Identifying Historical Occupations in Unstructured Texts: A User-Centric Approach

Guilherme Rodrigues Arashiro[1], Jaan Joosep Puusaag[2], and Stacey Koolman[2]

[1] Vrije Universiteit (VU), De Boelelaan 1105, 1081 HV Amsterdam
g.rodriguesarashiro@student.vu.nl
[2] Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam
{joosep.puusaag,stacey.koolman}@student.uva.nl

January, 2023

Abstract. The difficulties that arose with the desire to study social mobility in history have led to the creation of a Historical International Standard of Classification of Occupations, HISCO, which is an attempt to codify historical occupations based on ISCO68. While already useful on its own, accessing and assigning these numbers to occupations in historical documents, such as marriage certificates, historical censuses, or other unstructured texts, is not yet a standardised procedure. Moreover, there are often other types of information in these texts that can help with the original task of studying social mobility, their ranking in the profession they're in, or their relation to other people, objects, or places. This paper concerns itself with the process of creating a workflow for researchers to access HISCO and other occupation codes within unstructured texts, along with the additional information stated earlier, and the challenges that arose during it. By the end of the project, the group was able to formulate a working MVP which, although not optimized, correctly flags and presents users with occupation matches, and stores correct ones in a separate file.

Keywords: Digital Humanities · HISCO · Natural Language Processing · History

1 Introduction

Historical International Standard of Classification of Occupations (HISCO) is one of the most prominent theoretical models which assigns a five-digit code to historical occupations, in the format of x-xx.xx (Appendix A). HISCO is based on the ISCO standard set in 1968, as opposed to the earlier or later models, because (1) the ILO already had added some historical occupations that were absent in the others, (2) it already existed in many different languages, (3) it had already been used in a historiographical context previously (van Leeuwen, 2004).

HISCO was created in 2002 (van Leeuwen, 2004) to simplify the study of social mobility in the past on an international scale. The codes help researchers gain access to information regarding historical occupations in documents even if

they do not speak the language the document is written in. However, the process of assigning these codes to occupations found in historical manuscripts and other texts is not yet standardised. This problem makes doing such research difficult because they all have to invent independent ways to achieve the same goal. This also increases the chances of errors finding their way into the research, because all the researchers may be performing their tasks differently. Therefore, the aim of this project is to create a workflow for researchers to find HISCO codes in unstructured texts. This paper will discuss similar attempts at creating similar workflows, the approach the group took to create a new workflow, its application on a test dataset, as well as the problems encountered during this project.

It has to be kept in mind that alongside HISCO there are other projects such as the North Atlantic Population Project (NAPP) which is a standardized and codified dataset collection from various countries, and the The Cambridge Group for the History of Population and Social Structure (Cambridge) which uses a similar yet different method for codifying historical occupations but considering HISCO's international application and its basis on another internationally used classification this project uses that.

2 Related Work

Over the years, the digitization of historical text sources facilitated research in the humanities and social sciences by simplifying how information can be sorted and filtered. One frequently used method to extract specific information from these digitized text has been Named Entity Recognition (NER) (Karsvall and Borin, 2018), for instance, used this method to identify place names in medieval charters and link them to geographical locations using information from historical maps. Working with historical texts, however, comes with its difficulties. In this case, NER failed to recognize place names that no longer existed or treated those who have new names as two separate locations. Thus, NER is not perfect by itself but presents a pathway to resolve methodological challenges with historical texts. Plu et al. (2015) proposed a combination of semantic- and linguistic-based methods to improve these imperfections. Namely, training an algorithm with a subset of data to allow it to learn patterns that can be used to identify named entities and their links. Implementing such a hybrid approach requires large amounts of data for the accuracy to be high and time to train the algorithm. As this project has time constraints, the focus will be on using NER to first create an operative workflow that can be built up on in the future.

3 Dataset

The dataset used in this project was sourced from the Biographical Dictionary of Gelderland (Het Biografisch Woordenboek Gelderland). The data consists of biographies of people who have been significant in the history of Gelderland from the 13th to the 20th century. Each biography contains information on, for

instance, the family, date of birth and death, and occupations of the individuals (Huysman, I., Kloek, E., n.d.). Although HISCO mainly focuses on the 19th century, the dataset still contains multiple instances of historical occupations, and thus should provide matches for the workflow itself.

