Research Methods: Literature Review

Moukangwe Katlego : MKNKAT003

May 15, 2014

Abstract

In this document I looked at different methods and strategies that were looked at when implementing African language information retreival. The African language investigated was Zulu and problems inherent to Zulu information retrieval are considered and evaluated.

Introduction

Zulu is one of South Africas indigenous languages spoken by about 10-11 million people. Multiple attempts have been made to digitally preserve South African languages and cultures, however not a lot of progress has been made. There is currently little research and development performed on South African language information retrieval. One of the pioneering works performed on South African language information retrieval was made by Erica Cosijn, Ari Pirkola, Theo Bothma and Kalervo Jrvelin from the University of Pretoria in a paper titled *Information access in indigenous languages: a case study in Zulu.* Their paper focused on the digital accessibility of indigenous knowledge with Zulu being the main language. Cross-Lingual Information Retrieval (CLIR) and metadata were discussed as possible means of facilitating access. Popular CLIR approaches and their resource requirements were analysed. The critique concluded by showing that there are multiple problems and difficulties encountered when implementing CLIR for african languages. These difficulties present unique research opportunities for research in CLIR.

Zulu is morphologically different from English and a lot of European and Asian languages. This poses unique challenges when indexing and searching for Zulu text. This paper outlines current developments made in Zulu information retrieval, and how information retrieval is different for resource scarce languages. The initial problem to solve is harvesting Zulu information on the Internet. This process is accomplished by using a focused crawler. This is the first problem to solve when implementing a Zulu search engine. This document will review the necessary literature required when implementing a Zulu search engine, which is the task at hand.

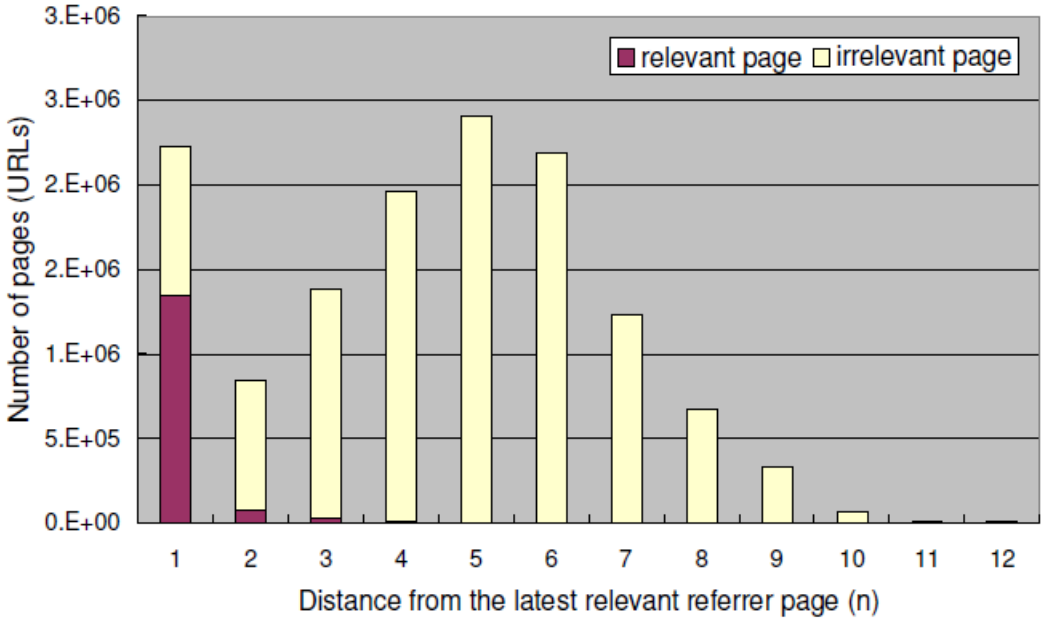
Focused Crawling

A general purpose Web crawler iteratively visits links and downloads the visited webpages. Focused crawling selectively downloads pages that are relevant to a specific topic. Focused crawling assumes topical locality. Topical locality means that pages that are linked together have similar content. A focused crawling system has three main components: a classifier, which determines the relevance of crawled pages to decide on link expansion; a distiller, which identifies hubs, i.e. pages with large lists

of links to relevant Web pages; and a crawler, which is controlled by the classifier and the distiller[1].

The domain or topic of interest is crawling for Zulu web pages. This means a Web page is relevant if it contains Zulu text. If a page contains Zulu text then it is assigned a relevance score of 1, otherwise 0. Language-specific Web crawling usually employs 2 strategies, which are *The referers relevance based strategy* and *a distance based strategy.*[1] *Referers relevance based strategy* follows links from relevant pages before visiting other pages. This strategy assumes that pages that are linked together are likely to have the same language[1]. The distance based strategy follows links extracted within a minimum distance from the last relevant page. The distance of a link is the number of pages visited from the last relevant page.

Experiments have been done on crawling Thai pages. Crawling was started from some well known Thai Web-directories. Both *referers relevance* and *distance based strategies* were employed and the results compared. A total of 14,240,725 URLs were visited. Below is a graph showing the ratio of relevant to irrelevant pages obtained through the distance based strategy for different distance values.



