Arabic Citation Parsing using Part of Speech and Named Entity Recognition

Youssef Karout and Hadi Hamoud and Fadi A. Zaraket
Arab Center for Research and Policy Studies, Doha
fadi.zaraket@dohainstitute.edu.qa

Abstract

This paper introduces an industry level citation element extractor from Arabic text. Citation element extraction enables editorial task automation for publishing houses, creation of citation networks, and automatic citation analytics for impact analysis firms. Citation library tools help users manage their citations. However, for Arabic, these tools lack basic support to identify and extract citation elements. Consequently, researchers, editors and reviewers manually manage Arabic citations tasks. We present a novel Arabic citation element dataset, use it to train a citation element extraction model, and use named entity recognition, morphological analysis, and keyword detection to improve the results for practical use. The paper reports industry ready performance with F1 scores ranging between .80 and .95 for interesting citation elements.

1 Introduction

Citation element extraction (CEE) considers an input text t and identifies citation elements such as author, title, and publisher in t. Then, it decides whether t is a citation. Figure 1 illustrates the extraction of author, title, location and other elements from an Arabic citation text.

Tools such as AnyStyle-Parser (Keil, b) and ParsCit (Kan), perform well with CEE for English. Digital citation tools and libraries such as Zotero (Zot) and Mendeley (men) help researchers manage and retrieve references. Yet, they lack adequate Arabic support.

This paper concerns the development of ACEX, an Arabic Citation Element eXtraction tool, to help researchers, editors, and reviewers at Arabic publishing houses automate the laborious and error prone tasks of generating, editing, and checking citation quality and styles. A key use case is to automatically detect and extract citations from anywhere within a document. ACEX retrieves the relevant citation elements and restyles them as needed,

جورج أنطونيوس، يقظة العرب: تاريخ حركة العرب القومية، ترجمة ناصر الدين الأسد وإحسان عباس (بيروت: دار العلم للملايين، ١٩٧٨)، ص ٤٣٦-٤٣٢.
ğwrğ ạnṭwnyws, yqẓh ālʿrb: tārỵẖ ḥrkh ālʿrb ālqwmyt, trğmh nāṣr āldyn ālʿasd wʿiḥsān ʿbās (byrwt: dār ālʿlm llmlāyyn, 1978), ṣ 432-436.
(George Antonious, "Awakening of Arabs: the history of the national Arabic Movement", Translated by Nasser Elddine Alasad and Ihsan Abbar (Beirut: Dar Al Ilm Lil Malayeen, 1978), pp. 432-436)

Author:	جورج أنطونيوس	Location:	بيروت	Date: 1978	Type: Book
Title:	يقظة العرب: تاريخ حركة العرب القومية			Pages: 436-432	
Translator:	ناصر الدين الأسد وإحسان عباس			Publisher:	دار العلم للملايين

Figure 1: Arabic citation and its citation elements.

reducing the time and effort required by editors who would otherwise have to manually scan, verify, and reformat each citation. This automation improves efficiency and ensures consistency and accuracy in citation processing. Moreover, ACEX proves valuable in cases where a document lacks a reference section as it efficiently scans the document and its footnotes, identify citations therein, and compiles a properly styled reference section. Another important application of the tool is its ability to support the construction of citation networks, which are essential for conducting impact analysis of Arabic research. This is especially valuable given that only a few emerging institutions in the Arab world are currently addressing this challenge and they lack automated parsers (arc; alm) .

Online citations increased significantly across the last decades. At least 114 million English-language scholarly documents are available online (Tkaczyk et al., 2018). Arabic is widely used in scholarly work across all 22 Arab countries with 138,283 publications reported between 2015 and 2020 (El-Ouahi, 2023). However, Arabic scholarly work is less cited in international (Western) publications and indexing efforts. Impact and quality control automation are partly to blame as they are not Arabic mature yet (Al-Shorbaji, 2022).

Citations come in different styles, such as the American Psychological Association (APA), Chicago, and Modern Language Association (MLA) styles. Different styles require different inference rules and methods for processing citation

text.

Arabic has a rich morphology, with multiple forms (up to four) per character depending on its position in the word, and is written most often with omitted diacritics (short vowels). These, among other language specifics, require special preprocessing steps such as normalization, stemming, and lemmatization when addressing Arabic natural language processing (NLP) tasks such as information extraction.

