Georgia Tech

Revisiting non-English Text Simplification: A Unified Multilingual Benchmark

Michael J. Ryan, Tarek Naous, Wei Xu

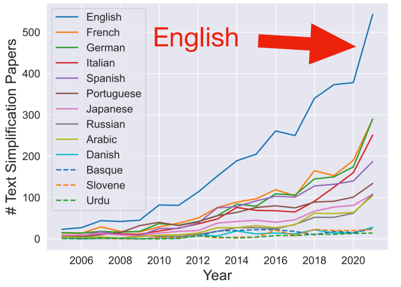
We released a text simplification dataset spanning 12 languages and over 1.7 million sentence pairs!

Code + Data

https://github.com/XenonMolecule/MultiSim

1. Motivation

- English text simplification research is growing faster than any other language.
- Lack of an accessible benchmark for text simplification in many languages.



2. Literature Survey

2 Translated Resources

French | Russian

5 Automatically Collected Resources

English | German | French | Italian

8 Target Audience (Company Driven) Resources

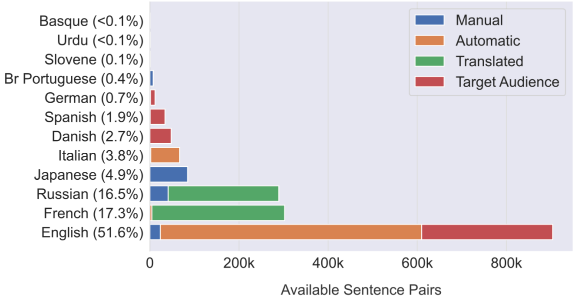
Arabic | English | Slovene | Japanese | Danish | German | Spanish

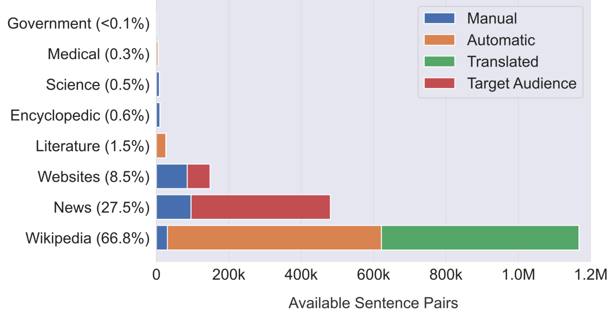
19 Manually Simplified Resources

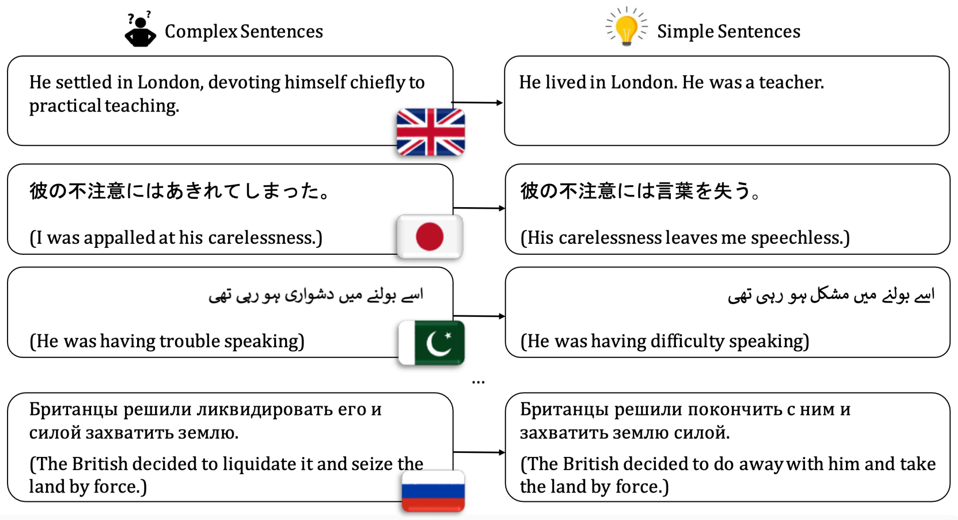
Basque | Brazilian Portuguese | English | French | German | Italian | Japanese | Urdu | Russian | Spanish

3. MultiSim Benchmark

- 12 Languages
- 8 Domains
- 27 Resources
- 1.7+ million sentence pairs







Complex Sentences

Simple Sentences

He settled in London, devoting himself chiefly to practical teaching. → He lived in London. He was a teacher.

彼の不注意にはあきれてしまった。
(I was appalled at his carelessness.)
→ 彼の不注意には言葉を失う。
(His carelessness leaves me speechless.)

