NKOSANA MALUMBA

LITERATURE REVIEW
MAY 15 2014

INTRODUCTION

Africa has 54 countries that have an estimated over 2000 languages in total. Some languages are endangered due to the assimilation of other dominant groups and the adoption of Western cultures (Mukami, 2013). In Africa, the language is an element of its culture as it presents the philosophy, history, stories, and medicinal practices of that particular culture. Therefore, an extinction of language will inevitably result in the loss of the diversity of the community that is Africa (Mukami, 2013).

Although there are a large amount of spoken languages in African, many of these, especially the Bantu languages of South Africa, still form part of the less researched languages in the world (Pretorius & Bosch, 2003). As a result, technologies that are crucial in the advancement of information retrieval research, such as corpora and dictionaries are still either undeveloped or incomplete. In the case of IsiZulu, which is the focal language of the Afriweb project, many researchers in linguistics have provided different perspectives which has resulted in a distributed and non-cohesive body of knowledge (Madondo & Muziwenhlanhla, 2000).

The first section will cover existing works in information retrieval that focus on uncommon languages such as Arabic, IsiZulu, and Swahili. This will include some of the strategies that were used in the development of these IR[1] systems and any challenges that were outlined in the development of these systems. The second section will provide an overview into the technologies that aid in information retrieval with relevance to the Afriweb project.

LANGUAGE BASED INFORMATION RETRIEVAL

Information retrieval is a field in computer science that is concerned with organization, storage, and displaying of information. The retrieved information is usually of a ranked nature that is sensitive to the user submitted query. (Manning, et al., 2008).

An interest in language-based information retrieval has been on the rise due to failure of the more popular search engines failing to provide language specific results for the smaller languages

Although websites such as Wikipedia do contain some pages that are in South African languages such as IsiXhosa and IsiZulu, it is close to impossible to find this information.

[1] Information retrieval

According to the census conducted in South Africa in the year 2011, there are about 8.1 million Xhosa and about 10.6 million Zulu speaking people. Therefore, a significant fraction of the population has difficulty in finding information that is written in their language. One area that is influenced by this is research, because research is largely based on finding information that is of relevance and usefulness to the topic at hand.

As the digital age advances, a vital necessity for tools that are language specific is increasing as this is aiding many areas such as teaching, learning, and research. The following sections will discuss topics in information retrieval of small languages. These topics include Arabic IR[2], Swahili IR, Afaan Oromo IR, and IsiZulu CLIR[3]

Arabic Information Retrieval

Arabic is a Semitic language that belongs to a family of languages dominant in parts of North Africa and Western Asia. Semitic languages are known for their nonconcatenative morphology were the words are created by adding in vowels to a set of root consonants. Arabic has distinctive morphological features and its writing orientation is from right-to-left (Abdelila, et al., 2004).

The effectiveness of an information retrieval system depends on the system's capacity to conform to the language in use. Therefore, understanding of characteristics of the language is very critical in building high quality information retrieval systems (El-Khair, 2007). As described by Nwesri et, al. (2007), Arabic is an inflectional language that requires the use of morphological analysis in the retrieval of Arabic text. This is to ensure that the various inflected forms of a word can be mapped to their root equivalent. Once converted to the root form, the search engine has a better chance of matching these words in a search through their relation to the root form. In the case of cross language informational retrieval, translations were achieved through morphological analysis by reducing Arabic text to its stem form. This made it easier to then translate the Arabic text to English. An additional post-processing step was required to correct an grammatical errors that may have occurred during the conversion to root form (Nwesri, Tahaghoghi & Scholer, 2007).

The major limitation in the development of the Arabic IR system has been the fact that there is a lack of adequate resources available to evaluate the system's performance in real world settings. These include tools such as complete morphological analyzers, a language corpus, and a machine-readable lexicon. For research purposes, a language corpus was derived from a newspaper that is printed in the language. Although this did not provide an exhaustive list of the words that are part of the language, it provided a good estimation of the words that are commonly used in the Arabic language. On evaluation of the corpus, inconsistencies in the transliteration of proper names were discovered, which could affected the accuracy of the information retrieval system. This inconsistency was pointed out in the TREC[4] evaluation

[2] Information retrieval
[3] Cross language information retrieval
[4] Text Retrieval Conference

of the LDC[5] corpus, which has been central to most research conducted on Arabic IR systems (Abdelila, et al., 2004).

