# Named Entity Recognition in Pre-modern Arabic Biographical Texts

**Ishida, Yuri**
ishidayuri@okayama-u.ac.jp
Okayama University, Japan

**Baba, Kensuke**
k-baba@fit.ac.jp
Fukuoka Institute of Technology, Japan

**Baba, Takahiro**
baba.takahiro@kurume-it.ac.jp
Kurume Institute of Technology, Japan

**1. Introduction**

This study aims to extract people's names from biographies written by Muslim scholars in the 17th and 18th centuries. Arabic is one of the six official languages of the United Nations and has a native population of approximately 500 million, making it the fourth-largest language in the world. However, the peculiarity of its grammatical features, such as homonyms, has made Natural Language Processing (NLP) in this language difficult. Research on this topic includes studies by Zayed & El-Beltagy (2015), who performed Named Entity Recognition (NER) on tweets in Arabic, and Aldumaykhi et al. (2022), who performed NER on newspaper articles published in Saudi Arabia. However, these studies focus on Modern Standard Arabic (MSA), which was established in the late 19th and early 20th centuries. Thus, the application of NER to pre-modern texts requires the development of a new suitable system. Most Arabic NLP research has been conducted on MSA, with the exception of the studies by Salah & Binti Zakaria (2018) and Alsaaran & Alrabiah (2021), who considered classical Arabic, from the 7th to 11th century. However, given that the center of science in the pre-modern period was an Islamic cultural area, the development of NLP techniques for pre-modern Arabic texts has the potential to offer new insights into the current historical perspective, which is primarily based on Western language literature.

2. Research Target

Biographies composed by Muslim scholars in the 17th and 18th centuries were selected as pre-modern Arabic texts to be investigated in this paper. These were selected for their potential to provide insights into the formation and development of Islamism, which had a major impact on contemporary society, dating back to the period of the Arabian Peninsula in the Haramayn (the two cities of Makkah and Madinah). Romanov (2013) studied the places of origin of scholars between the 7th and 12th centuries and concluded that, despite being the birthplace of Islam and the destination of its pilgrimages, Haramayn was no longer a center of scholarship after the 9th century. Muslim scholars traveled to Damascus, Baghdad, Isfahan, and Cordoba after this point in time. However, the works of Voll (1975) pointed out that this tradition changed in the 17th and 18th centuries—Muslim scholars gathered once again in the Haramayn, and it formed the prototype of Islamism. In other words, the movement of scholars has changed from traveling around to going direct. Explicit statistical demonstration of this change is expected to reinforce the argument that the altered nature of the scholarly movement was a major factor behind the emergence of highly exclusive Islamic extremism.

3. Method

Three biographical corpora were used in this study. For scholars of the 11th century Hijrah (1592–1688 AD), (1) Khulāṣat al-athār composed by Muḥammad Amīn ibn Faḍl Allāh al-Muḥibbī (1651-1699), and (2) Fawā'id al-irtiḥāl composed by Muṣṭafā Ibn-Fatḥallāh al-Ḥamawī (1633/34–1711-12) were selected. For scholars of the 12th century Hijrah (1689–1785 AD), (3) Silk al-durar composed by Muḥammad Khalīl al-Murādī (d. 1791) was selected.

Abdallah et al. (2012) showed the supremacy of integrating the rule-based system and the Machine-learning approach. Focusing on the formatted text for biographies, we set some rules. For example, dates follow after the word "dafn (buried)." Regarding machine learning, by manual annotation of the three biographies above, we improve the recent successful Arabic NERs; Stanza, CAMeL, and hatmimoha/Arabic.

**4. Conclusions**

The primary contributions of this study are twofold. Firstly, our NER helps historical studies by filling the lack between classical Arabic corpora, represented by CANERCorpus and KSUCCA, and MSA corpora, represented by the ANERCorp corpus and the AQMAR corpus. The three biographies we deal with are rich in information about historical events and personal networks. The analysis of the results of NER enables accurate tracking of the migration of scholars who shaped Islamism.

Secondly, the integrating of the rule-based system and the Machine-learning approach is a challenge for higher performances for Arabic NER. Aldumaykhi et al. (2022), the latest studies in MSA, adopt the merging and voting system based on the results of Stanza, CAMeL, and hatmimoha/Arabic. However, we utilize the character of well-formatted biography text by introducing a rule-based system.

# Bibliography

**Abdallah, Sherief** / **Shaalan, Khaled** / **Shoaib, Muhammad** (2012): "Integrating Rule-Based System with Classification for Arabic Named Entity Recognition", in: Gelbukh, Alexander (eds): 13th International Conference *Computational Linguistics and Intelligent Text Processing*, New Delhi, India, March 11-17, 2012: Part 1, 311–322. DOI: 10.1007/978-3-642-28604-9_26.

**Aldumaykhi, Abdullah** / **Otai, Saadand** / **Alsudais, Abdulkareem** (2022): "Comparing Open Arabic Named Entity Recognition Tools", in: *arXiv*. DOI: 10.48550/arXiv.2205.05857.

**Alsaaran, Norah** / **AlRabiah, Maha** (2021): "Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT" in: *IEEE Access* 9: 91537–91547. DOI: 10.1109/ACCESS.2021.3092261.

**Omnia Zayed and Samhaa El-Beltagy** (2015): "Named Entity Recognition of Persons' Names in Arabic Tweets", in: *Proceedings of the International Conference Recent Advances in Natural Language Processing*: 731–738 https://aclanthology.org/R15-1093 [28. 04. 2023].

**Romanov, Maxim G.** (2013): Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunni World (661–1300 CE). Ph.D. thesis, University of Michigan https://hdl.handle.net/2027.42/102300 [28. 04. 2023].

**Salah, Ramzi Esmail** / **Zakaria, Lailatul Qadri Binti** (2018): "Building the Classical Arabic Named Entity Recognition Corpus (CANERCorpus)", in: *2018 Fourth International Conference on*

*Information Retrieval and Knowledge Management (CAMP)*: 1–8. DOI: 10.1109/INFRKM.2018.8464820.

**Voll, John** (1975): “Muḥammad Ḥayyā al-Sindī and Muḥammad ibn ‘Abd al-Wahhāb: An Analysis of an Intellectual Group in Eighteenth-Century Madīna”, in: *Bulletin of the School of Oriental and African Studies* 38, 1: 32–39. DOI: 10.1017/S0041977X00047017.