This work considers AnyStyle (Keil, b), an open source machine learning (ML) approach for parsing citation elements. We directly apply it to Arabic citations as a baseline. We then make the following contributions.

- AnyStylePre, this model takes an Arabic citation, pre-processes it, and passes it to AnyStyle without retraining the base model.
- We build the AraCiteD dataset by manually annotating 867 Arabic citations for CEE. AraCiteD will be available online for the research community.
- We present ArAnyStyle, a CEE model trained on the original open source AnyStyle dataset augmented with AraCiteD.
- We build ACEX that takes the output of ArAnyStyle, elements extracted using regular expression techniques, named entity (Jarrar et al., 2022) and part of speech (POS) tags (Obeid et al., 2020) and returns improved CEE results. We apply post-processing techniques to handle unlabeled elements of the citation, referring to them in the sequel as "leftovers."

The integration of multiple tools proved both effective and complementary. By leveraging the strengths of each approach to offset the limitations of others, ACEX achieved consistently high F1 scores, ranging from 0.80 to 0.95, across key citation elements.

The rest of this paper is organized as follows. Section 2 provides definitions and review of citation elements. Section 3 presents related work. We then illustrate how we built AraCiteD in Section 4, followed by ACEX methodology in Section 5. We introduce our results, compare and discuss all approaches in Section 6 and finally conclude.

2 Background

Citation elements (CE) are categorized into main, keyword, and additional citation element types. Author, title, publisher, location, and date are main CE types. These tend to be specified in almost all citations. Keyword citation elements may be less specified and include page, volume, edition, editor, and translator. These are typically preceded by a keyword or an abbreviation of the keyword. For instance, keyword صفحة *ṣfḥh*(page) or its abbreviation ص. *ṣ.* precedes page numbers. Additional elements include Document Online Identifier (DOI), and unified resource locator (URL).

Simple regular expression techniques directly apply to detect boundaries of well structured citation elements based on prior knowledge of style and structure. This technique assumes that some citations elements follow predictable patterns, allowing for the identification of key elements. By leveraging these structural cues, regular expressions can isolate elements such as pages, editions, locations, and publication years. These detected elements benefit data driven models later as they provide ample data for training.

Named Entity Recognition (NER) concerns identifying named entities in text such as person names, organizations, facilities, geo-political entities, locations, dates, events, cardinals and ordinals. They help identify relevant CE types. We use Wojood-NER which identifies 21 named entity types for Arabic with more than 85% precision and recall (Jarrar et al., 2022).

POS tagging specifies the role of a word in a sentence such as a noun, verb, adverb and particle. It is a sub-task of morphological analysis and disambiguation. They help identify noun phrases, word parts and segments that map to clues for some CE types. We use the CAMeL toolkit to compute POS tags (Obeid et al., 2020) for Arabic.

3 Related Work

A knowledge based approach to citation analysis attempts to unify heterogenous citation styles into one INFOMAP system (Day et al., 2005).

Efforts emerged lately to create online databases with augmented services for Arabic scholarly work including publications and citations. EMarefa (ema), Manduma (dar), and AlManhal (alm) are examples with such services. EMarefa launched the Arab Citation and Impact Factor (ARCIF) (Al-Shorbaji, 2022) to keep track of Arabic scholarly articles across disciplines. ARCIF issued a 2023 report (arc) with impact factors covering major Arab research venues. However, they lack

automated citation analysis tools and rely on manual work.

Evaluation of various methods including trained, expert based, regular expression bases, and rule based suggested that a combination of these methods is needed to tackle bibliographic analysis tasks (Tkaczyk et al., 2018). They further identify the heterogeneous nature of bibliographic data as a major challenge, noting that publishers and authors tend to present and structure their information differently, each adhering to distinct stylistic conventions.

GROBID (Lopez, 2009) is a conditional random field (CRF) system designed to extract meta information from scientific papers in PDF format, including bibliographic references.

Anystyle (Keil, b) is an open-source ML CEE model. It works as a citation parsing system designed to segment citations and label their segments with corresponding CE types. It relies on a sequence-labeling model trained on annotated citation strings and combines statistical learning with layout and token level features to support a wide range of citation styles and languages. It provides an interface to retrain the model with user data.

