==

Language in India www.languageinindia.comISSN 1930-2940 Vol. 22:5 May 2022

==

Crawler and Its Linguistic Challenges in the Arabic Language Sites

A Case study of Syrian Newspapers

Asmaa Alhaj Badran

Jawaharlal Nehru University, New Delhi 110067

asmaaalhajbadran@gmail.com

==

Abstract

Crawler, a Web indexing program or an Internet robot/bot (Spetka, 2004), is a software application that runs automated scripts over the Internet. The Web engines use it to update the content and sites via copying all the accessed pages and processing them into indexes so that the users can search much more sufficiently. Crawling is the first stage that downloads Web documents, which the indexer indexes for later use by searching module, with feedback from other backgrounds. This module could also provide on-demand crawling services for search engines. Yet, with the massive amount of data that has been fed on the web, we still encounter some problems and challenges while crawling data. Subsequently, through the wide-open access to all search engines, Arabic content is hitherto scantily accessible. This paper descriptively details the stances and challenges that the Arabic language, the fifth most spoken language, might grapple with while crawling data.

Keywords: Web crawler, Arabic Language Sites, search engine, methodology, challenges, limitations, newspapers' logistic expressions.

1. Introduction

The fast-paced growth of the Internet magnetizes the researchers to facilitate the load of data from different fields, providing metadata for all users regardless of their mother tongue. The web-searching engine confronts various problems that may improve or detract from the crawling of online data. The problems are either novel, like those I faced through my crawling process, or deep-rooted; they have been dealt with before but never solved. This project aims to raise awareness of these problems that, consequently, could benefit the maintenance of the crawler server to be improved in indexing the data. Therefore, a Web crawler starts with a set of URLs to get them indexed; they are called *seeds*. These URLs are recursively visited according to a series of policies, which may allow, or allow you not, to access data. The initial step is to choose a Web page from the URL, which must then be processed by extracting the text and links. These extracted links are added to the URLs frontier to be crawled. (Bal, S., 2012) Best of the most significant crawlers, covered by the enormous pages on the Internet, prove inadequate for presenting a complete index. Generally, the crawler has two options for maintaining its index. Essentially, the crawler examines the web until the collection reaches several pages and then stops viewing pages. When it is time to reset the pool, the crawler creates a new one using the same procedure outlined above and then substitutes the old one with the new one. Conversely, the crawler may continue visiting sites after the collection has reached its desired size to modify the current collection progressively. (Cho, J. and Garcia-Molina, H., 2000)

2. Crawler and the Search Engine

Some Web engines use the spidering software to keep their Web indices and content updated on other sites. All the pages visited by the Web crawler can be copied for later processing using a search engine that indexes the pages to be available for the users to search more efficiently. A crawler can visit pages without approval, *yet the problem is that not all the pages allow crawling their data*. Moreover, as the number of Web pages is tremendously significant, the crawler falls short of a complete index. Indisputably, not all web pages are equally appealing to the crawler's client. For example, if the user creates a customized database on a specific topic, pages pertaining to that subject matter are more significant and should be accessed as soon as possible. Alternatively, a search engine may prioritize user query results proportional to the number of Web URLs that refer to a page, known as the backlink count. If the crawler fails to visit all pages, it is preferable to visit those with a higher backlink score since this will provide the end-user with higher-ranking results. (Cho, J. et al., 1997)

3. Methodology of the Work

3.1 Steps of Crawling Data

To extract data from an entire website, you need a web crawler. A web crawler can systematically extract data from multiple web pages for you. It does so by crawling or accessing web pages and returns the data you need for each one of them.

Setting the crawler off:

1. Choose a site to crawl.

2. Open the site and then copy the URL.

3. Consequently, paste the URL in a txt file within the URL Folder.

4. Open the language file, and then write the suitable Unicode of the given language of the site you chose.

5. Go to the crawler batch and get it launched.

6. A note will appear asking you to add the timing, the hour, and then the minutes.

7. Add them accordingly.

8. Press Enter button, then your data starts to crawl.

9. When the crawling process is completed, go to the output file and copy the extracted data.

10. After that, go to the sanitizer to sanitize the collected data.

11. A note will ask you to give a specific number of the sanitized data you need.

12. Take the sanitized data and copy them in your final copy in the dictionary file.

4. Challenges for the Crawler

As a key objective, the focus was to get some crawled data from the Arab press sites to analyze them in sequence, but what I faced throughout the process of indexing was auspiciously not up to the expected level. These problems can be divided into two parts:

A. Before crawling data: more precisely, technical problems

B. After crawling data: lexical problems, accumulatively, the indexed data in the dictionary. (Each problem will be explained separately)

A. Before crawling data:

First, the developers or editors blocked most of the sites that I tried to crawl from; they never permitted index.

