning (in input-controlled NLG) refer to cases where the information conveyed by the output differs from the information conveyed by the input. They are defined relative to the input, hence can only be identified with reference to the input. Errors of form in NLG in general refer to flaws or mistakes in how the word sequence in the output is put togehter (rather than what it means), e.g. grammatical errors, disfluencies, or inappropriate style.

5.1 Minimally merged error taxonomy

As our starting point we collated all error categories along with their definitions where available from all of our 15 papers (see Appendix). We removed those categories that relate to errors in the form, rather than the content, of outputs. Furthermore, we removed highly task or domain-specific categories, e.g. Missing Programming Language Information in code-to-summary generation (Mahmud et al., 2021b), and Toxicity-introducing Error in catastrophic error detection (Al Sharou et al., 2021; Specia et al., 2021b).

For the remaining error categories we then grouped those together that we took to refer to the same error phenomenon, and arranged the resulting groups in superset/subset relations. This gave us what we refer to as our minimally merged taxonomy, shown in Figure 1. Each node in the hierarchy in Figure 1 shows the original names of the error categories and the papers we extracted them from. For the definitions provided in the original paper for each of these error categories, see Appendix. We added two error categories (Content Substitution and Other) to ensure completeness and balance in the taxonomy.

For space reasons, in the diagram we are not showing subcategories that refer purely to (i) whether the error relates to a single word vs. multiple words (Caseli and Inácio, 2020); (ii) whether the error was major/critical vs. minor; (iii) which syntactic category the error related to (e.g. part of speech); and (iv) whether the error concerns function word(s) or content word(s). We return to these four sets of subcategories in the next section.

As can be seen from Figure 1, there is considerable consensus about the higher up categories, where we found up to ten papers using the same error category, albeit often under different names. In the next section, we develop the consensus further, generalising and creating single labels for sets of error names, to create a maximally merged version of the taxonomy.

5.2 Maximally merged consensus error taxonomy

Building on the process of alignment and consensus identification in the previous section, in the next stage our overall goal was to create a single generic error annotation taxonomy that would work across task construals and application domains. More specifically, our objectives were as follows:

1. To normalise the different names used in the source papers for the same error type using single error category names;
2. To ensure that names and definitions are general enough to work for text generated under both data-to-text and text-to-text tasks, the latter including at least summarisation, paraphrasing and machine translation; and
3. To extract the orthogonal error type dimensions and incorporate them separately, rather than duplicating them across different parts of the taxonomy as previously in Figure 1, e.g. for the meaning deviation subtypes towards the top right of the diagram (NEs, Pos/neg, Numerical, Other).

The extraction criterion for orthogonal error type dimensions was that any of the primary error categories can additionally be annotated with them, i.e. they necessarily result in duplication in the taxonomy if included there. We identified the following:

1. Type of deviation in meaning between input and output (Sharou and Specia, 2022; Thomson and Reiter, 2020; Tang et al., 2022) resulting from one of the primary error types (listed at the end of this subsection):
   (a) NE Deviation: Deviation in named entities.

