Amazigh Search Engine: Tifinaghe Character Based Approach

F. Ataa Allah[1]**, and S. Boulaknadel**[1]
[1]CEISIC, Royal Institute of Amazigh Culture, Rabat, Morocco

Abstract - *The technology underlying text search engines has advanced dramatically in the last decades. However up today, to the best of our knowledge, there is only the search engine "Google" that could support Amazigh language script (Tifinaghe). Nevertheless, google is based only on amazigh keyword matching, without using any natural language processing.*

Aware of such tool utility and effectiveness in improving the diffusion and the promotion of Amazigh language, we have proposed to develop a search engine that could support the language characteristics. Thus, the proposed search engine is designed to crawl and index the amazigh web pages written in Tifinaghe efficiently, and explore if using amazigh natural language processing such as stop words removal and light stemming in retrieval task produce much more satisfying search results.

Keywords: Amazigh Language, Tifinaghe Script, Multilingual Information Processing, Search Engine, Natural Language Processing, Indexing.

1 Introduction

Every day, the web creates new challenges, especially for information retrieval and web searching. The amount of information on the web is growing, as well as the number of the new scripts adopted by the less commonly taught languages, such as Tifinaghe for the Moroccan Standard Amazigh language.
Although the interest in Amazigh language is growing, unlike other less commonly taught languages such as Urdu [14], Thai [18], Hungarian [19], Bangla [10], and Punjabi [12], there is not any published work regarding Amazigh information retrieval or search engine.

Known by its complex morphology based on internal stem structure changing, and its oldest Tifinaghe script that was officially adopted, the Moroccan standard Amazigh language needs a specific information retrieval system that could support its particularities. In this perspective and in the context of promoting the Amazigh language, we have planned, in the Royal Institute of Amazigh Culture (IRCAM), to develop an Amazigh search engine supporting Amazigh morphology and Tifinaghe script, especially to avoid transliteration problems and to keep the Tifinaghe character semantics in document indexing and retrieval. Before adopting Tifinaghe as an official script in Morocco, most writing was in Latin alphabet supported by diacritics and phonetic symbols, or in Arabic script. This fact posed a problem for an internet user, who would like to look for some information. Which script he/she will use to write his/her query? How he/she could express a phoneme that has more than one correspondence in the Latin script, such as the word ⵜⴰⴼⵓⴽⵜ *the sun* that could be transliterated as "*tafuct*" or as "*tafukt*". Moreover, how could a child or an inexperienced person use diacritics and phonetic symbols? Thus, the Unicode encoding scheme, which encodes the set of Tifinaghe alphabet, is used as a common encoding platform to deal with the multilingual web pages in a uniform manner.

In this paper, we investigate, in developing an Amazigh search engine, the potential of indexing documents by using Tifinaghe characters, and the use of natural language processing to enhance Amazigh information retrieval. The components of this engine are mainly a data crawling, indexing, and searching system. The remaining of this paper is organized as follows: in Section 2, we give an overview about the Moroccan Amazigh standard language, and describe its characteristics and morphology. In section 3, we describe the structure of our web search engine. While, in Section 4, we present the Amazigh natural processing integrated into this engine. In the last section, we outline our conclusion and some perspectives.

2 Moroccan Amazigh language

The Amazigh language, known as Berber or Tamazight, is one of the oldest languages of humanity. Nowadays, it covers the Northern part of Africa which extends from the

Red Sea to the Canary Isles, and from the Niger in the Sahara to the Mediterranean Sea. In Morocco, this language is divided, due to historical, geographical and sociolinguistic factors, into three mean regional varieties, depending on the area and the communities: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas. Even though 50% of the Moroccan population is Amazigh speaker, the Amazigh language was exclusively reserved for familial and informal domains [7]. However in the last decade, thanks to the royal benevolence this language has become institutional, and integrated into the Moroccan education system.

2.1 Amazigh language characteristics

The Amazigh language is a branch of the Afro-Asiatic (Hamito-Semitic) language, a family in which languages like old Egyptian, Cushitic, or Chadic are found [9, 13]. Since the ancient time, it has its own writing that has been undergoing many slight modifications. In 2003, it has also been changed, adapted, and computerized by the Royal Institute of the Amazigh Culture, in order to provide the Amazigh language an adequate and usable standard writing system. This system is called Tifinaghe-IRCAM.

