3P-6

Indexing Tigrinya language documents

Omer Osman Ibrahim[†] Yoshiki Mikami[†]
Nagaoka University of Technology[†]

1. Introduction

The Web has become one of the most valuable source of information to many people around the world. However, due to its huge size and dynamic changes, finding information is not a straightforward task. The most feasible solution for finding information on the Web is to use a Search Engine. However, not all Search Engines fully support minority languages search on the Web. Tigrinya is one of the languages which is not fully and efficiently accessible through the available Search Engines. Tigrinya is a Semitic language spoken in Eritrea and Ethiopia. It is estimated to be spoken by over 6 million people both countries [1].

In this paper, we report the indexing issues involved in the design and development a search engine for the Tigrinya language. Indexing facilitates Information Retrieval by putting document terms in a data structure which makes it easy and efficient to search queries. Indexing Tigrinya language documents, various normalization steps that are specific to the language should be performed. The results of our work came up with an original analyzer for specific for Tigrinya language documents that can be used for indexing Tigrinya web or any other applications that may require efficient Tigrinya search.

2. The Tigrinya Indexer

The indexer is an important component of a language specific search engine and it should be aware of the specific features of the target language. It processes the fetched document collection taking in to account the unique features language before it finally stores the terms in an internal data structure called inverted index. The Tigrinya indexer has two components:

- Tigrinya Analyzer component which considers the unique features of the Tigrinya language.
- Indexer component that stores terms with additional information into the inverted index.

Since existing analyzers for other languages do not consider the Tigrinya language features, we propose a Tigrinya analyzer that considers the unique characteristics the language. The analyzer helps improve the search results.

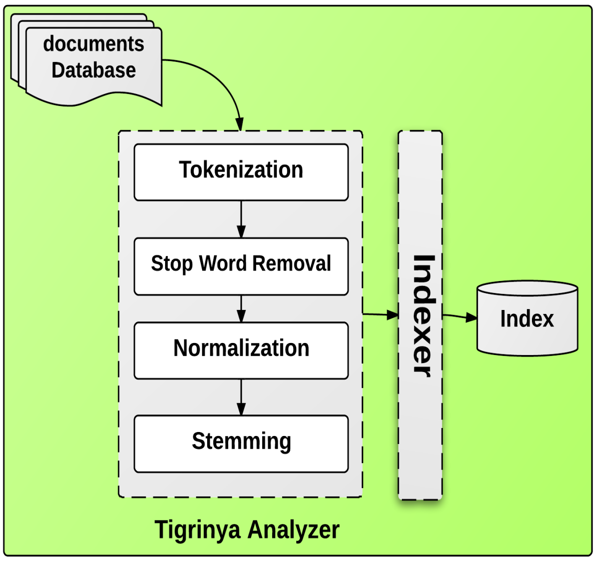

Figure 1. The Tigrinya Indexer

The Tigrinya analyzer first breaks text into tokens. Then, removes stop words from the text. Next, the text is normalized for short forms and same words with different alphabets. Finally, the words are stemmed to reduce them to their common form.

3. Language Specific Issues in Indexing Tigrinya documents

As shown in Figure 1. the Tigrinya analyzer performs pre-processing tasks before it documents are forwarded to the indexer. Language specific features considered by our analyzer include tokenization, stop word removal, stemming, normalization of short form of words (abbreviations) and normalization of same sound alphabet words.

3.1 Tokenization

Tokenization is the processes of determining word boundaries in a text. It defines the terms suitable for indexing. Tigrinya uses Ethiopic script. The Ethiopic script has its own unique punctuation marks such as ። ፡ ፤ ፧ , etc. The Tigrinya tokenizer identifies word boundaries using spaces and Ethiopic punctuation marks as split points.

3.2 Stop Word Removal

Stop words, by definition, are those words that appear in the texts frequently but do not carry significant information for the purpose of Information Retrieval [2]. They have little value in selecting documents that satisfy user's main query need. We have used a list of 1176 stop words that we have constructed from our corpus. Examples include ኣብ ,ካብ, ናይ, ምስ , ከም, ዘሎ, ናብ ... etc. The Tigrinya analyzer checks each token in the list of stop words. A token is removed if found on the list. The remaining tokens are forwarded to the next analysis step.

3.3 Short form of words handing

The Tigrinya analyzer also normalizes short form of words. In Tigrinya language, some words are commonly written in short form using forward slash("/").For example, ቤ/ት represents the word ቤት ትምህርቲ meaning school. Such forms are expanded before indexing using a Tigrinya short forms database.

In Tigrinya language, each alphabet represents a separate sound. However, there are a few alphabets that represent the same pronunciation. For example, the 'ሰ ' and 'ሠ' (se) series. The analyzer normalizes such alphabets into a common form.

3.4 Stemming

Stemming is a normalization step that reduces the morphological variants of words to a common form. Tigrinya is a morphologically rich language. Thus, stemming have positive effect on the search quality. The Tigrinya analyzer reduces inflected words before they are finally indexed. The analyzer uses a Tigrinya stemmer [3] that we have developed for the purpose of this work.

4. Implementation

Lucene is an open source, scalable and high performance information retrieval library. It has analyzers for many languages. Thus, we have implemented our own analyzer with Tigrinya language features and integrated it with Lucene.

5. Conclusion

Indexing increases the speed and performance of search engines. We have used our Tigrinya indexer for the purpose of developing a Tigrigna search engine. We hope our work will positively contribute in narrowing the digital divide among languages in the web.

References

[1] Ethnologue: http://www.ethnologue.com

[2] Zoe,F.et al.,(2006),"Automatic construction of Chinese stop word list", Research in computer science.

[3] Omer Osman, Yoshiki Mikami,(2012):" Stemming Tigrinya Words for Information Retrieval".The 24th International Conference on Computational Linguistics.