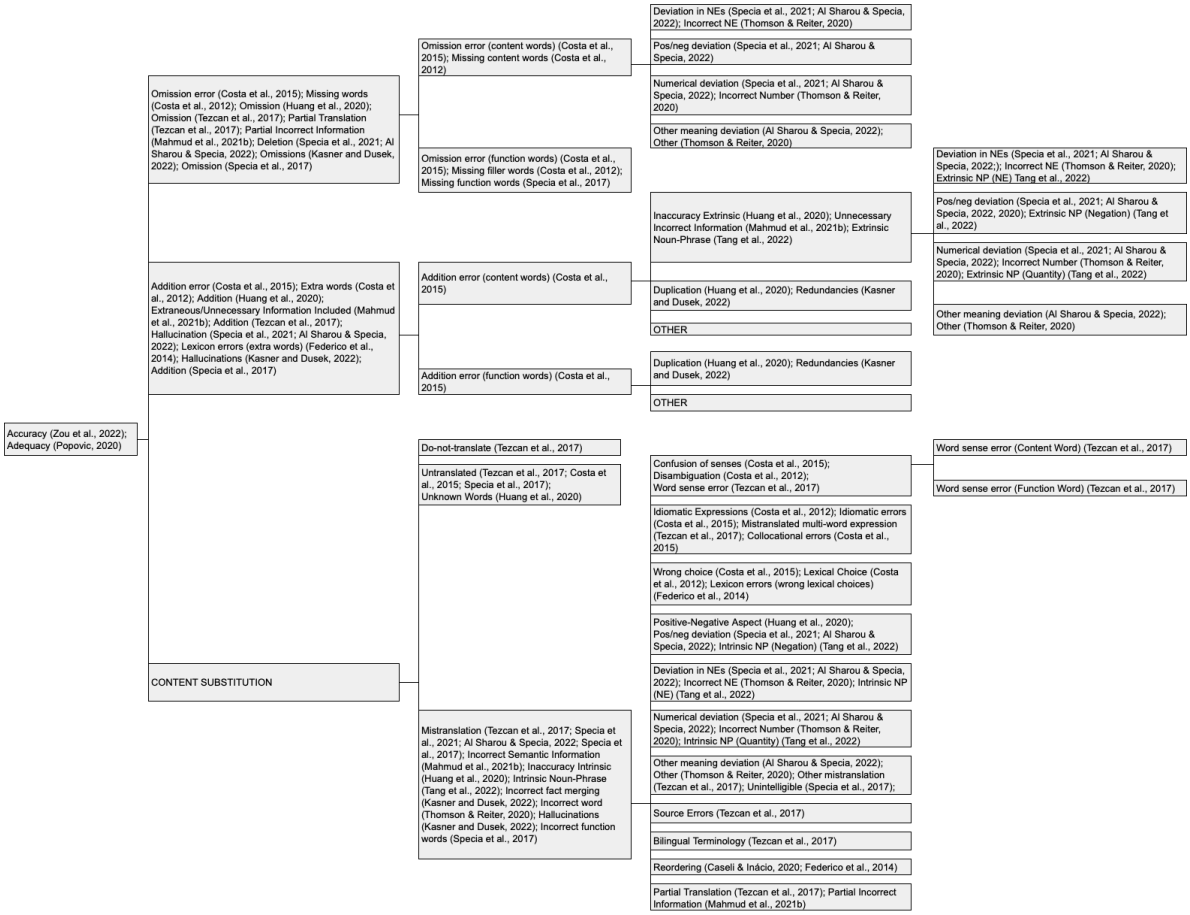

Figure 1: The minimally merged taxonomy of categories of errors in data-to-text and text-to-text generation (see Appendix for definitions of error categories). Note that we have left off some subclasses (see in text for details).

(b) Pos/Neg Deviation: Deviation in negation, polarity or positive/negative sentiment.

(c) Numerical Deviation: Deviation in numerical content.

(d) Other Meaning Deviation.

2. Number of words involved in a given error (Caseli and Inácio, 2020): Single Word and Multiple Words.
3. Severity of the error: Major and Minor (Zou, 2022; Popović, 2020; Specia et al., 2017, 2021a).
4. Degree to which words in the error contribute to the content/meaning of the output: Content Word(s) vs. Function Word(s) (Costa et al., 2012, 2015; Specia et al., 2017).

Note that our aim was to extract all error categories that met the extraction criterion precisely because, if systematically applied, they cause unnecessary duplication in the hierarchy. Conversely, the remaining error categories do not cause such duplication. In other words, this is a fundamental difference between, on the one hand, the error categories in the taxonomy which are in natural subsumption relationships with each other, and, on the other, the orthogonal error types which are not, and can apply to any categories at any level of the hierarchy. We believe it is therefore right to account for them differently.

