An Open-Source Evaluation Framework for Multilingual Lexical Simplification

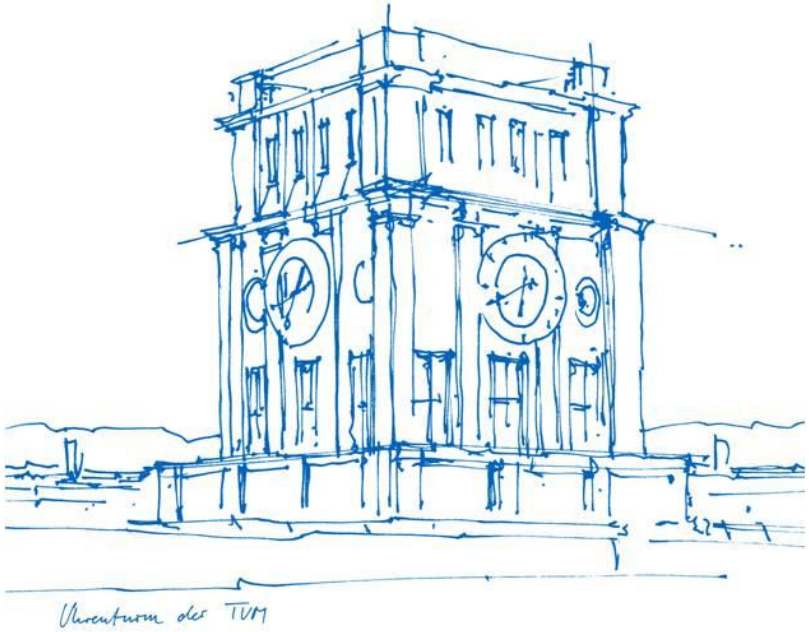
Tobias Lindenbauer

Carolin Ganahl

Luis Wiedmann

Machine Learning for Natural Language Processing

10. Juni 2024



Outline

Background → Approach → Preliminary Results → Next steps

Context-Aware Word Embedding Models [2]

Use Part-of-Speech (POS) tags as context

Train with word2vec

Technically involved pipeline

LSBert [3]

Use BERT

Separate BERT for CWI and SG

Simple pipeline, most work in SR

UniHD [4]

Prompt GPT-4

Extremely simple setup

Minor post-processing and parsing

Context-Aware Word Embedding Models [2]

LSBert [3]

	LexMTurk			BenchLS			NNSeval		
	PRE	RE	F1	PRE	RE	F1	PRE	RE	F1
[2]	0.177	0.140	0.156	0.180	0.252	0.210	0.118	0.161	0.136
[3]	0.306	**0.238**	**0.268**	0.244	**0.331**	**0.281**	**0.194**	**0.260**	**0.222**

UniHD [4]

TSAR-2022

	Acc@k@Top1 K=1	MAP@k K=3
Ensemble (Ours)	**0.8096**	**0.5834**
LSBert	0.5978	0.4079

We alter the prompt to: “Given the above context, list ten alternative **Spanish** words for ‘complex_word’ that are easier to understand.”

[4]

Run	ACC@1	Acc@k@Top1 $k=1$	Acc@k@Top1 $k=2$	Acc@k@Top1 $k=3$	MAP@k $k=3$	MAP@k $k=5$	MAP@k $k=10$	Potential@k $k=3$	Potential@k $k=5$	Potential@k $k=10$
Ensemble (Ours)	**0.6521**	**0.3505**	**0.5108**	**0.5788**	**0.4281**	**0.3239**	**0.1967**	**0.8206**	**0.8885**	**0.9402**
Single (Ours)	0.5706	0.3070	0.3967	0.4510	0.3526	0.2449	0.1376	0.6902	0.7146	0.7445
PresiUniv-1	0.3695	0.2038	0.2771	0.3288	0.2145	0.1499	0.0832	0.5842	0.6467	0.7255
UoM&MMU-3	0.3668	0.1603	0.2282	0.269	0.2128	0.1506	0.0899	0.5326	0.6005	0.6929
LSBert	0.2880	0.0951	0.1440	0.1820	0.1868	0.1346	0.0795	0.4945	0.6114	0.7472
TUNER	0.1195	0.0625	0.0788	0.0842	0.0575	0.0356	0.0184	0.144	0.1467	0.1494