Second, my laptop system does not support the crawler program, which is an issue that needs to be diagnosed before indexing data.

Third, Java script issues you may face through the process; Google does index JavaScript and AJAX. Nevertheless, these languages are not as quickly indexable as HTML. Therefore, if you are incorrectly configuring your AJAX pages and JavaScript execution, Google will not index the page.

***Fourth*, Connectivity or host issues:**

Google's spiders could not access the server when I tried to crawl; all that I got was that my host was doing maintenance. Therefore, no data will be indexed if the crawler cannot reach the site. Connectivity is a severe issue for indexing data; the site is not crawled whenever the host has an outage.

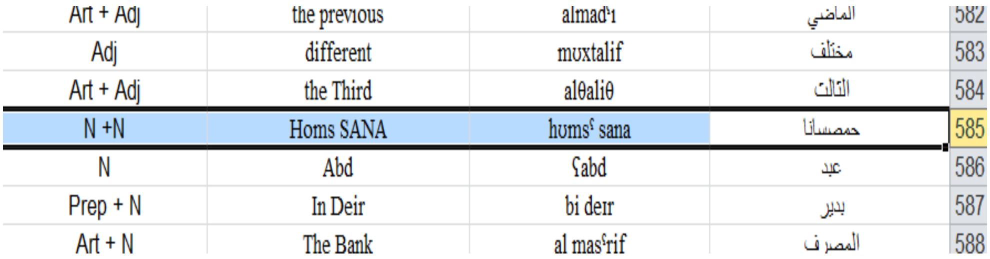
B. After crawling the data:

After I mined some data (crawled from http://www.w3newspapers.com/syria/- Al-Thawra Newspaper/a Syrian newspaper), we got them indexed on a different laptop; I started working on my dictionary, which consists of 2000 words. The dictionary is divided into the following columns:

1. Arabic words
2. Transcription
3. Translation
4. Parts of speech (POS).

We have been faced innumerable challenges while working on the crawled dictionary of the Arabic language; these challenges are:

First, many words got agglutinated in one cell, where you have to separate them manually. Table 1 shows cell number 585, where two words (Homssana, Homs Sana) were combined in one word.



Art + Adj	the previous	almadˤɪ	الماضي	582
Adj	different	mʊxtalif	مختلف	583
Art + Adj	the Third	alθaliθ	الثالث	584
N +N	Homs SANA	hʊmsˤ sana	حمصسانا	585
N	Abd	ʕabd	عبد	586
Prep + N	In Deir	bi deɪr	بدير	587
Art + N	The Bank	al masˤrif	المصر ف	588

Table. 1 The crawled dictionary of a Syrian newspaper highlighting the cell where words got agglutinated

Second, diacritics/ħarakāt (literally, motions; functionally, marks) or taškīl (forming) are one of, yet optional in representation, the primary aesthetical glyph features of the Arabic language help to indicate parts of speech and preclude the ambiguity resulting from their omission. It is challenging to find a digital text with its diacritics since, in Arabic, it is ubiquitous to drop them in publication unless in ambiguous situations. Computing the correct diacritics in a written sentence where diacritics are missing is also a challenge. Recently, this field has grabbed the attention of researchers to develop a specific algorithm for automatizing the process of diacritization for Modern Standard Arabic (MSA) corpus (Almanea, 2021). The inflectional and derivational structure of the Arabic language made it full of twists and turns for crawling data, on the one hand. On the other hand, the absence of capitalization in Arabic is another limitation tagged in proper noun recognition. Computing the correct diacritics in a written sentence where diacritics are missing is also a challenge.

For example, the following word could mean

وضَعَ /wadˤaʕa/ →*put/* PRES PROG (PP) with two fatħas on the root of the verb /wdˤʕ /, one is placed on /dˤ/ represents a short /a/, while the other is placed on /ʕ/ to create the same representation

While وضْعْ → /wadˤʕ/ with sukün (a tiny circle-shaped marker, placed on a letter to indicate that the placed-on consonant is not to be followed by a vowel) on /dˤ/ and /ʕ/ means *situation*

*More crawled examples with diacritics ambiguity :

1. / tadmur/ → تَدْمُرْ

A name of a city in Syria/ the Latin alternative is *Palmyra* .

