Automatic Readability Assessment for Chinese Text

Prof. Meichun LIU
Dr. John LEE

Prof. Ding-Xuan ZHOU
Dr. Zijun ZHANG

CityU
Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

CityU
School of Data Science
香港城市大學
City University of Hong Kong

Readability Assessment

- Task: Estimate the difficulty level of a text

Input text 1	Input text 2
A mother bird sat on her egg. The egg jumped. “Oh oh!” said the mother bird. “My baby will be here! He will want to eat.” “I must get something for my baby bird to eat!” she said. “I will be back!” So away she went.	But Tom’s energy did not last. He began to think of the fun he had planned for this day, and his sorrows multiplied. Soon the free boys would come tripping along on all sorts of delicious expeditions, …
Estimate: suitable for K1	**Estimate: suitable for Grades 6-8**

COMMON CORE STATE STANDARDS FOR

English Language Arts
&
Literacy in
History/Social Studies,
Science, and Technical Subjects

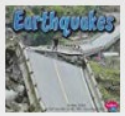
Readability Assessment

- Difficulty scales:
  - School grades (1-12); Lexile numbers; HSK 汉语水平考试 (levels 1-6); etc.
- Applications:
  - *Language learners*: search for extra-curricular reading material
  - *Language teachers*: adaptation of pedagogical material

LEXILE

Examples

Lexile



Earthquakes

by: Schuh, Mari C.
Describes earthquakes, how they occur, and the damage they cause.

470L

Add to Reading List

Pages: 24 ISBN13: 9781429634366

Find This Book

Pteranodon

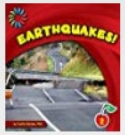
by: Riehecky, Janet
Simple text and illustrations present the life of pteranodon, how it looked, and its behavior.

440L

Add to Reading List

Pages: 24 ISBN13: 9780736853552

Find This Book



Earthquakes!

by: Minden, Cecilia
Teaches young readers about earthquakes, their causes and effects.

430L

Add to Reading List

Pages: 24 ISBN13: 9781602798649 Ages: 6 to 9

Find This Book

NTNU Chinese Readability Index Explorer

Solution (1)

- Readability formula
- E.g., Dale-Chall formula (1948)
  - Considers only sentence length and proportion of "difficult words"
  - Disregards complexity in sentence structure and meaning

$$\text{DCRF} = 0.1579\left(\frac{difficultWords}{totalWords} * 100\right) + 0.0496\left(\frac{totalWords}{totalSentences}\right)$$

Solution (2)

- Statistical classifiers
  - Trained on text features
  - E.g., 75% accuracy on 6-way classification on Chinese textbook passages with SVM (Sung et al. 2015)

Lexical	Syntactic	Semantic
Average # strokes	Average # prepositional phrases	# complex semantic categories
# adverbs	# complex sentences	# content words
…	…	…

Research goals

- Goal: improve readability assessment accuracy with:
  - Corpus with Hong Kong language data
  - Deep learning methods
  - Semantic frame features

啟思中國語文網

Corpus: Mainland data

- 2,621 graded textbook passages
  - From Prof. Xu Dekuan, Ludong University

Grade	# Passages	# words per passage	Grade	# Passages	# words per passage
1	235	109.0	7	199	1202.1
2	320	198.6	8	142	1176.9
3	386	329.5	9	134	1443.8
4	321	425.4	10	140	1617.1
5	282	569.8	11	89	1900.9
6	252	660.9	12	121	1930.7

Corpus: Hong Kong data

- 982 graded textbook passages
  - From《啟思中國語文》《啟思語文新天地》《現代中國語文》《現代普通話》

Grade	# passages	# words per passage
1	169	40.4
2	164	94.9
3	166	161.3
4	159	231.6
5	168	273.9
6	156	323.2

