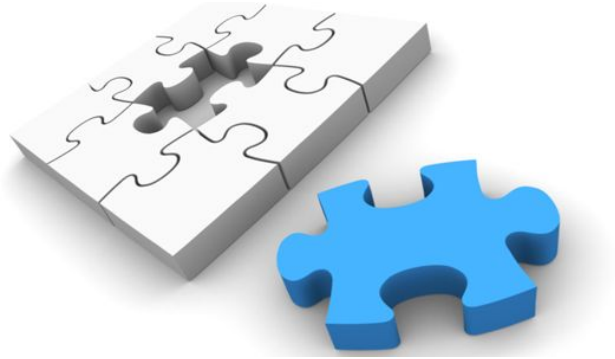
Explainable Prediction of Text Complexity: The Missing Preliminaries for Text Simplification



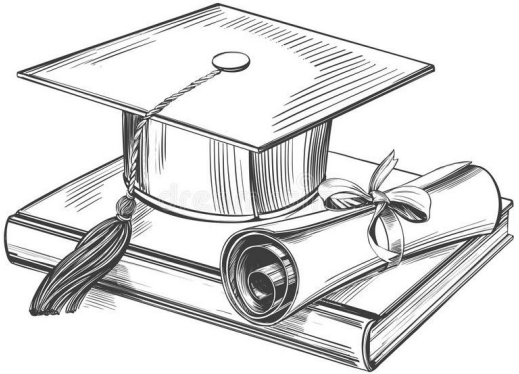
Cristina Garbacea[1], Mengtian Guo[3], Samuel Carton[4], Qiaozhu Mei[1,2]

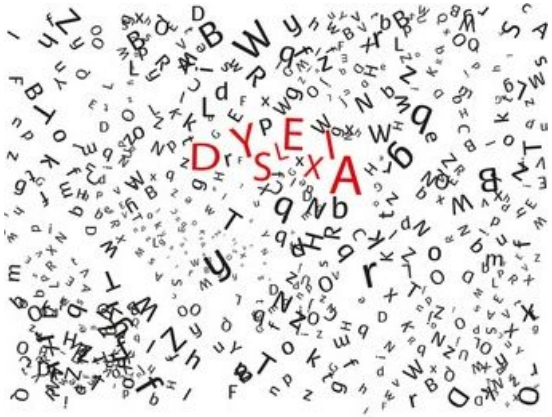
[1]Department of EECS, [2]School of Information, University of Michigan, Ann Arbor, MI, USA
[3]School of Information and Library Science, University of North Carolina, Chapel Hill, NC, USA
[4]Department of CS, University of Colorado, Boulder, CO, USA

Text Simplification Matters for Real People in Real Life!

Healthcare **Education** **Kids**

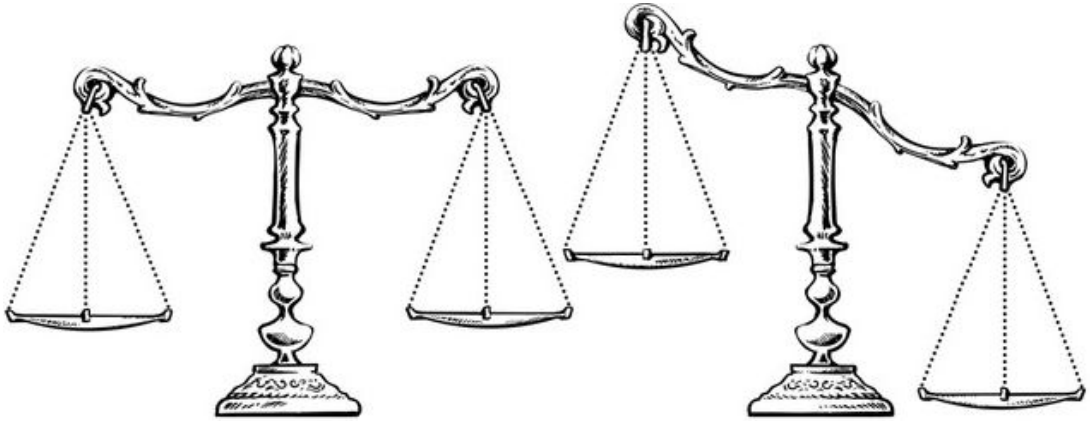
Text Simplification Matters for Real People in Real Life!





