Simple or Complex? Learning to Predict Readability of Bengali Texts

Susmoy Chakraborty[1*], **Mir Tafseer Nayeem**[1*], **Wasi Uddin Ahmad**[2]

[1]Ahsanullah University of Science and Technology, [2]University of California, Los Angeles

susmoyaust36@gmail.com, mir.nayeem@alumni.uleth.ca, wasiahmad@ucla.edu

*Equal contribution, listed by alphebetical order

AAAI 2021

What is Readability?

- Measures how much energy the reader will have to expend in order to understand writing at optimal speed and find interesting
- First step of Text Simplification

Research Goal

Readability analysis of low-resource language **Bengali**: 7th most spoken language in the world with 230 million native speakers

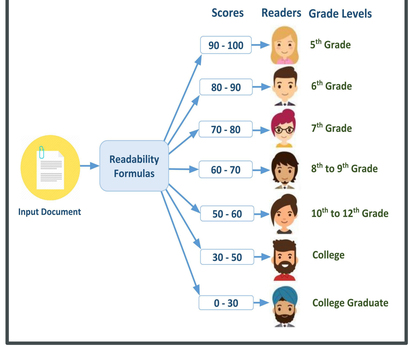
Motivation

- Importance of **Readability Measurement** in education, health care, government, etc.
- Languages having readability analysis tool: English (e.g., **Grammarly** and **Readable**), Arabic, Italian, Japanese
- **Bengali language:** No such tool available, previous works are narrow and sometimes faulty due to the lack of resources

Our Contributions

- Age-to-age comparison to adapt U.S. education system based readability formulas
- **Document-level dataset:** 618 documents with 12 different grade levels
- **Sentence-level dataset:** 96,335 sentences with simple and complex labels
- Neural architectures, which will serve as baseline for future works
- Consonant conjunct count algorithm and a human-annotated corpus comprising 341 words to evaluate the effectiveness of this algorithm
- Updated pronunciation dictionary with more than 67k words
- 3,396 Bengali easy words list
- Bengali readability analysis tool

Readability Prediction task



Dataset

- Documents from several published textbooks, popular sources from Bangladesh and India for children and adults
- Dividing documents into sentences to create large-scale dataset due to long range dependencies of RNNs, Insufficient document-level dataset for training supervised neural models

Dataset	#Docs	Avg. #sents	Avg. #words
NCTB	380	66.8	585.8
Additional	238	391.2	3045.0

Statistics of document-level dataset

	Train	Dev	Test
Simple Sentences			
#Sents	37,902	1,100	1,100
Avg. #words	8.16	7.97	8.31
Avg. #chars	44.71	43.85	45.57
Complex Sentences			
#Sents	54,033	1,100	1,100
Avg. #words	8.04	8.08	8.16
Avg. #chars	44.01	44.65	44.63

Statistics of sentence-level dataset

Document-level dataset to experiment with **formula-based approaches**

Sentence-level dataset to train **supervised neural models**

Experiments: Formula-based Approaches

We apply **6** U.S. education system based readability formulas to our Bengali documents with proper **age-to-age comparison**

Document	BN age	ARI	U.S. age	FE	U.S. age	FK	U.S. age	GF	U.S. age	SM OG	U.S. age	DC	U.S. age
Class 1	6	1	**5-6**	40.9	18-22	9	14-15	6	11-12	N/A	-	5.9	10-12
Class 2	7	1	5-6	30.6	18-22	10	15-16	10	15-16	9	14-15	5.3	10-12
Class 3	8	3	**7-9**	21.9	≥21	12	17-18	11	16-17	10	15-16	7.2	14-16
Class 4	9	3	**7-9**	34.1	18-22	10	15-16	9	14-15	9	14-15	7.3	14-16
Class 5	10	6	11-12	11.0	≥21	13	18-19	15	20-21	12	17-18	7.4	14-16
Class 6	11	4	9-10	21.1	≥21	12	17-18	14	19-20	11	16-17	8.2	16-18
Class 7	12	6	**11-12**	13.1	≥21	13	18-19	13	18-19	11	16-17	7.2	14-16
Class 8	13	6	11-12	16.2	≥21	13	18-19	13	18-19	12	17-18	8.5	16-18
Class 9/10	14-15	12	17-18	-8.6	-	18	≥20	20	≥21	17	≥19-20	7.3	**14-16**
Class 11/12	16-17	11	**16-17**	-2.6	-	18	≥20	19	≥21	16	≥19-20	8.1	**16-18**
Children 1	6-10	1	**5-6**	32.0	18-22	10	15-16	8	13-14	8	13-14	5.0	**10-12**
Children 2	6-10	2	**6-7**	33.8	18-22	10	15-16	9	14-15	9	14-15	6.1	12-14
Adults 1	≥18	12	**17-18**	-22.8	-	21	**≥20**	24	**≥21**	19	**≥19-20**	11.5	**≥21**
Adults 2	≥18	3	7-9	27.3	**≥21**	11	16-17	10	15-16	9	14-15	7.1	14-16

Performance of formula-based approaches, bold values: correct prediction, **ARI** formula performed well