There are numerous other crawling methodologies and improvements being researched. The most notable to date is the use of two or multiple *classifiers*. The first *classifier*(called the baseline learner) collects training data for the second *classifier*(called the apprentice learner). The priority of an arbitrary node, v is specific to the features associated with the link, (u, v) that leads to it. The features used by the apprentice are derived from the Document Object Model(DOM) of node u[9]. The baseline learner specifies what kind of content is desired and the apprentice learner specifies how to obtain pages to feed into the baseline learner[9]. This approach has been proven to improve crawling speeds enoumously and improve assigning visiting priorities for unvisited nodes. All these methodologies can be applied to Zulu, depending on the ease that Zulu documents can be specified

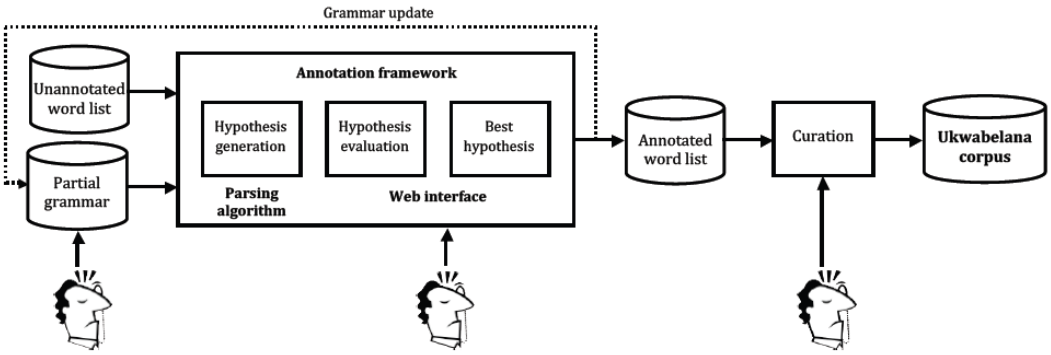
or recognized.

Cross language information retrieval

Cross language information retrieval(CLIR) is a subfield of Information retrieval concerned with finding information in a language different from the language of the user query. Kazuaki Kishidas paper on *Translation disambiguation for cross language information retrieval using context-based translation probability*[3] outlines multiple problems that were faced during their CLIR systems implementation. These problems include: cognate matching, translation types (query, document, interlingual), dictionary-based mapping, disambiguation of multiple translations, machine translation, phrasal translation,parallel and comparable corpus-based methods for probabilistic translation, mining Web resources for translation, merging issues for retrieval against multilingual corpora, use of pivot languages for indirect translation, language-specific issues such as tokenization and segmentation for Asian languages, stopword lists, stemming, decompounding and part-of-speech tagging.[2]. As mentioned earlier, the University of Pretoria performed a case study on CLIR for Zulu. They investigated the applicability and viability of CLIR in accessing Zulu indigenous knowledge databases and problems faced regarding resource-scarce languages. The paper investigates three methods that can be used for CLIR for Zulu. These methods are outline below.

Corpus-based approach
In a corpus-based approach the source language query is translated into the target language query via a parallel corpus. Parallel corpora contain text in the source language alongside its translation or translations. A parallel corpus need not have a word-by-word translations of the source language and target language. It needs to contain words associated with the source language query. The case study done by the University of Pretoria concluded that using parallel corpora wouldnt be viable because of lack of readily-available Zulu corpora. A corpus-based approach is now possible due to corpora being developed for South African languages. The most well developed corpus for Zulu to date is called *Ukwabelana*, built by the University of Witwatersrand linguistic department.

Ukwabelana is an open source morphological Zulu corpus. Zulu is an agglutinative language. This means that most words could be obtained by combining various prefixes, suffixes and other words. The Ukwabelana project gathered a list of common Zulu terms, defined a partial grammar and parsed Zulu words through an algorithm that proposed possible parses or meanings based on the partial grammar. This Ukwabelana corpus employed its own labelling scheme. They labelled different words, prefixes and suffixes and categorised the labels according to classes. The annotation process they followed is presented in the diagram below(Figure 1) . This approach sped up the labelling scheme by a factor of about 3-4.



The unannotated words in the diagram above were obtained from the Zulu bible and Zulu literature. A simple context free grammar was used to define the partial grammar. Definite Clause Grammars(DCGs) were used as a formalism for representing the morphological Zulu grammar. Thier annotation framework caried out hypothesis generation, hypothesis evaluation and selected the best hypothesis. Hypothesis generation was required to account for missing verbs and noun-roots. Logic programming and abductive reasoning was used for the hypothesis generation component. Abductive reasoning is a reasoning method used when there is insufficient information or data. It generates possible hypotheses (parses) for an observation (word) and a given theory (grammar)[5]. The hypothesis is then evaluated by considering all possible explanations and then the best hypothesis was selected. In the case of Ukwabelana, a Web interface was used to crowdsource the selection of the best hypothesis.