Non-native speakers **Low education** **Medical conditions**

Text Simplification Matters for Real People in Real Life!



- Mismatch between language complexity and literacy skills is a critical source of ***bias*** and ***inequality***
- *18 years of education* on average for a reader to properly understand the clinical trial descriptions on ClinicalTrials.gov (Wu et al., 2016) => ***self-selection bias in participants***

Text Simplification Matters for Real People in Real Life!

- Text simplification can improve the ***fairness*** and ***transparency*** of text information systems
- It is critical to explain the rationale behind the simplification decisions

Ironically, in the literature text simplification lacks transparency!

Approaches to Text Simplification

Lexical Simplification (Devlin, 1999) • replaces complex words/phrases with simpler alternatives	**Semantic Simplification** (Kandula et al, 2010): • paraphrases text portions into simpler variants
Syntactic Simplification (Siddhartan, 2006) • alters the syntactic structure of a sentence	**End-to-end Simplification** (Zhang and Lapata, 2017; Guo et al., 2018; Bercken et al., 2019) • MT in a monolingual setting regardless of the type of simplifications

Approaches to Text Simplification

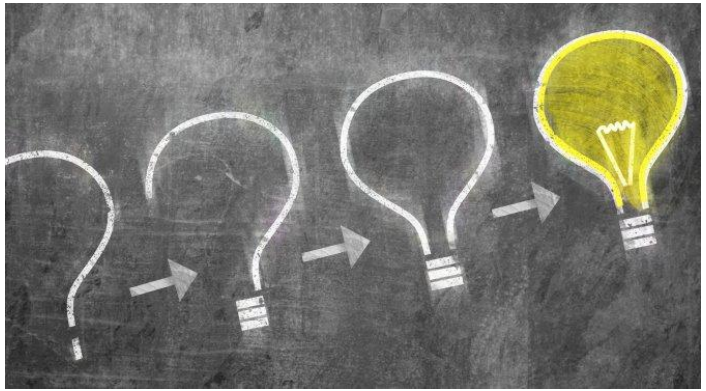
Lexical Simplification (Devlin, 1999) • replaces complex words/phrases with simpler alternatives	**Semantic Simplification** (Kandula et al, 2010): • paraphrases text portions into simpler variants
Syntactic Simplification (Siddhartan, 2006) • alters the syntactic structure of a sentence	**End-to-end Simplification** (Zhang and Lapata, 2017; Guo et al., 2018; Bercken et al., 2019)  MT in a monolingual setting regardless of the type of simplifications

End-to-end Text Simplification Challenges

- absence of large-scale parallel complex → simple training data
- lack of black-box interpretability (Alva-Manchengo et al, 2017)
- evaluation can favour shorter, not simpler, variants (Napoles et al, 2011)
- success is often compromised in out-of-sample, real world scenarios (D'Amour et al, 2020)

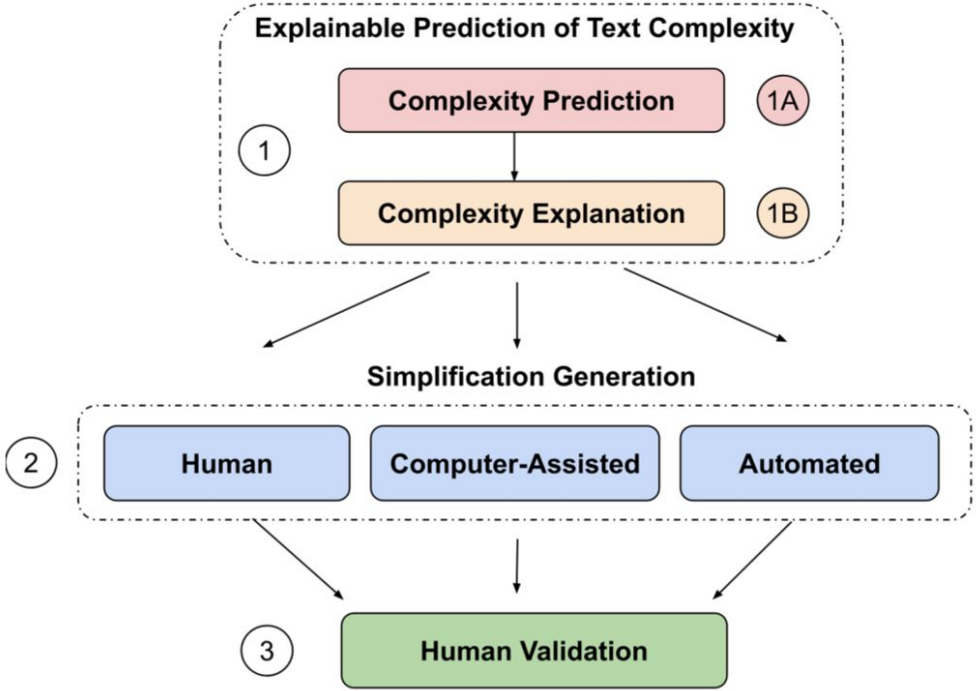
Our Work Fills in the Gap in End-to-end Text Simplification