Spanish

Table 4 [4]

Run	ACC@1	Acc@k@Top1 $k=1$	Acc@k@Top1 $k=2$	Acc@k@Top1 $k=3$	MAP@k $k=3$	MAP@k $k=5$	MAP@k $k=10$	Potential@k $k=3$	Potential@k $k=5$	Potential@k $k=10$
Ensemble (Ours)	**0.7700**	**0.4358**	**0.5347**	**0.6229**	**0.5014**	**0.3620**	**0.2167**	**0.9171**	**0.9491**	**0.9786**
Single (Ours)	0.6363	0.3716	0.4625	0.5160	0.4105	0.2889	0.1615	0.7860	0.8181	0.8422
GMU-WLV-1	0.4812	0.2540	0.3716	0.3957	0.2816	0.1966	0.1153	0.6871	0.7566	0.8395
Cental-1	0.3689	0.1737	0.2433	0.2673	0.1983	0.1344	0.0766	0.524	0.5641	0.6096
LSBert	0.3262	0.1577	0.2326	0.286	0.1904	0.1313	0.0775	0.4946	0.5802	0.6737
TUNER	0.2219	0.1336	0.1604	0.1604	0.1005	0.0623	0.0311	0.2673	0.2673	0.2673

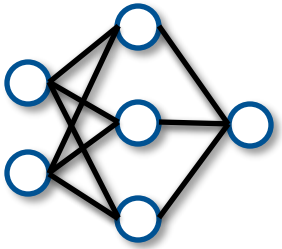
Portoguese

Table 5 [4]

