Resource for Error Analysis in Text Simplification: New Taxonomy and Test Collection

Benjamin Vendeville[1,2,3], Liana Ermakova[2], Pierre De Loor[3,4]

[1] Université Bretagne Occidentale,[2] HCTI, [3] Lab-STICC, [4] ENIB

Introduction

Text simplification is a growing field in NLP. The goal is to make scientific text more accessible by non experts. Many advances have been made in building simplicity datasets, simplification metrics, and simplification systems. The SimpleText track at CLEF proposes for multiple years tasks centered around that objective, including a task on simplifying scientific texts. However, we observed that many generated simplifications are invalid, not because they lack simplicity, but because they distort or omit essential information. We refer to these as errors.

Problem. The lack of a comprehensive framework for defining, detecting, and evaluating error detection methods in ATS limits progress, new resources are needed.

Contributions. In this resource paper, we make the following contributions to support researchers in building better error detection methods for ATS:

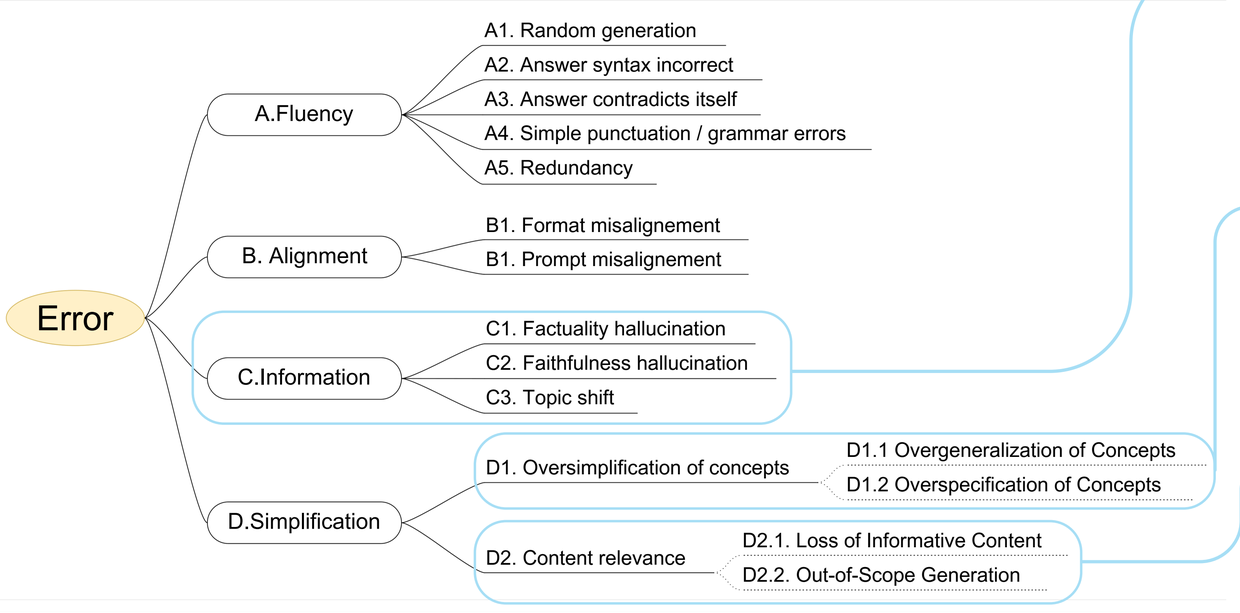
- *Taxonomy*: A new taxonomy of errors in ATS with a focus on information distortion.
- *Test Collection*: A test collection based on our taxonomy, accompanied by a detailed annotation scheme.

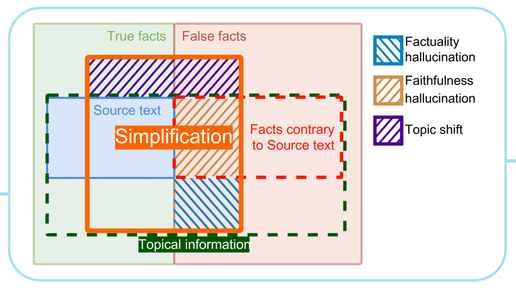
Taxonomy

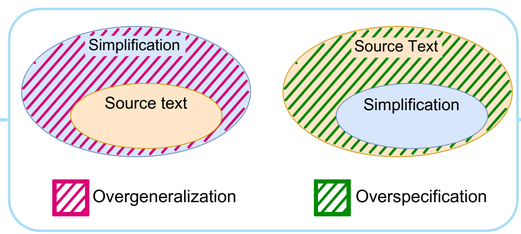
We identify 4 groups of errors:

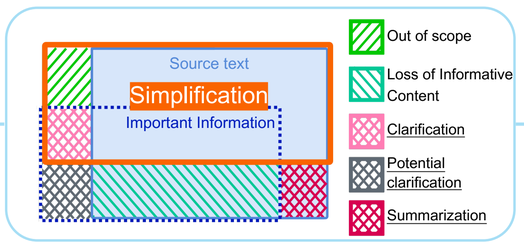
- A. Fluency: Is the answer provided in a correct form that a fluent speaker would speak?
- B. Alignment: Is the format of the answer correct?
- C. Information: Is the information provided accurate and relevant to the input?
- D. Simplification: Does the response focus on simplification and understand the task?

Each regrouping multiple errors, for a total of 14 error classes, plus a "*No Error*" class.









