

A system to simplify Arabic texts by reducing the lexical and syntactic complexity using a hybrid method combining Machine Learning and Rule-Based techniques

1 Text Simplification

Definition

Simplifying a text is the process of reducing its linguistic complexity, while maintaining its meaning and original information .

A man carrying a large number of books entered the room

دخّل رجل يحمل عدد كبير من الكتب إلى الغرفة.

Complex

دخّل رجل الي الغرفة مع الرجل كتب كثير.

Simple

A man walked into the room. The man had many books.

One Complex sentence transformed to two simple ones

Why?

2 Importance

Usage in designing and simplifying the language curriculum for both second language and first language learners

Make text easy-to-read for first language users with cognitive impairments and low literacy language level

A fundamental pre-process in NLP applications such as text retrieval, extraction, summarization, translation

3 Challenges

Highly morphologically rich language

Flexible word order

Multifunctionality of Arabic nouns

Lack of vocalisation diacritics

Lack of Arabic resources:
Datasets Corpora Arabic NLP tools



Stage I

4 Assessing Readability

To measure and annotate the difficulty of the Arabic text.

Adopting CEFR [A1,A2,B3,B4,C1,C2]

Based on average of complex lexical items and complex sentence structure.

6 Lexical Simplification

Not all the words in the text needs to be simplified

Identify the Complex Word

Generate series of substitutions

Substitution Ranking based on context

Select the new simple Substitute

Stage II

Complex Grammatical structure

C2 A1 A2 A1 A2 B2 A1 A2 B2 C2 A1 B2 A2
لذا ينبغي على البرلمان أن يبعث برسالة لأن هذه هي رغبة الأغلبية الساحقة

The Parliament should send a message, because that is the majority of people want.

Assign Level for each word

Identify complex grammatical structure

Assign Readability Level for Sentence

7 Syntactic Simplification

Syntactic Dependency Parsing

Identify complex structure

Transformation rules

Generation/ Regeneration

Stage III

8 New Arabic Frequency List

New classified Arabic frequency list consisting of 8834 unique Lemmas from Buckwalter, Al-kitaab and KELLY's