2.1.1 Tifinaghe-IRCAM graphical system

Since February, 11th, 2003, Tifinaghe-IRCAM has become the official graphic system for writing Amazigh in Morocco. This system contains:

- 27 consonants including: the labials (ⵀ, ⵃ, ⴼ), dentals (ⵜ, ⴷ, ⵟ, ⴹ, ⵏ, ⵍ, ⵕ, ⵝ), the alveolars (ⵙ, ⵣ, ⵚ, ⵥ), the palatals (ⵛ, ⵊ), the velar (ⴽ, ⵅ), the labiovelars (ⴽⷯ, ⵅⷯ), the uvulars (ⵇ, ⵅ, ⵖ), the pharyngeals (ⵄ, ⵃ) and the laryngeal (ⵀ);

- 2 semi-consonants: ⵢ and ⵡ;

- 4 vowels: three full vowels ⴰ, ⵉ, ⵓ and neutral vowel (or schwa) ⴻ which has a rather special status in Amazigh phonology.

2.1.2 Punctuation and numeral

No particular punctuation is known for Tifinaghe. IRCAM has recommended the use of the international symbols: " " (space), ".", ",", ";", ":", "?", "!", "…", for punctuation markers; and the standard numeral used in Morocco (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) for Tifinaghe writing [1].

2.1.3 Directionality

Historically, in ancient inscriptions, Amazigh language was written horizontally from left to right, from right to left, vertically upwards, downwards or in boustrophedon. However, the orientation most often adopted in Amazigh language script is horizontal and from left to right, which is also adopted in IRCAM-Tifinaghe writing.

2.1.4 Tifinaghe Encoding

In order to allow a large diffusion to the Amazigh language, and to contribute in its promotion, the Tifinaghe script computerization was a basic and essentially step. In this context, several efforts have been undertaken to encode Tifinaghe in Unicode/ISO 106461.

Actually, the Amazigh encoding system is composed of four Tifinaghe character subsets: the basic set of IRCAM, the extended IRCAM set, other Neo-Tifinaghe letters in use, and modern Touareg letters. The first subset constitutes the set of characters chosen by IRCAM to arrange the orthography of the different Moroccan Amazigh dialects while preserving most characters of the historical Tifinaghe script. This subset is classified in accordance in the range U+2D30..U+2D65, U+2D6F with the order specified in Tifinaghe-IRCAM alphabet. While, the Tifinaghe block is the range U+2D30..U+2D7F [2].

2.2 Amazigh morphology

In contrast with English, Amazigh is a highly inflected language. It has three main syntactic categories: noun, verb, and particle.

2.2.1 Noun

Nouns distinguish two genders, masculine and feminine; two numbers, singular and plural; and two cases, expressed in the nominal prefix.

- The feminine is used for female persons and animals as well as for small(er) objects. The productive derivation masculine feminine is quite regular morphologically, using noun prefixes and suffixes.

- The plural has three forms: the external plural consisting in changing the initial vowel, and adding suffixes; the broken plural involving changes in the internal noun vowels; and the mixed plural that combines the rules of the two former plurals.

- The annexed (relative) case is used after most prepositions and after numerals, as well as when the lexical subject follows the verb; while, the free (absolute) case is used in all other contexts.

2.2.2 Verb

The verb has two forms: basic and derived forms. The basic form is composed of a root and a radical, while the derived one is based on a basic form in addition to some prefix morphemes. Whether basic or derived, the verb is conjugated in four aspects: aorist, imperfective, perfect, and negative perfect. Person, gender, and number of the subject are expressed by affixes to the verb. Depending on the mood, these affixes are classed into three sets: indicative, imperative, and participial.

2.2.3 Particles

Particles contain pronouns; conjunctions; prepositions; aspectual, orientation and negative particles; adverbs; and subordinates. Generally, particles are uninflected word. However in Amazigh language, some of these particles are flectional, such as the possessive and demonstrative pronouns (For more details c.f. [1, 6]).

3 Search engine

Creating a search engine scaling to today's web presents many challenges. Such as system should supports a fast crawling technology to gather web documents and keep them up to date. Its storage space must be used efficiently to store indices and, optionally, the documents themselves. Its indexing system must support multilingual information, and process hundreds of gigabytes of data efficiently. Moreover, its query engine must handled queries quickly on retrieving relevant web pages.

Taking into our consideration these measures, we have tried to develop an Amazigh search engine that is mainly structured on three parts: data crawling, indexing, and searching. Most of the programs used in this search engine are writing in Perl, SQL, PHP and C language.