Experiments: Supervised Neural Approaches

Binary sentence classification problem, classes: simple and complex

Baseline: BiLSTM, BiLSTM + Attention

Ablation study: BiLSTM with Global Average Pooling and Global Max Pooling

- **CL** (Character Length) and **CC** (Consonant Conjunct) feature fusion: Extraction of CL (with white spaces) and CC from input sentences to concatenate with pooling layers

Simple: আমরা এই সব পোশাক প্রতিদিন পরি
[We wear all these clothes everyday]
CL: 30

CC: আমরা এই সব পোশাক প্রতিদিন পরি = 1

Complex: তাহার ওষ্ঠাধরের উভয় প্রান্ত ঈষৎ প্রসারিত হইল মাত্র
[Only the ends of his lips were slightly extended]
CL: 50

CC: তাহার ওষ্ঠাধরের উভয় প্রান্ত ঈষৎ প্রসারিত হইল মাত্র = 5

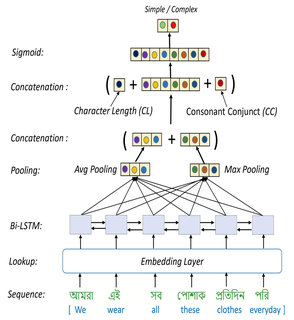
Algorithm 1: Consonant conjunct count algorithm.

```
Procedure ConsonantConjunctCount(W)
    Data: Input word W, which is an array of Bengali characters.
    Result: Return the number of consonant conjuncts in input word W.
    A ← Bengali sign VIRAMA (Wikipedia 2020);
    cc_count ← 0;
    l ← length(W);
    for k ← 0 to l − 1 do
        if W[k] == A then
            if k − 1 ≥ 0 and k + 1 < l then
                if k − 2 ≥ 0 then
                    if W[k − 1] and W[k + 1] is a Bengali Consonant and W[k − 2] != A then
                        cc_count ← cc_count + 1;
                    end
                end
                else if W[k − 1] and W[k + 1] is a Bengali Consonant then
                    cc_count ← cc_count + 1;
                end
            end
        end
    end
    return cc_count;
```

We use all Bengali **pretrained language models** available to date

Our code, data and all other resources:
https://github.com/tafseer-nayeem/BengaliReadability

Experiments: Supervised Neural Approaches (contd.)



Our readability prediction model

Baseline Models				
Models	**A**	**R**	**P**	**F1**
BiLSTM	77.5	69.4	82.8	75.5
BiLSTM + Attention	76.4	65.9	83.3	73.6
Ablations				
Models	**A**	**R**	**P**	**F1**
BiLSTM with Pooling	81.3	78.8	83.0	80.8
+ Word2vec	85.5	80.2	89.7	84.7
+ CL + CC	85.7	80.9	89.5	85.0
+ GloVe	86.1	79.3	**91.9**	85.1
+ CL + CC	86.1	81.3	89.9	85.4
+ fastText	86.0	80.1	90.8	85.1
+ CL + CC	**86.4**	82.9	89.1	85.9
+ BPEmb	86.2	81.5	90.0	85.6
+ CL + CC	86.0	81.2	89.8	85.3
+ ULMFiT	85.5	77.6	**92.0**	84.2
+ CL + CC	86.2	80.4	91.0	85.4
+ TransformerXL	86.3	82.7	89.0	85.8
+ CL + CC	**86.7**	83.5	89.3	**86.3**
+ LASER	**86.4**	84.3	88.0	86.1
+ CL + CC	86.3	**84.6**	87.6	86.1
+ LaBSE	86.0	80.3	90.8	85.2
+ CL + CC	**86.7**	**86.5**	86.8	**86.7**

Performance of Baseline and our ablations, best: green, second best: blue. A: Accuracy, R: Recall, P: Precision, and F1: F1 score

Significant impact of CL and CC

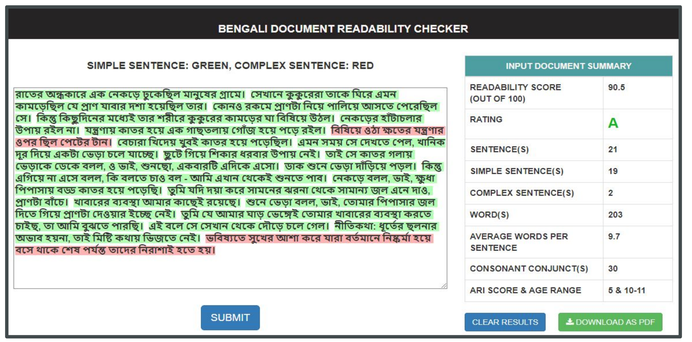
Experiments: Supervised Pretraining

fastText supervised text classification techniques

Models	A	R	P	F1
fastText Unigram	86.0	82.8	88.4	85.5
fastText Bigram	86.6	84.9	87.9	86.4
fastText Trigram	**87.4**	**85.0**	**89.2**	**87.1**

Performance of Supervised Pretraining

Our **Bengali readability analysis tool**, demo video: https://youtu.be/U05Pf9Y4tCQ



Future Works

Increasing sentence-level dataset, our tool-based user study, Bengali-English code-mixed texts