After taking out the orthogonal error types, the remaining error categories in the taxonomy are as shown on the left of Figure 2. The corresponding definitions are the following:

1. **Content/Meaning Error**: The highest level error category subsuming all errors in outputs that relate to the content/meaning of the output rather than its form (see also start of Section 5.2 on content vs. form).
2. **Omission**: Some content that is present in the input and should be rendered in the output is not present in the output. Moreover there is no content in the output that is intended to render it, but does so wrongly. That is, this type of

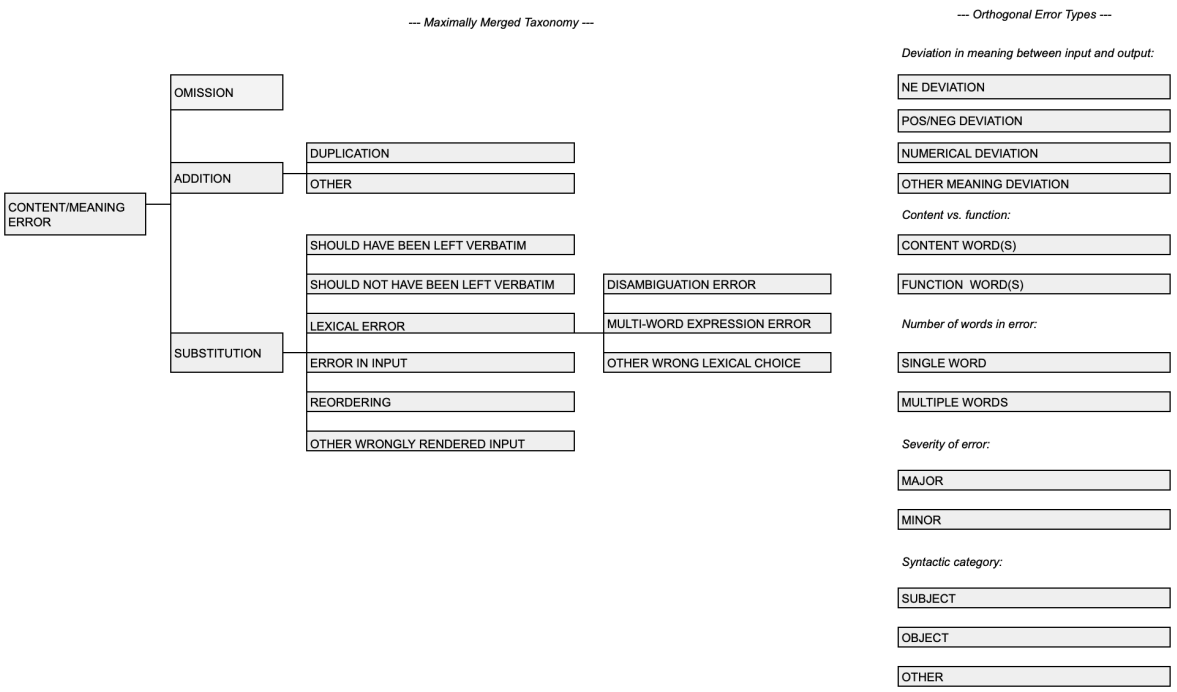

Figure 2: The maximally merged consensus taxonomy of categories of errors in data-to-text and text-to-text generation, with orthogonal error types.

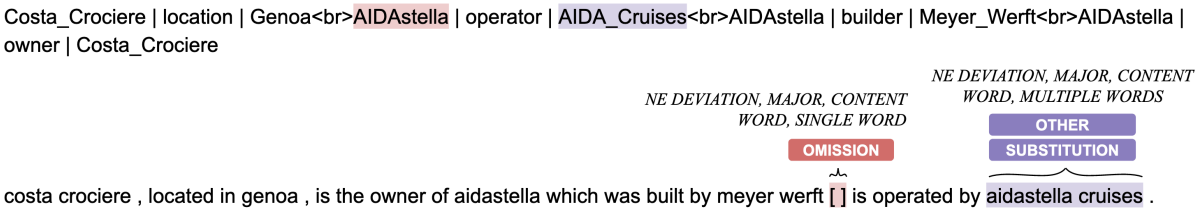

Figure 3: Input/output pair from WebNLG dataset: input 'triples' at the top, verbalisation beneath, both with linked annotations for two errors, using maximally merged consensus taxonomy.

error can be fixed by adding something to the output.