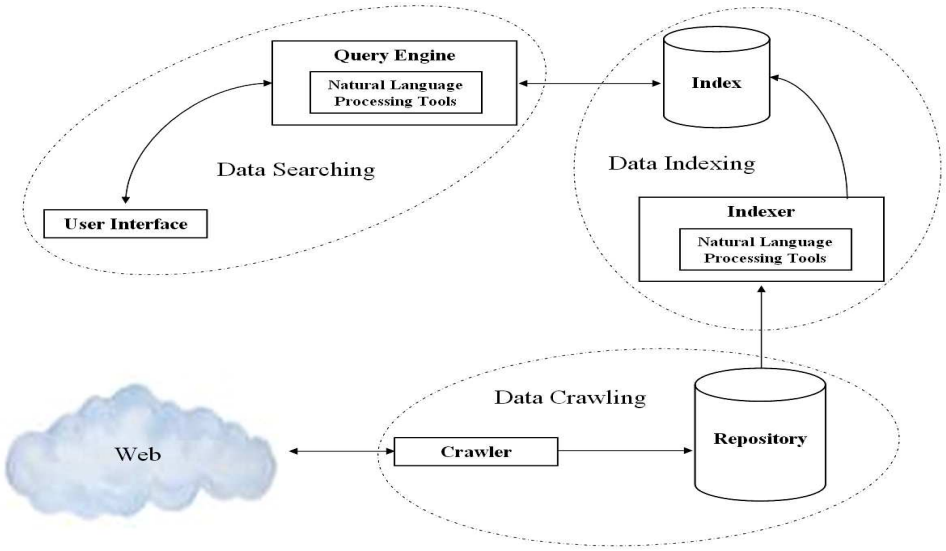

Figure 1. Amazigh search engine architecture

3.1 Data crawling

Data crawling is a process behind the search engine, based on spiders or web robots, to fetch web pages. In this context, we have developed a spider that collects automatically and daily all the pages containing Tifinaghe script from the IRCAM web site. All the fetched pages are stored in a web pages' repository, and associated to an ID number. To this ID, called also docID, a web page is assigned whenever a new URL is parsed.

3.2 Data indexing

Based on vector space model [16], our data indexing system creates an index for a set of documents and assigns a weight to each term-document association, based on the Okapi BM-25 formula [15]. The Okapi formulas, especially the BM-25 scheme, attack the problem of higher term frequencies in long documents, and the chances of retrieving long documents over shorter ones. Moreover, they have proven their performance and efficiency over other schemes

in information retrieval systems [3], and on web retrieval [8]. For this reason, we have opted to integrate the Okapi BM-25 to our system.

The data indexing system performs four steps on each document, including word identification, stop-word removal, light stemming, and indexing. First, we identify individual words in the document. Second, all the "stop words" in a document are removed based on a pre-defined list. Then, we reduce the words to their light stem. Finally, we record the information about the relationships between these words and the documents to support searching.

3.3 Data searching

The data searching system includes query engine and a web user interface. The query engine accepts search queries from users, applies the light stemming, represents the query as a vector in term-document space, and assigns a weight to each term-query. Then, it calculates scores between the user query and the set of documents. After retrieving search results, the query engine ranks the search results according to content analysis scores, generates a summary for each search result, based on the web pages' repository, and renders its link. Whereas, the web user interface allows users to submit their search queries and view the search results. When a user performs a search through the web interface, the query is passed to the query engine, which retrieves the search results and passes them back to the user, who can specify the number of retrieved web pages per each result page.

4 Natural language processing for Amazigh

4.1 Stop word identification

Amazigh language, as all other natural languages, is composed of two types of words: content words and functional ones. Generally, functional words that constitute the stop word list are a part of Amazigh particles. Usually, this list is used to identify words that don't need to be indexed because their use in a query word will return a large number of documents, possibly not relevant ones. In fact, the keywords that would be chosen for building the index should discriminate between documents by not occurring too often, or too seldom [5]. Thus, we have undertaken a deep research in the Amazigh morphology in order to extract the stop word list. This list includes aspectual, orientation and negative particles; disjunctive, interrogative, possessive, demonstrative, prepositional and indefinite pronouns; direct and indirect object pronouns; adverbs; prepositions; conjunctions; and subordinates.

4.2 Light stemming

Light stemming is a standard information retrieval task [11, 17]. It refers to a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognizing patterns and finding roots. This process maps several terms onto one base form, which is then used as a term in the vector space model. Our method is mainly based on the composition of words that is usually formed in the Amazigh language as a sequence of prefix, core, and suffix [4], without using any stem dictionary or exception list. Our algorithm is merely based on an explicit list of prefixes and suffixes that need to be stripped in a certain order. This list is derived from the common inflectional morphemes of gender, number and state for nouns; personal markers, aspect and mood for verbs; and affix pronouns for kinship nouns and prepositions. While, the derivational morphemes are not included in order to keep the semantic meaning of words.