Arabic information retrieval systems have also attempted to make use of some of the strategies available to improve the performance of the retrieval system, such as stemming, stop lists, and term weighting. These techniques are all aimed at improving the relevancy of the results.

This section has outlined the benefits of using the language's morphology in developing an IR system. Additionally, lack of resources such as a language corpus has been shown to have a limiting effect on the effectiveness of the IR system. The next section will discuss Swahili IR system. Swahili is a Bantu language that is closely related to IsiZulu in its origins.

Swahili Information Retrieval

Swahili is a language of Bantu origin, which is part of the language group that South African languages such as IsiZulu and IsiXhosa belong. The language also has borrowed words from Arabic, which is because of the prevalence of Muslim groups in the areas that Swahili is dominant.

In 1992, Hurskainen described the first morphological analyzer for the Swahili language. The motivation for the development of this parser was that, in agglutinating languages such as Swahili, it is not convenient to use a direct string matching search method. This is due to the various inflected forms that a word can have because of its use in a piece of text (Hurskainen, 1995). As discussed in section 2.2.1, morphological analysis makes it possible to map the various inflected forms to a root word, which allows the search engine to pick up all these related terms given a query. The information retrieval system, SWATWOL[6] was designed to analyze the Standard Swahili language based on a two level formalism, where each character has a lexical and surface representation. A few morphological inconsistencies were found in this process. These were resolved using rules that are parsed in to the system as input (Hurskainen, 1995).

SWATWOL was implemented in the Swahili Language Manager (SALAMA), which is a computational system that facilitates many kinds of applications that are based on written Swahili text. Some of the abilities of SALAMA include producing the full vocabulary of a given text, translation of a particular text and syntactic analysis (Hurskainen, 1999). In the information retrieval space, SWATWOL was used as the morphological analyzer for unrestricted and non-encoded Swahili text. According to Hurskainen, there was a notable improvement in the accuracy of the search because of the integration. This is because the morphological characteristics of a language can be far more reliable as a key for information retrieval as opposed the direct search method (Hurskainen, 1995).

The application of morphological analysis in the development of information retrieval systems for languages that have two-directional word formation has proved to be a powerful

[5] Linguistic data consortium
[6] Swahili Two Level

method (Hurskainen, 1992). These formations are affected by the affixes that are applied to a word and change in morphemes that occur as a result of these applications. As Swahili and IsiZulu belong to the same language group, an assumption can be made that the use of the language's morphology will aid in the development of an effective IR system.

The next section will discuss cross language cross language information language in a Semitic language, which has many similarities with IsiZulu, in terms of its morphology.

Cross Language Information Retrieval for the Afaan Oromo Language

Afaan Oromo is a language that is widely spoken in Ethiopia. It belongs to the group of Semitic languages similar to Arabic. Similar to IsiZulu, the grammatical information of the language is conveyed by prefixes and suffixes. Afaan Oromo is the instructional medium in junior and secondary schools. Additionally, a number of works such as newspapers, magazines, education resources, official documents, and religious languages have been published in the language (Tune, et al., 2007).

In Afaan Oromo, the lack of a rich language source did not present a barrier as in the IsiZulu case. An Oromo-English dictionary was available that had been developed from hard copies of a bilingual dictionary. It was enhanced by incorporating additional entries, other language reference sources (Tune, et al., 2007).

Two popular information retrieval techniques, stopword lists and stemming, were adopted because of a rich language source. This significantly reduced the number of words in Afaan Oromo topics as the search engine could omit the words given as stopwords (Tune, et al., 2007). A light stemmer was used to automatically remove frequent inflectional suffixes that were attached to base form words. This was to ensure that the translation process could be simplified by translation of root words. The stemmed words were then translated using the bilingual dictionary by taking into account all possible keywords. Unmatched words were manually added to the dictionary (Tune, et al., 2007).

In the evaluation of the system, Afaan Oromo queries were translated into the English language and tested on the LA Times and GH Herald newspaper sources, which totaled about 169 477 documents (Tune, et al., 2007). Table 1 denotes the results of the experiment.

