An Open Multilingual System for Scoring Readability of Wikipedia

The paper introduces a new dataset and multilingual system to score the readability of the text of Wikipedia articles

Authors

Mykola Trokhymovych
mykola.trokhymovych@upf.edu

Indira Sen
indira.sen@uni-konstanz.de

Martin Gerlach
mgerlach@wikimedia.org

Introduction

Readability measures how easy it is to read a given text. Previous studies have shown that the readability of English Wikipedia is often poor, creating barriers for many readers and editors. Moreover, measuring readability beyond English is an open problem due to a lack of scalable formulas, multilingual models, and ground-truth data. In this paper, we address these challenges by collecting new data and building a multilingual readability scoring model.

Challenges

- Adapting readability formulas to each language not scalable/feasible
- Lack of multilingual models/tools
- Scarcity of open multilingual ground-truth data

Contributions

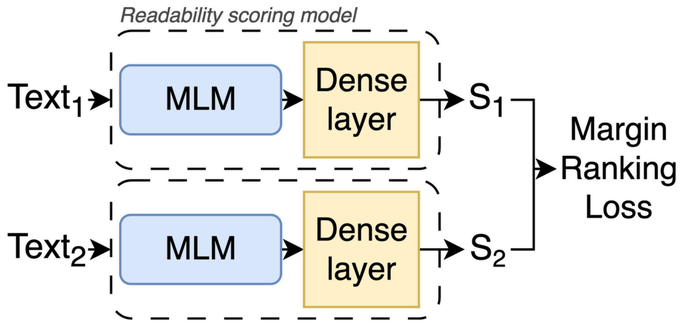
- A new open multilingual dataset covering 14 languages;
- A multilingual model for text readability scoring, working in zero-shot cross-lingual transfer;
- The first systematic overview of the state of readability of Wikipedia articles beyond English;
- A public API endpoint of the model for use by readers, editors, and researchers.

Approach

- Match Wikipedia articles with a simplified version to build the dataset of text pairs of two readability levels;
- Use Siemease architecture for training pairwise multilingual text readability ranking model;
- Use a new multilingual model to define the state of readability of Wikipedia across languages.

System sketch

A new readability scoring model



- xlm-roberta-longformer as the base MLM
- ~100 languages support
- Siemease architecture for training:
  - Training is based on pairwise ranking, but still able to score individual texts
- Loss: Margin Ranking Loss
- Fine-tuning only using data in English:
  - No language-specific fine-tuning for other languages! (zero-shot)

A new multilingual dataset

WikiReaD (Wikipedia Readability Dataset)

What? Pairs of encyclopedic articles in two readability levels (simple, hard) for 14 languages.

How? Matching Wikipedia articles with a simplified version.

- Simple English Wikipedia (>100K pairs)
- Children encyclopedias (10-10K pairs)

Model performance

Setup 1: Evaluation on English.
- Simple Wikipedia (training corpus): RA=0.976
- Vikidia (out-of-corpus): RA=0.991

Setup 2: Evaluation on Non-English data (zero-shot cross-lingual transfer)
- RA > 0.8 for all languages
- RA > 0.9 for 10/15 datasets

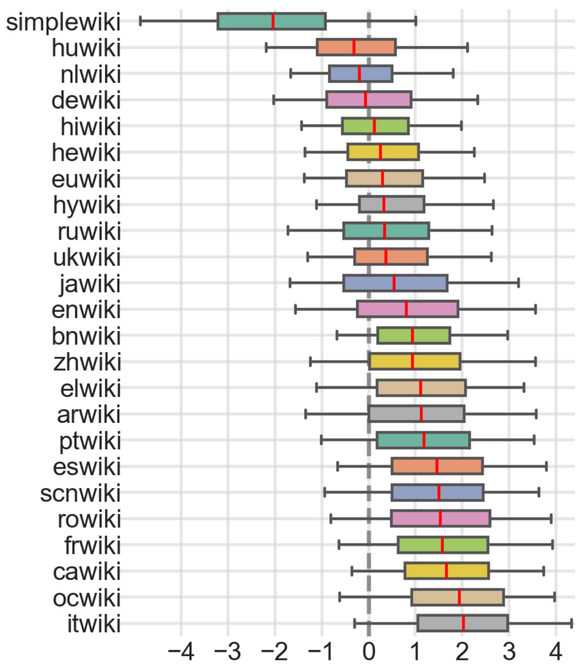
Setup 3: Evaluation on reference benchmarks (e.g. Lee&Vajjala, ACL 2022)

Outperform reference benchmarks
- VikidiaFr: 0.978 (ours) vs 0.811 (NPRM)
- OneStopEnglish: 0.974 (ours) vs 0.878 (NPRM)

Dataset statistics (easy/hard)

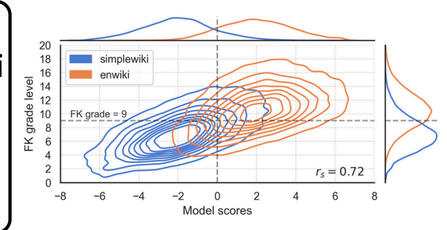
	Dataset	#Pairs	Avg. #Sen.	Avg. #Char.
English	simplewiki-en	112,342	6.2/7.9	84.6/130.9
English	vikidia-en	1,991	6.4/14.3	83.3/142.8
Catalan	vikidia-ca	234	5.2/9.7	79.3/145.2
German	vikidia-de	260	6.4/11.2	75.8/131.0
Greek	vikidia-el	39	6.0/11.8	96.8/134.9
Spanish	vikidia-es	1,915	5.7/7.7	109.0/179.4
Basque	vikidia-eu	571	6.5/8.7	114.6/129.5
French	vikidia-fr	12,221	5.7/7.3	106.9/152.1
Armenian	vikidia-hy	485	14.3/11.4	105.3/115.1
Italian	vikidia-it	1,662	4.5/6.0	84.6/152.6
Occitan	vikidia-oc	12	4.2/7.1	77.0/105.6
Portuguese	vikidia-pt	809	5.7/11.8	97.3/157.9
Russian	vikidia-ru	125	5.8/11.2	83.8/110.6
Sicilian	vikidia-scn	10	3.8/4.7	50.9/86.3
German	klexikon-de	2,255	17.7/8.9	73.9/136.9
Basque	txikipedia-eu	1,162	7.3/8.4	107.4/126.4
Dutch	wikikids-nl	12,090	8.0/7.5	83.7/112.0

The state of readability of Wikipedia across languages



A model score of approx. 0 separates texts from simplewiki (easy) and enwiki (hard).

This corresponds to Flesch-Kincaid grade level (FKGL) ~ 9.



Overall readability in most Wikipedias are similar to English Wikipedia

Dataset

Gitlab repo

API endpoint