A Semantic Search Engine for Historical Handwritten Document Images

Vuong M. Ngo[1] ✉, Gary Munnelly[1], Fabrizio Orlandi[1], Peter Crooks[2], Declan O'Sullivan[1], and Owen Conlan[1]

[1] ADAPT Centre, SCSS, Trinity College Dublin, Ireland
[2] Department of History, Trinity College Dublin, Ireland
{vuong.ngo, gary.munnelly, fabrizio.orlandi}@adaptcentre.ie,
{pcrooks, declan.osullivan, Owen.Conlan}@tcd.ie

Abstract. A very large number of historical manuscript collections are available in image formats and require extensive manual processing in order to search through them. So, we propose and build a search engine for automatically storing, indexing and efficiently searching the manuscript images. Firstly, a handwritten text recognition technique is used to convert the images into textual representations. In the next steps, we apply the named entity recognition and historical knowledge graph to build a semantic search model, which can understand the user's intent in the query and the contextual meaning of concepts in documents, to return correctly the transcriptions and their corresponding images for users.

Keywords: handwriting transcription, named entity, knowledge graph

1 Introduction

Every year, the great collections of historical handwritten manuscripts in museums, libraries and other organisations are digitised as electronic images. The digitisation makes the manuscripts available to a wider audience, and preserves the cultural heritage. The automatic recognition of textual corpora and named entities generated from medieval and early-modern manuscript sources with high accuracy is a challenge (Ahmed et al. 2017; Nozza et al. 2021). Manuscript images are often processed through keyword spotting or word recognition to be accessed and searched, such as Cheikhrouhou et al. (2021), Martınek et al. (2020) and Kang et al. (2021). There are some papers build a search system for handwritten images, such as Lang et al. (2018), Colutto et al. (2019), Stauffer et al. (2020) and Vidal and et al. (2020). However, their systems only offer keyword search.

Unlike keyword search, semantic search improves search precision and recall by understanding the user's intent and the contextual meaning of concepts in documents and queries (Ngo and Cao 2011; Jiang 2020). This paper proposes a semantic search engine for full-text retrieval of historical handwritten document images based on named entity (NE), keyword (KW) and knowledge graph (KG). This would help not only in processing, storing and indexing automatically, but also would allow users to access quickly and retrieve efficiently manuscripts.

2 System Architecture

The Public Record Office of Ireland (PROI) was destroyed on 30 June 1922, resulting in the loss of 700 years of Irish history. The Beyond 2022 Project (`https://beyond2022.ie`) is combining historical research, archival discovery, and technical innovation to create a virtual reconstruction of the PROI. There are over 300 volumes of surviving and collected handwritten copies of lots documents, with some 100,000 pages containing 25 million words of text.

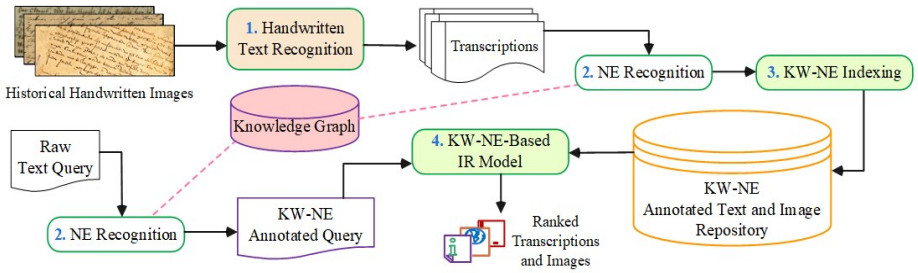


Fig. 1. The system architecture

Our system architecture of the search engine is illustrated in Figure 1 which has four separate processing modules being `Handwritten Text Recognition`, `NE Recognition`, `KW-NE Indexing` and `KW-NE-Based IR Model`. Firstly, the historical handwritten document images are digitised to transcriptions through the `Handwritten Text Recognition` module. Then, the transcriptions are annotated by NEs through the `NE Recognition` module. This module needs to connect to the `Knowledge Graph` to extract the classes and identifiers of NEs. Next, KWs and NEs of the annotated transcriptions and the respective original images are presented and indexed by the `KW-NE indexing` module and stored in `KW-NE Annotated Text and Image Repository`. The raw text query is also annotated NEs through the `NE Recognition` module to become a KW-NE annotated query. Finally, the `KW-NE-Based IR Model` module compares the annotated query and the annotated documents to return the ranked transcriptions and images.

