On Arabic Stop-words: A Comprehensive List and a Dedicated Morphological Analyzer

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi
ALELM Team
Stop-words

Definition

- Common words that frequently appear
- Isolated
  - no information
- In a sentence
  - grammatical and syntactical function
Stop-words exploitation & removal

- Text categorization
- Information extraction
- Text summarization
- Search engine indexing
- ...

**Removal**

- corpus size --
- Saves space in indexes
- Increases accuracy without damaging retrieval efficiency

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Stop-words identification

Static stop-words list:
- Corpus-dependent
- Rely on the frequency feature
- Don’t consider all clitized forms

Morphological analysis:
- The grammatical category default
- Don’t consider all clitized forms
  - detect هو (POS=particle)
  - ignore الذي (POS=noun)

To remove these words we must identify them

Contains هو excludes clitized forms ... وهو، فهو،...
Objectives

01
A comprehensive Arabic rule-based stop-words list

02
A contextual stop-words analyzer
State of the art

Stop-word lists

• Shereen Khoja (168)
• Ibrahim Abu El-Khair (1529)
• Walaa Medhat et al. (1061)
• Alajmi et al. (200)
• Stop words project (162)
• Ranks NL (102)

Morphological analysis

• BAMA
• ALKHALIL
• MADAMIRA
• CALIMA-star

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Proposed approach

01
Simple Stop-words list
A
- إلّي
- 84 clitized forms
- أَلِيَّكَ، فَإِلَيْكَ، لَإِلَيْكَ، إِلَيْكُمْ، إِلَيْنِيّ، إِلَيْهِ، إِلَيْهَا، إِلَيْهَا...

02
Context-free analysis
A
- جاء الولد من المدرسة
- جاء الولد [من، من] المدرسة

Contextual analysis
B
- جاء الولد من المدرسة
- جاء الولد [من، من] المدرسة

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Proposed approach

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
1.A - Simple Stop-words list design

Simple Arabic stop-words Lexicon
معجم الكلمات المستبِطة العربية البسيطة

Native particles
حروف خالصة
- Conjonction حروف العطف
- Conditional حروف الشرط
- ... حروف الاستفهام
- Interrogative particles حروف الاستفهام

Special nouns
حروف بصفات الأسماء
- Relative pronoun الأسماء الموصلة
- Personal pronoun ضمائر الرفع المنفصلة
- ... أسماء الاستفهام
- Interrogative pronouns أسماء الاستفهام

Special verbs
حروف بصفات الأفعال
- Certainty verbs أفعال اليقين
- Transposition verbs أفعال التحويل
- ... أفعال الشروع
- Beginning verbs أفعال الشروع

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Proposed approach

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
1.B- Complex Stop-words list design

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
1.B- Complex Stop-words list design

**ProClitic**
- Conjunction
- Prepositions
- Accusative Particles
- Future particle
- Definition particle
- Arabic rules

**EnClitic**
- Connected pronouns
- Prevention
- Ascribing
- Arabic rules

**Word type**
- Particle
- Noun
- Verb
- Arabic rules

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Proposed approach

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
1- Arabic Stop-words List (ASL)

diacritized stop-words list named ASL (Arabic Stop-words List) :

<table>
<thead>
<tr>
<th>Enclitic</th>
<th>Simple stop-word</th>
<th>Proclitic</th>
<th>Complex stop-word</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>maybe</td>
<td>And</td>
<td>And maybe</td>
<td>NP</td>
</tr>
<tr>
<td>-</td>
<td>that</td>
<td>like</td>
<td>like that</td>
<td>SN</td>
</tr>
<tr>
<td>-</td>
<td>you are</td>
<td>will</td>
<td>we will be</td>
<td>SV</td>
</tr>
<tr>
<td>you عَلَيْكَ</td>
<td>no more than</td>
<td>-</td>
<td>no more than you</td>
<td>SN</td>
</tr>
<tr>
<td>you عَلَيْكَ</td>
<td>maybe</td>
<td>-</td>
<td>Maybe you</td>
<td>SV</td>
</tr>
<tr>
<td>they هُمْ</td>
<td>to</td>
<td>And is</td>
<td>And is to they</td>
<td>NP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Native Particles</th>
<th>Special nouns</th>
<th>Special verbs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple list</td>
<td>84</td>
<td>263</td>
<td>1,784</td>
<td>2,131</td>
</tr>
<tr>
<td>Complex list</td>
<td>1,590</td>
<td>9,627</td>
<td>27,936</td>
<td>39,153</td>
</tr>
</tbody>
</table>

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer

Khoja (168)
Abu El-Khair(1529)
Alajmi (200)
Proposed approach

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
2- Stop-words analyzer design

**Context-free analysis**

- Clitics Lexicon
- ASL

**Pre-processing**

- Potential clitics

**Potential stop-words**

**Contextual analysis**

- HMM
- Smoothing
- Viterbi

**Model**

**Analyzed text**

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
2- Stop-words analyzer design

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Stop-words list evaluation

Tagged evaluation corpus

Accuracy

Compute

Exist

78,477 tokens
15,684 unique

<table>
<thead>
<tr>
<th></th>
<th>Khoja</th>
<th>Abu El-Khair</th>
<th>List 1</th>
<th>Ranks NL</th>
<th>ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>68</td>
<td>144</td>
<td>552</td>
<td>98</td>
<td>1,247</td>
</tr>
</tbody>
</table>
Context-free evaluation

Corpus (manually annotated)
1,628 words
419 stop-words

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAMA</td>
<td>91.09</td>
</tr>
<tr>
<td>Alkhalil 1</td>
<td>88.70</td>
</tr>
<tr>
<td>Alkhalil 2</td>
<td>93.98</td>
</tr>
<tr>
<td>CALIMA-star</td>
<td>93.12</td>
</tr>
<tr>
<td>ASL analyzer</td>
<td>99.94</td>
</tr>
</tbody>
</table>

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Contextual evaluation

Corpus
(manually annotated )
1,628 words
419 stop-words

Accuracy
82.35%

<table>
<thead>
<tr>
<th>MADAMIRA</th>
<th>ASL analyzer</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.17</td>
<td>97.85</td>
</tr>
</tbody>
</table>

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Conclusion

Identification
- Static list
- Morphological analysis

Default
- grammatical category
- clitized forms

Designed
- comprehensive Arabic rule-based stop-words list
- contextual stop-words analyzer

Evaluation
- 1,247 Vs 552 (2nd list)
- context-free 99.94%
- contextual 97.85%

On arabic stop-words: a comprehensive list and a dedicated morphological analyzer
Thank you

On Arabic Stop-words: A Comprehensive List and a Dedicated Morphological Analyzer

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi
ALELM Team

ICALP 2019
16-17 Octobre- Nancy