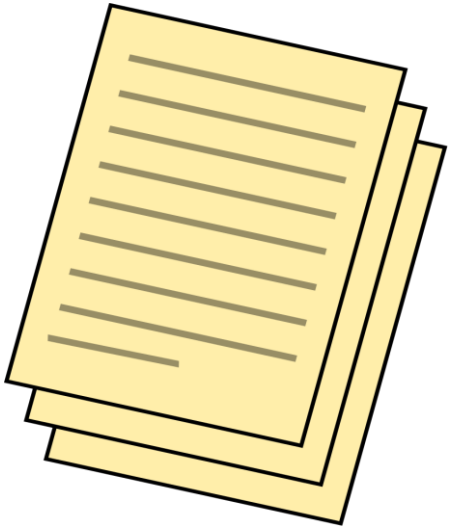
Automatic Assessment of Text Readability

Hypothesis: Text predictability is a measure of its readability

Why is it important?

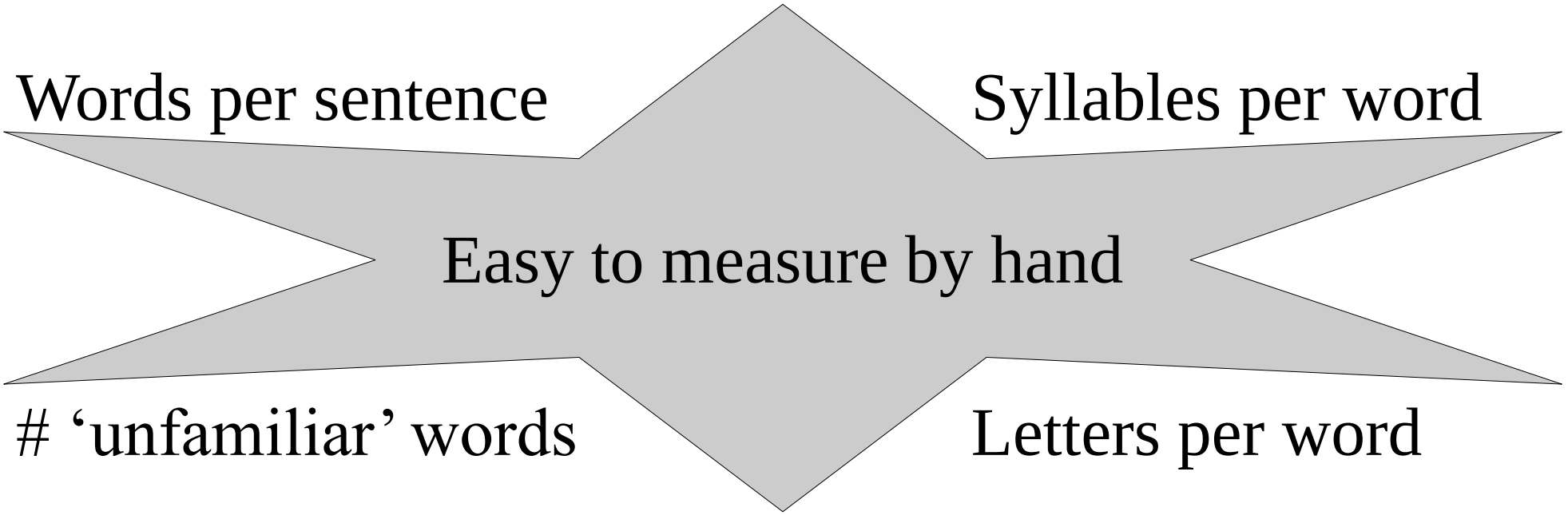
Authors can improve quality of writing

Produce better search results



Organizations can ensure writing standards

Early Formula-Based Methods



Supervised learning based methods

Features

- Sentence length
- Frequency of words
- Discourse relations
- Parse tree depth
- Number of noun phrases

Ground truth

- Texts classified by grade level
- Pairs of phrases ranked by humans

~~Features~~ => Deep learning

~~Ground truth~~ => Our central hypothesis

“Predictability (or perplexity) of a text, as determined by a modern language model, is a good measure of its readability”

The Algorithm

- Train language models on various corpora
- Each corpus represents a target reader's linguistic background
- Compute perplexity on given text as the measure of readability

Outputs:

1) Minimum perplexity among all models

2) Perplexity of model trained on target reader's corpus (eg. medical text for a doctor)

Why is this important?

