Strategies for Arabic Readability Modeling: Comprehensive Approach

Supervised by Dr. Goudjil Mohamed.
Students:Mechache ElhadjeAhmed and Bellouti Redhouane.

Abstract

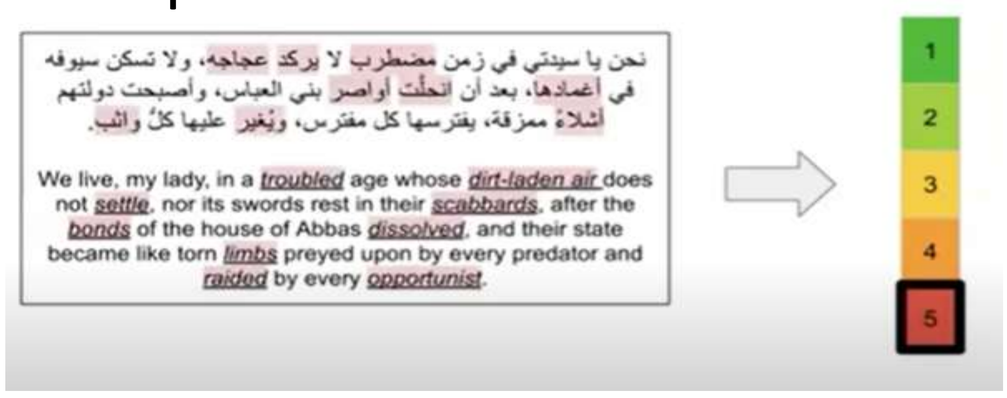
This poster explores the challenges and strategies for Arabic readability modeling, focusing on both word-level and fragment-level readability assessment. We utilize a diverse range of approaches, from rule-based methods to pre-trained language models, and evaluate their performance on the SAMER Arabic Text Simplification Corpus. Our results demonstrate that combining different modeling techniques yields the best results, achieving high accuracy in readability assessment. This work contributes to the development of NLP applications for education, content analysis, and accessibility in Arabic.

Introduction

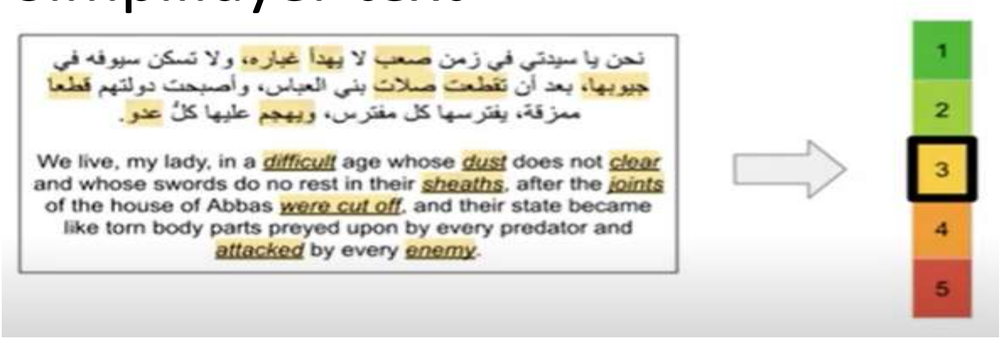
Arabic readability assessment is essential in NLP applications for education, accessibility, and content simplification. The Arabic language presents unique challenges due to its morphological richness, orthographic ambiguity, and dialectal diversity. This research addresses these complexities through a comprehensive modeling approach using rule-based and deep learning techniques, supported by the SAMER Corpus.

Example

Complex text



Simplifayer text



Arabic Linguistics Facts

Arabic inflects for gender, number, person, case, state, aspect, mood, and voice, resulting in 22,400 possible POS tags (compared to 48 POS tags in English). Each word can have three different core meanings (i.e., lemmas). Orthographic ambiguity is present due to optional diacritics used to specify short vowels and consonantal doubling, leading to 7 diacritizations per word.

