Broken Plural Detection for Arabic Information Retrieval

Abduelbaset Goweder
University of Essex
Dept. of Computer Science
Wivenhoe Park,
Colchester CO4 3SQ, UK
agowed@essex.ac.uk

Massimo Poesio
University of Essex
Dept. of Computer Science
Wivenhoe Park,
Colchester CO4 3SQ, UK
poesio@essex.ac.uk

Anne De Roeck
The Open University
Dept. of Computing
Walton Hall, Milton Keynes
Buckinghamshire, MK7 6AA, UK
A.DeRoeck@open.ac.uk

ABSTRACT
Due to the high number of inflectional variations of Arabic words, empirical results suggest that stemming is essential for Arabic information retrieval. However, current light stemming algorithms do not extract the correct stem of irregular (so-called broken) plurals, which constitute ~10% of Arabic texts and ~41% of plurals. Although light stemming in particular has led to improvements in information retrieval [5, 6], the effects of broken plurals on the performance of information retrieval systems has not been examined.

We propose a light stemmer that incorporates a broken plural recognition component, and evaluate it within the context of information retrieval. Our results show that identifying broken plurals and reducing them to their correct stems does result in a significant improvement in the performance of information retrieval systems.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and indexing – *Dictionaries, Indexing methods, Linguistic processing.*

General Terms
Algorithms, Performance, Experimentation, Languages.

Keywords
Light stemming, root stemming, broken plurals.

1. INTRODUCTION
In information retrieval, documents are retrieved by matching the terms of a query with the index terms. The match between query terms and document terms can be problematic due to morphological variation. One way to tackle this problem is to make use of stemming which conflates morphological variants of a word into a common form. It increases recall and reduces the number of index terms, saving storage space and processing time. It is considered particularly effective for languages with more complex morphology [8] such as Arabic, Finnish, and Turkish. Research in the area of Arabic information retrieval has shown that stemming does result in improved performance of Arabic information retrieval systems [1, 3, 5, 6].

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. STEMMING FOR ARABIC
There are some orthographic aspects that make Arabic difficult to stem. For instance, short vowels are not written (missing information), and there is a risk of mistaking root consonants for affixes and removing them (information loss). Particularly challenging is the high incidence of morphological forms where roots have been changed. The majority of these cases are broken plurals.

Two types of stemming approaches have been developed for Arabic: ***light stemming*** and ***heavy stemming***. Light stemming (or prefix-suffix removal) is the process of stripping off prefixes and suffixes to produce the stem of a word, so it is also called surface affix removal. For instance, applying light stemming to the Arabic word ***w****aqlam****hm*** (واقلامهم , "**and their** pens") gives the stem *aqlam* (اقلام, "pens"): in this example the prefix *w* (و ,"and") and the suffix *hm* (هم , "their") were removed. Heavy stemming (or root stemming), is the process of stripping off prefixes, suffixes and infixes to produce the root of a word, and is also called deep affix removal. For instance, applying heavy stemming to the word ***w****a****q****l****a****m****hm*** (واقلامهم , "**and their** pens") produces the trilateral root *qlm* (قلم , "pen"). In this case, the prefix *w* (و ,"and") and the suffix *hm* (هم , "their") were removed, as were two infixes *Alef* (ا), in the second and fifth positions of the original word (underlined in the English transliteration).

The comparative benefits of each type of stemming for information retrieval are unclear. Heavy stemming is preferred in linguistic processing-based applications, while light stemming is considered more useful in information retrieval-based systems [2].

Several stemming algorithms have been designed, implemented, and tested for Arabic. None of these algorithms has been universally accepted. Current light-stemming algorithms can correctly handle regular forms of a word, but they fail to handle irregular forms (e.g., broken plurals) [6].

3. A LIGHT STEMMER WITH BROKEN PLURAL IDENTIFICATION
As a first step in our study of the effect of broken plural recognition on information retrieval, we developed a basic light stemmer for Arabic that only removes prefixes and/or suffixes attached to a word and ignores any infixes encountered. The basic light stemmer in effect reduces a word to its stem for an easier identification of broken plural stems. For this purpose, we amended an existing root stemmer, developed by Khoja [7].
We then developed different methods for identifying broken plurals, including: simple matching, restricted matching, and the

dictionary approach. In the simple matching method, a word is light-stemmed, and the resulting stem is compared against a set of broken plural patterns found in traditional grammars of Arabic. An evaluation of the performance of the simple matching method showed that it achieves low precision (~13% on a test set of about 187,000 words). We also considered two techniques for improving performance: affix information and proper name detection. Both resulted in a slight improvement.