A descriptive analysis (El-Ouahi, 2023) highlights the importance of the Arabic Citation Index (ARCI) that started in 2020 by Clarivate and the Egyptian Knowledge Bank (EKB) covering 2015-2020. Deeper and more informative analysis requires automated CEE tools that allow for coverage of a huge body of legacy research. Citation and personal bibliography management tools like Mendeley (men) and Zotero (zot) lack proper support for Arabic citations. Researchers and scholars manually manage their citations with such tools. Providing CEE resources for Arabic empowers such platforms to improve Arab research productivity.

Scispace (SciSpace : Science in the Age of AI) emerged lately leveraging AI and LLM advances to support research writing and citation management. It offers automatic manuscript formats, integration with known journal templates, and insight extraction from research content. Similarly, ScholarAI (ScholarAI : AI Chat for Scientific Papers) leverages ChatGPT-4 abilities, provides a plugin and a specialized ChatGPT to interact with research content and citations. It performs reasonably well for Arabic CEE, yet it requires paid access fees with OpenAI, and also exposes IP and content for authors and publishing houses.

VOSViewer (VOSViewer :Tool for Bibliometric network) helps constructing and visualizing bibliometric networks, including citation, co-citation, and co-authorship relations. Arabic CEE tools are essential to enable use of such tools.

Table 1: AraCiteD citation elements

author	795	title	859	container	71
number	2	translator	132	location	646
publisher	630	date	837	url/website	39
genre	74	reviewe(d,r)	2	journal	128
volume	211	note	93	affiliation	4
collection	24	edition	64	editor	35
director	36	pages	3	newspaper	32

4 AraCiteD Dataset

Heterogenous citation styles present a generalization problem for CEE techniques. Our experience unveils larger irregularities when dealing with varying Arabic citation styles versus Western citations. Aside from few rigorous research oriented publishing houses, Arabic citation styles border on free style citations even within the same document sometimes.

Standard edited books provide more consistent citation styles, yet include diverse cited manuscript types, and vary across discipline. We considered the books published by a research publication house (omitted for blind review) that employs a version of the Chicago-17 style modified and adapted for Arabic. These books span scientific disciplines including social sciences, humanities, public administration, history, political sciences, psychology, philosophy and economy.

Two research assistants manually annotated the reference lists of 111 academic articles, comprising of 867 Arabic references (English references were excluded) with 4,714 individual reference elements. They were given the reference lists and they then tagged each CE using the corresponding CE type labels such as author, title, location, and date. The annotators received guidelines and training that included definitions of reference elements, and they worked with an expert editor on a small set of example references. The expert editor from the publishing house then reviewed the annotated references and their elements for compliance with the publisher's reference style and corrected errors, which occurred in fewer than 6% of the elements.

The resulting dataset will be provided to the research community following the AnyStyle JSON file formats. Table 1 illustrates counts of CE types

in AraCiteD.

5 ACEX Methodology

Given an Arabic text detailing a citation, ACEX preprocesses the text, runs ArAnyStyle, NER, POS tagging, and regular expressions over the normalized text, aggregates the results, and then performs post processing and handles the leftover items in the text. The following steps describe the process.

Step 1- Preprocessing: ACEX preprocesses the Arabic text and normalizes Arabic letters, white spaces, numerals, and punctuation marks. It replaces Arabic commas ، ,, period, parentheses, and quotations (“ ”) with their English counterparts. Other normalizations merges Arabic appearing characters from Urdu and Farsi as they appear in Arabic text due to font and keyboard variations. Normalization also covers different forms of a letter, e.g. alef and hamza have several representations أ إ ا ء ئ ؤ *ʾa ʾi ā ʾ ʾy ʾw*. Punctuation marks are richer in the Arabic unicode set, and several of them have similar forms.

Step 2- ArAnyStyle: is a version of AnyStyle trained AraCiteD on top of the original AnyStyle dataset (Keil, a). We used a pre-processed version of AraCiteD in training, and we feed ArAnyStyle a similarly pre-processed input. we performed the training with the interface provided by the AnyStyle framework, with no modifications to the training configuration.