3. **Addition**: Some content that is not present in the input and should not be rendered in the output is present in the output. Moreover there is no content in the input that it is intended to render, but renders wrongly. I.e. this type of error can be fixed by removing something from the output.
   (a) Duplication: Some content is repeated verbatim in the output, but there is no corresponding repetition in the input.
   (b) Other.
4. **Substitution**: Some content in the output, that is intended to convey some content that is present in the input, does it wrongly. This definition means that a substitution cannot equally be construed as the combination of an omission and an addition. This type of error can be fixed by replacing something in the output.
   (a) Should Not Be Verbatim: Some part of the input has been copied verbatim to the output, but should have been rendered differently.
   (b) Should Be Verbatim: Some part of the input should have been copied verbatim to the output, but has been rendered differently.
   (c) Lexical Error: An error that can be fixed by replacing one lexical item in the output with another.
   (d) Error In Input: An error that is caused by an error in the input.
   (e) Reordering: An error that can be fixed by reordering parts of the output.
   (f) Other Wrongly Rendered Output.

5.3 Using the consensus taxonomy for manual error annotation

Figure 3 shows an input/output pair from the WebNLG Shared Task data annotated with the (maximally merged) consensus taxonomy, including annotations for the orthogonal error types. The input meaning representation (known as a set of triples in WebNLG terminology) is shown at the top, with a verbalisation for it produced by one of the participating systems.

The steps in annotating the output text for errors are as follows (shown here for manual annotation by marking up and labelling character spans; alternatively labels can be attached to default spans, such as sentences or whole inputs/outputs):

1. Compare input and output identifying and marking up word spans in the output text that contain some error, and the corresponding span in the input; in the case of Omission errors, the span in the output will be an empty string in the approximate place where the verbalisation of the omitted content would be, had it been rendered, and in the case of Addition errors, conversely an empty string is annotated span in the input;
2. For each linked annotation, a label is attached from the top level in the taxonomy (Omission, Addition, Substitution), then from the second level, until leaf nodes are reached;
3. Finally, the orthogonal error type labels are attached, one from each type.

Note that this is intended as an illustration of how the consensus taxonomy would be used for manual annotation. See following section re expanding the taxonomy with further error categories, and using it for automatic error annotation.

6 Discussion

Error analysis identifying different types of errors plays an important role in NLP system development, providing information about specific strengths and weaknesses and their frequencies of occurrence, for different approaches, rather than a global quality assessment. For this, whether manually or automatically carried out, error categories need to be defined, at multiple levels of granularity.

The current situation is that many different sets of error categories are in use, certainly for different application tasks (MT, Paraphrasing, data-to-text, etc.), but very much also the same tasks, as can be e.g. seen from the ten different MT sets we have included in this paper. Creating a consensus taxonomy incorporating and standardising existing taxonomies means both being able to create annotations and counts that are directly comparable across different research efforts, and, through maximising consensus increasing the taxonomy's acceptability.

The consensus taxonomy as presented incorporates only error categories as used in previous work. The taxonomy can be expanded in various ways at the leaf nodes to increase granularity, notably in the Substitution category, and particularly to reflect domain and task-specific distinctions. In principle, the taxonomy can be used for both manual and automatic error annotation.

In standardising the error categories we have tried to make them applicable across all input-controlled forms of text generation. However, the judgment in particular of whether there is an Omission is a different one in tasks where not all of the input needs to be rendered in the output, such as Summarisation. The task we will use the taxonomy for is data-to-text generation as indicated below.

7 Conclusion

We have presented work where we took 15 papers with error annotation schemes and derived a consensus taxonomy from them in two stages. The first was directly forming a taxonomy from error categories and hierarchical relations between them in their original forms; the second stage was maximally standardising and merging error categories and identifying and treating separately what we called orthogonal error categories that are not in any subsumption relations with other categories.

An important aim is to create a basis for error annotation that is comparable across different research efforts through a single, standardised taxonomy, moreover enhancing acceptability to different practitioners by maximising the consensus embodied in the taxonomy.