Baseline: SVM

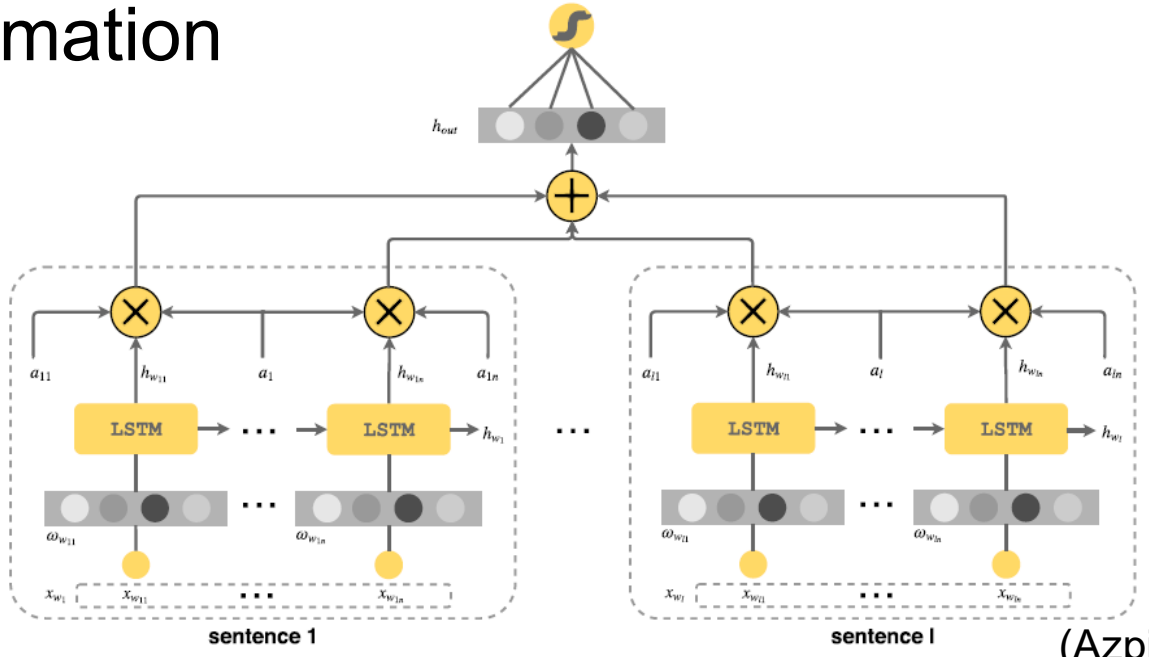
- Use SVM to predict grade of a passage
  - Trained with features in Sung et al. (2015)
  - Outperformed logistic regression, random forest

Accuracy metric	Mainland dataset (Grades 1…12)	HK dataset (Grades 1…6)
top-1	36.63%	46.14%
top-2	62.38%	71.78%
top-3	76.51%	84.97%
Adjacent-grade	55.98%	61.21%

Contribution by Dr. Jinshan Zeng, Liu Bie Ju Centre for Mathematical Sciences

Deep learning: Vec2Read

- Bidirectional long short-term memory (BiLSTM) neural network
  - Multiattentive: weighted attention for semantic, part-of-speech and morphologic information



(Azpiazu & Pera 2019)

Deep learning: Vec2Read

- State-of-the-art readability results for 7 languages (Azpiazu & Pera 2019)
- Preliminary results on Chinese
  - In progress: experiments on different embeddings for Chinese words

Accuracy metric	Mainland dataset (Grades 1…12)	
	SVM	Vec2Read
top-1	36.63%	30.97%
adjacent-grade	55.98%	68.86%

Contribution by Dr. Jinshan Zeng, Liu Bie Ju Centre for Mathematical Sciences

Semantic frame features

- Based on Mandarin VerbNet from Prof. Liu
- Features that correlate with text difficulty
  - More non-core frame elements
  - Omission of frame elements that occupy subject position
  - More NPs than clauses as verb argument
  - Use of metaphor
- In progress: automatic feature detection

Research output

- Lee, J, Liu, M et al. (2020) Using Verb Frames for Text Difficulty Assessment. *Proc. International FrameNet Workshop*

- Liu, M et al. (2020) Mandarin Physical Contact Verbs: a Frame-based Constructional Approach. *Proc. 21st Chinese Lexical Semantics Workshop*

External grants

- **GRF**
  - "Semantic Modeling for Sentence-level Readability Assessment" ($401K, 2021-22)
  - John Lee, Meichun Liu (LT) and Weiwei Sun (Cambridge University)
- Under review: **Language Fund**
  - "A Text Difficulty Analysis Tool for Developing Extra-Curricular Reading Materials"
  - John Lee, Meichun Liu (LT)