Step 3- Arabic NER: named entity recognition ML model which identifies 21 named entities for Arabic (Jarrar et al., 2022). Wojood-NER plays a crucial role in disambiguating key bibliographic elements such as author names, locations, publishers, and dates. It helps address missing Arabic field types in AnyStyle by supplementing them with outputs from ArAnyStyle. Additionally, Wojood-NER’s accurate segmentation of organizations, events and person names significantly enhances citation extraction by distinguishing authors, individuals, groups, and organizations. This capability is essential for applications such as citation restyling, citation database construction, and citation network creation.

Wojood-NER was instrumental in enhancing CEs such as authors (persons or organizations appearing at the beginning of the citation), publication venues, locations, numbers, and dates.

Step 4 - Arabic POS: specifies the role of a word in a sentence such as a noun, verb, adverb, adjective and particle. It also specifies POS tags for prefixes and suffixes of the Arabic words. Verb and noun POS tags help identify verb and noun phrases, and adverbs help identify patterns for temporal and location entities.

Step 5 - Regular expressions: take the input text and split it into several tokens/segments where each segment may be part of a citation element. It splits the text using delimiters and other Arabic citation element specifiers and keywords that may play the same separation role, but are full Arabic words such as حرّره *ḥrrh* (edited it) or قدّمه *qdmh* (wrote a preface for it). These are important as often times Arabic writing omits punctuation and delimiter marks and uses them as reading scopes. This characteristic also applies to citation texts. Regular Expressions proved effective in detecting CE types such as the editor, translator, publisher, pages, and volumes, since these elements are typically linked to specific keywords that aid in their identification.

These steps result in a matrix of citation elements and features extracted using these various techniques as illustrated, which we then pass to aggregate results for additional processing. The Figure 2 provides a high level representation of the steps to perform the citation extraction.

There are several possible ways to aggregate the citation element matrix. ACEX makes decisions for each CE using a rule-based aggregation strategy. For all elements, priority is given to outputs from ArAnyStyle, the fine-tuned AnyStyle model. If a CE type is not detected with ArAnyStyle, NER, POS and REGEX outputs are used to infer the missing element.

Specifically, We use NER for authors and locations, as it performs well in identifying such named entities. We use regular expressions for publisher, editor, translator, date, and volume, since these elements rely on local grammars and patterns with specific keywords.

After the initial decision round, some entities remain unlabeled, we refer to them as “leftovers". We introduce a post-processing approach that compares the embeddings of the leftovers with those of known citation elements from the AraCiteD, identifying potential matches. Leftovers are thus classified using majority voting among the k nearest neighbors (from AraCiteD) for each leftover segment. If the neighbors’ votes indicate, with sufficient confidence, that a segment belongs to a

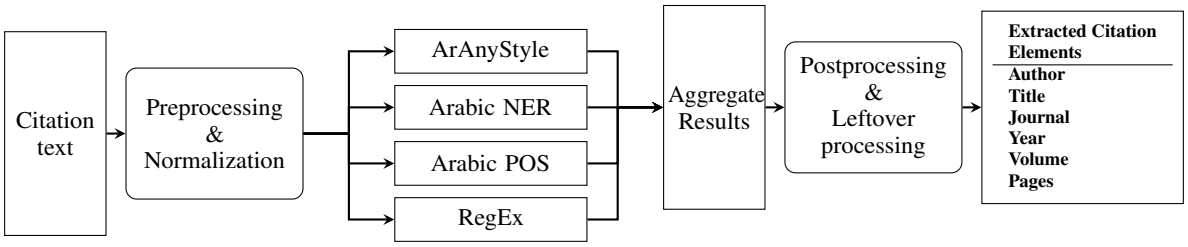

Figure 2: Flow diagram for ACEX

specific citation type that was not originally identified, we append the segment to that type. Implementation tests show that this method is particularly effective at extracting publisher, editor, and translator citation elements. These elements exhibit well-defined and relatively stable semantic patterns, typically consisting of a small set of indicative keywords followed by a human or organizational name.

Owing to this low structural complexity, reliable extraction can be achieved without requiring large amounts of training data.

6 Results

In this section, we discuss the ArAnyStyle and ACEX results. We also discuss details from experiments related to different approaches discussed in Section 5.

6.1 Performance Evaluation Metrics

To assess performance, we compute recall, precision, and F1 score by measuring the exact match between predicted citation elements and the ground truth from the testing dataset. It is important to note that the testing dataset is not used during the training phase of any model, ensuring an unbiased evaluation.