3 Image Representation and Knowledge Graph

Transkribus (Kahle et al. 2017) is used for training and deploying Handwritten Text Recognition (HTR) models to derive text transcription from image scans. Given the rate at which transcriptions can be generated, NE Recognition (NER) and Entity Linking (EL) are required to automated annotate all instances of entities occurring in the transcription text. We used SpaCy (Honnibal et al. 2020) for NER and had highly results on 18^{th} century English text. To provide flexibility, an NLP pipeline has been implemented as a thin layer over a number of standard NLP tools. The output of the pipeline is a NLP Interchange Format (Hellmann et al. 2013) in which a NER tool has annotated classes of entities and, where possible, an EL tool has connected the recognized entities to KG.

The KG collects structured data from various historical sources. Part of the data is manually curated by historians through spreadsheets. Other data sources (e.g. geographical data from OSi (Debruyne et al. 2017)) are imported automatically as RDF for direct insertion into KG. The schema (or ontology) used to structure KG, is mainly based on the popular CIDOC-CRM ontology (Doerr 2003). A short excerpt of KG is depicted in Figure 2. It shows a few main entities and relationships related to a person (of type `CIDOC-CRM:E21_Person`) named "William Sutton", who was member of a few relevant offices in Ireland.

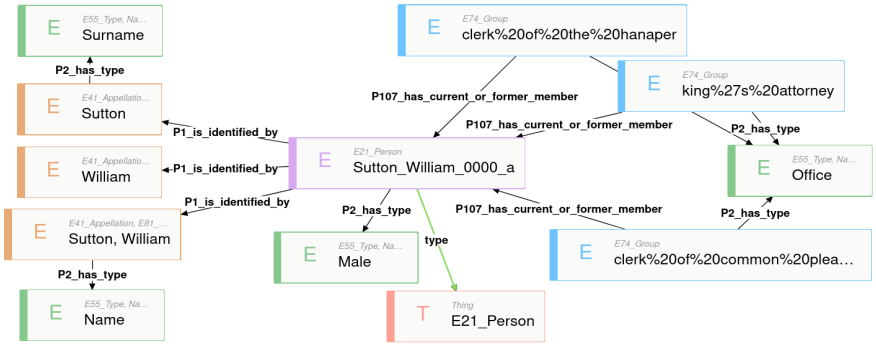
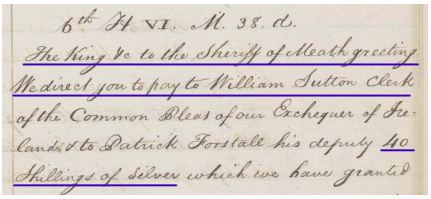


Fig. 2. A portion of our historical KG about "William Sutton".

4 Information Retrieval Model and Demo



"... The King &c to the Sheriff of Meath greeting We direct you to pay to William Sutton clerk ... 40 Shillings of silver ..."

d = {..., king, &c, occu_sheriff, coun_meath, greeting, direct, pay, william_sutton/person, occu_clerk, 40, shilling, silver, ...}

Fig. 3. An example about NE and KW annotation of a medieval historical manuscript

A search engine needs to not only return the best documents, but also be fast. We implemented the index and search functions based on Elasticsearch to have a real-time search engine (Gheorghe et al. 2015). The Okapi BM25 model was proposed to find and rank the relevant handwritten manuscripts for queries. In the model, documents and queries are presented by sets of concepts being NEs or KWs. Figure 3 presents an image of a handwritten medieval historical manuscript, its transcription and its concept set d, applied in the model. In the transcription, there are three kinds of words determined by our NER tool: (1) stop-words being *the*, *to*, *of*, *we* and *you*; (2) NEs being *sheriff*, *Meath*, *clerk* and *William Sutton*; and (3) KWs being *king*, *&c*, *greeting*, *direct*, *pay*, *shilling* and *silver*. The stop-words are not added into the concept set d.

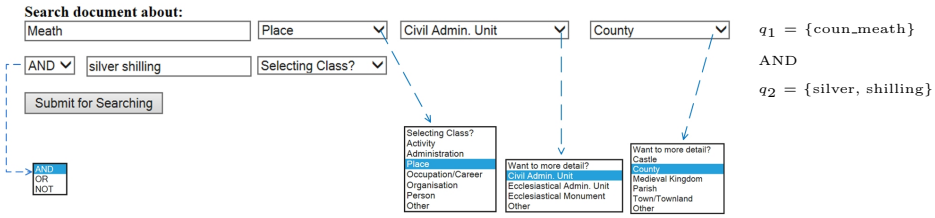


Fig. 4. User interface of our deployed search engine

Figure 4 presents the interface of our search engine[1], and the concept sets of q_1 and q_2. In that, *coun_meath* is the identifier of an entity named *Meath* and classed *Country*, which is determined by our NER algorithm. While, *silver* and *shilling* are keywords. To exploit the features of NEs for semantic search, a NE needs to be presented by its most specific meaning in the concept set d. It means that, with a NE in the transcription,

