Chapter 29
Combining Indexing Units for Arabic Information Retrieval

Souheila Ben Guirat
LISI: Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia & Jarir: Joint Group for Artificial Reasoning and Information Retrieval, Tunisia

Ibrahim Bounhas
LISI: Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia & Jarir: Joint Group for Artificial Reasoning and Information Retrieval, La Manouba University, Tunisia

Yahya Slimani
LISI: Laboratory of Computer Science for Industrial Systems, Carthage University, Tunisia & Jarir: Joint Group for Artificial Reasoning and Information Retrieval, La Manouba University, Tunisia

ABSTRACT

Using either stems or roots as index terms offered considerable performance to Arabic Information Retrieval (IR) systems compared to the use of surface words for indexing. Many comparative works tried to find out the best from these two indexing approaches but until then, no of the two methods widely overtook the other. Each of the two index types performed better under different test circumstances in terms of recall and precision. In this paper, the authors propose a hybrid approach combining the two indexing units in a way they take the advantages from both of them and try to overcome their shortcomings. Then, based on some combining techniques, the authors assign a weight for each indexing unit and try to find out the best weighting values.

1. INTRODUCTION

Choosing the indexing unit is yet a challenging problem in Arabic IR (Elayeb & Bounhas, 2015). When using surface words, precision reaches high levels, but we will report low recall rates because of the high derivation and agglutination of Arabic. Furthermore, this choice requires more resources in terms of storage space and processing time (Aljlayl & Frieder, 2002). Thus, other types of indexing units gave better performance to Arabic IR systems like stems and roots. To better understand this problem, we present an example in section 1. We summarize our contribution in section 2.

DOI: 10.4018/978-1-5225-5191-1.ch029

1.1. Motivation

Let consider the Arabic root "kassama-مسق" and some related words (cf. Table 1). In one side, root-based methods will relate many derived words to the same form and this will cause ambiguity and reduce precision. If we consider the example, "inkassama-مسقنا" (was divided) and "akssama-مسقا" (swear) will be represented by the same index i.e. "kassama-مسق" (divide).

Stem indexes reduce ambiguity (Ayed, 2014; Bounhas et al., 2015), but it will from the other side, reduce recall, since some studies (Al-Kabi et al., 2011) showed that in most cases, morphological variants of words have somehow similar semantic interpretations and are not completely dissimilar and different word forms may bear similar meaning. When we apply a light stemming to the same example, "alinkissamat-تامساقنالا" (the divisions) and "inkissam-مساقنا" (division) will have the same stem but their semantic relation with "kisma -قمسق" (division) will be ignored.

Indeed, light stemming methods offer less recall and more precision. However, lemmatization helps to achieve better recall rates, but reduces precision. Thus, combining these techniques seems promising as it realizes some compromise between precision and recall, in a way we ensure that "alinkissamat -تامساقنالا" (the divisions) is equal to "inkissam-مساقنا" (division), not so far from "inkassama-مسقنا" (was divided), a little bit different from "iktassama-مستقا" (share somebody in something), but not totally different from "kaassama-مساق" (share something with somebody) and that "akssama-مسقا" (swear) and "istakssama-مسقتسا"(conjure) are somehow related (cf. Figure 1).

1.2. Related Work

Many root extraction stemmers (A-Kabi et al., 2011; Al-Shawakfah et al., 2010) were proposed in literature like the famous stemmer of Khoja & Garside (1999), which extracts roots by removing affixes then checking the remaining letters against a root list to avoid invalid roots. Khoja algorithm was the base of many other later works. Compared to surface based retrieval (Aljlayl & Frieder, 2002), root based techniques performed better in terms of recall and resources economy.

Larkey (Larkey & Connell, 2001) and many other researchers (Aljlayl & Frieder, 2002; Chen & Gey, 2002) focused on light stemming, which aims to extract stems through affixes truncation. These methods allowed better precision and reduced storage size and processing time, compared to indexing by surface words (Aljlayl & Frieder, 2002).

A comparison between Khoja stemmer and Larkey light stemmer proved that these techniques still have some weaknesses (Handi et al., 2012). In (A-Kabi et al., 2011), the authors compared four root stemmers, e.g. Al-Mustafa (Al-Kabi & Al-Mustafa, 2006), Taghva et al. (2004), Al-Sarhan et al. (2003) and Rabab'ah et al. (2005).

Table 1. Some Arabic words related to the root "kassama-مسق"

مسقنا **Inkassama**	مسقتسا **istakssama**	مستقا **iktassama**	مسقا **Akssama**	مساق **kaassama**	ماسقنا **inkissam**	قمسق **kisma**	تامساقنالا **alinkissamat**
Was divided	Conjure	Share somebody in something.	Swear	Share something with somebody	Division	Division	The divisions