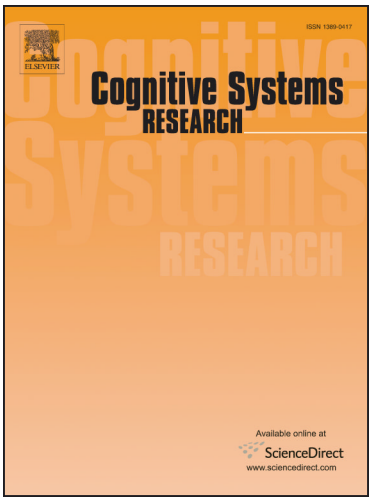

A Cognitive Inspired Unsupervised Language-Independent Text Stemmer for Information Retrieval

Fahd Saleh Alotaibi, Vishal Gupta

PII: S1389-0417(18)30060-3
DOI: https://doi.org/10.1016/j.cogsys.2018.07.003
Reference: COGSYS 654

To appear in: *Cognitive Systems Research*

Received Date: 16 February 2018
Revised Date: 30 May 2018
Accepted Date: 5 July 2018

Please cite this article as: Saleh Alotaibi, F., Gupta, V., A Cognitive Inspired Unsupervised Language-Independent Text Stemmer for Information Retrieval, *Cognitive Systems Research* (2018), doi: https://doi.org/10.1016/j.cogsys.2018.07.003

A Cognitive Inspired Unsupervised Language-Independent Text Stemmer for Information Retrieval

Fahd Saleh Alotaibi[1], Vishal Gupta[2]

[1]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, Email: fsalotaibi@kau.edu.sa

[2] Department of Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India, Email: vishal_gupta100@yahoo.co.in

Abstract: In Information Retrieval systems, stemming handles the words that can occur in different morphological forms, and hence matches the terms of the documents and the queries that are related in meanings. In this article, we have proposed a cognitive inspired language-independent stemming that learns group of morphologically related words from the ambient corpus without any linguistic knowledge or human intervention and it behaves in a way the human brain works. The main idea of our proposed algorithm is to determine only those variants of the words from the ambient corpus that match the original intent of the query terms. We conducted ad-hoc retrieval experiments in a number of languages of varying morphological complexity using standard TREC, FIRE, and CLEF document collection. The results indicate that stemming improves the retrieval accuracy and the effectiveness of stemming algorithm increases with the increase in the morphological complexity of algorithm. The results also indicates that the performance of our proposed algorithm is better than the stemmers based on linguistic knowledge and other state-of-the-art statistical stemmers in almost all the languages under study. In multi-lingual setup these results are quite encouraging.

Keywords: Morphology, Stemming, Stemmer, Language-Independent Stemming, Information Retrieval, Corpus-Based Stemming

1. INTRODUCTION

In Information Retrieval systems that manage large document collections, indexing involves generation of index words that best describes the documents and queries of the collection. The documents and queries in the natural languages contain a large number of variant words that generally refers to the same concept. The key idea is that the words which are quite similar in