Motivation: increasing the transparency and explainability of a machine learning procedure may help its generalization into unseen scenarios (Doshi-Velez and Kim, 2018)

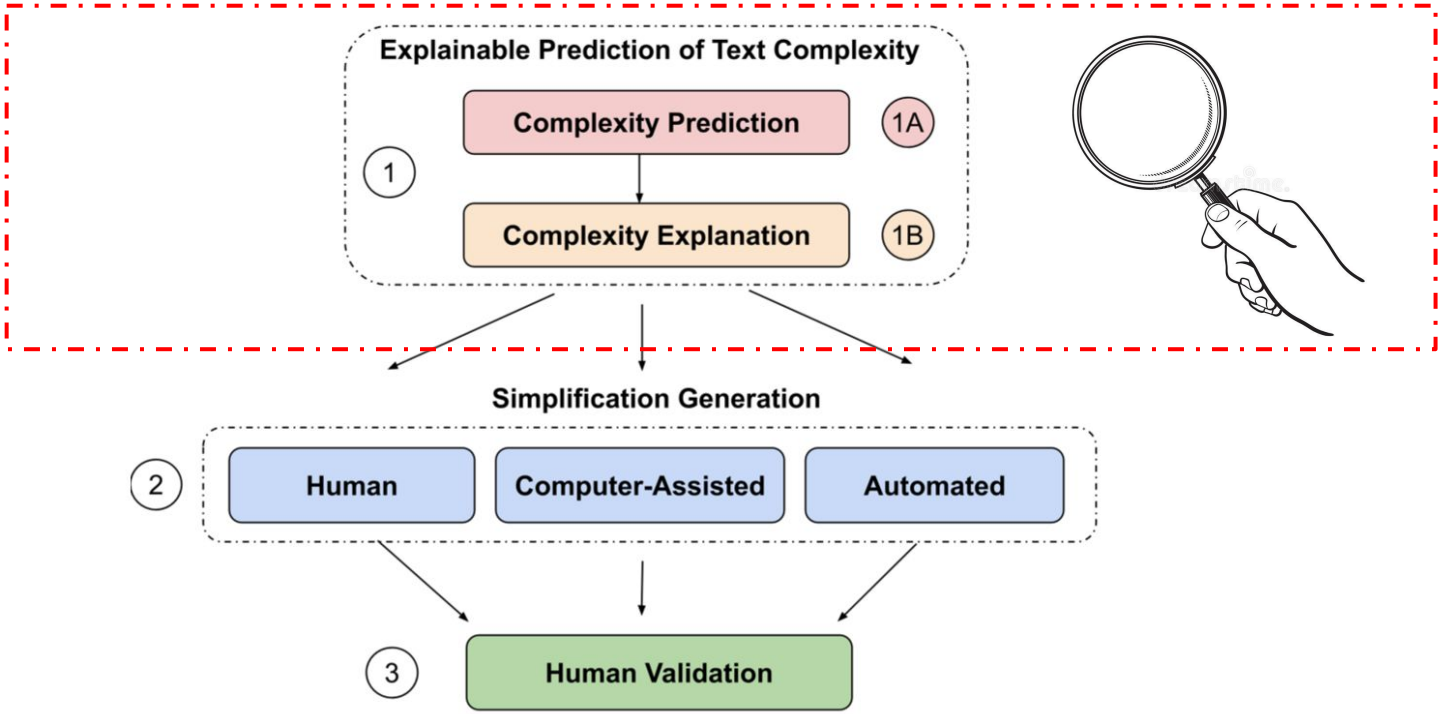


Contribution: the general problem of text simplification can be formally decomposed into a compact and transparent pipeline of modular tasks => improved performance of black-box simplification models