اسے بولنے میں دشواری ہو رہی تھی
(He was having trouble speaking)
→ اسے بولنے میں مشکل ہو رہی تھی
(He was having difficulty speaking)

...

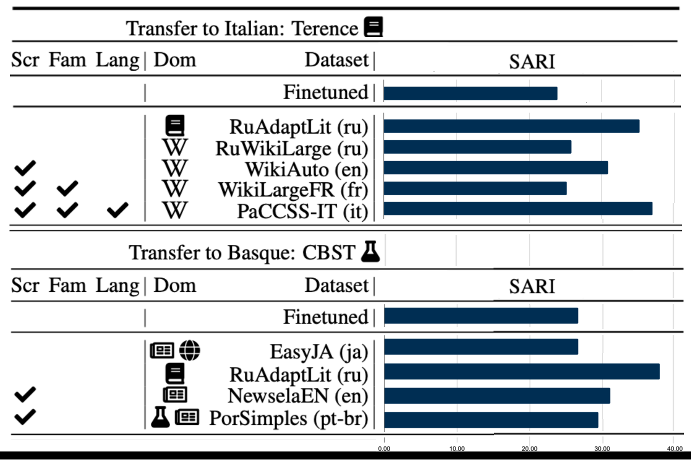
Британцы решили ликвидировать его и силой захватить землю.
(The British decided to liquidate it and seize the land by force.)
→ Британцы решили покончить с ним и захватить землю силой.
(The British decided to do away with him and take the land by force.)

4. Finetuning Experiments

- Train mT5 to simplify text on **single** dataset, all data in one **language**, or **all** of MultiSim.
- **Joint-All training helps all languages besides English** which already has a wealth of in-language training data.

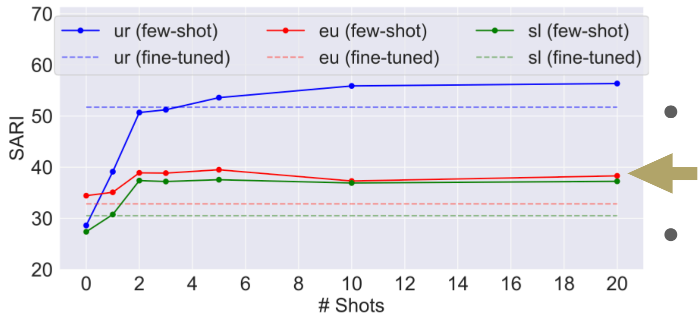
Lang	Dataset	SARI Single	 Lang	 All
es	Simplext	—	19.91	**32.68**
	NewselaES	29.89	28.56	**35.36**
da	DSim	31.40	31.40	**38.44**
it	Simpitiki	—	20.10	**24.27**
	Teacher	—	29.98	**30.97**
	AdminIT	—	34.72	**36.21**
	Terence	—	**37.77**	36.92
	PaCCSS-IT	**57.30**	55.98	54.43
ja	EasyJA	67.36	**70.95**	70.11
	EasyJAExt	43.15	50.26	**53.49**
ru	RuAdaptFairy	—	23.77	**26.55**
	RuAdapt Ency	—	**34.73**	34.40
	RSSE	—	29.49	**35.08**
	RuAdapt Lit	41.75	**42.03**	42.01
	RuWikiLarge	32.01	34.95	**37.59**
fr	CLEAR	34.86	30.85	**35.37**
	WikiLargeFR	35.20	38.22	**39.23**
en	ASSET	35.98	**42.77**	41.56
	NewselaEN	38.60	**40.18**	38.80
	WikiAuto	42.46	**42.48**	42.00

5. Cross-lingual Transfer Experiments

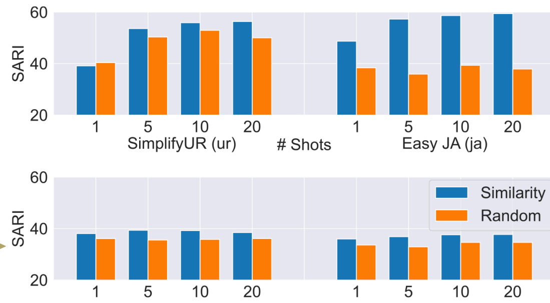


- Train mT5 on data in one language and evaluate on another.
- Matching **script** and **language** help improve transfer performance.
- Matching **domain** can help regardless of script.
- **Russian** is a good candidate language for cross-lingual transfer.

6. Few-shot Experiments



- Prompt BLOOM with example sentences from training set.
- **Better than fine-tuning** for **low resource languages**.
- Picking examples by **semantic similarity** search works better than random sampling.



We validate all findings through human evaluation!