- If our NER can determine its identifier, the NE will be presented by its identifier in d. For example, *occu_sheriff*, *coun_meath* and *occu_clerk* are identifiers of entities named *sheriff*, *Meath* and *clerk*, and added into d.
- If our NER only determines its most specific class, the NE will be presented by a combined information including its name and class. For example, the entity named *William Sutton* does not exist in our historical KG, so its identifier cannot be extracted. However, the NER determines its most specific class being *Person*. So *william_sutton/person* is added into d.

5 Conclusion

We proposed a novel semantic full-text search system for images of historical handwritten manuscripts. Unlike the existing approach only using KW extracted from images, we exploited NE, KW and KG of increase search performance. In that, NER and HTR tools were built to recognise transcriptions and NEs from the manuscript images. Besides, to increase the precision of our NER tool, the historical KG was designed and proposed. Then, we implemented the index and search functions for transcriptions based on Elasticsearch and Okapi BM25 to search images in real-time. Finally, the semantic search engine was also implemented and deployed.

Acknowledgment

Beyond 2022 is funded by the Government of Ireland, through the Department of Culture, Heritage and the Gaeltacht, under the Project Ireland 2040 framework. The project is also partially supported by the ADAPT Centre for Digital Content Technology under the SFI Research Centres Programme (Grant 13/RC/2106_P2).

[1] https://by2022.adaptcentre.ie/conf_demo

References

Ahmed, R., Al-Khatib, W., and Mahmoud, S. (2017). Survey on handwritten documents word spotting. *International Journal of Multimedia Information Retrieval*, 6(1):31–47.

Cheikhrouhou, A., Kessentini, Y., and Kanoun, S. (2021). Multi-task learning for simultaneous script identification and keyword spotting in document images. *Pattern Recognition*, 113:107832.

Colutto, S., Kahle, P., Guenter, H., and Muehlberger, G. (2019). Transkribus. a platform for automated text recognition and searching of historical documents. In *Proceedings of the 15th Int. Conf. on eScience (eScience)*, pages 463–466.

Debruyne, C., Meehan, A., Clinton, É., McNerney, L., Nautiyal, A., Lavin, P., and O'Sullivan, D. (2017). Ireland's authoritative geospatial linked data. In *International Semantic Web Conference*. Springer.

Doerr, M. (2003). The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3).

Gheorghe, R., Hinman, M., and Russo, R. (2015). *Elasticsearch in Action*. Manning Publications Co., USA, 1st edition.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *The Semantic Web – ISWC 2013, LNCS, Vol. 8219, Springer*, pages 98–113.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). SpaCy: Industrial-strength Natural Language Processing in Python.

Jiang, Y. (2020). Semantically-enhanced information retrieval using multiple knowledge sources. *Cluster Computing*, 23:2925–2944.

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR Int. Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24.

Kang, L., Riba, P., Villegas, M., Fornés, A., and Rusiñol, M. (2021). Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition*, 112:107790.

Lang, E., Puigcerver, J., Toselli, A. H., and Vidal, E. (2018). Probabilistic indexing and search for information extraction on handwritten german parish records. In *Proc. of 16th Int. Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 44–49.

Martınek, J., Lenc, L., and Kral, P. (2020). Building an efficient OCR system for historical documents with little training data. *Neural Computing and Applications*, 32:17209–17227.

Ngo, V. and Cao, T. (2011). Discovering latent concepts and exploiting ontological features for semantic text search. In *Proc. of the 5th Int. Joint Conference on Natural Languag Processing (IJCNLP-2011), ACL*, pages 571–579.

Nozza, D., Manchanda, P., Fersini, E., Palmonari, M., and Messina, E. (2021). Learningtoadapt with word embeddings: Domain adaptation of named entity recognition systems. *Information Processing & Management*, 58(3):102537.

Stauffer, M., Fischer, A., and Riesen, K. (2020). Filters for graph-based keyword spotting in historical handwritten documents. *Pattern Recognition Letters*, 134:125–134.

Vidal, E. and et al. (2020). The carabela project and manuscript collection: Large-scale probabilistic indexing and content-based classification. In *the 17th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, pages 85–90.