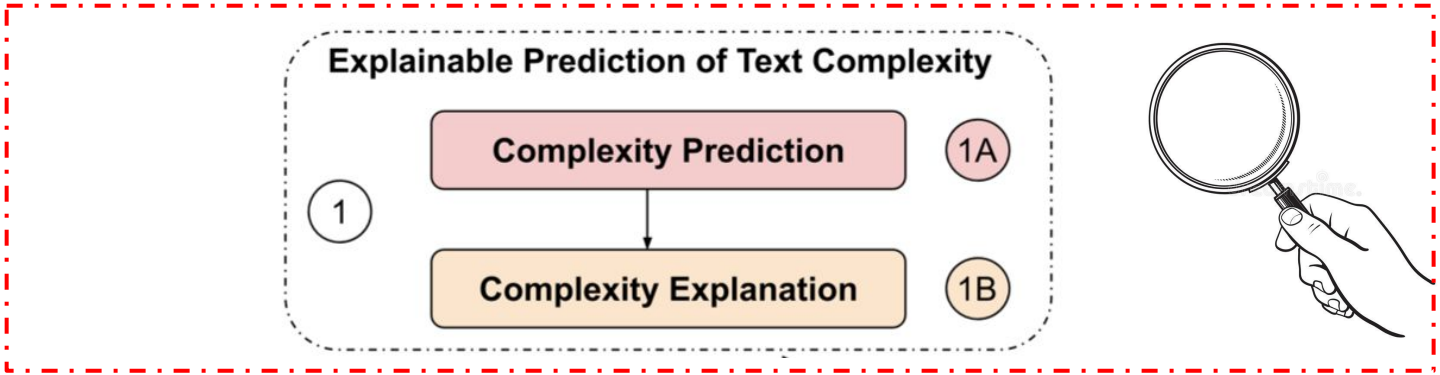
Explainable Pipeline for Text Simplification



Explainable Pipeline for Text Simplification



Explainable Prediction of Text Complexity



It is critical to explain the rationale behind simplification decisions, especially behind a black-box model**:**

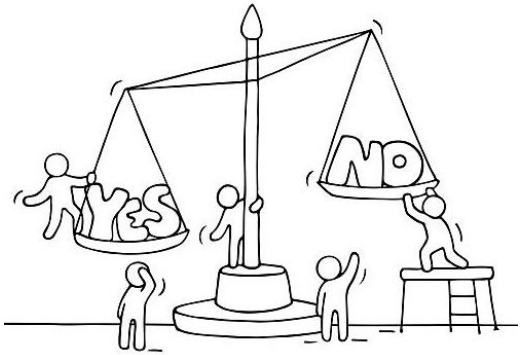
1) **Complexity Prediction:** classify a piece of text into two categories, needing simplification or not

2) **Complexity Explanation:** highlight parts of the text that need to be simplified

No prior work has addressed the explainability of text complexity prediction!

Explainable Prediction of Text Complexity

Complexity Prediction



Candidate Models:

Shallow Classifiers: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF)

Deep Classifiers: LSTM & CNN (word level and char-level), Extractive Adversarial Networks (Carton et al, 2018)

Pre-trained Classifiers: ULMFit (Howard et al, 2018), BERT (Devlin et al, 2019), XLNeT (Yang et al, 2019)

Evaluation Metric: classification accuracy

Explainable Prediction of Text Complexity

Complexity Explanation

Candidate Models:
LIME (Ribeiro et al, 2016), SHAP (Lundberg et al, 2017)
Extractive Adversarial Networks (Carton et al, 2018)

Evaluation Metrics:
Tokenwise Precision, Recall, F1
Word-level Edit distance (ED) (Levenstein, 1966)
Translation Edit Rate (TER) (Snover et al, 2006)

Baselines:

1) **Random Highlighting:** randomly draw the size and the positions of tokens to highlight
2) **Lexicon based highlighting:** Age-of-Acquisition (AoA) lexicon (Kuperman et al., 2012)
3) **Feature highlighting:** most important features of best LR models

Text Simplification Datasets

Datasets from three different domains:

- **news**: Newsela (Xu et al, 2015) with splits from (Zhang and Lapata, 2017)
- **Wikipedia**: WikiLarge (Kauchak, 2013) with splits from (Zhang and Lapata, 2017)
- **scientific papers:** Biendata (2019) - corpus of research papers matched with press releases

Table 1: Aligned complex-simple sentence pairs.

