Enhancing State-of-the-Art NLP Models for Classical Arabic

Tariq Yousef **Lisa Mischer** **Hamid Reza Hakimi** **Maxim Romanov**

"The Evolution of Islamic Societies (c. 600-1600): From Algorithmic Analysis into Social History"
Emmy Noether Junior Research Group, Hamburg University
{firstname.lastname}@uni-hamburg.de

Abstract

Like many other historical languages, Classical Arabic is hindered by the absence of adequate training datasets and accurate "off-the-shelf" models that can be readily used in processing pipelines. In this paper, we discuss our ongoing work to develop and train deep learning models specially designed to manage various tasks related to classical Arabic texts. We specifically concentrate on Named Entity Recognition, classification of person relationships, toponym classification, detection of onomastic section boundaries, onomastic element classification, as well as date recognition and classification. Our efforts aim to confront the difficulties tied to these tasks and to deliver effective solutions for analyzing classical Arabic texts. Though this work is still under development, the preliminary results presented in the paper suggest excellent to satisfactory performance of the fine-tuned models, successfully achieving the intended objectives for which they were trained.

1 Introduction

Arabic chronicles and biographical collections preserve a plethora of information on long-term environmental and societal processes that shaped and molded Islamic societies. Numerous and extensive, these written texts are the richest "mine" of information and are particularly valuable for the period before the 15th century, for which exceptionally few archival documents are available.

Our work focuses on constructing the social history of the Islamic world from historical and biographical texts that constitute a significant part of the Arabic written tradition. The project studies a vast corpus of digitized texts and relies on a series of computational methods for identifying and linking relevant information from the corpus. Currently, the project is at the stage of fine-tuning relevant NLP-based approaches. The work described in this paper is the methodological foundation for the creation of the network of knowledge. This network will serve as the main research framework of the project for the study of the social history of the Islamic world.

Classical Arabic, like all other historical and ancient languages, lacks adequate training datasets and accurate "off-the-shelf" models that can be directly employed in the processing pipelines. In light of this, our objective is to make a valuable contribution to the field by creating comprehensive training datasets for various tasks related to classical Arabic text processing and analysis. Furthermore, we aim to develop, train, and fine-tune models that can be easily integrated and shared with fellow researchers in the community, facilitating their work and promoting further advancements in the field of classical Arabic language processing.

In the following sections, we present our in-progress work in developing and training deep learning models tailored for handling diverse tasks relevant to classical Arabic texts. Specifically, we focus on NER, person relationships classification, toponyms classification, onomastic section boundaries detection, onomastic entities classification, as well as date recognition and classification.

2 Related Work

Recent advancements in deep learning and language modeling have significantly propelled the development of Natural Language Processing (NLP) models for the Arabic language. Several transformer-based language models are currently available and provide state-of-the-art performance in various downstream tasks.

ARABERT (Antoun et al., 2020) is the first transformer-based language model for the Arabic language. The CAMEL LAB (Obeid et al., 2020) introduced a collection of pre-trained models for several Arabic NLP tasks such as Part-of-speech (POS) tagging, named entity recognition (NER), sentiment analysis, and text classification.

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 160–169,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_019

Figure 1: An example illustrating a typical biography.