6.2 Performance Evaluation of Anystyle Parser Configurations

We evaluated the performance of the Anystyle parser using three different configurations :

1. Default AnyStyle Model.
2. AnyStylePre: Anystyle with preprocessing of the citations and with punctuation normalization.
3. ArAnyStyle: AnyStyle trained on a preprocessed version of AraCiteD.

6.3 Performance Evaluation

Table 2 shows the results of our evaluation on a test set that consists of 59 citations (319 different citations elements) not seen during the training phase. The evaluation is done across 8 main citation element types. It should be noted that, at this stage, the evaluation is conducted on texts composed exclusively of citations written in different styles. This setting allows us to assess the performance of ACEX without interference from auxiliary textual content. In the following sections, a use case is presented in which ACEX is applied to full texts containing both citations and non-citation content.

We summarize the results as follows.

1. AnyStyle: we passed the citation without any processing and punctuation normalization, and used AnyStyle default model. This model showed poor performance, and it fails to capture certain citation elements. The model performed poorly on the elements Publisher, Editor, and Translator. This is understandable, as these elements have semantics that the model had never encountered in Arabic, making them challenging to detect.
2. AnyStylePre: same as AnyStyle, but we applied punctuation normalization before passing the citation to the default model. It showed slight improvement in the recall for the date element, but it still fails to capture other elements such as editor and translator.
3. ArAnyStyle: AnyStyle trained on AraCiteD with punctuation normalization improved consistently across most citation elements. While precision decreased for Location, recall increased. AnyStyle and AnyStylePre performed well for Authors and Titles, likely because these elements exhibit consistent positional features, suggesting they are less complex. Nevertheless, fine-tuning still resulted in measurable gains.
4. ACEX clearly outperforms all approaches as it improves the decisions of ArAnyStyle. This demonstrates the practicality of hybrid models that leverage both data driven approaches and domain expertise. Furthermore, comparing

Table 2: Citation extraction performance across citation elements

	AnyStyle			AnystylePre			ArAnyStyle			ACEX		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Author	.87	.98	.92	.87	.98	.92	.92	.99	.95	.92	.99	.95
Title	.78	.99	.87	.78	.99	.87	.91	.99	.95	.91	.99	.95
Location	1.00	.68	.81	1.00	0.68	0.81	.9	.82	.86	.87	.98	.92
Publisher	0	0	0	0	0	0	.85	.85	.85	.93	.96	.95
Date	.94	.72	.81	.78	.88	.83	.92	.97	.94	.91	.98	.94
Editor	0	0	0	0	0	0	.78	.39	.52	.65	1	.79
Translator	0	0	0	0	0	0	.86	.89	.87	.86	.89	.87
Volume	.37	.83	.51	.35	.82	.49	.72	.87	.79	.72	.96	.82

the ArAnyStyle and ACEX columns reveals the improvements gained by integrating POS and NER into the system. For example, we note performance gains in the Location elements due to Wojood's contribution and in Publisher and Editor elements thanks to the leftover processing. It is worth noting that future work could enhance the extraction process, particularly the translator component, by incorporating additional data and contributing to more accurate leftover item classification and processing. An interesting improvement can be done by incorporating morphological analysis (Obeid et al., 2020; Abdelali et al., 2016; Darwish et al., 2014)

6.4 Use-case 1: Deployement of ACEX for automatic citation detection and database creation

As previously discussed, a key application of this tool is the automatic detection and re-styling of citations within the reference sections of unedited documents. To accomplish this, the algorithm systematically scans each paragraph of the document, applying the proposed model to identify and extract potential citation elements. For each candidate segment, a confidence score is computed based on the presence and consistency of these elements, it allows to determine whether the text represents a valid citation or not. To further refine detection—particularly since the focus is on reference sections—a density-based filter is applied to remove outliers that are distant from citation-dense regions, improving precision and contextual relevance.

This method was applied to six different books, successfully extracting all citations from their respective reference sections. Furthermore, it enabled the creation of several supporting databases—such as author, publisher, translator, and editor databases, as well as a structured citation element database—which contribute to the ongoing enhancement of the model.