Dataset	Training	Validation	Test
Newsela	94,208 pairs	1,129 pairs	1,077 pairs
WikiLarge	208,384 pairs	29,760 pairs	59,546 pairs
Biendata	29,700 pairs	4,242 pairs	8,486 pairs

Inferring Ground-truth Labels

For **complexity prediction**:

- **Label 1**: complex sentence != aligned simple sentence
- **Label 0**: complex sentence == aligned simple sentence

For **complexity explanation**:

- ideally, ground truth annotations at the token level are available
- as a proxy, all tokens w_i in complex sentence *d* which are absent in simple sentence *d'* are candidate words for deletion or substitution

Complexity Prediction Results

Does an input sentence need to be simplified?

Difficulty of complexity prediction task varies:

across models:

pre-trained classifiers yield superior performance

across domains:

~80% accuracy for Newsela and WikiLarge,

>95% accuracy for Biendata

Classifier	Newsela	WikiLarge	Biendata
NB n-grams	73.10 %	62.70 %	84.30 %
NB enriched features	73.10 %	63.10 %	86.00 %
LR n-grams	75.30 %	71.90 %	89.60 %
LR enriched features	76.30 %	72.60 %	91.70 %
SVM n-grams	75.20 %	71.90 %	89.50 %
SVM enriched features	77.39 %	70.16 %	88.60 %
RF n-grams	71.50 %	71.50 %	84.60 %
RF enriched features	74.40 %	73.40 %	87.00 %
LSTM (word-level)	73.31 %	71.62 %	89.87 %
CNN (word-level)	70.71 %	69.27 %	89.05 %
CNN (char-level)	78.83%†	74.88 %	88.00 %
CNN (word & char-level)	75.90	74.00 %	92.30 %
Extractive Adversarial Networks	72.76 %	71.50 %	88.64 %
ULMFiT	**80.83%****	74.80 %	94.17 %
BERT	77.15 %	**81.45%****	94.43 %
XLNeT	78.83%†	73.49 %	**95.48%****

Complexity Explanation Results

How well can complexity classification be explained?

Evaluation: TER, ED 1.5 metrics (the lower the better) capture variations among explanations

Baselines: AoA lexicon, LR features are strong baselines

Models:
LIME & LSTM, SHAP & LR strong on all datasets
Extractive Adversarial Networks (high ED 1.5): jointly making predictions and generating explanations helps

Dataset	Explanation Model	P	R	F1	TER	ED 1.5
Newsela	Random	0.515	0.487	0.439	0.985	13.825
	AoA lexicon	**0.556**	0.550	0.520	0.867	12.899
	LR Features	0.522	0.250	0.321	0.871	12.103
	LIME & LR	0.535	0.285	0.343	0.924	12.459
	LIME & LSTM	0.543	**0.818**	**0.621**	0.852	11.991
	SHAP & LR	0.553	0.604	0.546	0.848	12.656
	Extractive Networks	0.530	0.567	0.518	**0.781**	**11.406**
WikiLarge	Random	0.412	0.439	0.341	1.546	17.028
	AoA lexicon	0.427	0.409	0.357	1.516	16.731
	LR Features	0.442	0.525	0.413	0.993	17.933
	LIME & LR	0.461	0.509	0.415	**0.988**	18.162
	LIME & LSTM	**0.880**	0.470	0.595	1.961	25.051
	SHAP & LR	0.842	**0.531**	**0.633**	1.693	22.811
	Extractive Networks	0.452	0.429	0.359	1.434	**16.407**
Biendata	Random	0.743	0.436	0.504	1.065	12.921
	AoA lexicon	0.763	0.383	0.475	1.064	13.247
	LR Features	0.796	0.257	0.374	0.979	10.851
	LIME & LR	**0.837**	0.466	0.577	0.982	**10.397**
	LIME & LSTM	0.828	0.657	0.713	**0.952**	16.568
	SHAP & LR	0.825	0.561	0.647	0.979	11.908
	Extractive Networks	0.784	**0.773**	**0.758**	0.972	10.678

Complexity Explanation Results

How well can complexity classification be explained?