Table 1: Afaan Oromo experiment results

Run-label	Relevant-tot.	Rel. Ret.	MAP	R-Prec
OMT	1,258	870	22.00%	24.33%
OMTD	1,258	848	25.04%	26.24%
OMTDN	1,258	892	24.50%	25.72%

OMT refers to the Oromo title, OMTD referees to Oromo title and description and OMTDN refers to Oromo title, description, and narration. This was for assessing the overall performance of the Oromo-English cross language information retrieval system .The Relevant-tot is the total number of relevant documents, the Rel. Ret refers to the relevant retrieved documents, the MAP is the mean average precision, and R-Prec is the non-interpolated average precision.

OMTD was shown to have a better performance, which was attributed to the fact that most title fields in the CLEF[7] topics where very short. Therefore, the description of the document provided a better sense of what the content of the document was, which resulted in increased relevancy. Including the narrative lead to the increase of data to be considered by the matching, which negatively affected precision (Tune, Varma & Pingali, 2007).

The Afaan Oromo CLIR system has outlined the various effects that search optimization methods such as stemming and the use of stopwords have on the efficiency of retrieval systems. The next section will discuss the retrieval of IsiZulu indigenous knowledge. This will include the challenges that were encountered when attempting to use the CLIR approach to the retrieval of IsiZulu text

Indigenous Knowledge Retrieval in IsiZulu

Indigenous knowledge is the local knowledge that is unique to cultures and societies. It is a body of knowledge that enables communities to survive, and it commonly held by the people (Cosjin, et al., 2002). Indigenous knowledge is based on the ideas, experiences, practices, and information that has been generated either locally or elsewhere and has gone through a certain amount of transformation in other to be incorporated in the way of life of a particular culture. Indigenous knowledge is not confined to rural areas; it is present throughout all types of communities (Njiraine, et al., 2010). Indigenous knowledge has been originally collected and stored in paper archives, however, the digitization of these achieves has necessitated the development of an information retrieval system.

Due to government legislature, efforts have been increased to collect indigenous knowledge from various language groups, to prevent loss. A study that was carried out in Kenya shows that there has been a significant increase the publications in South Africa (Njiraine, et al., 2010). In the case of Kenya, the slow growth has been attributed to the possible absence of legislature regulations in the publications of indigenous knowledge.

Cross Language Information Retrieval Systems were the domain of research concerning indigenous knowledge. This is due to the South Africa situation, where there are 11 official languages. As seen in Fig. 2, the publications in South Africa increased rapidly from the year 2000 and reached its peak in 2005. The rate of publication then decreased sharply in the next few years.