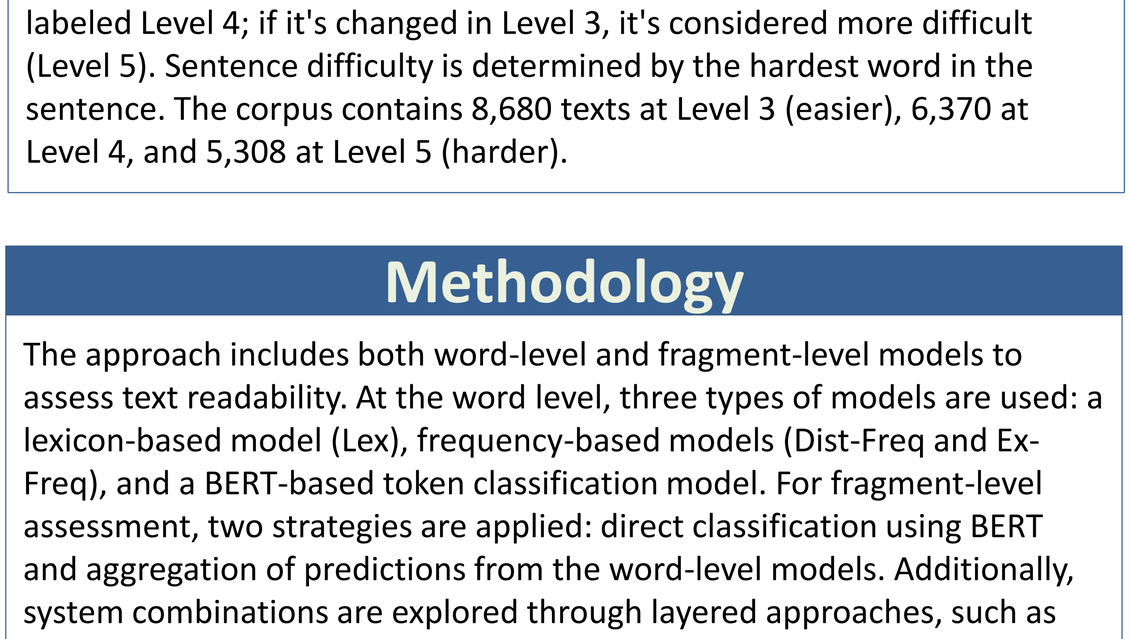
Exemple

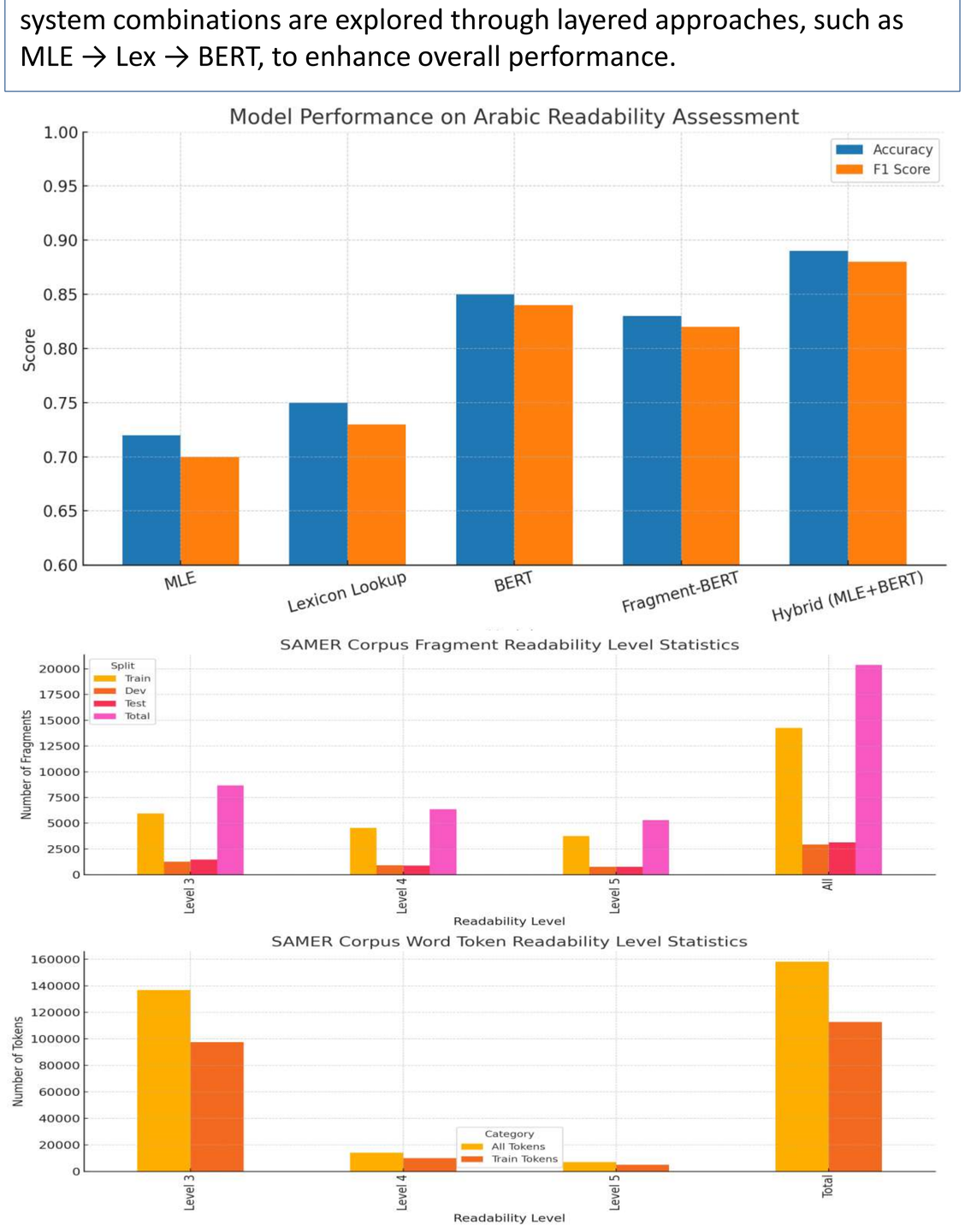
Data: The SAMER Corpus

The SAMER Corpus is a collection of Arabic texts designed for readability research. It includes original texts from 15 Arabic novels, each rewritten at two simpler levels: Level 4 and Level 3. Words are labeled based on how they change across levels—if a word remains unchanged in Level 4, it's labeled Level 4; if it's changed in Level 3, it's considered more difficult (Level 5). Sentence difficulty is determined by the hardest word in the sentence. The corpus contains 8,680 texts at Level 3 (easier), 6,370 at Level 4, and 5,308 at Level 5 (harder).

Methodology

The approach includes both word-level and fragment-level models to assess text readability. At the word level, three types of models are used: a lexicon-based model (Lex), frequency-based models (Dist-Freq and Ex-Freq), and a BERT-based token classification model. For fragment-level assessment, two strategies are applied: direct classification using BERT and aggregation of predictions from the word-level models. Additionally, system combinations are explored through layered approaches, such as MLE → Lex → BERT, to enhance overall performance.





Discussion

Combining rule-based and deep learning approaches proved effective for Arabic readability modeling. Challenges remain in handling dialectal variation and orthographic ambiguity. The availability of the SAMER Corpus and the use of pretrained models like AraBERT and CAMeLBERT opened new avenues for better contextual understanding and generalization.

Conclusions

This research contributes to Arabic NLP by offering a robust framework for readability assessment. The release of the SAMER Corpus and model resources enables further research. Future directions include dialectal Arabic modeling, online readability tools, and integration with educational platforms.

Contact

Dr. Goudjil mohamed : mohamed.goudjil@gmail.com
mechache elhadje : elhadjemechache@gmail.com
bellouti Redouane :redhouane761@gmail.com

References

- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic
- Al-Khalifa, H. S., & Al-Ajlan, A. A. (2010). Automatic readability measurements of the Arabic text: An exploratory study.
- Alhafni, B., Hazim, R., Liberato, J. P., Al Khalil, M., & Habash, N. (2024). The SAMER Arabic text simplification corpus
- Habash, N. (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies