Arabic Stop Words: Towards a Generalisation and Standardisation

Karim Bouzoubaa[1], Hicham Baidouri[1], Taoufik Loukili[1], Taoufik El Yazidi[2]

Emails: karim.bouzoubaa@emi.ac.ma, baidourihicham@gmail.com, tloukili@gmail.com, ya_touf_12@yahoo.fr

[1] Ecole Mohammadia d'Ingénieurs (EMI), Med V^th University - Agdal, Avenue Ibn Sina B.P. 765 Rabat Maroc
[2] Institut d'Etudes et de Recherche sur l'Arabisation (IERA), Med V^th University - Souissi, B.P. 765 Rabat Maroc

1. Introduction

Language resources are an essential component in all natural language processing (NLP) applications. Language resources are divided into written and oral resources. In the current work, we are interested in written resources. There exist different types of written resources:

- Character : Graphic symbol used as a unit in writing
- Lexicon: the set of words of a given language
- Dictionary: List of terms usually arranged in alphabetical order, providing definitions, explanatory information or descriptive data for each item
- Glossary: Specialized list of terms relating to a particular field of study or interest, which may contain explanatory or descriptive information on the items listed. Example: glossary of terms employed in the standardization of geographical names.
- Corpus: set of documents grouped for a certain goal
- etc.

Furthermore, the structural diversity of linguistic resources available has made their exchange difficult and complex. Clearly, a unified and standardized representation is required prior to any exploitation of these resources outside their own context of design. Some efforts are being developed. For example, two main proposals for standardization of electronic dictionaries have emerged, namely the TEI (Text Encoding Initiative) (Sperberg-McQueen, Burnard, 2005) and LMF[1] (Francopoulo, George, 2008).

However, the work and achievements that are interested in the automatic processing of the Arabic language still need more organization and structure given the specificity of the Arabic language (short vowels, absence of capital letters, complex morphology, etc.). The diversity of representations makes difficult the dissemination and exchange between different sub communities, making it impossible to have a powerful and relevant consultation tools that could be developed around these resources.

With the advent of information retrieval oriented applications such as search engines and question/answering systems, appeared a new kind of resources that are named stop words. A stop word is a very common word that appears very frequently in a text and that carries little meaning, it serves only a syntactic function but does not indicate subject matter. In Arabic, the

[1] Lexical Markup Framework

following are examples of stop words: (… ‰)." ~Stop words are generally grouped as a list in a single file or database.

Stop words are filtered out prior to, or after, processing of texts. Hans Peter Luhn[2], one of the pioneers in information retrieval, pointed that it is better to ignore such words in the indexing phase in order to get better results during a search. The stop words are also useful during a morphological analysis of a text since there is no need to analyse them.

Stoplists can be divided into two categories; domain independent stoplists and domain dependent stoplists. They can be created using syntactic classes or using corpus statistics, which is a more domain dependent approach, used for well-defined fields. They can also be created using a combination of the syntactic classes and corpus statistics to obtain the benefits of both approaches. Whatever the language, there is no definite list of Arabic stop words (stoplist) which all NLP tools incorporate.

Here are examples of some already developed stoplists:

- Khoja (Khoja, Garside, 1999) developed a relatively short stoplist (168 words) that has been used for an Arabic stemmer.
- Chen and Gey (Chen, Gey, 2002) used a list they created by translating an English list and augmenting it with high frequency words from the corpus creating a rather large list 1,131 words.
- Abu El-Khair (Abu El-Khair 2006) created three stoplists. The first one is a general stoplist based on the Arabic language structure. All possible words or articles that may be considered a stop word were collated from the different syntactic classes in Arabic (Adverbs, Conditional Pronouns, Prepositions, etc) in a systematic way to ensure the completeness of the list. The resulting list consisted of 1,377 words. The second list is a corpus-based list which was containing 359 words. These words were occurring in the corpus more than 25,000 times. The third stoplist is a merge of both the general and corpus-based stoplists.

The objective of the current work is not to create a new stop list. Rather, we would like to unify all previous efforts in a more general framework:

1. Propose a new categorisation of Arabic stoplists
2. Propose these lists as standards for the Arabic NLP community
3. Propose a standard structure for the representation of these resources

2. A domain-independent stoplist

This stoplist should be domain independent. This means that it should include stop words that are related only to the structure and syntax of the Arabic language. This is partly what has been done by Abu El-Khair (Abu El-Khair 2006). However, due to the specific characteristics of the Arabic language, Abu El-Khair put in his list not only basic stop words (such as) but also stop words generated from the concatenation of basic stop words and Arabic prefixes and suffixes (such as). This way, the developer does not exactly know if the stop word in hand is a basic or a composite one. In order to give more flexibility, we propose to develop a list containing only basic words and specify for every group of

[2] http://en.wikipedia.org/wiki/Hans_Peter_Luhn

those words (for example ضمائر) the list of affixes that this group can concatenate to.

