Post-doc Internship offer 2023-2024

Recovering hidden text in Byzantine seals

1 Information about the internship

- **Supervisors:**
  - Victoria Eyarabide, STIH Laboratory, Sorbonne Université
  - Laurence Likforman-Sulem, Département IDS, Telecom Paris
- **Location:** Masion de la recherche, Sorbonne Université - 28 rue Serpente, 75006 Paris.
- **Duration:** 12 months
- **Keywords:** Deep Nets, Character recognition, Natural Language Processing, Document images.

2 Detailed topic

Byzantine seals (see Figure 1) are small circular objects used to identify the sender of letters. They carry a large part of the knowledge on the administration and the Byzantine aristocracy, but also on the cult of the saints. Seals have two sides: an observe side which most often includes iconography, and a reverse side including text such as the sender's name, his/her social position, and elements of prayers. Seals have been altered over time, so characters may be damaged or even erased. In addition, since the surface of a seal is small (between 10 to 50 mm in diameter), engravers have gained room by removing word spaces, omitting or fusing characters, and omitting even entire words. Consequently, the text is often abbreviated. The objective of this postdoctoral internship consists in combining deep learning approaches[2] and natural language processing(NLP)[1] tools to fully recover the text on seals despite its abbreviated form and damaged characters.

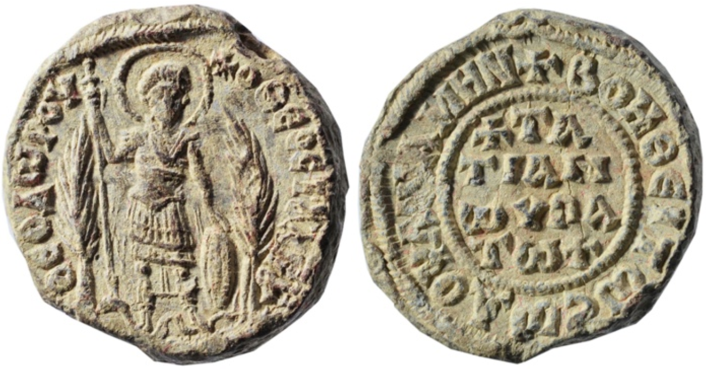

Figure 1: An example of Byzantine Seal (Tatianos hypatos [3, p. 225]

This research will be developed within the framework of the ANR BHAI project. In previous research [7, 4], we obtained transcripts of seal reverse sides using a two-step neural-based approach, localizing first characters by a deep object detector, then recognizing characters by a convolutional net. We plan to continue this research by splitting this task into several steps. We will first develop a Bayesian approach that predicts word (complete or abbreviated) boundaries [6, 5]. Then word hypotheses will be possibly expanded by using text normalization and machine translation transformer approaches. To train the systems, we will rely on Greek corpus and dictionaries [8]. The candidate will work in Paris, at Sorbonne University (V. Eyharabide), in close collaboration with Telecom Paris (L. Likforman).

3 Profile of applicant

Applicants are required to have:

- A PhD in Computer Science.
- Advanced skills in Python programming are mandatory.
- A strong background in Machine Learning & Deep Learning on images using related libraries (scikit-learn, Tensorflow, Pytorch, etc.).
- Fluency in written and spoken English is essential.
- Communication skills in French are a plus but not required.
- A good publication record will be a plus.

The position is open immediately. Review of applications will begin as soon as applications are received and continue until the position is filled.

4 Application

Applicants should send an email to Victoria Eyharabide maria-victoria.eyharabide@sorbonne-universite.fr and Laurence Likforman-Sulem Laurence.likforman@telecom-paris.fr with:

- A full curriculum vitae including a complete list of publications
- A transcript of higher education records
- A one-page research statement discussing how the candidate's background fits the proposed topic
- Two support letters of persons that have worked with them.

References

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases, 2022.

[2] Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283, 2022.

[3] Jean-Claude Cheynet. *Les sceaux byzantins de la collection Yavuz Tatiş. İzmir.* Privately published, Izmir, 2019.

[4] Victoria Eyharabide, Laurence Likforman-Sulem, Lucia Maria Orlandi, Alexandre Binoux, Theophile Rageau, Qijia Huang, Attilio Fiandrotti, Béatrice Caseau, and Isabelle Bloch. Study of historical Byzantine seal images: the BHAI project for computer-based sigillography. In *ICDAR 2023 International Workshop on Historical Document Imaging and Processing (7th edition) (HIP'23)*, San José, California, United States, August 2023.

[5] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.

[6] Shu Okabe, Laurent Besacier, and François Yvon. Weakly supervised word segmentation for computational language documentation. In *Annual meeting of the Association for Computational Linguistics*, 2022.

[7] Rageau Théophile. Deep learning approach for character recognition in byzantine seal images. Rapport de stage de master mva, Telecom Paris, 2022.

[8] François Yvon. Rewriting the orthography of sms messages. *Natural Language Engineering*, 16(2):133–159, 2010.