In our own future work, we will next use the consensus taxonomy to annotate system outputs from the WebNLG 2020 Shared Task, and then create automatic methods for performing the annotation task. Assessing inter-annotator agreement as part of the manual annotation and performance in automatic annotation will serve as two aspects of testing the taxonomy in action.

Limitations

The process of selecting and filtering papers we employed runs the risk of missing some papers due to the search terms and other criteria for paper selection.

The taxonomies presented in this paper in Section 5 of this paper have not been empirically tested. We acknowledge that so far, we have not verified the following: (1) the degree of comparability of annotations based on our taxonomies, (2) the feasibility of annotating the error types in the taxonomies, and (3) the usability across different error annotation tasks has not been tested.

Ethics Statement

This paper is based on a survey type approach where we work up from the original papers in our literature survey to develop consensus taxonomies, on the basis of these original papers. Therefore, it carries minimal ethical risk.

Acknowledgements

We thank all the reviewers for their valuable feedback and advice. Huidrom's work is supported by the Faculty of Engineering and Computing, DCU, via a PhD studentship. Both authors benefit from being members of the SFI Ireland funded ADAPT Research Centre.

References

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62.

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*, pages 171–179.

Jan Dirk Blom. 2010. *A dictionary of hallucinations*. Springer.

Helena Caseli and Marcio Inácio. 2020. Nmt and pbsmt error analyses in english to brazilian portuguese automatic translations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3623–3629.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes, and Luísa Coheur. 2012. An english-portuguese parallel corpus of questions: translation guidelines and application in smt. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2172–2176.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? *arXiv preprint arXiv:2010.04529*.

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)*, pages 383–398, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zdeněk Kasner and Ondrej Dusek. 2022. Neural pipeline for zero-shot data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3914–3932, Dublin, Ireland. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.

Elliott Macklovitch. 1991. Evaluating commercial mt systems. In *Evaluators' Forum on MT systems, organized by ISSCO at Ste. Croix, Switzerland*.

Junayed Mahmud, Fahim Faisal, Raihan Islam Arnob, Antonios Anastasopoulos, and Kevin Moran. 2021a. 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)*, pages 1–16.

Junayed Mahmud, Fahim Faisal, Raihan Islam Arnob, Antonios Anastasopoulos, and Kevin Moran. 2021b. 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)*, pages 1–16, Online. Association for Computational Linguistics.

Débora Beatriz de Jesus Martins and Helena de Medeiros Caseli. 2015. Automatic machine translation error identification. *Machine Translation*, 29(1):1–24.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Rinaldo Livio Perri, Donatella Spinelli, and Francesco Di Russo. 2017. Missing the target: the neural processing underlying the omission error. *Brain topography*, 30(3):352–363.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović and Aljoscha Burchardt. 2011. From human to automatic error classification for machine translation output. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Irene Rivera-Trigueros. 2021. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, pages 1–27.

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*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021a. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. 2021b. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadin, Matteo Negri, and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 55–71.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Arda Tezcan, Véronique Hoste, and Lieve Macken. 2017. Scate taxonomy and corpus of machine translation errors. *Trends in E-tools and resources for translators and interpreters*, pages 219–244.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

David Vilar, Jia Xu, D'Haro Luis Fernando, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *LREC*, pages 697–702.

Deyan Zou. 2022. Multi-dimensional consideration of cognitive effort in translation and interpreting process studies. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 416–426, Orlando, USA. Association for Machine Translation in the Americas.

Longhui Zou, Michael Carl, Masaru Yamada, and Takanori Mizowaki. 2022. Proficiency and external aides: Impact of translation brief and search conditions on post-editing quality. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research)*, pages 60–74.

A Original Definitions of Content Error Categories from Papers

This section presents definitions, to the extent provided in the original papers, of the content error categories incorporated as the nodes in our minimally merged taxonomy (Figure 1). In those cases where no definition is provided in the original paper, we list just the name of the error category.