[7] Conference and Labs of the Evaluation Forum

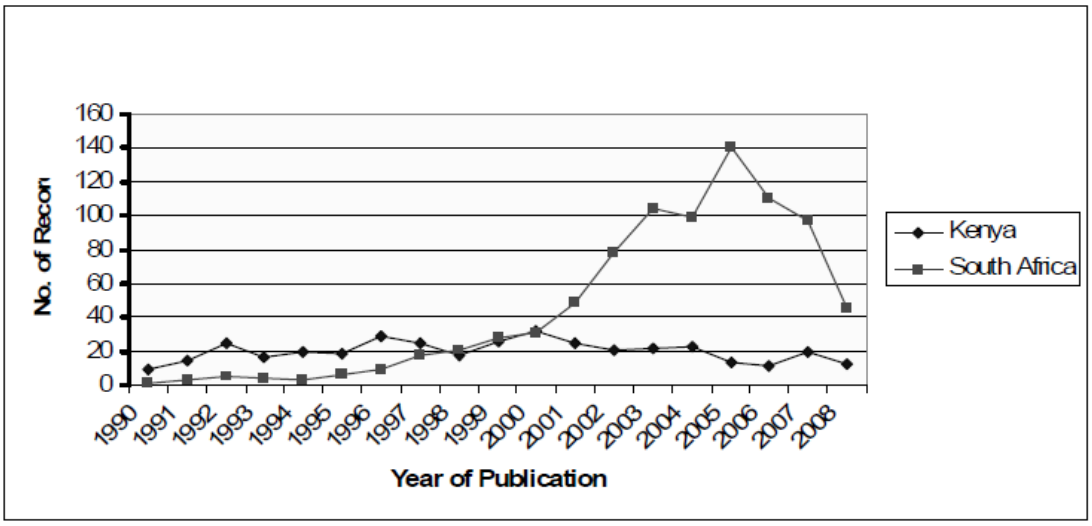

Figure 1: Trend of publications of IK Literature

In IsiZulu, grammatical information is conveyed by the morphology of the language. This presented several challenges when attempting to submit user queries in English to retrieve indigenous knowledge in IsiZulu. For example, the word “esesabekayo” comes from the root “esabeka” which can either mean capable, fearful, feared, awe inspiring, or wonderful. However, in English, the words capable, fearful and wonderful all have different semantics. This resulted in a loss of semantics in the translated meanings, which affected the precision of the returned results (Cosijn, et al., 2007). Similar problems were found when attempting to translate from an IsiZulu term to the English language, when the results had been retrieved from a query.

The next section will discuss web crawling, which is the process of harvesting web pages from the web page. The relevance of this top stems from the fact that the IsiZulu search engine will need to index web pages that will be aiding in providing search results given a user’s query.

WEB CRAWLING

A web crawler is an Internet software application that systematically browses the World Wide Web to index the contents of websites or the Internet as a whole. Given a set of URLs as input, the crawler visits these URLs and, based on a set of rules, indexes the page and also scans for other URLs within the same page, which it can next visit (Spiegler, Van Der Spuy & Flach, 2010).

As the focus of this project is to create an African Language based search engine, there is need for a focused crawler that is able to harvest documents that contain some IsiZulu text.

This section will discuss two types of focused crawlers (topic and language based) and language corpora.

Topic Based Crawler

A topic-based crawler generally seeks, acquires, indexes, and maintains pages on a set of specific topics that represent a relatively narrow segment of the web. A topic-based crawler has two main components, which is a classifier and a distiller. The classifier makes relevance judgments on the crawled pages based on the selected topics that are given by the user as input. The distiller ranks the importance of the crawled pages and aims to identify hubs for a particular topic to improve the rate of harvesting (Chakrabarti, et al., 1999).

Ideally, a focused crawler should be able to download webpages that are only relevant to the topic being that is required. The probability that a link to a particular page is relevant can be predicted before downloading the page. This can be done with the use of anchor text analysis, which checks to see if the anchor text has a correlation to the topic being investigated. Another approach would be to determine the relevance of a page after its content has been downloaded (Yamana & Chan, 2010). Once found, relevant pages would be harvested by the crawler and any links found on the harvested page can be added to the crawl frontier. The frontier is responsible for determining the next URI to be visited by the crawler.

Relevance is enforced on to the crawler with the use of a hypertext classifier that uses a categorized taxonomy to judge which categories a certain document belongs to. There are two strategies used, the soft focused and hard focused strategy. In the soft crawler strategy, the crawler uses the relevancy score of each crawled page as a priority value for all the unvisited pages that are added to the crawler's frontier (Chakrabarti, et al., 1999). In hard focused crawling, the classifier is invoked on a newly crawled page and checks the taxonomy for the best matching category. The URLs found on the page as also checked against this category, if they are nodes on the best matching path, they are added to the crawler's frontier.

Focused crawlers have been found to provide highly specialized fields with highly relevant information using a topic based classifier. However, there is need for human input in training the crawler to become highly specialized through examples and manual classification of the results that are received (Chakrabarti, et al., 1999).

The next section will discuss language-based crawlers. The language based crawler differs with the topic based crawler in that its classifier requires a language model to determine if a web page should be harvested or not.

Language Based Crawler

As the World Wide Web has expanded over the past few years, the CJK (Chinese, Japanese, and Korean) web has seen an increase in its websites. Focused crawlers have been developed to capture these unique languages as the English crawlers have given unsatisfactory results in the harvesting of non-English web pages (Yamana & Chan, 2010). The interconnectivity of the Web has led also to several issues, as it is possible to find a CJK webpage that links to an English one or vice versa.

Two strategies were used in ensuring that the crawler was able to identify the language of the Web page. The first method was to extract the domain name from the hyperlink's URL and then determine the top-level domain, for example, ".js" for Japanese Web pages. Once the URL has been determined to be of the targeted domain, the next step was to enqueue the URL for crawling. Alternatively, if the anchor text is also in the target language, then it is also added to the queue of URLs to be visited (Yamana & Chan, 2010).