Barriers to usage of proprietary LLMs

Comparison across papers difficult

Few and small datasets

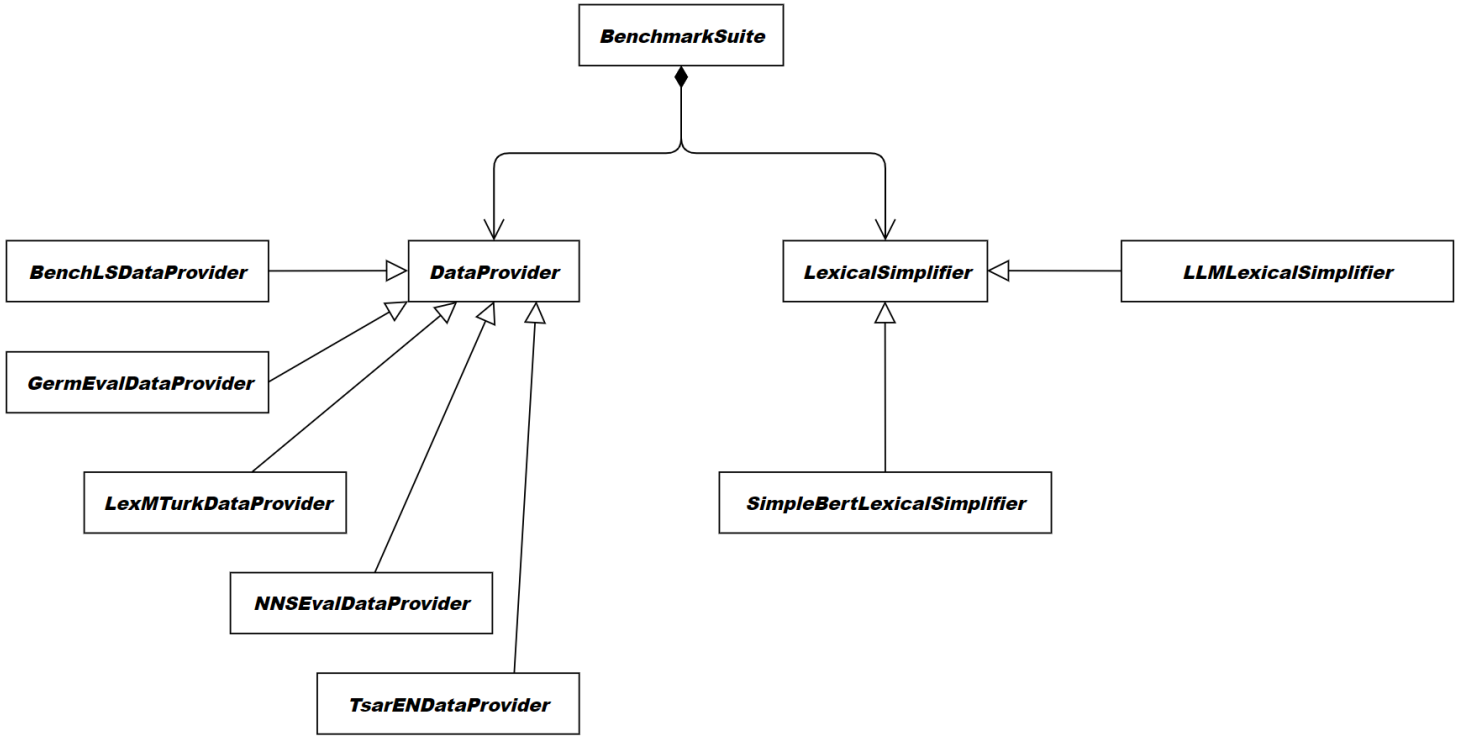
Dataset	Architecture
	

Competitiveness of Open-Source in Multilingual Lexical Simplification?

Open-Source Evaluation Framework

+

Study and development in low-resource environments


BenchmarkSuite
BenchLSDataProvider
DataProvider
LexicalSimplifier
LLMLexicalSimplifier
GermEvalDataProvider
LexMTurkDataProvider
SimpleBertLexicalSimplifier
NNSEvalDataProvider
TsarENDataProvider

Multilingual									
	Google BERT			DistilBERT			RoBERTa		
Dataset	Potential	Precision	Recall	Potential	Precision	Recall	Potential	Precision	Recall
EN-BenchLS	0.7578	0.1588	0.2749	0.7094	0.1516	0.2564	0.7847	0.1767	0.2956
EN-LexMTurk	0.906	0.217	0.2311	0.87	0.212	0.229	0.916	0.2386	0.2472
EN-NNSeval	0.6444	0.1084	0.2172	0.6025	0.105	0.1925	0.6485	0.1113	0.2102
EN-TsarEN	0.7098	0.1526	0.1682	0.614	0.122	0.1348	0.7098	0.1433	0.1522
DE-GermanEval	0.1587	0.0199	0.036	0.1375	0.0158	0.0287	0.2144	0.0279	0.0507

German									
	Google BERT			DistilBERT			DbmdzBERT		
Dataset	Potential	Precision	Recall	Potential	Precision	Recall	Potential	Precision	Recall
EN-BenchLS	0.0441	0.0050	0.0088	0.1389	0.017	0.0262	0.2443	0.0326	0.0565
EN-LexMTurk	0.0640	0.0074	0.0071	0.23	0.0282	0.0299	0.358	0.0484	0.0507
EN-NNSeval	0.0293	0.0029	0.0073	0.0669	0.0079	0.0174	0.1423	0.0213	0.0578
EN-TsarEN	0.0259	0.0026	0.0023	0.0699	0.0078	0.008	0.114	0.013	0.0127
DE-GermanEval	0.3788	0.0596	0.1032	0.324	0.0484	0.0855	0.351	0.0547	0.0997