Evaluation: TER, ED 1.5 metrics (the lower the better) capture variations among explanations

Baselines: AoA lexicon, LR features are strong baselines

Models:
LIME & LSTM, SHAP & LR strong on all datasets
Extractive Adversarial Networks (high ED 1.5): jointly making predictions and generating explanations helps

Dataset	Explanation Model	P	R	F1	TER	ED 1.5
Newsela	Random	0.515	0.487	0.439	0.985	13.825
	AoA lexicon	**0.556**	0.550	0.520	0.867	12.899
	LR Features	0.522	0.250	0.321	0.871	12.103
	LIME & LR	0.535	0.285	0.343	0.924	12.459
	LIME & LSTM	0.543	**0.818**	**0.621**	0.852	11.991
	SHAP & LR	0.553	0.604	0.546	0.848	12.656
	Extractive Networks	0.530	0.567	0.518	**0.781**	**11.406**
WikiLarge	Random	0.412	0.439	0.341	1.546	17.028
	AoA lexicon	0.427	0.409	0.357	1.516	16.731
	LR Features	0.442	0.525	0.413	0.993	17.933
	LIME & LR	0.461	0.509	0.415	**0.988**	18.162
	LIME & LSTM	**0.880**	0.470	0.595	1.961	25.051
	SHAP & LR	0.842	**0.531**	**0.633**	1.693	22.811
	Extractive Networks	0.452	0.429	0.359	1.434	**16.407**
Biendata	Random	0.743	0.436	0.504	1.065	12.921
	AoA lexicon	0.763	0.383	0.475	1.064	13.247
	LR Features	0.796	0.257	0.374	0.979	10.851
	LIME & LR	**0.837**	0.466	0.577	0.982	**10.397**
	LIME & LSTM	0.828	0.657	0.713	**0.952**	16.568
	SHAP & LR	0.825	0.561	0.647	0.979	11.908
	Extractive Networks	0.784	**0.773**	**0.758**	0.972	10.678

Complexity Explanation Results

How well can complexity classification be explained?

Evaluation: TER, ED 1.5 metrics (the lower the better) capture variations among explanations

Baselines: AoA lexicon, LR features are strong baselines

Models:
LIME & LSTM, SHAP & LR strong on all datasets
Extractive Adversarial Networks (high ED 1.5): jointly making predictions and generating explanations helps

Dataset	Explanation Model	P	R	F1	TER	ED 1.5
Newsela	Random	0.515	0.487	0.439	0.985	13.825
	AoA lexicon	**0.556**	0.550	0.520	0.867	12.899
	LR Features	0.522	0.250	0.321	0.871	12.103
	LIME & LR	0.535	0.285	0.343	0.924	12.459
	LIME & LSTM	0.543	**0.818**	**0.621**	0.852	11.991
	SHAP & LR	0.553	0.604	0.546	0.848	12.656
	Extractive Networks	0.530	0.567	0.518	**0.781**	**11.406**
WikiLarge	Random	0.412	0.439	0.341	1.546	17.028
	AoA lexicon	0.427	0.409	0.357	1.516	16.731
	LR Features	0.442	0.525	0.413	0.993	17.933
	LIME & LR	0.461	0.509	0.415	**0.988**	18.162
	LIME & LSTM	**0.880**	0.470	0.595	1.961	25.051
	SHAP & LR	0.842	**0.531**	**0.633**	1.693	22.811
	Extractive Networks	0.452	0.429	0.359	1.434	**16.407**
Biendata	Random	0.743	0.436	0.504	1.065	12.921
	AoA lexicon	0.763	0.383	0.475	1.064	13.247
	LR Features	0.796	0.257	0.374	0.979	10.851
	LIME & LR	**0.837**	0.466	0.577	0.982	**10.397**
	LIME & LSTM	0.828	0.657	0.713	**0.952**	16.568
	SHAP & LR	0.825	0.561	0.647	0.979	11.908
	Extractive Networks	0.784	**0.773**	**0.758**	0.972	10.678

Complexity Explanation Results

How well can complexity classification be explained?