Note we do not include syntactic, discourse-level and other error categories not relating to content/meaning errors, as included in some of the work cited here.

The error categories listed are the lowest(most specific) level of the original error hierarchy in each case.

In one or two cases, the original work additionally provides syntactic labels (e.g. Huang et al.) which we omit if they can apply to any of the error categories (are orthogonal to them).

A.1 Top-level error categories

1. *Accuracy* (Zou et al., 2022)
   (a) *Critical*.
   (b) *Minor*.
2. *Adequacy* (Popović, 2020)[5]
   (a) *Major*.
   (b) *Minor*.

A.2 Omission-type error categories

1. *Omission error* (Costa et al., 2015): "omission errors happen when the translation of a word present in the source text is missing in the resulting translation."
   (a) *Omission error (content words)*.
   (b) *Omission error (function words)*.
2. *Missing words* (Costa et al., 2012): "when one or more words are missing in the translation."
   (a) *Missing filler words*.
   (b) *Missing content words*.
3. *Omission* (Huang et al., 2020): "Key point is missing from the output."
4. *Missing context* (Mahmud et al., 2021b):
   (a) *Missing Prog. Language Information*: "Missing Attributes that refer to PL specific information."
   (b) *Missing Database Information*: "Missing database attributes that provide needed context to method functionality."
5. *Missing information* (Mahmud et al., 2021b):
   (a) *Missing conditional information*: "Misses code branching information."
   (b) *Missing critical information*: "Comment is missing critical semantic information."
   (c) *Missing Task Elaboration*: "Did not describe what code was doing properly."
   (d) *Missing Non-Critical Information*: "Useful comment but non-critical info missing."
   (e) *Missing Web-Related Information*: "Comment failed to mention web-related identifier."
   (f) *Failed to Mention Identifiers*: "Does not mention specific variable/attribute names, often using a generic identifier."
   (g) *Missing Identifier*: "No identifier mentioned at all."
   (h) *Missing Data Structure Information*: "Does not capture relevant data structure info."
   (i) *Missing Syntax Information*: "Important syntactic information (e.g. code ordering) is missing."[6]
   (j) *Missing Exception*: "Does not mention relevant exception info."
6. *Absent word* (Caseli and Inácio, 2020).
7. *Absent n-gram* (Caseli and Inácio, 2020).
8. *Deletion* (Specia et al., 2021a; Al Sharou and Specia, 2022): "critical content that is in the source sentence is not present in the translation."the translation."
   (a) *TOX* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in toxicity (hate, violence or profanity)."
   (b) *SAF* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in health or safety risks."
   (c) *NAM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in named entities."
   (d) *SEN* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in sentiment polarity or negation."
   (e) *NUM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in units/time/date/numbers."
   (f) *INS* (Al Sharou and Specia, 2022): "Deviation in instructions."
   (g) *OTH* (Al Sharou and Specia, 2022): "Other critical meaning deviation."
9. *Omission* (Specia et al., 2017).
10. *Missing function words* (Specia et al., 2017).
11. *Incorrect Number* (Thomson and Reiter, 2020): "This includes numbers which are spelled out as well as digits."
12. *Incorrect Named Entity* (Thomson and Reiter, 2020): "This includes people, places, organisations, and days of the week."

[5]The other error type, Comprehensibility, is not included here, as it is more to do with understanding content that has been correctly included.

[6]This refers to programming language syntax, rather than linguistic.

13. *Other* (Thomson and Reiter, 2020): "Any other type of mistake."
14. *Omissions* (Kasner and Dusek, 2022).

