Information Processing and Management 50 (2014) 315–341

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Crime profiling for the Arabic language using computational linguistic techniques

Meshrif Alruily [a,*], Aladdin Ayesh [b], Hussein Zedan [b]

[a] Information and Computer Science, Al Jouf University, P.O. BOX 2014, Al Jouf, Sakaka, Saudi Arabia
[b] Software Technology Research Laboratory, De Montfort University, Bede Island Building, The Gateway, Leicester LE1 9BH, UK

ARTICLE INFO

*Article history:*
Received 27 July 2011
Received in revised form 1 September 2013
Accepted 6 September 2013
Available online 31 October 2013

*Keywords:*
Arabic language
Crime domain
Pattern recognition
Clustering
Information extraction
Syntactic analysis

ABSTRACT

Arabic is a widely spoken language but few mining tools have been developed to process Arabic text. This paper examines the crime domain in the Arabic language (unstructured text) using text mining techniques. The development and application of a Crime Profiling System (CPS) is presented. The system is able to extract meaningful information, in this case the type of crime, location and nationality, from Arabic language crime news reports. The system has two unique attributes; firstly, information extraction that depends on local grammar, and secondly, dictionaries that can be automatically generated. It is shown that the CPS improves the quality of the data through reduction where only meaningful information is retained. Moreover, the Self Organising Map (SOM) approach is adopted in order to perform the clustering of the crime reports, based on crime type. This clustering technique is improved because only refined data containing meaningful keywords extracted through the information extraction process are inputted into it, i.e. the data are cleansed by removing noise. The proposed system is validated through experiments using a corpus collated from different sources; it was not used during system development. Precision, recall and F-measure are used to evaluate the performance of the proposed information extraction approach. Also, comparisons are conducted with other systems. In order to evaluate the clustering performance, three parameters are used: data size, loading time and quantization error.



## 1. Introduction

Nowadays, the volume of data in electronic form is increasing rapidly, whether it is structured or unstructured data. According to McKnight (2005), between 85% and 90% of data is held in unstructured form. Therefore, text mining is necessary to manage and extract useful information from unstructured sets of data, such as web pages and emails, using text mining techniques. Hence, text mining has become an important and active research field. It is well known that text mining techniques have been mostly developed for the English language because most electronic data are in English. However, the recent advances in localisation and expansion of non-English electronic text make it more imperative than ever to develop techniques to process other languages, such as Arabic.

Moreover, in the Arabic language, like other languages, there is a multitude of specialist texts in areas such as the biomedical sciences, finance, politics and crime, some of which have not been investigated by researchers. This study focuses on the crime domain in the Arabic language for which no information system can be found in the literature. Therefore, there is a need to investigate the crime context using a text mining approach.

* Corresponding author.
*E-mail addresses:* mfalruily@ju.edu.sa (M. Alruily), aayesh@dmu.ac.uk (A. Ayesh), zedan@dmu.ac.uk (H. Zedan).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.09.001

Furthermore, text mining research relies on the availability of a suitable corpus. However, there is no available crime corpus in the Arabic language, and it is one of the contributions of this study to create such a corpus. The content of this corpus has been collected from the crime sections of different Arabic newspapers published in a number of Arabic countries.

In our previous works, we developed a system to extract crime information from texts (Alruily, Ayesh, & Zedan, 2009). The approach is based on a dictionary that is created manually. Also, we were able to develop three dictionaries, for crime type, crime location and nationality, that were automatically built in order to be used in the information extraction (Alruily, Ayesh, & Zedan, 2010). Moreover, we developed the first system that was able to cluster Arabic crime news reports using the Self Organising Map (SOM) technique (Alruily, Ayesh, & Al-Marghilani, 2010). The current research represents a combination of our previous works (Alruily et al., 2010; Alruily et al., 2010). As a result, the system will be able to recognise phrases that contain information related to crime in a given document in order to extract the type of crime, crime location and nationality of persons involved in the event, and in order to generate a summary. Moreover, during the information extraction stage, dictionaries containing the crime type, crime location and nationality will be automatically created, which will also assist in the information extraction process. In this research, the extracted keywords (summary) will be utilised by a Self Organising Map (SOM) to perform the clustering and visualisation tasks. As well known, the extracted entities are keywords, which play an important role in many text mining tasks, such as clustering, summarisation and document retrieval, because they reveal the essential idea of the document or article (TeCho, Nattee, & Theeramunkong, 2008).

The rest of the paper is organised as follows. In Section 2, a background to the topic and a review of related work are presented. Section 3 provides an introduction to the Arabic language, focusing on the syntactic information that is related to this research. Section 4 presents the crime domain syntactic analysis. Section 5 provides an overview of the framework's architecture. Section 6 presents the experimental results. In Section 7, the performance evaluation results are given. Finally, the conclusion of this work is presented in Section 8.

## 2. Background

This section is on information extraction, and it includes descriptions of its early development and the approaches utilised. Also, related works are discussed. It is divided into three parts, as in the subsections below.

### 2.1. Information extraction approaches

Generally, the information extraction approaches can be divided into the following categories:

- Handcrafted rules, known as linguistic approaches.
  This approach is usually used for extracting information from specific domains. It is simple to build because its goal is only to fill out templates, i.e. it is not document understanding. Moreover, it can be trained on annotated or unannotated corpora. Although it can achieve reasonable levels of success, there are some disadvantages (Toral & Munoz, 2006; Riloff, 1993). For instance, it is time consuming because it is slow to build, and it is difficult to scale to new domains (Riloff, 1993). Toral and Munoz (2006) explained that this approach is applied using rules and gazetteers. The Gate system was developed at Sheffield University, and is a type of software that follows this approach. The task of this system is to extract named entities (Cowie & Lehnert, 1996).
- Machine learning approaches.
  Because of the difficulties that faced researchers when building hand crafted rules, a need for automatically learning extraction rules emerged. There are three types of machine learning approach: supervised, semi-supervised and unsupervised. According to Nadeau (2007), most developed systems have been designed based on handcrafted rule based systems or supervised learning based systems. In both approaches, a corpus must be studied and analysed by hand to gain sufficient pertinent knowledge to build the rules or to feed the machine learning algorithms. However, the supervised learning techniques, which include Hidden Markov Models (HMM), Maximum Entropy (ME), Support Vector Machines (SVM) and Conditional Random Fields (CRF) need a large annotated corpus for designing the systems. As a result, this disadvantage of supervised machine learning led to the emergence of semi-supervised and unsupervised machine learning (Nadeau, 2007). The semi-supervised (weakly supervised) is implemented with little supervision. The idea of this type of machine learning is to use a set of seeds to provide a system with a little external support to start learning how to extract. For example, finding the names of the diseases to extract can be done by providing the system with seeds, such as five disease names. First, the system seeks out the sentences that contain these seeds in order to understand the contexts in which they appear. Then, the system tries to find other disease names that exist in the same context (Sekine, 2004). Nadeau (2007) developed a semi-supervised Named Entity Recognition (NER) technique that learned to recognise 100 entity types with little supervision. With regards to unsupervised machine learning, clustering is considered the most typical approach where, for example, named entities can be gathered based on the similarity of context from clustered groups (Nadeau, 2007; Sekine, 2004).
  Moreover, there is an approach called hybrid, which is a combination of the handcrafted recognition and machine learning approaches.