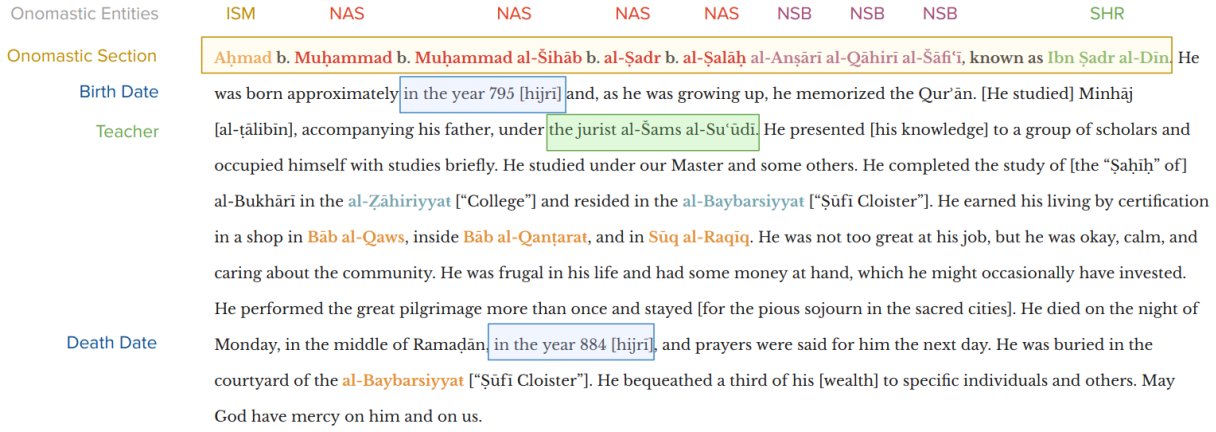

Figure 2: Translation of the example in Figure 1.

These models have been trained using different corpora, namely, classical Arabic CA, dialectal Arabic DA, modern standard Arabic MSA, and the MIXED corpus which comprises all available corpora. FARASA[1] (Abdelali et al., 2016) offers diverse solutions and models for Arabic text processing. It also provides a RESTful API, allowing users to access its functionalities and leverage language-independent solutions.

Further, several pre-trained models have been trained for different downstream tasks such as ARAT5 (Nagoudi et al., 2021) and ARAGPT2 (Antoun et al., 2021b) for Arabic language generation and understanding; ARAELECTRA (Antoun et al., 2021a), ARBERT, and MARBERT (Abdul-Mageed et al., 2021) for language representation.

The majority of the models mentioned in this context have been trained primarily on modern Arabic texts. However, their applicability to classical Arabic texts varies in terms of performance. Notably, the CAMELBERT-CA model[2] is the only model that is specifically trained on classical Arabic texts. It offers the highest initial performance, if compared to all the other models. Serving as a cornerstone for our research, this model formed the basis for our initial annotations and subsequent fine-tuning, allowing us to adapt it to our specific tasks and requirements.

3 Corpus

Texts utilized in our project are a sub-corpus of the OpenITI corpus (Nigst et al., 2023).[3] At the moment, our sub-corpus includes 101 multi-volume texts (c. 71 million tokens), which include approximately 495 thousand biographical records. Most of these texts—about 70 of them—come from the period of 1000–1600 CE and from all the major regions of the Islamic world, spanning from Spain (al-Andalus) and North Africa (al-Maġrib), to Egypt

[1] https://farasa.qcri.org/.

[2] CAMeL-Lab/bert-base-arabic-camelbert-ca

[3] https://github.com/openiti/, Open Islamicate Texts Initiative.

(Miṣr) and Syria (al-Šām), to Iraq (al-ʻIrāq), Iranian provinces (Fārs, Khurasān, etc.) and Central Asia (Mā-warā'-l-nahr).

Figure 1 illustrates a typical structure of biographies in our corpus (Figure 2 offers a translation for additional clarity). The onomastic section, which provides details about the biographee's name, genealogy, origins, as well as some social and religious background, is typically located at the beginning of the biography. The onomastic section may also mention members of the immediate and extended family. This section is usually followed by information about the biographee's education: with whom they studied, in which cities, and, sometimes, what specifically they studied. The section on teachers is often followed by a section on biographee's students, who are listed in a structurally similar manner. In some biographies, descriptions of the biographee's characteristics are given as well, either as the opinion of the main text's author or as opinions of other earlier biographers. In the middle, biographies often include other important facts from the life of biographees. Usually, concluding sections of biographies provide information on the date and place of biographee's death, and, occasionally, the location of biographee's burial.