A.3 Addition-type error categories

1. *Addition error* (Costa et al., 2015): "the translation of a word that was not present in the source text and was added to the target text."
   (a) *Addition error (content word).*
   (b) *Addition error (function word).*
2. *Extra words* (Costa et al., 2012): "cases where the translation engine generates sentences containing words, most commonly filler words, that should be removed in order to obtain a correct sentence."
3. *Addition* (Huang et al., 2020): "Unnecessary and irrelevant snippets from the source are included in the summary."
4. *Inaccuracy Extrinsic* (Huang et al., 2020): "The summary has content not presented in the source and factually incorrect."
5. *Duplication* (Huang et al., 2020): "A word or longer portion of the text is repeated unnecessarily."
6. *Extraneous/Unnecessary Information Included* (Mahmud et al., 2021b):
   (a) *Unnecessary Data Structure Info*: "Adds unnecessary data structure info to comment."
   (b) *Unnecessary File Information*: "Adds unnecessary file information to comment."
   (c) *Unnecessary Incorrect Information*: "Adds information to comment that is both incorrect and unnecessary."
7. *Extra word* (Caseli and Inácio, 2020).
8. *Extra n-gram* (Caseli and Inácio, 2020).
9. *Addition* (Tezcan et al., 2017): "refer[s] to target words not represented in the source."
10. *Omission* (Tezcan et al., 2017): "refer[s] to source words not represented in the target text."
11. *Hallucination* (Specia et al., 2021a): "critical content that is not in the source is introduced in the translation, for example, profanity words are introduced that were not in the source."
    (a) *TOX* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in toxicity (hate, violence or profanity)."
    (b) *SAF* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in health or safety risks."
    (c) *NAM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in named entities."
    (d) *SEN* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in sentiment polarity or negation."
    (e) *NUM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in units/time/date/numbers."
    (f) *INS* (Al Sharou and Specia, 2022): "Deviation in instructions."
    (g) *OTH* (Al Sharou and Specia, 2022): "Other critical meaning deviation."
12. *Addition* (Specia et al., 2017).
13. *Extraneous function words* (Specia et al., 2017).
14. *Incorrect Number* (Thomson and Reiter, 2020): "This includes numbers which are spelled out as well as digits."
15. *Incorrect Named Entity* (Thomson and Reiter, 2020): "This includes people, places, organisations, and days of the week."
16. *Other* (Thomson and Reiter, 2020): "Any other type of mistake."
17. *Hallucinations* (Kasner and Dusek, 2022).
18. *Redundancies* (Kasner and Dusek, 2022).
19. *Extrinsic Noun-Phrase* (Tang et al., 2022): "A model introduces word(s) not from the source text that function(s) in a summary as subject, object, or prepositional object but cannot be verified from the source."
    (a) *Named Entity* (Tang et al., 2022).
    (b) *Quantity* (Tang et al., 2022).
    (c) *Negation* (Tang et al., 2022).

A.4 Substitution-type error categories

1. *Untranslated error* (Costa et al., 2015): "when the engine cannot find any translation candidate for a given source word, [and] cop[ies] it to the translation output 'as is'."
2. *Confusion of senses* (Costa et al., 2015): "is the case of a word that was translated into

something representing one of its possible meanings, but, in the given context, the chosen translation is not correct."