Google Bert									
	MAP @ K			Potential @ K			Accuracy @ K top 1		
Dataset	K = 3	K = 5	K = 10	K = 3	K = 5	K = 10	K = 3	K = 5	K = 10
EN-BenchLS	0.1834	0.1588	0.1313	0.7169	0.8288	0.7578	0.4015	0.5328	0.4898
EN-LexMTurk	0.2608	0.2247	0.1852	0.854	0.92	0.906	0.472	0.606	0.58
EN-NNSeval	0.1912	0.1649	0.1257	0.5481	0.728	0.6444	0.3054	0.4477	0.41
EN-TsarEN	0.2974	0.2565	0.1942	0.829	0.8705	0.7098	0.386	0.487	0.3782
DE-GermanEval	0.091	0.0668	0.1144	0.0029	0.0048	0.1587	0.001	0.001	0.0654

LLM

```
Benchmarking model on DE ...
Benchmarking model on GermanEvalDataProvider...
Benchmarking:   0%|          | 0/1040 [00:00<?, ?it/s]/usr/local/lib/python3.10/dist-packages/transformers/generation/configuration_utils.py:515: UserWarning: `do_sample` is set to `Fal
  warnings.warn(
Benchmarking:   0%|          | 3/1040 [00:10<1:15:13,  4.35s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation:  In der Legislaturperioc
Benchmarking:   0%|          | 4/1040 [00:13<1:01:02,  3.54s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation:  AtmosphäreReturning emp
Benchmarking:   0%|          | 5/1040 [00:15<54:42,  3.17s/it]  Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 6/1040 [00:21<1:10:18,  4.08s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation:  Die einfachere Version
Benchmarking:   1%|          | 7/1040 [00:24<1:06:48,  3.88s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 8/1040 [00:28<1:06:56,  3.89s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 9/1040 [00:33<1:08:28,  3.99s/it]You seem to be using the pipelines sequentially on GPU. In order to maximize efficiency please use a dataset
Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 10/1040 [00:37<1:11:55,  4.19s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 11/1040 [00:42<1:15:37,  4.41s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 12/1040 [00:48<1:20:56,  4.72s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 13/1040 [00:54<1:27:31,  5.11s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 14/1040 [01:00<1:33:41,  5.48s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
Benchmarking:   1%|          | 15/1040 [01:07<1:39:54,  5.85s/it]Failed to parse the output from the LLM: The provided string is not a valid list representation: Returning empty list.
```

… so far no results

Next Steps

- First Results with LLMs
- Integrate Datasets from Different Languages
- Deployability Analysis
- Investigate Context and Difficulty Aware Filling
- Finetune Small Models

Tasks - People

	Tobias	Luis	Caro
Literature Review	X	X	X
Dataset Approach	X		
Architecture Approach		X	
Implement Eval Datasets processing		X	X
System Design	X		
Implement wrapper for Models & Datasets	X		
Implement BERT & LLM model wrapper and BenchmarkSuite	X		
Evaluation Metrics		X	X
Colab Configuration			X
Presentation	X	X	X

References

[1] K. Tan, K. Luo, Y. Lan, Z. Yuan, und J. Shu, „An LLM-Enhanced Adversarial Editing System for Lexical Simplification“. arXiv, 22. März 2024. doi: 10.48550/arXiv.2402.14704.
[2] G. Paetzold und L. Specia, „Unsupervised Lexical Simplification for Non-Native Speakers“, *Proceedings of the AAAI Conference on Artificial Intelligence*, Bd. 30, Nr. 1, Art. Nr. 1, März 2016, doi: 10.1609/aaai.v30i1.9885.
[3] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, Y. Shi, und X. Wu, „LSBert: Lexical Simplification Based on BERT“, *IEEE/ACM Trans. Audio Speech Lang. Process.*, Bd. 29, S. 3064–3076, 2021, doi: 10.1109/TASLP.2021.3111589.
[4] D. Aumiller und M. Gertz, „UniHD at TSAR-2022 Shared Task: Is Compute All We Need for Lexical Simplification?“ arXiv, 5. Januar 2023. doi: 10.48550/arXiv.2301.01764.