The second approach to identifying broken plurals we considered was the restricted matching method, in which the broken plural patterns are used to detect broken plurals according to sets of rules that govern their applicability. We first attempted to find restrictions by hand; the results of the performance of this method showed a noticeable improvement, particularly in precision (~53%). Next, we attempted to find restrictions automatically, using supervised machine learning techniques (decision tree classifiers). The results of the overall performance of the restricted matching approach using decision tree classifiers showed an increase in precision, to reaching about 75%, but with a slight decrease in recall (~95%).

Having developed the methods above, and used them to analyse broken plural instances in a large corpus [4], it became very easy to try a third approach for identifying broken plurals: using a dictionary which lists broken plural stems. The dictionary was first constructed by automatically extracting all instances of broken plural stems that match broken plural patterns. Secondly, we manually went through the output to identify and list the actual broken plural stems. Finally, the list was revised in collaboration with a linguist, who is an Arabic native speaker. An evaluation of the performance of the dictionary approach found that a significant improvement in precision (~92%) over the other two approaches. We developed a new light stemming algorithm using this broken plural detection method, that reduces both regular and broken plurals to their singular form.

4. EVALUATION ON INFORMATION RETRIEVAL

The performance of the new light stemming algorithm was evaluated in an information retrieval task. The new algorithm was incorporated in a new indexing method referred to as "stem+BP". This new indexing method was compared with the three standard indexing methods (full word, root, and 'basic' light stem). We used the Greenstone digital library, developed at the University of Waikato in New Zealand, as an information retrieval system for our experiment. A collection of 385 documents, extracted from a large corpus [4], and a set of 50 queries with their relevance judgments, created to search for particular information, was used to evaluate the four indexing methods.

The results of Figure 1 clearly suggest that the proposed "stem+BP" indexing method outperforms the three standard indexing/stemming methods. The significance of the results shown in Figure 1 was assessed using two non-parametric statistical tests: the Sign test and the Wilcoxon signed-rank test. They clearly indicate that the "stem+BP" indexing method significantly outperforms the three standard indexing methods (p (1-tailed) < .01). This suggests that stemming has a substantial effect on information retrieval for highly inflected languages such as Arabic, confirming the results obtained by [1, 3, 5, 6].

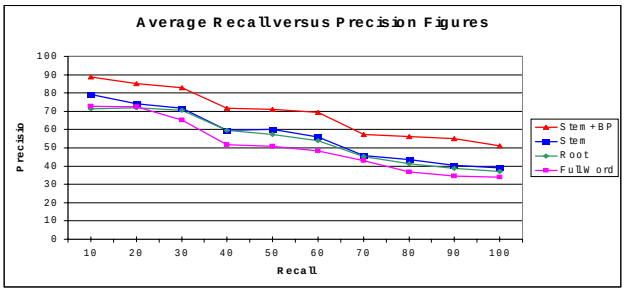

Figure 1: The Average Recall vs. Precision Figures of the Four Indexing Methods for 50 Queries.

5. CONCLUSION

Our work provides evidence that identifying broken plurals results in an improved performance for information retrieval systems. We found that any form of stemming improves retrieval for Arabic; and that light stemming with broken plural recognition outperforms standard light stemming, root stemming, and no form of stemming.

6. ACKNOWLEDGMENTS

We would like to thank Shreen Khoja [7] for providing her root stemmer, and permission to modify it to obtain a light stemmer. Our thanks also go to the Libyan Secretariat of Education for supporting this work.

7. REFERENCES

[1] Abu-Salem, Hani; Al-Omari, Mahmoud; and Evens, Martha W. (1999). "Stemming Methodologies over Individual Query Words for an Arabic Information Retrieval System." JASIST, 50(6):524-529.

[2] Al-Kharashi, I. and Al sughaiyer, I. (2002). "Data Set for Designing and Testing an Arabic Stemmer." In Workshops of LERC 2002.

[3] Al-Kharashi, I. and Evens, Martha W. (1994). "Comparing words, stems and roots as index terms in an Arabic Information retrieval system." JASIST, 45(8):548-560.

[4] Goweder, Abduelbaset and De Roeck, Anne (2001). "Assessment of a Significant Arabic Corpus." ACL 2001. Arabic language Processing. pp. 73-79, 2001.

[5] Larkey, L. S. and Connell, M. E. (2001) "Arabic information retrieval at UMass in TREC-10." In TREC 2001. Gaithersburg: NIST, 2001.

[6] Larkey, L.; Ballesteros, L.; and Connell, M.E (2002). "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis." In SIGIR'02, August 11-15, 2002, Tampere, Finland, pp 275–282, 2002.

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

[8] Popovic, M. and Willett, P. (1992). "The effectiveness of stemming for natural language access to Slovene textual data." JASIST, 43: 384-390.