To evaluate the quality of the extracted citation elements, a random sample of 100 citations (comprising 354 individual citation elements) was manually annotated to serve as a ground truth dataset. Using this annotated data, standard evaluation metrics such as precision, recall, and F1-score were computed. The results, presented in Table 3, demonstrate the high accuracy of the model and its strong potential for various downstream applications.

Building on this case study, we extended our evaluation to assess the benefits of automatic citation restyling. In this follow-up, we conducted a study in which 100 citations were automatically reformatted according to the publishing house's style using the proposed solution. The restyled references were presented as tracked changes in a Word document, allowing the editor to accept, reject, or further modify each suggestion. We then asked the editor to restyle the reference section without using these proposed modifications. The study demonstrated a 86% reduction in editing time (approximately 2 minutes using ACEX vs approximately 15 minutes with the traditional approach), highlighting the efficiency gained through automated citation restyling.

6.5 Use-case 2: Deployment of ACEX for automatic footnote processing

Another important application of this tool is the automated processing of footnotes. It enables scanning the document to identify citations within footnotes, restyle them according to the required format, verify their accuracy if necessary, and generate

Table 3: Use-case 1: Automatic citation database creation performance

	ACEX		
	P	R	F1
Author	1	.99	.99
Title	1	.99	.99
Location	.94	.87	.90
Publisher	1	.92	.95
Editor	1	1	1
Translator	1	1	1

a bibliography section that meets the publisher's guidelines. Performing this task manually is highly time-consuming, as the editor must read through all the footnotes, extract multiple citations, reformat them, and then compile the bibliography. The model performed satisfactory for automation on this application reducing required manual time to extract the citations and their elements. It also successfully resolved confusing cases that usually take significant expert time where the footnote contained multiple citations with other textual elements as shown in the example in Figure 3.

While advanced LLM models such as GPT or Gemini can produce strong results, they come with certain drawbacks: subscription costs, potential privacy concerns imposed by some publishers, and occasional failures when handling complex Arabic structures or author styles. We also observed performance drops for long titles, which are sometimes truncated, and for complex volume and issue combinations, which can be misparsed. Additional errors may occur when author names resemble those of well-known figures or when titles are similar to each other.

7 Conclusion

ACEX offers a robust method for citation extraction, fusing the results of AnyStyle, NER, POS tagging, Regular expressions and complenting them with a heuristic to decide on leftover elements.

This combined approach demonstrates high recall and precision. Overall, the proposed Approach, proves to be effective for extracting Arabic citation elements making it well-suited for deployment in real-world applications to automate editing tasks as well as citation network extraction tasks. In practice, ACEX is also complemented with a large library of preprocessed citations with predefined

سامية بيبرس، مسيرة التعاون العربي الأفريقي: رؤية عربية، آفاق أفريقية، السنة ٩، العدد ٣٢ (٢٠١٠) ص ٨٢. مما يجدر ذكره هنا، أن المشاركين في ملتقى التعاون العربي - الأفريقي، توصلوا في ختام اجتماع لهم في الشارقة، في ٩ كانون الأول ١٩٩٧، إلى إقامة منطقة تجارية تفضيلية عربية أفريقية. لكن لم يُترجم هذا الأمر إلى واقع ملموس. انظر: عبد السلام إبراهيم بغدادي، الجماعات العربية في أفريقيا: دراسة في أوضاع الجاليات والأقليات العربية في أفريقيا، جنوب الصحراء (بيروت: مركز دراسات الوحدة العربية، ٢٠٠٥)، ص ٧٤٨؛ والجمهورية (بغداد)، 7-16/12/1997.

sāmyh bybrs, msyrh ālt'āwn al'rby al'afryqy: rwyh 'rbyt, 'āfāq 'afryqyt, ālsnh 9, al'dd 32 (2010) ṣ 82. mmā ygdr ḏkrh hnā, 'an ālmšārkyn fy mltqā ālt'āwn al'rby - āl'afryqy, twṣlwā fy ḫtām ā̄ğtmā' lhm fy ālšārqt, fy 9 kānwn āl'awl 1997 'ilā 'iqāmh mnṭqh tğāryh tfḍylyh 'rbyh 'afryqyt. lkn lm yutrğm hḏā āl'amr 'ilā wāq'mlmws. ānẓr: 'bd ālslām 'ibrāhym bğdādy, ālğmā'āt al'rbyh fy 'afryqyā: drāsh fy 'awḍā' ālğālyāt wāl'aqlyāt al'rbyh fy 'afryqyā, ğnwb ālṣḥrā' (byrwt: mrkz drāsāt alwḥdh al'rbyt, 2005), ṣ 748, wālğmhwryh (bğdād), 7-16/12/1997.