- Feature engineering requires effort, separate for each language
- Ground truth is often unavailable and expensive to create. If it isn't needed, we can:
  1. Produce scores for various languages
  2. Produce domain specific scores
- Current methods don't generalize to all kinds of text
- Entropy is a canonical measure

Experiments

Baselines

- We compare our approach to six formula-based metrics and one ML based scheme..
- We use unigram models provided by Kevyn Collins-Thompson Jamie Callan as well as those produced from our own dataset in our evaluations.

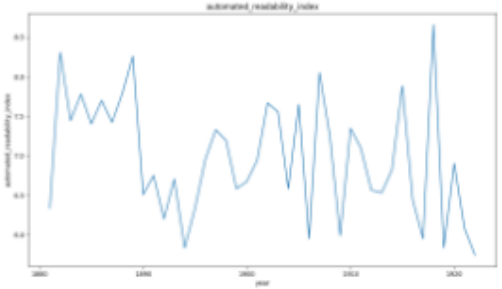
WSDM Dataset	12,728 pairwise comparison annotated by 624 human readers	To check how our method appropriates human readers
NCERT Dataset	Mathematics books from grades 6-12, 1000 passages each of 50-100 words	To evaluate performance on short and large passages.
Simple vs. Standard Wikipedia	Articles from simple wikipedia and corresponding standard wikipedia	To evaluate on the same topic with different levels of difficulty

	WSDM	Textbook	Wikipedia
ARI	79%	52%	97%
FRE	75%	49%	94%
FKGL	75%	50%	96%
GFI	71%	53%	97%
SI	63%	54%	98%
CLI	79%	54%	94%
UNI-WSDM	91%	47%	60%
UNI-Textbook	54%	95%	55%
This Paper	81%	76%	95%

Accuracy of prediction of pairwise readability for various schemes

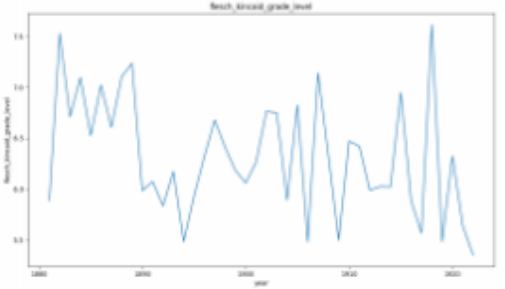
2. Short Term Language Change

Readability seen as a function of time, for novels published between 1881 and 1922 in the Oxford “Corpus of English Novels”