5 Conclusion

This paper gives an overview about the Moroccan standard Amazigh language idiosyncrasies, and presents a search engine architecture supporting Tifinaghe-IRCAM script. In order to improve the Amazigh web pages searching quality, we have integrated Amazigh tools based on stop word identification system and a light stemming process; and investigated the vector space information retrieval framework, within the Okapi BM-25 scheme. However, there are several promising directions that deserve further study. First, it is noted that our crawling system fetches only web pages from the IRCAM web site. Whereas, we are looking to extend this system to fetch from other web sites that their pages include Tifinaghe script. Second, considering that the number of Amazigh web pages is growing rapidly, we will investigate some clustering algorithms. Finally, we plan to use other linguistic processing such as the lemmatization and compound noun extraction.

6 References

[1] Meftaha Ameur, Aicha Bouhjar, Fatima Boukhris, Ahmed Boukous, Abdallah Boumalk, Mohamed Elmedlaoui, El Mahdi Iazzi, and Hamid Souifi. "Initiation à la langue amazighe". The Royal Institute of Amazigh Culture, 2004.

[2] Patrick Andries. "La police opentype Hapax berbère"; La typographie entre les domaines de l'art et de l'informatique, IRCAM, Rabat, Maroc, 13—36, 2007.

[3] Fadoua Ataa Allah. "Information retrieval: applications to English and Arabic documents". Ph.D. Dissertation at the University of Mohammed V-Agdal, Rabat, Morocco, 2008.

[4] Fadoua Ataa Allah, and Siham Boulaknadel. "Light Morphology Processing for Amazigh Language"; Language Resources and Human Language Technologies for Semitic Languages: Status, Updates, and Prospects. LREC 2010, May 2010.

[5] Richard K. Belew. "Finding Out About: Search Engine Technology from a Cognitive. Perspective". Cambridge University, USA, 2000.

[6] Fatima Boukhris, Abdallah Boumalk, El Houssain El Moujahid, and Hamid Souifi. "La nouvelle grammaire de l'Amazigh". The Royal Institute of Amazigh Culture, 2008.

[7] Ahmed Boukous. " Société, langues et cultures au Maroc : Enjeux symboliques". Najah El Jadida. Casablanca, Maroc, 1995.

[8] Jianfeng Gao, Guihong Cao, Hongzhao He, Min Zhang, Jian-Yun Nie, Stephen Walker, and Stephen Robertson; "TREC-10 Web track experiments at MSRA". The 10th Text Retrieval Conference, Gaithersburg, Maryland, USA, 384—392, Nov 2001.

[9] Joseph Greenberg. "The Languages of Africa". The Hague, 1966.

[10] Nafid Haque, M. Hammad Ali, Matin Saad Abdullah, and Mumit Khan; "Infrastructure for Bangla Information retrieval in the context of ICT for development"; Advances and Innovations in Systems Computing Sciences and Software Engineering, Springer Netherlands, 325—330. Aug 2007.

[11] Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. "Improving Stemming for Arabic Information Retrieval : Light Stemming and Cooccurrence Analysis"; The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 275—282, Aug 2002.

[12] Gour Mohan, Ankur Garg, Pramod K. Gupta, Sunita Arora, and Karunesh Arora. "Adaptation of On-line information retrieval framework to support Punjabi language"; The 2008 Annual Seminar of C-DAC Noida Technologies, Noida, India, Feb 2008.

[13] Omar Ouakrim. "Fonética y fonología del Bereber". Survey at the University of Autònoma de Barcelona, 1995.

[14] Kashif Riaz. "Concept Search in Urdu"; The Second Ph.D. Workshop in the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, 33—40, Oct 2008.

[15] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. "Okapi at TREC-3"; The 3rd Text Retrieval Conference, Gaithersburg, Maryland, USA, 109—126, Nov 1994.

[16] Gerard Salton. "Automatic Information Organization and Retrieval". McGraw-Hill, 1968.

[17] Jacques Savoy. "Stemming of French words based on grammatical categories"; Journal of the American Society for Information Science, Vol. 44, n° 1, 1—9, 1993.

[18] Thanaruk Theeramunkong, Wirat Chinnan, Thanasan Tanhermhong, and Virach Sornlertlamvanich. "Full-text Search for Thai Information Retrieval Systems"; The 5th International Workshop on Information retrieval with Asian languages, Hong Kong, China, 75—80, Sep-Oct 2000.

[19] Anna Tordai, and Maarten de Rijke. "Hungarian Monolingual Retrieval at CLEF 2005", Working Notes for the CLEF 2005 Workshop, Vienna, Austria, Sep 2005.