The second method that was used is similar to the topic-based crawling concept, but required replacing the topical classifier with a language identifier that would indicate if the downloaded page is written in the target language (Yamana & Chan, 2010). This required a language model to be built, considering the language and character encoding schemes from the occurrence frequency of each n-gram in each language's text corpus (Yamana & Chan, 2010). A language model, usually known as a statistical language model, assigns a probability to a set of words using a probability distribution, to estimate the probability that a particular text is in a particular language.

The next section will discuss language corpora and their effects on language based focused crawlers.

LANGUAGE CORPUS

A corpus is a large set of structured texts that are used to do statistical analysis, hypothesis testing and rule validation. In the case of information retrieval, corpora provide information about a language that is used to create faster and efficient retrieval systems (Spiegler, et al., 2010).

Corpora have been found to affect the performance of focused crawlers and cross language information retrieval systems. In the English-Zulu CLIR[8] system, a parallel text corpus was created to translate from one language to the other (Cosjin, et al., 2002). Several ambiguities resulted in some translations being incorrect and thus affecting the quality of the queries that are submitted. In language based crawling, a language corpus is used to create a language model which is a statistic model which approximates if a given stream of text is of a particular language. This allows the crawler to classify if the page should be harvested or not.

[8] Cross language information retrieval

In the Afaan Oromo case, it was found that the quality of the corpus affects the stopword lists that are generated by the stopword algorithm, which has several effects on the words that are omitted. Additionally, some terms were not found in the dictionary source, which reflects on the quality of the language corpus being an issue (Tune, et al., 2007).

SUMMARY

The purpose of this chapter has been to provide an overview into the field of information retrieval and an investigation into existing works in the field. This included several strategies that were used in improving the efficiency of the retrieval systems.

BIBLIOGRAPHY

El-Khair, l. . A., 2007. Arabic Information Retreival. In: *Annual Review of Informatiomn Science and Technology.* Egypt: John Wiley and Sons , pp. 505 - 533.

Abdelila, A., Cowie, J. & Soliman, H. S., 2004. Arabic Information Retrieval Perspectives. *Arabic Language Processing.*

Chakrabarti, S., van der Berg, M. & Dom, B., 1999. Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks,* Volume 31, pp. 1623-1640.

Cosijn, E. et al., 2007. RETRIEVING INFORMATION IN ONE LANGUAGE VIA ANOTHER.

Cosjin, E., Pirkola, A., Bothma, T. & Jarvelin, K., 2002. Information access in indigenous languages: a case study in Zulu. *South African Journal of Libraries and Information Science,* 68(2), p. 94.

Hurskainen, A., 1995. Information Retreival and Two-directional Word Formation. *Nordic Journal of African Studies,* 4(2), pp. 81-92.

Hurskainen, A., 1995. Information Retrieval and Two Directional Word Formation. *Nordic Journal of African Studies,* 4(2), pp. 81-92.

Hurskainen, A., 1999. Swahili Language Manager. *Nordic Journal of African Studies,* 8(2), pp. 139-157.

Manning, D. C., Prabhakar, R. & Schutze, H., 2008. *Boolean Retrieval.* Volume 1 ed. Cambridge: Cambridge University Press.

Mukami, L., 2013. *African Review.* [Online]
Available at: http://www.africareview.com/Special-Reports/Africas-endangered-languages/-/979182/2008252/-/12yos0s/-/index.html
[Accessed 14 May 2014].

Njiraine, D., Ocholla, D. & Onyancha, O. B., 2010. Indigenous knowledge research in Kenya and South Africa: an informetric study.. *Indilinga African Journal of Indigenous Knowledge Systems: Indigenous Knowledge and Poverty Eradication,* 9(2), pp. 194-210.

Spiegler, S., van der Spuy, A. & Flach, P., 2010. s.l., s.n.

Tune, K. T., Varma, V. & Pingali, P., 2007. *Evalutation of Oromo-English Cross Language Information Retreival.* Hyderabad, India, Cross Language Evaluation Forum.

Yamana, H. & Chan, S.-B., 2010. *The method of improving the specific language focused crawler. In Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing.* s.l., s.n.