Vers une évaluation universelle du niveau de complexité des textes

Marc Benzahra
Superviseur **François Yvon**
Equipe **TLP**
Bureau **A203**
Financement **Glose**

JDD - 06/06/2019

Text complexity definition

Readability is what makes some **texts easier to read than others** at the **content** aspect of a text

e.g. Word length, sentence length, number of syllables, paragraph length, words POS tags and dependency tree

Whereas **lisibility** is about the **form** of a text

e.g. Words spacing, text font, character size

Glose

Text complexity usage

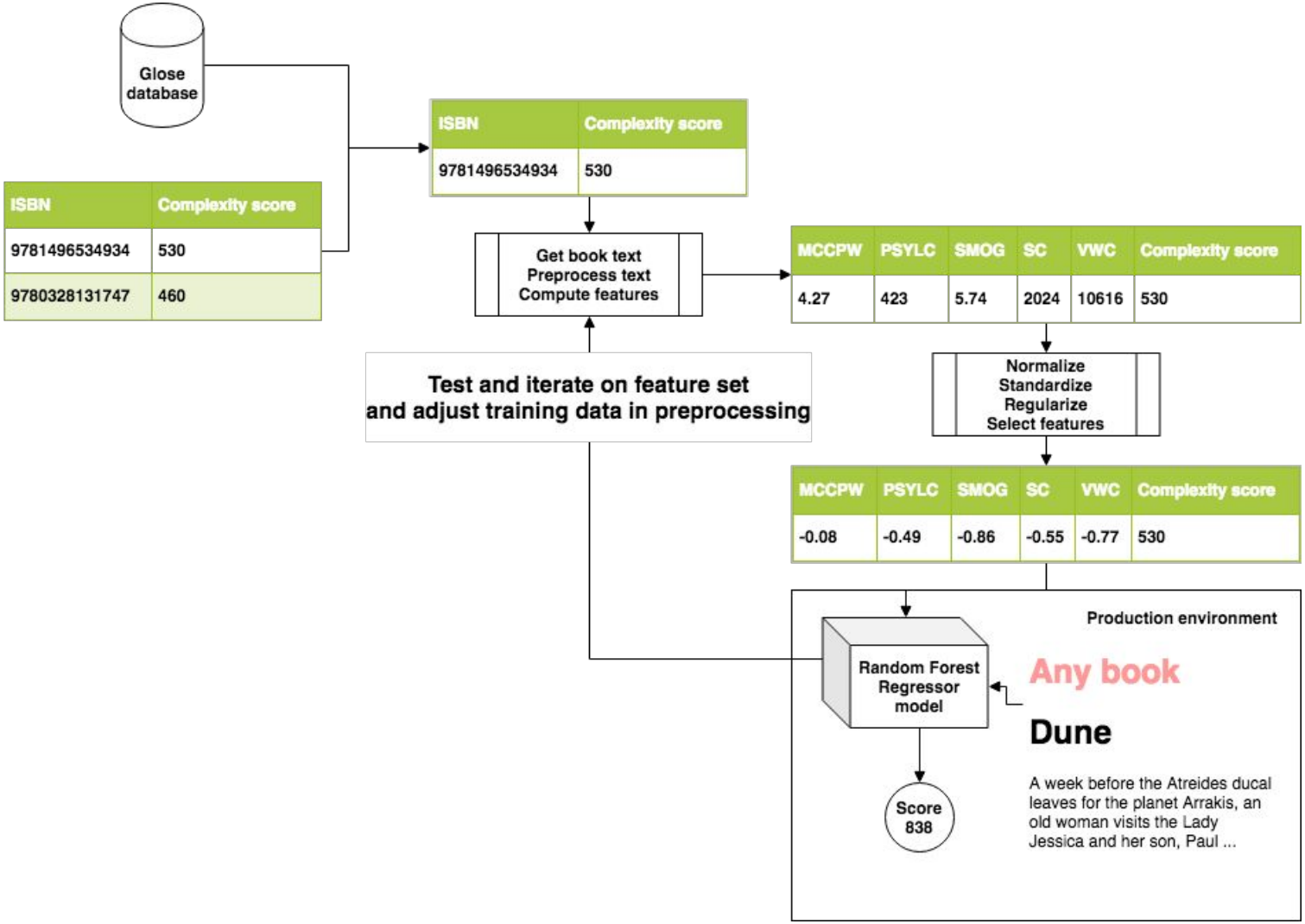
Convey information to most readers

e.g. Drugs leaflets, news information, administrative and legal texts

Engage readers through content that is **gradually more difficult**

e.g. First and second language learners, ideally to any domain learning condition

Glose


Glose database
ISBN
Complexity score
9781496534934
530
9780328131747
460
ISBN
Complexity score
9781496534934
530
Get book text
Preprocess text
Compute features
MCCPW
PSYLC
SMOG
SC
VWC
Complexity score
4.27
423
5.74
2024
10616
530
Test and iterate on feature set
and adjust training data in preprocessing
Normalize
Standardize
Regularize
Select features
MCCPW
PSYLC
SMOG
SC
VWC
Complexity score
-0.08
-0.49
-0.86
-0.55
-0.77
530
Production environment
Random Forest Regressor model
Any book
Dune
A week before the Atreides ducal leaves for the planet Arrakis, an old woman visits the Lady Jessica and her son, Paul ...
Score
838