4 Methodology

Manually created data plays a crucial role in the training process of machine learning models. Large and accurate training datasets are particularly important to the development of more precise models with improved performance. Historical and classical languages pose a unique challenge as there is often a lack of readily available training datasets. Creating such datasets requires domain experts with specialized knowledge to perform accurate data annotation. Given the limited resources available, we have decided to employ the active learning process (Wang and Hua, 2011) as a solution to generate training data for the various tasks we aim to tackle. Figure 3 illustrates the active learning paradigm we adapted as an efficient strategy to produce accurate training data to be used for model training. Figure 4 illustrates a lightweight annotation scheme that we developed to increase the easiness, speed, and accuracy of manual annotation.[4]

[4]Inspired by markdown, our scheme relies on short opening tags, where the end of the tagged entity is determined by a number indicating the number of tokens. For example, a tag `P3T` can be placed at the beginning of a 3-token entity indicating a person, who was a teacher of the biographee.

After preparing our sub-corpus, we began working with a random subset of biographies. Initially, we utilized CAMeL LAB models for lemmatization, Named Entity Recognition (NER), and Part-of-Speech (POS) tagging to perform the initial annotation steps. Also, in the initial stage, we employed our own rule-based models for date recognition and classification. Following this, we proceeded with the first round of manual refinement and correction conducted by domain experts, resulting in the successful correction of 1,100 biographies. To ensure consistency and accuracy, annotators performed cross-validation to ensure the correctness and consistency of the annotations. Using this refined dataset, we fine-tuned the NER model and trained a model specifically designed to detect boundaries of the onomastic section. Subsequently, we employed these models to annotate a subset of biographies and repeated the cycle of manual correction and fine-tuning. This iterative process will continue until we achieve a stable performance that meets our expectations, enabling the models to accurately perform the intended tasks.

5 NER for Classical Arabic

Named Entities Recognition aims to identify entities within the biographies and classify them into three main categories, namely, TOPONYM, PERSON, and MISC. This task can be viewed as a token classification task, where each token in the text is assigned a specific label. There is a notable distinction between modern Arabic names and traditional Arabic (Islamic) names. Modern names often follow a similar structure to Western names, including a given name and a surname or family name. Traditional Arabic names typically consist of up to six different elements, though not all of these elements have to be present in each case and they may appear in any order. Traditional Arabic names and their elements are explained in more detail in the section 6.2.

In the initial phase, we employed `CAMeL-BERT-CA-NER` model to create the first round of annotations. Then, annotators refined the automatic annotations and added the missing annotations according to our annotation scheme. The first manual correction round resulted in a training dataset comprising 1,100 sentences, including 3,244 persons, 692 toponyms, and 198 miscellaneous entities. Next, we used this dataset to fine-tune `CAMeLBERT-CA-NER` model and used the

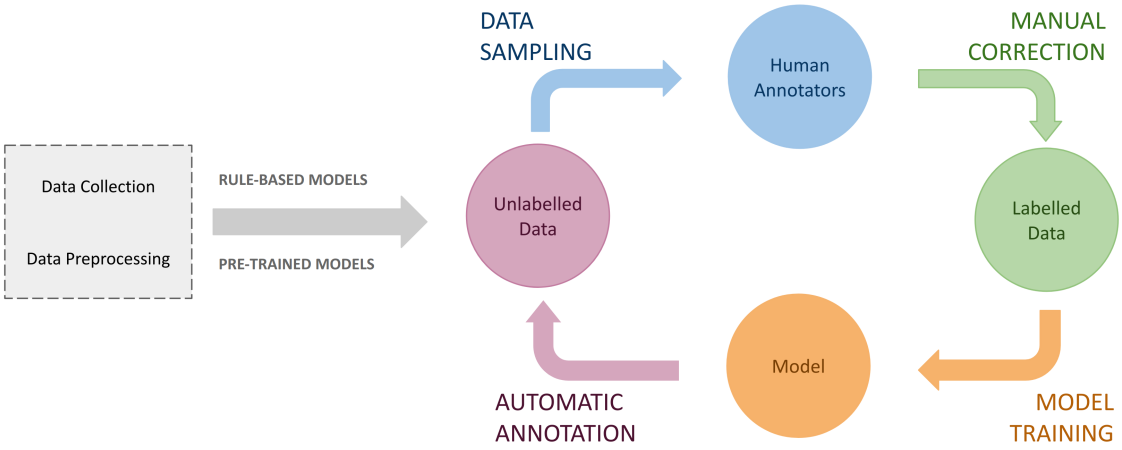

Figure 3: Development Process.