Samia Bibers, Arab African Collaboration Path: Arab Vision, African Horizons, year 9, number 32, (2010), p. 82. It is worthy here to note that the participants in the Arab African Collaboration Convention achieved at the closure of oenof their meeting in Sharjah in 9 December 1997 to establish an Arab African commercial preference zone. But that never became a concrete reality. Look: Abd Alsalam Ibrahim Baghdadi, Arabic Communities in Africa: A Study in the Status of the Arab Communities and Minorities in Africa, South of the Desert, (Beirut: Arab Unity Studies Center, 2005), p. 748, and AlJumhouria (Baghdad) 7 and 16/12/1997.

Figure 3: Arabic footnote (transliterated using ArabTeX). citation 1 in blue, citation 2 in red.

styled citation text from the specific publishing house for the editing tasks. For citation network analysis, ACEX excels at detecting the important citation element types such as author(s), title, collection publisher, location and dates. In future work, we will focus on refining ArAnyStyle with additional data so that it can handle the leftover items directly with less need for the second pass.

Limitations

Limitations are mainly in collecting and annotating data, as well as finding already existing tools for annotating and parsing Arabic citations and bibliographic references. The data collection part is both labor and time consuming. The limited resources of Arabic annotated datasets restricts the performance of our models and approaches, raising the need for customization and specification.

In addition, the complexity of the Arabic language and the variations of citation styles complicates the extraction process.

To address these issues, we need to continue developing more sophisticated models, enhance annotation tools, and expand the availability of high-quality annotated Arabic citation datasets.

Ethics Statement

The data was collected and used with the appropriate approvals of the intellectual property owners. All results are reported following best academic standards and practices.

References

AlManhal: Authoritative arabic scholarly content. https://www.almanhal.com/. Last accessed: 2024-05-17.

E-Marefa: The leading complete arab online databases. https://emarefa.net. Last accessed: 2024-05-17.

Mandumah: The pioneers for arabic datasets. https://www.mandumah.com/. Last accessed: 2024-05-17.

Mendeley: Reference management software. https://www.mendeley.com. Last accessed: 2024-05-15.

Zoteri: Your personal research assistant. https://www.zotero.org/. Last accessed: 2024-05-17.

Zotero: Your personal research assistant. https://www.zotero.org. Last accessed: 2024-05-15.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Najeeb Al-Shorbaji. 2022. Measuring knowledge production in arabic using arcif: Statistical indicators and impact factor. In *Higher Education in the Arab World: Research and Development*, pages 113–140. Springer.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using stem-templates to improve Arabic POS and gender/number tagging. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2926–2931, Reykjavik, Iceland. European Language Resources Association (ELRA).

Min-Yuh Day, Richard Tzong-Han Tsai, Cheng-Lung Sung, Cheng-Wei Lee, Shih-Hung Wu, Chorng-Shyong Ong, and Wen-Lian Hsu. 2005. *A knowledge-based approach to citation extraction*.

Jammal El-Ouahi. 2023. The arabic citation index – toward a better understanding of arab scientific literature. *Quantitative Science Studies*, 4(3):728–755.

Mostafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. Marseille, France.

Min-Yen Kan. PatsCIT: An open-source crf reference string parsing package. https://github.com/knmnyn/ParsCit. Last accessed: 2024-05-15.

Sylvester Keil. a. ANYSTYLE dataset. https://github.com/inukshuk/anystyle/blob/main/res/parser/core.xml. Last accessed: 2024-05-15.

Sylvester Keil. b. ANYSTYLE: Fast citation reference parsing. https://github.com/inukshuk/anystyle. Last accessed: 2024-05-15.

Patrice Lopez. 2009. Grobid: combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

ScholarAI : AI Chat for Scientific Papers. https://typeset.io/. Last accessed: 2024-05-17.

SciSpace : Science in the Age of AI. https://scholarai.io/. Last accessed: 2024-05-17.

Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers.

VOSViewer :Tool for Bibliometric network. Last accessed: 2024-05-17.