

EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages

Text Simplification

Aim	How to Simplify	German Example	Evaluation										
<ul style="list-style-type: none"> Adaption of a given text to improve the text comprehension for a target group: <ul style="list-style-type: none"> children, foreign language learners, or people with reading problems. 	<ul style="list-style-type: none"> Sentence Split, Complex Word Substitution, Reordering, Word Deletion or Addition, Compound Segmentation, etc. 	<p>complex: Die Gewerkschaft setzt sich dafür ein, dass zum Beispiel höhere Löhne gezahlt werden.</p> <p>simple: Die Gewerkschaft setzt sich zum Beispiel für höhere Löhne oder mehr Urlaub ein.</p> <p>plain EN: The trade union advocates, for example, for higher wages or more vacation.</p>	<table border="1"> <thead> <tr> <th>Manual</th> <th>Automatic</th> </tr> </thead> <tbody> <tr> <td>Simplification</td> <td>SARI & BERT-Score</td> </tr> <tr> <td>Readability</td> <td>FRE</td> </tr> <tr> <td>Meaning</td> <td>BLEU</td> </tr> <tr> <td>Preservation</td> <td></td> </tr> </tbody> </table>	Manual	Automatic	Simplification	SARI & BERT-Score	Readability	FRE	Meaning	BLEU	Preservation	
Manual	Automatic												
Simplification	SARI & BERT-Score												
Readability	FRE												
Meaning	BLEU												
Preservation													

EASSE (original) vs. EASSE-multi vs. EASSE-DE

EASSE-original	EASSE-multi	EASSE-DE
Language(s) : EN BLEU, BERTScore (RoBERTa),	Language(s) : 70 BLEU, BERTScore (BERT-base-multilingual),	Language(s) : DE BLEU, BERTScore (BERT-base-multilingual),
metrics : SARI, SAMSA, FKGL, basic quality estimation	metrics : FKGL + language specific, extended quality estimation	metrics : SARI, FRE, Vienna non-fictional, extended quality estimation
tokenization : 13a, moses, penn, whitespace,	tokenization : 13a, moses, penn, whitespace, + SpaCy	tokenization : 13a, moses, penn, whitespace, + SpaCy
baselines : reference, truncate	baselines : reference, truncate, + identity	baselines : reference, truncate, + identity
test sets : ASSET, HSplit, PWKP, QATS, TURKcorpus, WikiSplit	test sets : X	test sets : ABGB, APA-LHA, BiSECT, DEplain-APA, DEplain-web, TextComplexityDE, Geolino, Simple-German-Corpus
system outputs : 11 English TS models	system outputs : X	system outputs : 11 German TS models
human assessment : X	human assessment : X	human assessment : ✓
evaluation reports (generated) : X	evaluation reports (generated) : X	evaluation reports (generated) : ✓
reproduction scripts : X	reproduction scripts : X	reproduction scripts : ✓

Relevance of Language-specific Evaluation

Scores (S) and Ranks (R) of German TS models on the DEplain-APA test set.

EASSE-original (lower-casing and 13a tokenizer)							EASSE-DE (no lower-casing and SpaCy tokenizer)						
	BLEU↑		SARI↑		BS-P↑			BLEU↑		SARI↑		BS-P↑	
	S	R	S	R	S	R		S	R	S	R	S	R
hda LS	23.77	4	26.82	12	0.38	7	hda LS	22.30	5	26.06	12	0.55	7
socketeye-APA-LHA	12.42	11	40.27	3	0.13	12	socketeye-APA-LHA	11.84	11	40.16	3	0.37	12
socketeye-DEplain-APA	20.97	7	44.89	1	0.36	9	socketeye-DEplain-APA	19.58	7	44.14	1	0.53	9
mbart DEplain APA	30.01	1	39.12	5	0.47	1	mbart DEplain apa	28.49	1	38.72	5	0.64	1
mbart DEplain APA web	29.62	2	34.44	10	0.47	1	mbart DEplain apa web	28.03	2	33.81	10	0.64	1
mT5-DEplain-APA	23.70	5	39.80	4	0.46	3	mT5-DEplain-APA	22.32	4	39.41	4	0.61	4
mt5-simple-german-corpus	8.92	12	38.2	6	0.29	11	mt5-simple-german-corpus	8.12	12	37.92	6	0.48	11
BLOOM-zero	17.23	10	35.19	9	0.36	9	BLOOM-zero	16.14	9	35.43	9	0.53	9
BLOOM-10-random	19.23	8	35.52	8	0.38	7	BLOOM-10-random	17.97	8	35.93	8	0.57	5
BLOOM-10-similarity	22.21	6	41.21	2	0.39	6	BLOOM-10-similarity	20.97	6	41.27	2	0.57	5
custom-decoder-ats	1.29	13	36.65	7	-0.13	13	custom-decoder-ats	1.24	13	36.42	7	0.16	13
Identity baseline	28.50	3	15.88	13	0.45	4	Identity baseline	26.89	3	15.25	13	0.63	3
Truncate baseline	18.94	9	28.31	11	0.41	5	Truncate baseline	16.11	10	27.20	11	0.55	7

Key Messages

- EASSE-multi facilitates evaluation of sentence simplification for **multiple languages**
- language-specific** tokenization, metrics, and (BERT-) models is helpful
- naming the settings** of evaluation metrics is crucial (i.e., language setting, tokenizer, lower casing, and BERT-Score model)

- EASSE-DE facilitates evaluation of sentence simplification for **German**
- compare only models evaluated on **same test set** and with **same evaluation settings**
- new German TS Benchmark (see below)

Benchmarking German Sentence Simplification Models Using EASSE-DE (evaluated with EASSE-DE using no lower-casing and SpaCy tokenizer)