	TOPONYM		PERSON		MISC	
	zero-shot	fine-tuned	zero-shot	fine-tuned	zero-shot	fine-tuned
Precision	83.45%	92.95%	56.56%	94.29%	3.28%	71.21%
Recall	91.10%	95.94%	64.25%	95.35%	2.50%	74.21%
F1	87.11%	94.42%	60.16%	94.82%	2.84%	72.68%

Table 1: NER Model Performance.

trained model to automatically annotate a set of biographies. Subsequently, the next step resulted in a bigger training dataset 3,826 sentences containing 10,333 persons, 1,906 toponyms, and 612 miscellaneous entities. Table 1 provides a comparative analysis of the performance between the `CAMeLBERT-CA-NER` model (zero-shot) and the most recent fine-tuned model. Notably, there is a significant improvement in the identification of persons' names, with an increase of approximately 34.6% in the F1 score. This improvement can be attributed to the fact that the initial model is trained on modern Arabic person names which are structurally different if compared to traditional Arabic names.

Furthermore, our MISC entities did not overlap much with the MISC entities of the original `CAMeLBERT-CA-NER` model. Our fine-tuned model started to learn from our annotations, resulting in a promising performance with an F1 score of 72.68%.

5.1 Person Relationships Classification

The NER model achieved great performance in detecting persons in the biographies. Further, we wanted to determine the relationships between these detected persons and the biographee, a person in whose biography they appear. For this purpose, we defined six main classes (as shown in table 2). Detailed classification of the roles of individuals mentioned in biographies will allow us to model and generate complex networks, which will help the project to study the social organization of communities of Muslims across different periods and regions of the growing Islamic world. Table 2 illustrates the dataset utilized for training. The role classification information was manually added to person tags, which were generated automatically using the fine-tuned NER model.

Based on our reading of annotated biographies, we have singled out several consistently present classes of persons. Some of these classes can be described as unimportant, especially those that do not indicate any kind of direct contact to biographees.[5] We ended grouping them into the UNDEFINED class. Aiming for better performance and reliable classification, we trained two models with a reduced number of classes. The first model classifies the detected persons in the biography into three main classes TEACHER, STUDENT, and UNDEFINED. The second model classifies persons into four classes FAMILY, OPINION, CONTACT, and UNDEFINED. Then we merge the classes from both

[5]Lots of persons appear in so-called "chains of transmission" (Ar. *isnād*); most of these individuals never met the biographee and therefore are not part of that biographee's immediate social network.

Classes	# Entities
TEACHER (T)	2,891
STUDENT (S)	2,117
OPINION (O)	905
FAMILY (F)	603
CONTACT (C)	562
UNDEFINED (X)	3,372
Total	10,325

Table 2: Training Dataset for the *Persons Classification* Task.

models allowing persons to be associated with two classes simultaneously. For instance, a person may be classified as both STUDENT and FAMILY, in cases, for example, when a biographee is a student of his father.

Table 3 shows the classification results revealing that the TEACHER and STUDENT classes achieved good performance compared to other classes because the dataset contains a substantial number of entities belonging to these specific types. Currently, we are working to expand the training dataset by correcting and refining the automatic annotations created by the models. We believe that having a bigger training dataset would enhance the performance, especially for the labels of the second model.

5.2 Toponym Classification

In addition to the recognition of toponyms in the biographies, we are also interested in the relation between the biographee and the place mentioned. We defined six main classes, namely, places of: BIRTH, DEATH, BURIAL, KNOWLEDGE transfer, RESIDENCE, and UNDEFINED places. For the training of the preliminary model, we used a dataset of 1,047 biographies; for the subsequent fine-tuning of the preliminary model, we used 824 biographies. Table 4 illustrates the datasets utilized for the training.