Evaluation: TER, ED 1.5 metrics (the lower the better) capture variations among explanations

Baselines: AoA lexicon, LR features are strong baselines

Models:
LIME & LSTM, SHAP & LR strong on all datasets
Extractive Adversarial Networks (high ED 1.5): jointly making predictions and generating explanations helps

Dataset	Explanation Model	P	R	F1	TER	ED 1.5
Newsela	Random	0.515	0.487	0.439	0.985	13.825
	AoA lexicon	**0.556**	0.550	0.520	0.867	12.899
	LR Features	0.522	0.250	0.321	0.871	12.103
	LIME & LR	0.535	0.285	0.343	0.924	12.459
	LIME & LSTM	0.543	**0.818**	**0.621**	0.852	11.991
	SHAP & LR	0.553	0.604	0.546	0.848	12.656
	Extractive Networks	0.530	0.567	0.518	**0.781**	**11.406**
WikiLarge	Random	0.412	0.439	0.341	1.546	17.028
	AoA lexicon	0.427	0.409	0.357	1.516	16.731
	LR Features	0.442	0.525	0.413	0.993	17.933
	LIME & LR	0.461	0.509	0.415	**0.988**	18.162
	LIME & LSTM	**0.880**	0.470	0.595	1.961	25.051
	SHAP & LR	0.842	**0.531**	**0.633**	1.693	22.811
	Extractive Networks	0.452	0.429	0.359	1.434	**16.407**
Biendata	Random	0.743	0.436	0.504	1.065	12.921
	AoA lexicon	0.763	0.383	0.475	1.064	13.247
	LR Features	0.796	0.257	0.374	0.979	10.851
	LIME & LR	**0.837**	0.466	0.577	0.982	**10.397**
	LIME & LSTM	0.828	0.657	0.713	**0.952**	16.568
	SHAP & LR	0.825	0.561	0.647	0.979	11.908
	Extractive Networks	0.784	**0.773**	**0.758**	0.972	10.678

Benefit of Complexity Prediction

A smart end-to-end simplification model should not further simplify input if already simple

Select current best pre-trained models from literature:

- ACCESS (Martin et al, 2020)
- DMLMTL (Guo et al, 2018)

Out-of-sample simple sentences are changed:
> 90% DMLMTL, >70% ACCESS

Benefit of Complexity Prediction

A smart end-to-end simplification model should not further simplify input if already simple

Select current best pre-trained models from literature:

- ACCESS (Martin et al, 2020)
- DMLMTL (Guo et al, 2018)

Out-of-sample simple sentences are changed:
> 90% DMLMTL, >70% ACCESS

Blindly simplifying texts regardless of complexity level (erroneously) over-simplifies simple sentences!

Out-of-Sample Simple Sentences Incorrectly Simplified by State-of-the-Art Text Simplification Models

- In Ethiopia, HIV disclosure is low → In Ethiopia , HIV is low (ACCESS)
- Mustafa Shahbaz , 26 , was shopping for books about science . → Mustafa Shahbaz , 26 years old , was a group of books about science . (ACCESS)
- Healthy diet linked to lower risk of chronic lung disease → Healthy diet linked to lung disease (DMLMTL)
- Social workers can help patients recover from mild traumatic brain injuries → Social workers can cause better problems . (DMLMTL)

Semantic, syntactic and logical errors abound in the simplified simple sentences!

Benefit of Complexity Prediction

Without Complexity Prediction:
evaluation loss when simple sentences are over-simplified

With Complexity Prediction:
evaluation loss for an imperfect complexity predictor

Q: Which loss is higher?
- complexity predictors are > 80% accurate
- considerable drop of errors (30-70%) in all evaluation metrics

with complexity prediction; accurate predictor, higher benefit

Benefit of Complexity Prediction

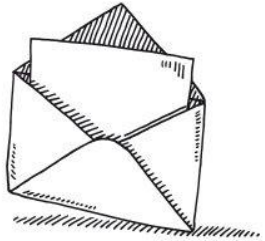
Without Complexity Prediction:
evaluation loss when simple sentences are over-simplified

With Complexity Prediction:
evaluation loss for an imperfect complexity predictor

Q: Which loss is higher?
- complexity predictors are > 80% accurate
- considerable drop of errors (30-70%) in all evaluation metrics with complexity prediction; accurate predictor, higher benefit

Out-of-sample performance substantially improves with complexity prediction!

Main Takeaways



Formally decompose text simplification into a compact pipeline of sub-tasks
Explainable Complexity Prediction preliminary step for text simplification

Complexity Prediction: the system should first decide if simplification is necessary
often neglected by existing end-to-end simplifiers => biased outputs, lack of generalization

Complexity Explanation: highlight differences between original and simplified sentence
validate and evaluate black-box models, enhance fidelity and fairness in real-world

Major motivation of text simplification: **improve fairness and transparency**
it is critical to explain the rationale behind the simplification decisions behind a black-box model