Zulu doesnt have a well defined grammar and hence a partial grammar was defined and iteratively improved. This in-turn improved the hypothesis generation of the parsing algorithm. Alongside sentences and certain words, parts of speech were tagged. This was achieved by analysing the left and right context of a word form and selecting the most probable part-of-speech(POS) from a given list of possible POS-tags. Ukwabelana contains 10,000 morphologically labeled words and 3,000 POS-tagged sentences. It contains around $10,000$ common zulu types and around $30,000$ sentences. All the software used in the analysis and corpus generation is available online for free at http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/Resources/.

Machine-translation approach
This approach uses machine-translation systems for query translation. Translation is based on translation dictionaries for the translation to be successful. For many language pairs, machine translation systems do not exist. In the South African context, there are machine translation systems from English to Afrikaans and several African languages. However, the quality of translations is poor for CLIR applications[4].

Dictionary based techniques Dictionary approaches use a bilingual dictionary for query translation. The basic strategy is to look up a single word and put them into the target query. Regular translation dictionaries contain normalized words or words in base form. If a query string appears in inflected form, then in cannot be readily translated. Matching words in inflected form can be achieved using a morphological parser or using approximate string matching. Currently there isnt

any morphological parser suitable for parsing Zulu. Paper [4] used approximate string matching to handle inflected words.

Various algorithms were used for approximate string matching like n-gram based matching, edit distance or Longest common subsequence(LCS). In an n-gram matching strategy, the query string and indexed entries are decomposed into n-grams. N-grams are substrings of length n of adjacent characters. The similarity between the query strings and indexed entries is computed by comparing the sets of n-grams[6]. Edit distance can be defined as the minimum distance it takes to convert one string to another. LCM of two strings is the longest character sequence that occurs in both strings.

Problems faced by Zulu CLIR

1. There is currently no comprehensive electronic dictionary for Zulu. There is progress in making an electronic dictionary for Zulu but it is not ready for production.
2. Zulu has no single word equivalents for most words in English. This problem was partially solved by paraphrasing or borrowing words from other languages.
3. Finding similarities of words in inflected form. Inflection tend to add more noise and distorts similarity calculations.

Monolingual information retrieval of resource-scarce languages

One of the pioneering works on monolingual information retrieval of resource-scarce languages was done for Arabic in a paper titled, *Arabic information retrieval.*[7] They performed three monolingual runs and a single cross-language run using different stemming, stop-words and dictionary based approaches to compare the viability, performance and feasibility of each option. In order to handle variations in the way text was represented in Arabic, several kinds of normalizations were performed on text in the corpus and in the queries. The paper compared using stemming and not using stemming in their search engine. Stemming provided a 40% improvement on returned results. The improvement was measured by the increase in precision. INQUERY[8] is the search engine framework they used for the testing which approach was better. The language model used is described by Leah S. Larkey paper[7]. The results are outlined in the table below.

CONDITION	Name of Run	Average Precision	Number of Queries at or Above Median	
			Average Precision	Rel Ret in top 1000
Inquery baseline	UMass4	.2104	10/25	10/25
Inquery stemmed	UMass1	.3129	18/25	24/25
LM baseline	not submitted	.1858		
LM stemmed	UMass2	.2597	16/25	20/25

The above results suggest that using stremming is a good approach for agglutinative languages. This implies it will be a good idea to use stemming when working with Zulu.

Conclusion

South Africa has a heritage rich society. Research needs to be done towards digitizing and preserving African languages. In this paper I discused current developments in focused crawling and how the research can be used to find and download Zulu Web documents. I looked into attempts in developing an African language search engines. There was a lot of pioneering work into Zulu corpus building and working with resource-scarce languages. All the work done showed that there can be multiple problems encountered when implementing African language search engines. All this show that the field of African language information retrieval is still young and there is still more room for research.

References

[1] Simulation Study of Language Specific Web Crawling, Kulwadee Somboonviwat, Takayuki Tamura, Masaru Kitsuregawa. 2005. hyperref[scale]http://ieeexplore.ieee.orgstamp/stamp.jsp?tp=&arnumber=16

[2] Cross-Language Information Retrieval: the way ahead Fredric C. Gey a, Noriko Kando, Carol Peters.Received 10 June 2004; accepted 14 June 2004

[3] Translation disambiguation for cross language information retrieval using context-based translation probability Keio University, Japan, kishida@slis.keio.ac.jp

[4] Information access in indigenous languages: A case study in Zulu. Erica Cosijn & AriPirkola & Theo Bothma & Kalervo Jrvelin Department of Information Science, University of Pretoria, South Africa

[5] Ukwabelana- An open-source morphological corpus for Zulu.Sebastian Spiegler, Andrew van der Spuy, Peter A. Flach

[6] Pfeifer et al., 1996, Robertson and Willett, 1998, Salton, 1989, Zobel and Dart, 1995.

[7] Arabic Information Retrieval at UMass in TREC-10, Leah S. Larkey and Margaret E. Connell Center for Intelligent Information Retrieval

[8] Callan, J. P., Croft, W. B., and Broglio, J. TREC and TIPSTER Experiments with INQUERY. Information Processing and Management, 31 (3), pp. 327-343, 1995.