This model was trained with two datasets. The first dataset was initially annotated with a rule-based model which classified toponyms based on certain keywords preceding them. This data, without any manual revisions, was then used to train our preliminary model. We then used this preliminary model to annotate the second dataset. We then manually corrected this second dataset, creating a revised set of training data.

The evaluation of the model is based on this second dataset that has been pre-annotated with the preliminary fine-tuned CAMELBERT-CA-based model and then manually corrected. We used the rule-based model as our baseline to compare the results of the fine-tuned CAMELBERT-CA-based model to. Table 5 shows the evaluation of the classification task results. For now, the achieved performance of both models—the rule-based one as well as the fine-tuned CAMELBERT-CA-based one—is not satisfying. The low results for the fine-tuned CAMELBERT-CA model are due to a lack of sufficient training data as our manually labeled dataset only contains 437 classified toponyms. Since we achieved high-accuracy results for all other token classification tasks with our fine-tuned CAMELBERT-CA models with bigger training datasets, we are currently expanding our training datasets.

6 Onomastic Entity Recognition for Classical Arabic

Each biography starts with a robust onomastic section on the biographee. The onomastic section is particularly valuable as it gives information on various backgrounds of the biographee. To identify the respective descriptors in the text, two different models are required. The first model identifies the boundaries of the onomastic section in the text of a biography. Applied to the identified onomastic section, the second model recognizes and classifies different onomastic elements.

6.1 Onomastic Section Boundaries Detection

We formulated this problem as a token classification task. In this approach, the model assigns labels to individual tokens within the given text. Specifically, we utilized three labels. 1) `B-ONOM` represents the beginning of the onomastic section, indicating that the token marks the start of the relevant section; 2) `I-ONOM` indicates that the token is inside the onomastic section, and 3) `O`, which is assigned to tokens that are outside the onomastic section.

The model was trained with 3,848 biographies labeled using with the active learning cycle over two rounds. The recent fine-tuned model achieved a precision 87.39%, a recall of 88.24%, and an F1 score of 87.81%. However, it is important to mention certain challenges and factors that influenced the numerical evaluation results. Largely, this is due to minor inconsistencies in the manu-

	TEACHER	STUDENT	OPINION	CONTACT	FAMILY	UNDEFINED
Precision	92.60%	94.16%	95.28%	60.26%	80.88%	90.19%
Recall	94.44%	94.65%	84.87%	56.88%	75.86%	93.54%
F1	93.51%	94.40%	89.78%	58.52%	78.29%	91.83%

Table 3: Person Relationships Classification Results.

Classes	# Entities
BIRTH (B)	23
DEATH (D)	60
BURIAL (G)	16
KNOWLEDGE (K)	77
RESIDENCE (R)	183
UNDEFINED (X)	78
Total	437

Table 4: Training Datasets for the *Toponym Classification* Task.

ally labeled training data, e.g. whether or not to include punctuation marks at the end of the onomastic section. Despite these challenges, however, it is important to stress that for the purposes of our project, the achieved level of accuracy is perfectly sufficient. The minor discrepancies, which are mainly due to closing tags being placed after an extra punctuation mark or an extra token, have no effect on the accuracy of the final outcome of the main research task.

6.2 Onomastic Entity Classification

Traditional Arabic (Islamic) names, as they appear in biographical collections, are quite different from their modern counterparts and are more akin to the short social profiles of individuals. With up to six different onomastic elements that may occur in any order and not all of them are always available, they give us the biographee's:[6] 1) "personal name" (ISM, Ar. *ism*); 2) the list of mainly male ancestors, which has the structure of "the son of ... the son of ... etc." (NAS, Ar. *nasab*); 3) "descriptive names", which describe tribal, religious, professional, geographical, and other affiliations (NSB, Ar. *nisba*); 4) a "patronymic" name that has the form of "The Father of ... / Abū Fulān" or "The Mother of ... / Umm Fulān" (KUN, Ar. *kunya*); 5) "honorific titles" (LQB, Ar. *laqab*); and 6) the "name of renown" (SHR, Ar. *šuhra*). The "descriptive names" (NSB, Ar. *nisba*) are the most valuable onomastic element for the goals of our project as they allow us to model different social and historical processes in the context of the development of the Islamic world.