→ While / tudammɪr/تُدَمِّر

Is the progressive form of the verb *destroy*

2. /xatˤtˤatˤə/ → he planned PT خَطَّطَ

/xitˤətˤ/ → plans خِطَط و

3. /dˤaharə/→ the PP of the verb *appear* ظَهَرَ

/dˤʊhʊr/ → *Afternoon* ظُهْر

4. / baħaθə/ → PP form of *search* بَحَثَ

/baħəθ/ → N/ *search* بَحْث

5. / ʕammar/→ عَمَّار

A proper noun *Ammar*

→ / ʕəmār / , N *construction* عَمَار

Third, Lexical ambiguity can cause negative impacts and unreliable feeds in the information field, which would fail most of the systems to perform adequately. Most of the search engines fell flat in addressing the limitations emerging from the linguistic peculiarities of the Arabic orthographic system. No proper engine is trained to interpret and disambiguate the doubts caused by the system. Subsequently, lexical ambiguity was one of the main problems in crawling and compiling my dictionary, in which you have to read the complete text to understand the exact suitable word, as in :

1. / saʕa/ ساعة

Might mean a *watch*, a *clock*, or an *hour*; it depends on the context .

2. / ʕʊquːd/ عقود

Could refer to *contracts*, or could denote the word *decades*.

3. /alʕanāsˤir/ العناصر

It could refer to (*the members*), or to the word (*the ingredients*).

4. /alθəwra/ الثورة

Would be either (*the Revolution*), the popular newspaper issuing in Syria (Al-Thawra) , or even to a street's name .

5. / alqadəm/ القدم

A local block in Damascus city, or the *foot* of the body.

6. / alkarāma/ الكرامة

A local *football team* in Syria, or *dignity* .

Fourth, Arabic is a very synthetic language; prefixes and suffixes are added to make the word incorporate subject, direct and indirect objects, their plurality, etc. According to Edward Sapir's language studies, this feature of the Arabic language makes it close to Roman-European languages, such as Latin or French. So the main problem while piling up a dictionary of a synthetic language is having the word indexed with its affixes without stripping out its morphological inflections. The indexed data in my dictionary have many inflected words, either with enclitic pronouns, a definite article, or prepositions. For instance :

1. /liquuwatına/ لقواتنا

/li/ → for (Prep)

/quuwa/ → army/ Force (N)

/t/ → plural marker

/ na/ → our (enclitic pronoun)

2. /wa taʕrifahunnə/ وتعريفهن

/wa/ → and (conj)

/taʕrif/ → defining / informing (V)

/hunnə/ → them/ Feminine (Enclitic Pro)

3. /juwasˤilün/ يواصلون

/ju/ a letter refers to the present progressive

/wasˤil/ → keep doing (V)

/ün/ they/ Masculine (Enclitic Pro)

5. Logistic Expressions of the Newspapers

One of the main challenges to present in this paper is the authoritarian logistics that monitor and ensure appropriate linguistic content imposed over the usage of specific terminologies by the (government-controlled) newspapers or mass media in a war-torn, politically divided country. Having a chaotic ground in Syria, dilapidated in sectarianism, suppression, conflicts, repression, a conspicuous unequal social hierarchy, and armed opposition, all these factors are echoed in the repertoire of the newspaper's discourse. Parallelly, out of the growing tension between the inside, the Syrian government, and outside, the opposition and the expatriates, many projects have been launched to establish new outlets for the people to have an accessible open forum abroad. Practically, unavoidable collateral damage will be consequently followed by language use, hate speech, dispute, and political rifts and preferences.

Given these points, this paper presents the differences between the local pro-government newspapers issued in Syria, and the expatriate journals, anti-government outlets, published out of the country. Both used the same political terms to refer to different concepts and situations or additional terms to refer to the same concepts as two sides of the same coin.

.

*** Limitations of the work:**

It is inevitable to get disappointed when you do not get what you seek, and it is occasionally contradictory to what it had been proposed. The major dissatisfaction in my work was reflected in the negative responses to not being granted a permit to access all of the Syrian newspapers issued out of the country, the opposition news media, they did not authorize me to index the data. Instead, I manually mined some selected terminologies commonly used by other local newspapers. Concomitantly, the potential problem relating to the current analysis is to obtain a permit, since accessing the data of the opposition's newspapers was unauthorized and censored, instead, it was proposed as a substitute to redirect the attention to crawl data from the *Arab Forums and the Israeli newspapers.* Having briefly considered some terms used by both sides as a case of study; how they would juxtapose, per se, the leading terms towards the Palestinian Intifada (literally, shaking off) as an *Uprising* or *Revolution*.

The events of the Palestinian-Israeli conflict generate a continuous stream of material, as well as an insatiable demand for stories, some examples that arose from this conflict are to be clarified hence:

The data were crawled from an article on (Zionist Terminologies Mistakenly Used by Arabs, 2010).