3. *Wrong choice* (Costa et al., 2015): "occur when a wrong word, without any apparent relation, is used to translate a given source word."
4. *Collocational errors* (Costa et al., 2015): as wrong choice, but for "blocks of words" rather than single words.
5. *Idiomatic errors* (Costa et al., 2015): "concern errors in idiomatic expressions that the system does not know and translates as regular text."
6. *Lexical Choice* (Costa et al., 2012): "the translation engine chose the wrong translation candidate word."
7. *Disambiguation* (Costa et al., 2012): "the system is not able to disambiguate the correct meaning of a source word in a given context."
8. *Idiomatic Expressions* (Costa et al., 2012): "expressions that should have not been translated literally."
9. *Inaccuracy Intrinsic* (Huang et al., 2020): "Terms or concepts from the source are misrepresented and thus unfaithful."
10. *Positive-Negative Aspect* (Huang et al., 2020): "The output summary represents positive statements whereas the source segment is negative, and vice versa."
11. *Unknown Words* (Huang et al., 2020): "words or expressions [...] for which the translation engine could not find any translation candidate and for that reason were kept in the source language and copied to the translation output.
12. *Incorrect Semantic Information*: (Mahmud et al., 2021b):
    (a) *Partial Incorrect Information*: "Semantically meaningful, with a few errors."
    (b) *Semantically Unrelated to Code*: "Does not capture code context whatsoever."
    (c) *Algorithmically Incorrect*: "Conveys a different algorithmic meaning as compared to the code."
13. *Over-Generalization*: (Mahmud et al., 2021b):
    (a) *Different Meaning*: "Comment overgeneralizes on the meaning of the code functionality."
    (b) *Algorithmically Incorrect*: "Overgeneralizes to the point of incorrectness."
    (c) *Missing Attribute Specification*: "Uses generic names such as var."
14. *Not translated word* (Caseli and Inácio, 2020).
15. *Incorrectly translated word* (Caseli and Inácio, 2020).
16. *Not translated n-gram* (Caseli and Inácio, 2020).
17. *Incorrectly translated n-gram* (Caseli and Inácio, 2020).
18. *Reordering* (Caseli and Inácio, 2020).
19. *Reordering errors* (Federico et al., 2014).
20. *Lexicon errors (including wrong lexical choices and extra words)* (Federico et al., 2014).
21. *Missing words* (Federico et al., 2014).
22. *Untranslated* (Tezcan et al., 2017): "refer[s] to words that are not translated in the target but are copied instead, when they should have been translated."
23. *Do-not-translate* (Tezcan et al., 2017): "refer[s] to source words that have been unnecessarily translated into the target."
24. *Mistranslation* (Tezcan et al., 2017).
    (a) *Multi-word expressions*: "The translation is incorrect (and often too literal) because the source sentence contains multi-word expression such as an idiom, a proverb, a collocation, a compound or a phrasal verb."
    (b) *Part of speech*: change in part of speech between source and target text.
    (c) *Word sense disambiguation*: "The target text fragment refers to different (and a wrong) sense of the corresponding source text fragment."
        i. *Content Word*.
        ii. *Function Word*.
    (d) *Partial Translation*: "The incorrect and partial translation of Dutch separable verbs."
    (e) *Other*.
25. *Bilingual Terminology* (Tezcan et al., 2017).
26. *Source Errors* (Tezcan et al., 2017): MT errors that do not originate from the MT system.

27. *Mistranslation* (Specia et al., 2021a): "critical content is translated incorrectly into a different meaning, or not translated (i.e. it remains in the source language) or translated into gibberish."
    (a) *TOX* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in toxicity (hate, violence or profanity)."
    (b) *SAF* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in health or safety risks."
    (c) *NAM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in named entities."
    (d) *SEN* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in sentiment polarity or negation."
    (e) *NUM* (Specia et al., 2021a; Al Sharou and Specia, 2022): "Deviation in units/time/date/numbers."
    (f) *INS* (Al Sharou and Specia, 2022): "Deviation in instructions."
    (g) *OTH* (Al Sharou and Specia, 2022): "Other critical meaning deviation."
28. *Mistranslation* (Specia et al., 2017).
29. *Untranslated* (Specia et al., 2017).
30. *Incorrect function words* (Specia et al., 2017).
31. *Unintelligible* (Specia et al., 2017).
32. *Not Checkable* (Thomson and Reiter, 2020): "A statement which can not be checked; either the information is not available or it is too time-consuming to check."
33. *Incorrect Number* (Thomson and Reiter, 2020): "This includes numbers which are spelled out as well as digits."
34. *Incorrect Named Entity* (Thomson and Reiter, 2020): "This includes people, places, organisations, and days of the week."
35. *Incorrect word* (Thomson and Reiter, 2020): "A word which is not [a number or noun phrase] and is incorrect."
36. *Other* (Thomson and Reiter, 2020): "Any other type of mistake."
37. *Incorrect fact merging* (Kasner and Dusek, 2022).
38. *Intrinsic Noun-Phrase* (Tang et al., 2022): "A model misrepresents word(s) from the source text that function(s) in a summary as subject, object, or prepositional object."
    (a) *Named Entity* (Tang et al., 2022).
    (b) *Quantity* (Tang et al., 2022).
    (c) *Negation* (Tang et al., 2022).