For C. Information and D. Simplification errors, we study the source and simplified texts as sets of facts defined as a tuplet of (*subject*, *relation*, *object*)
We study their intersections along 4 questions:

1. Truthfulness: Is the simplification true ?
   a. *Regarding to what ?*
2. Topicality: Is the simplification on topic ?
3. Reformulation: Does the simplification reformulate correctly ?
4. Importance: Is the simplification information important ?
   a. *Does it keep the important information ?*
   b. *Does it add important information ?*

Examples of errors:

- B1: *{"Current academic and industrial research is interested in autonomous vehicles."}*
- B2: *Current academic and industrial research is interested in autonomous vehicles. <Query> simplify this: <example> […]*
- D2.1:
  - *Source: Regular physical exercise not only improves cardiovascular health but also has been linked to better sleep quality, which in turn enhances cognitive function.*
  - *Simplification: Exercise is good.*

Dataset

	#Total	#True	%True
No Error	2,659	820	30.84
A. Fluency			
A1. Random generation	2,659	142	5.34
A2. Syntax Error	2,659	191	7.18
A3. Contradiction	2,659	23	0.86
A4. Punctuation/Grammar Error	2,659	241	9.06
A5. Redundancy	2,659	112	4.21
B. Alignment			
B1. Format misalignment	2,659	47	1.77
B2. Prompt misalignment	2,659	96	3.61
C. Information			
C1. Factuality hallucination	2,659	23	0.86
C2. Faithfluness hallucination	2,659	360	13.54
C3. Topic shift	2,659	152	5.72
D. Simplification			
D1.1. Overgeneralization	2,659	306	11.51
D1.2. Overspecification	2,659	136	5.11
D2.1. Loss of informative content	2,659	520	19.56
D2.2. Out-of-Scope generation	2,659	418	15.72

Statistics of our manually annotated dataset

	Fleiss' κ	Unanim. %	Cohen's Kappa scores for Annotator pair									
			AB	*AC*	*AD*	*AE*	*BC*	*BD*	*BE*	*CD*	*CE*	*DE*
No Error	0.34	38.9	0.49	0.30	0.16	0.48	0.66	0.28	0.73	0.23	0.57	0.16
A. Fluency	*0.38*	*67.3*	0.44	0.23	0.22	0.44	0.40	0.37	0.75	0.19	0.40	0.37
B. Alignment	0.45	76.8	0.37	0.58	1.00	0.38	0.25	0.37	0.79	0.58	0.27	0.38
C. Information	0.02	47.3	0.06	0.22	0.14	0.07	0.17	0.40	0.38	0.04	0.27	0.28
D. Simplification	0.26	25.2	0.26	0.24	0.14	0.40	0.49	0.19	0.45	0.20	0.45	0.12

Inter annotator agreement for our annotators, on 100 randomly selected simplifications

Simpletext 2025 results

We used this data for the S*impleText* track at *CLEF 2025*. We used this dataset as the test dataset on the task 2.2 on detecting errors in automatic text simplification, and created a syntetic dataset on this taxonomy, which we provided as training data. In the end, we had :

- 7 participants
- 31 submissions

	No Error		A		B		C		D	
	F1	*AUC*	*F1*	*AUC*	*F1*	*AUC*	*F1*	*AUC*	*F1*	*AUC*
DebertaLlmensemble	**0.76**	**0.56**	0.28	0.13	0.35	0.17	**0.30**	**0.15**	**0.37**	**0.22**
paraphrase_mpnet	0.75	**0.56**	0.25	**0.15**	0.28	0.11	0.13	0.08	0.14	0.16
mpnet	0.44	0.55	0.25	**0.15**	0.21	0.09	0.15	0.09	0.14	0.16
roberta	0.69	0.49	0.23	0.12	0.24	0.10	0.11	0.08	0.12	0.16
gpt4o	0.68	0.50	**0.35**	0.15	**0.38**	**0.19**	0.25	0.12	0.29	0.18
llama	0.68	0.48	0.28	0.13	0.32	0.18	0.26	0.14	0.30	0.19
OpenChat	0.64	0.42	0.15	0.07	0.14	0.06	0.14	0.08	0.22	0.15
MajorityVoting	0.63	0.41	0.15	0.07	0.11	0.04	0.17	0.08	0.23	0.16
Mistral	0.56	0.35	0.15	0.06	0.10	0.04	0.11	0.07	0.17	0.14

Some methods and results from participants at the SimpleText track of CLEF 2025

Conclusions

We introduced here the first taxonomy of errors in Automatic Text Simplification and built a test collection through the annotation of real-world ATS examples. Our findings show that errors are still prevalent in ATS and that existing methods fail to detect them reliably. While our taxonomy provides a structured approach to error classification, effective annotation requires careful selection and training of annotators. The test collection was used for the SimpleText shared task at CLEF 2025 with great success and will be published freely in a .csv format on GitHub.
The repository also provides annotation scheme and the code for analyzing the annotated dataset introduced in this paper.
By releasing our test collection, taxonomy annotation scheme, and analysis code,we aim to support further research in this area.

GitHub link

Acknowledgments

This research was funded by the French National Research Agency (ANR) under the projects ANR-22-CE23-0019-01 and ANR-19-GURE-0001 (program *Investissements d'avenir* integrated into France 2030)

Future works

In the future, we plan to source aditionnal annotations, with additionnal annotator training to improve aggreement. We also hope to develop better detection and classification systems.