1. **The Middle East**: is a concept used by the Israeli press to justify their occupation of the Palestinian land; the Palestinian press media, on the other hand, refuses to use this concept and prefers to use "the Arabic East."
2. **Israel vs. the Jewish Entity**:

Using the term "Israel" means that you are formally declaring the establishment of the state of Israel.

3. **Submitting vs. Normalization:**

Normalization, as an expression, is used to melt and marginalize the Arab identity and accept the Zionist Entity as a legal one.

4. **The Promised Land vs. The Palestinian Land**.

The Promised Land, ardhul mi‘ād, expresses a compellable surrender to the notion that the Zionist's descendants to be given the territory from the Euphrates to the River of Eygpt. The land was promised and subsequently passed by God to Abraham to inherit the right to re-establish their "national homeland."

5. **Al-Buraq Wall vs. The Wailing Wall.**

The narrow sense of using the Wailing wall over Al- Buraq Wall is considered limestone for Muslims worldwide. The Palestinian press refuses to use the wailing wall. It is relatively considered a derogatory term (Halkin,2001), which refers to the Jews' practice of weeping next to the wall.

6. **The Settlers vs. The Migrants:**

The idea of Israeli migration was not accepted by the Palestinian press as though it is preferable to call them the settlers instead of migrants.

7. **Resistance vs. Terror and Violence:**

The Israeli press describes the Palestinian confrontation as an act of terror and aggression against them, while it is merely a non-violent resistance.

8. **Captive vs. detainee:**

The Israeli press labels the Palestinian captives as detainees to blot out the war expressions, and they would treat the captives inhumanely as criminals while kept in prison.

9. **Rights vs. Demands:**

What is a right to the Palestinians becomes a demand for the Israelis. So, the settlements issue became a right for the Jewish people. In contrast, as a legal principle of the Palestinian refugees, the Palestinian right of return turns out to be a demand.

10. **Usurpers vs. Habitants:**

The word in*habitant* is used to legitimate the Jews as the authorized owners of the occupied Palestinian land; from Palestinian perspective, they are authentically the *usurper* of history, land, and rights.

11. **Zionist Entity vs State of Israel:**

The Arab Media would instead utilise the Zionist Entity to legitimate and recognise Israel as apartheid than a state. Besides, "Israel was no state at all, but an illegal colonialist excrescence." (Sundquist, Eric J., 2005). 'Entity' is a derogatory term for the Israelis; They would always refer to Israel as a nation and state.

6. Conclusion

Web Crawler is a very open program designed to view websites that have been registered as new or updated by their owners. Entire websites or selected pages can be accessed and indexed selectively. Crawlers are so-called because they crawl through a website one page at a time, following links to many other sections of the website until enough pages have been viewed. However, the server confronts various challenges that must be solved by the crawlers, including the fact that not all sites enable indexing of their data. Besides, the crawler has to adapt to the different languages of the available sites.

==

Appendix:

List of Abbreviations:

URL: Uniform Resource Locator

AJAX: Asynchronous JavaScript And XML

HTML: Hypertext Markup Language

POS: parts of speech

MSA: Modern Standard Arabic

PP: Present Progressive

PRES PROG: Present Progressive

N: Noun

PT: Past Tense

Prep: Preposition

Conj: Conjuction

V: Verb

Pro: Pronoun

===

References

1. Almanea, M. (2021). "Automatic Methods and Neural Networks in Arabic Texts Diacritization: A Comprehensive Survey". IEEE Access. 9: 145012–145032. doi:10.1109/ACCESS.2021.3122977. ISSN 2169-3536. S2CID 240011970.

2. Bal, S. (2012). The Issues and Challenges with the Web Crawlers. International Journal of Information Technology & Systems. 1. 1-10.

3. Cho, J., Garcia-Molina, H.,(1997). L.*Efficient Crawling Through URL Ordering, Computer Science Department*. Stanford University, Stanford, CA, USA.

4. Cho, J. and Garcia-Molina, H. (2000). *The evolution of the web and implications for an incremental crawler. In proceedings of the Twenty-sixth International Conference on Very Large Databases*. Cairo, Egypt.

5. Halkin, H. *(2001).* ""Western Wall" or "Wailing Wall"?". Forward. Retrieved September 28, 2015.

6. *Al-Thawra Newspaper* http://www.w3newspapers.com/syria/

7. Spetka, S.(2004) *"The TkWWW Robot: Beyond Browsing"*. NCSA. Archived from the original on 3 September 2004. Retrieved 21 November 2010.

8. Sundquist, Eric J. *(2005). Strangers in the Land: Blacks, Jews, post-Holocaust America. Harvard University Press.* ISBN 978-0-674-01942-3.

9. *Zionist Terminologies Mistakenly Used by Arabs* (2010). Islam Way Magazine. http://ar.islamway.net/article/6027/

==