Classroom levels scale

Glose

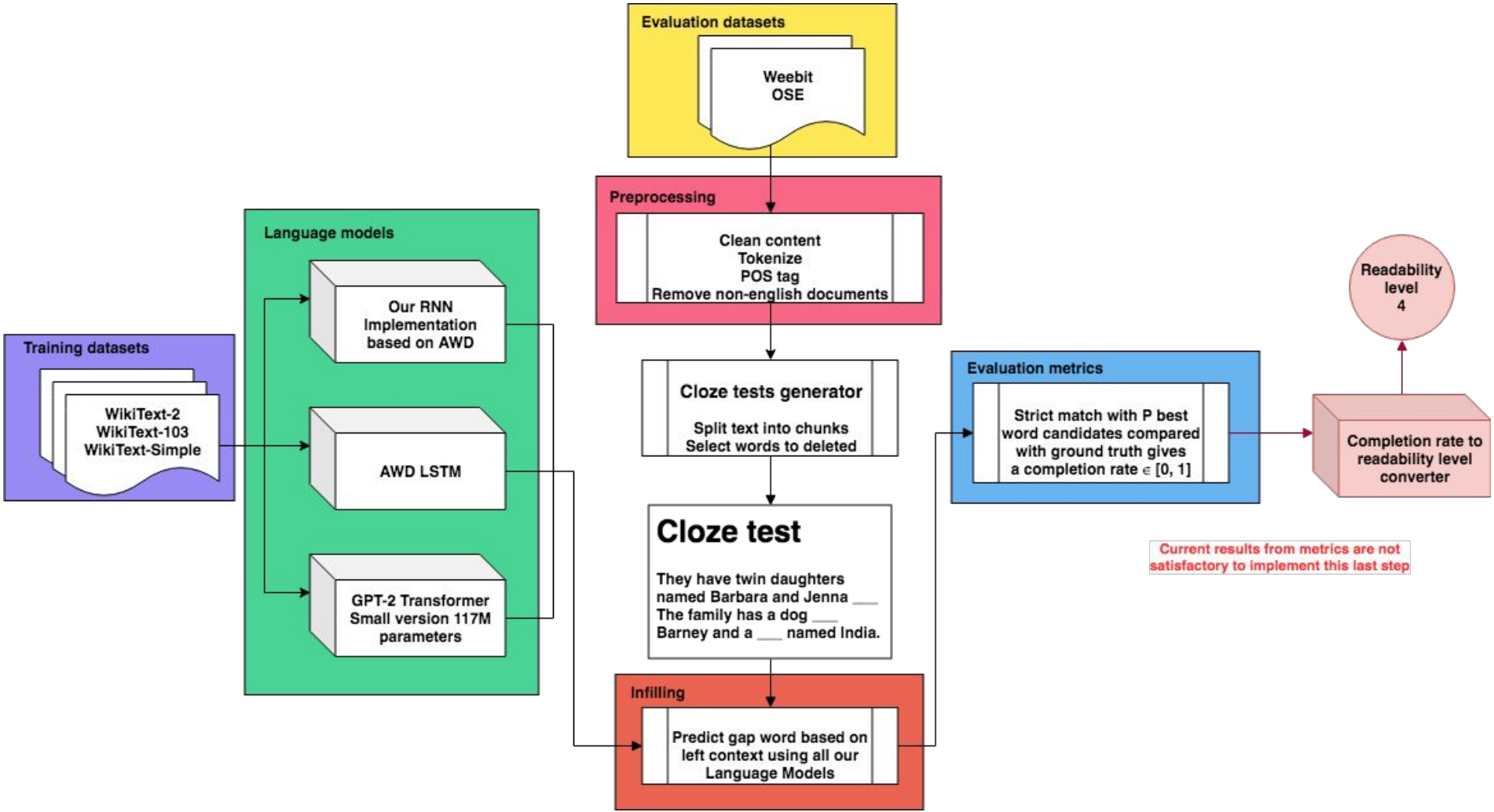
Novel approach: unsupervised task

Hypothesis
The simpler a text is, the better it should be understood by a machine

Idea
We should expect a strong correlation between readability levels and performance of language models at infilling Cloze tests

Current state
Correlations observed with our systems are still really small compared to other readability metrics such as Flesch-Kincaid or the previous model

Glose


Evaluation datasets
Weebit
OSE
Preprocessing
Clean content
Tokenize
POS tag
Remove non-english documents
Language models
Our RNN Implementation based on AWD
AWD LSTM
GPT-2 Transformer Small version 117M parameters
Training datasets
WikiText-2
WikiText-103
WikiText-Simple
Cloze tests generator
Split text into chunks
Select words to deleted
Cloze test
They have twin daughters named Barbara and Jenna ___
The family has a dog ___ Barney and a ___ named India.
Infilling
Predict gap word based on left context using all our Language Models
Evaluation metrics
Strict match with P best word candidates compared with ground truth gives a completion rate ∈ [0, 1]
Completion rate to readability level converter
Readability level 4
Current results from metrics are not satisfactory to implement this last step