ARI

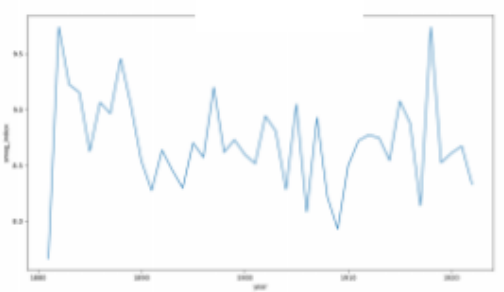
CLI



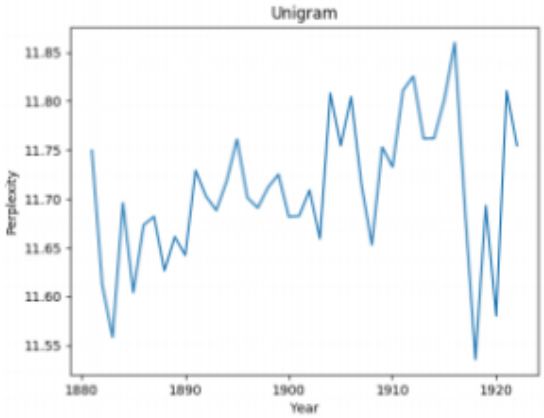
FKGL

FRE

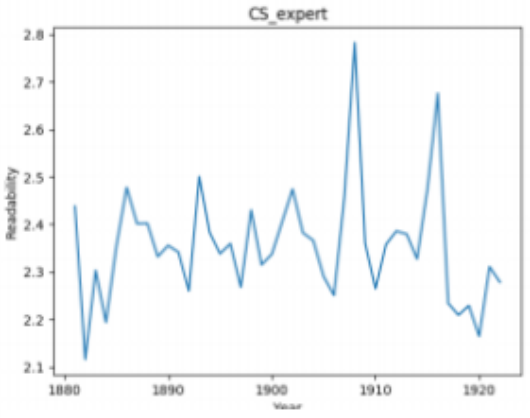
GFI



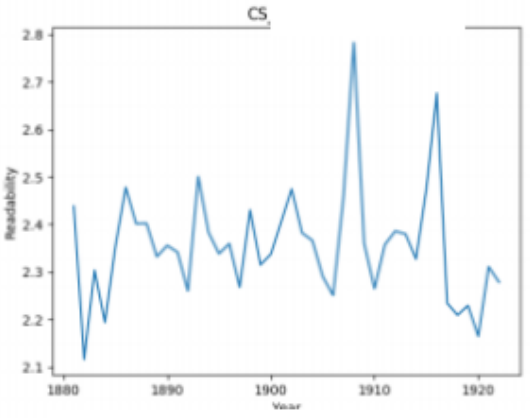
SMOG



Unigram Model



Biology expert corpus

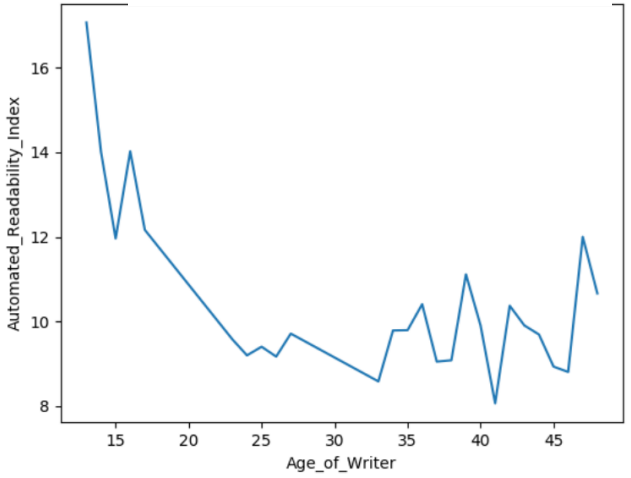


CS expert corpus

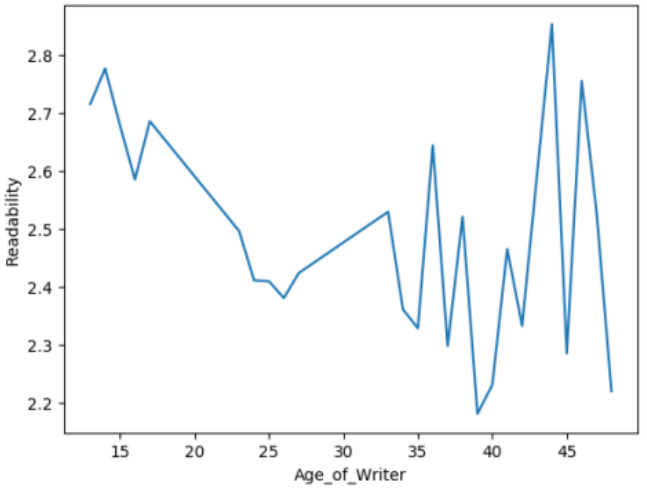
3. Effect of author's age on blog readability

To analyze the readability of written text as a function of the author's age and gender.

Staitical Formula



Biology expert Corpus



CS expert Corpus