Once the onomastic section within the biographical text has been identified, our next objective is to recognize and classify discrete onomastic elements present within it. For this purpose, we trained a token classification model that distinguished among the six main classes, described above (ISM, NSB, NAS, SHR, KUN, and LQB).

First, we pre-annotated the initial data set with our rule-based model (Table 6), which was built on data from the ONOMASTICON ARABICUM (Institute de Recherche et d'Histoire des Texts).[7] More specifically, we developed an onomastic gazetteer from data elements which were classified as *ism*, *kunya*, *laqab*, *nisba* and *šuhra* in the descriptions of persons collected in the ONOMASTICON ARABICUM. Further, the gazetteer included technical terms, used in texts to explain the spelling of rare names. We used our rule-based model to assign classes to all tokens inside onomastic sections; these assignments were then manually corrected. The model was trained with 2,011 biographies. Table 7 shows the training results. Notably, ISM achieved the best performance since it is unique in each onomastic section and it comes almost always as the first entity in the section. The training process adhered to the project's active learning cycle. The initial training data was generated by manually correcting labels derived from the rule-based model. Subsequently, the onomastic entity recognition model was trained using this corrected sample. The subsequent samples were then annotated using the onomastic entity recognition model.

7 Date Recognition, Classification, and Parsing

The aim of this model is threefold: 1) recognition of dates; 2) classification of dates; 3) parsing dates

[6] The main source of methodological guidance for this work is Malti-Douglas and Fourcade 1976, which summarizes the main research method of the ONOMASTICON ARABICUM Project.

[7] See, https://onomasticon.irht.cnrs.fr.

	rule-based			fine-tuned		
	Precision	Recall	F1	Precision	Recall	F1
BIRTH	78.57%	100%	88%	45.45%	100%	62.5%
DEATH	85.45%	78.33%	81.74%	50%	58.33%	53.85%
BURIAL	77.78%	77.78%	77.78%	0%	0%	0%
KNOWLEDGE	68.25%	59.72%	63.70%	96.55%	73.68%	83.58%
RESIDENCE	88.59%	52.38%	65.84%	57.14%	77.61%	65.82%
UNDEFINED	36.99%	75%	49.54%	41.67%	22.73%	29.41%

Table 5: Toponym Classification Results.

Classes	# Entities
ISM	1,888
NSB	3,260
NAS	3,856
KUN	910
LQB	285
SHR	179
Total	10,378

Table 6: Training Dataset for the *Onomastic Entities Classification* Task.

to numerical values. The research focus of the project is on the period of *c.* 600-1600 C.E. and information from the texts is assigned to a certain point in time somewhere within this period. We are particularly interested in what kind of information—historical events—can be associated with specific points in time.

First, the model searches the Arabic text with a regular expression (`regex`) which will match phrases reporting on dates. The regex matches days, days of the week, months, and years. Additionally, it captures ten preceding tokens, which are considered the context of a date. At the moment, we are primarily interested in years and their thematic contexts.[8] When the regex finds an occurrence of a date phrase, it usually returns two main groups of elements. One of the groups is the context; another one is a series of spelled-out numerals of the date statement, including ones, tens (decades), hundreds (centuries), and, in late texts, a thousand (for the first millennium). The model then uses a dictionary that returns numerical values of date statement element. Summing up these numerical values gives us the actual value of the date. Additional regex is then applied to the date context to check if it has any of the most common contextual vocabulary. For example, tokens like *wulida* (he was born) or *wulidat* (she was born) are used to classify dates as dates of birth. If the context contains more than one term from the date classification dictionary, we use the one closest to the date statement. We defined six main classes of dates, which are BIRTH, DEATH, KNOWLEDGE transfer, appointment or termination of an OFFICE, PILGRIMAGE, and UNDEFINED dates.