On the practical side, if one needs to use only basic stop words, there is no necessity to consider the list of affixes. On the contrary, if one needs to use both basic and composite stop words, it is better to create a new stop list (from the one we provide) containing basic words (by copying them) and composite ones (by writing a short program that concatenate basic words with the list of proposed affixes).

3. A type-oriented stoplist

In general, there is no fixed domain-dependant stoplist. This kind of list depends mainly on the corpus in hand and from where the stop words have been extracted. Any list of such kind is completely useless once we change the corpus. Therefore, we propose to develop not a corpus-dependant list or a domain-dependant list but rather develop a type-oriented list. For example, it is more intuitive to group most frequent terms used as currencies in a list and to consider them as a stoplist. These currencies stoplist can be considered as an actual stoplist in any finance or economic oriented corpus. It is rather not logical to consider it as a stoplist in corpora talking about biology or linguistics. It is then up to the programmer to decide for every type-oriented stoplist if yes or no the list is to be considered for the corpus in hand.

Examples of types that can be considered as stoplists are: currencies, scales, or numbers. Of course, these stoplists are not static. The structure and the representation of stoplists we suggest later allows one to add as many list as s(he) needs.

4. Representation of stoplists

In order to guarantee an exchange between interested people and in order to offer a standard structure and format that could be used by everyone; we propose an XML-based file[3] containing both kinds of stoplists we explained in previous sections. For each kind (domain-independent and type-oriented), we mention the stop word with its class (or category) and the list of affixes (prefixes and suffixes) this stop word could be concatenated to.

For example, in the domain-independent stoplist, "أب" is a stop word that belongs to the class "أسماء خمسة". Its list of affixes can be found in the list named "list1". This way, the word "كأبٌ" is also a stop word.

```
<stopwords>

<domain-independent stopwords>
<word>
       <desc> أب              </desc>
       <class> أسماء خمسة     </class>
       <affixes> list1        </affixes>

</word>
<word>
       <desc> أو              </desc>
       <class> حرف عطف منفصل </class>
       <affixes> list2        </affixes>

</word>
…
</domain-independent stopwords>

<type-oriented stopwords>
<word>
       <desc> دولار           </desc>
       <class> عملة           </class>
       <affixes> list3        </affixes>

</word>
…
</type-oriented stopwords>

</stopwords>

<list-of-affixes>
<list>
       <desc> list1   </desc>
       <prefixes> ك … </prefixes>
       <suffixes> …   </suffixes>
</list>
<list>
       <desc> list2   </desc>
```

[3] The file can be downloaded from www.emi.ac.ma/bouzoubaa/download.html

```
        <prefixes>من، في… </prefixes>
        <suffixes> …    </suffixes>
</list>
…
</list-of-affixes>
```

5. Experiments

In order to illustrate the usefulness of our approach and show the link between some type-oriented stoplists and some domains (for example the link between currencies stoplist and any finance-oriented corpus), we are currently conducting experiments[4] by the use of Arabic Wikipedia. Our intention is to calculate the frequency of every stopword in the domain-independ stoplist in the whole Arabic Wikipedia (to confirm they are really Arabic stop words regardless the corpus in hand) and calculate the frequency of every stopword in each type-oriented stoplist in every domain of Arabic Wikipedia (History, physics, chemistry, etc.).

6. Conclusion

Due to the large importance of linguistic resources in the NLP domain, it is necessary to provide as much as possible for the Arabic NLP community free, flexible and standard resources. In this article, we are proposing a framework for Arabic stop words. More particularly, we propose a generalisation of the content of that resource and propose a standard structure for its representation.

References

I. Abu El-Khair, (2006) Effect of stop words elimination for Arabic information retrieval: a comparative study, Dept of library and information science, faculty of arts, Minia University-Egypt, iabuelkhair@gmail.com, International Journal of Computing & Information Sciences, vol. 4 No. 3 December, On-Line.

A. Chen, F. Gey, (2002) Translation Term Weighting and Combining Translation Resources in Cross-language Retrieval, *Tenth Text REtrieval Conference (TREC 2001),* http://trec.nist.gov/pubs/trec10/papers/berkeley_trec10.pdf.

G. Francopoulo, M. George (2008), ISO/TC 37/SC 4N453 (N330 Rev.16), Language resource management — Lexical markup framework (LMF).

S. Khoja, R. Garside, (1999) Stemming Arabic text, Computing Department, Lancaster University, Lancaster, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

C. M. Sperberg-McQueen, L. Burnard, (2005), TEI P5 Guidelines for Electronic Text Encoding and Interchange (revised). The Association for Computers and the Humanities, http://www.tei-c.org/Guidelines/P5/

[4] Results will be available as soon as finished from www.emi.ac.ma/bouzoubaa/download.html