Merci pour votre attention

« Créer, c'est ainsi donner une forme à son destin. »

Albert Camus, *Le Mythe de Sisyphe*

Glose

www.glose.com

References

William H Dubay, 2007
Smart Language
In *Readers, Readability, and the Grading of Text. Costa Mesa: Impact Information.*

Thomas François and **Cédrick Fairon**, 2012
An AI readability formula for French as a foreign language
In *Proceedings of the 2012 Joint Conference on Empirical Methods in NAtural Language Processing and Computational Natural Language Learning, pages 466-477.*

Stephen Merity, **Nitish Shrirish Keskar** and **Richard Socher**, 2018
Regularizing and optimizing LSTM language models
In *Proceedings of the International Conference on Learning Representations, ICLR.*

References

Alec Radford, **Jeffrey Wu**, **Rewon Child**, **David Luan**, **Dario Amodei** and **Ilya Sutskever**, 2019
Language models are unsupervised multitask learners
Technical report, OpenAI.

Sarah E. Petersen and **Mari Ostendorf**, 2009
A machine learning approach to reading level assessment
Computer Speech & Language *23(1):89 - 106.*

Sowmya Vajjala and **Ivana Lucic**, 2018
Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification
In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Education Applications, pages 297-304, New Orleans, Louisiana.*