Table 9 shows the classes and their number of occurrences in the test dataset. The model was evaluated with 1,047 biographies. The numerical value extraction by this model achieved a mean average percentage error (MAPE) of 1.55% and a mean average error (MAE) of 4.76 years if the date phrase was correctly recognized as such.

Table 8 shows the results for this rule-based model. One of the reasons for a low precision for the class UNDEFINED is that this class is assigned whenever no other indicator was in the ten preceding tokens. Those ten tokens are not enough for every case, so sometimes the indicator was the 11th preceding token, and therefore the date was mistakenly classified as UNDEFINED. Overall, the results by this rule-based model show promising performance, especially since the parsing is already working very well for recognized dates. Still, date recognition presents a lot of obstacles.

So far, the model only returns information about a date if it's explicitly stated in the recognized phrase. However, not all information is always presented explicitly. A common instance is the omission of the century, as authors often expect readers to infer the exact century from the context. Consequently, we need to enhance our model to derive any missing data from other dates provided in the biography, headers of chapters containing the biography (especially when biographies are grouped into periods—a common practice in our

[8] Year statements are the most frequent type of date statements; more precise indications of time are significantly less frequent and, structurally, are more diverse and less consistent.

	ISM	KUN	NSB	NAS	LQB	SHR
Precision	99.07%	99.42%	97.15%	96.55%	75.32%	79.07%
Recall	97.72%	97.73%	97.99%	98.07%	80.56%	70.83%
F1	98.39%	98.57%	97.57%	97.30%	77.85%	74.73%

Table 7: Onomastic Elements Classification Results.

	BIRTH	DEATH	KNOWLEDGE	OFFICE	PILGRIMAGE	UNDEFINED
Precision	96.03%	94.64%	89.66%	66.67%	80.00%	52.41%
Recall	90.98%	84.57%	59.09%	16.67%	50.00%	81.46%
F1	93.44%	89.33%	71.23%	26.67%	61.54%	62.20%

Table 8: Date Recognition and Classification Results.

Classes	# Entities
BIRTH (B)	133
DEATH (D)	376
KNOWLEDGE (K)	44
OFFICE (O)	12
PILGRIMAGE (P)	8
UNDEFINED (X)	85
Total	658

Table 9: Evaluation Dataset for the *Date Classification* Task.

sources), or the scope of the historical source where the biography was found. We are still assessing the most efficient approach to implement this disambiguation.

Another challenge involves handling date statements that refer to periods or approximate years. For example, instead of exact numbers, we may find words like *ba'ḍ* and *nayyif*, which refer to an unspecified year within a specified decade. While these date statements cannot be translated into precise numerical values, this limitation does not severely impact our project. Given the extensive period we are studying, we typically operate on the granularity of decades, rounding exact years to the nearest decade when necessary.

Additionally, authors sometimes report alternative dates for the same event, either by detailing both dates fully or by abbreviating the second date. In such cases, we make an effort to collect and process both dates.

8 Conclusions and Future Work

In the preceding sections, we have outlined our efforts in adapting existing state-of-the-art Arabic NLP models to specific research tasks. We fine-tuned an NER model, specifically tailored for historical and biographical texts in classical Arabic, which exhibits excellent performance in detecting persons and toponyms. Moreover, we trained models to further classify detected persons, based on how they are related to the biographee. This model achieved good performance, particularly in the classification of teachers and students within the biographical context. The model for toponym classification is still under development as we are lacking sufficient training data. We are currently working on increasing our training dataset for this task.

Further, we trained a boundaries detection model to locate the onomastic section inside biographies, and yet another model that identifies onomastic elements within that section. The model for date recognition, classification, and parsing achieved promising results for the main goals of the project. Still, date recognition is not a trivial task and we are researching ways to overcome limitations such as the missing centuries.

Although this work is still in progress, the preliminary results reported in the paper indicate the excellent-to-satisfactory performance of the fine-tuned models, effectively meeting the intended goal for which they were trained. However, our ongoing efforts involve expanding the training datasets and further fine-tuning the models with the aim of achieving even better results.

Finally, our contribution extends beyond the trained models themselves. We have also developed and curated valuable training datasets that can serve as a resource for other researchers and contribute to the advancement of work in the field of classical Arabic. These datasets provide a foundation for further exploration and improvements in the current models.

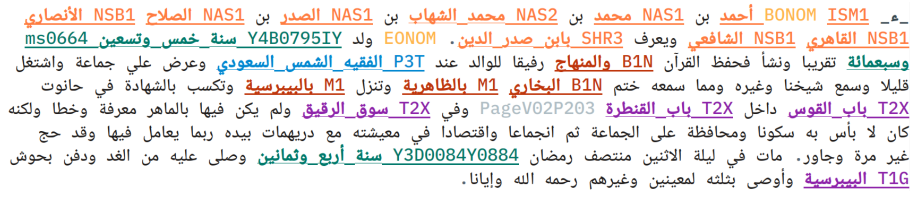

Figure 4: Example of automatically annotated biography from al-Saḫāwī's *al-Ḍaw' al-lāmi'*.

Our future work can be summarized as follows. In the short term, our primary focus lies in generating additional training data to facilitate further fine-tuning of the models. This iterative process is aimed at continuously improving the performance of the models until reaching a stage where they can effectively annotate the entire corpus. Figure 4 illustrates an exemplar output of our annotation pipeline, wherein the annotated text has undergone processing by all the discussed models. After annotating all biographies in our corpus and extracting all relevant metadata (onomastic elements, persons, toponyms, dates, etc.), our subsequent objective is to organize this information into networks comprised of overlapping thematic clusters. These thematic clusters will serve as an analytical framework, enabling us to explore and derive insights from the interconnections and relationships among various social, professional, and religious groups, within extensive historical and geographical contexts as they are recorded in our vast corpus. Additionally, the project explores the development of these networks through spatial and temporal analysis, which is grounded in the recognition of dates and toponyms. Overall, this network will serve as the main research framework of the project for the study of the social history of the Islamic world. Further, this project will help to identify weakly researched topics in the field of Arabic studies and at the same time provide a new research tool for fellow researchers to start working on these topics.

Limitations

Ancient and classical languages, including classical Arabic, face significant challenges in terms of the availability of adequate training datasets and pre-trained models. Creating such datasets is a non-trivial task, demanding considerable time and resources. It necessitates the involvement of domain experts possessing the requisite knowledge to perform annotations in accordance with prescribed guidelines or annotation schemes. The scale of the dataset and the complexity of classification tasks present additional challenges. To tackle these obstacles, we have adopted an active learning development cycle, which allows us to efficiently and rapidly generate training data. Furthermore, in certain cases, we decided to reduce the number of labels or split the labels into two sets and train two separate models instead of one model in order to get better performance.

Acknowledgement

This research is a part of the work within the Emmy Noether Research Group (№445975300), "The Evolution of Islamic Societies (c.600-1600 CE): Algorithmic Analysis into Social History" (EIS1600), funded by the German Research Foundation (DFG) and hosted at Universität Hamburg.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–16. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021a. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021b. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Institute de Recherche et d'Histoire des Texts. Onomasticon Arabicum.

Fedwa Malti-Douglas and Geneviève Fourcade. 1976. *The Treatment by Computer of Medieval Arabic Biographical Data: An Introduction and Guide to the Onomasticum [i.e., Onomasticon] Arabicum*. Number 6 in Série Onomasticon Arabicum. Editions du Centre national de la recherche scientifique.

E Moatez Billah Nagoudi, A Elmadany, and M Abdul-Mageed. 2021. Arat5: Text-to-text transformers for Arabic language understanding and generation. *arXiv preprint arXiv:2109.12068*.

Lorenz Nigst, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, and Peter Verkinderen. 2023. OpenITI: a Machine-Readable Corpus of Islamicate Texts.

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 Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Meng Wang and Xian-Sheng Hua. 2011. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):1–21.