In this project, the biographical data was utilized almost to its entirety to look for occupation matches, with the notable exception of the biographies' abstracts, as the purpose of testing the workflow is to analyse matching capabilities of HISCO with unstructured texts, and thus an abstract could interfere with such work. The Biographical Dictionary of Gelderland dataset is complete, as there are no null cells/entries within it and, given the fact that the data is both valid and unique, it can also be considered reliable. Finally, the dataset is mostly representative, given that all biographies are correctly filled-in and detailed, although there is always the chance of other relevant citizens of Gelderland not being added.

4 Method

As previously stated, the primary focus of this project is the formulation of a workflow in which occupations found in unstructured text are matched to their equivalent counterparts in databases such as HISCO, as a means to assist history researchers in grasping the professions found in specific texts. There are multiple ways to tackle this objective, and the most prominent ones will be analysed and tested throughout this section. Nevertheless, all approaches will consist of some form of entity recognition, which could potentially be expanded into entity linking should that be desired as well.

Entity recognition is an information extraction technique which classifies named entities in unstructured text (i.e., that does not have a pre-defined data model) into their correct categories, such as organizations, persons, geographic locations, etc. On the other hand, entity linking aims to assign a unique identity to a given entity, so that it has a specific and correct match (i.e., linking the word "Brazil" in the phrase "Brazil is one of the largest countries" to the country of Brazil instead of the city of Brazil in the state of Indiana). Both techniques differ from one another as, while entity recognition focuses on classifying a named entity to a given category, entity linking matches it to a specific identity.

The main benefit of both techniques is to aid in the automatic classification and identification of unstructured text, being heavily utilized in Natural Language Processing tools. Both entity recognition and linking enable researchers, data scientists and other interested parties to quickly categorise and structure text with a certain degree of confidence, allowing them to not waste time in manually organizing data, a task which can become impossibly long depending on the size of the required dataset.

In the particular case of this project, the main goal would be to establish the entity recognition between the HISCO database for occupations in any available language (which would preferably be chosen by the current user) and a dataset of historical texts provided by the researcher. The tool would then analyse the

dataset, and prompt the user with any meaningful matches it found with HISCO occupations, thus asking the user to confirm whether the items match and, if so, storing pairing for later use.

4.1 Preprocessing

Prior to any comparison, it is important to establish some form of standard to the data being utilized, and thus preprocessing the datasets is required. For the HISCO dataset, it was chosen to import it into the tool as a JSON file, in which all occupational entries can be loaded whilst also maintaining their unique attributes, such as language. With this, the users can choose to query only in their desired languages, and the tool will be able to filter this dataset accordingly. It is also worth noting that, since HISCO data will be the same for all users, this dataset can be already pre-inserted into the tool thus removing the necessity of having the researchers manually scrape it for their usage.

Although there is also a necessity for standardization in the historical texts, it is considerably harder to maintain it given that each researcher may have their datasets in a different format, or perhaps entirely unstructured, especially as sources may differ entirely; some may scrape websites while others scanned and digitized the texts through methods such as OCR. Due to this, structuring this data will be a limited endeavor.

In light of this, the only hard standard which will be applied is that the user presents their data in a JSON format, as the tool will be expecting a dataset in this format to begin the comparisons. However, the user will be prompted to acknowledge that the file being inputted is a historical text that will be analysed for profession matches, and that the HISCO dataset only contains occupations from a certain time period, meaning that it may not be perfected for the researcher's given data. With this, both datasets are added, and the comparison may start.

4.2 Solution approach and design

After the preprocessing step is completed, the user will be prompted to choose the language in which their data is written. After that, the tool will begin to parse the historical text to find string values that match one of the entries in the HISCO dataset, which will be already trimmed based on the aforementioned filter. This, of course, is the main aspect of the project: how to best compare both datasets. Despite there being many options to tackle this, it was decided that this project will utilize a Levenshtein distance method to compute the string matches. The usage of other NLP techniques was also heavily discussed and of the group's primary interest, but given the timeframe given for this project, it was decided that other methods will be left for future work and improvements.

The Levenshtein distance is an algorithm through which a pair of strings is compared and a numerical value is given as a result, with this number representing the difference in character between the two strings (Yujian and Bo, 2007). As an example: while the Levenshtein distance of 'bat' and 'bat' is 0, the distance

between 'bat' and 'bot' is 1. This is particularly useful, as a low result has a higher chance of being a match, although it would be a mistake to presume that only results of zero are matches, as there are other factors to compute, such as plurality ('teacher' and 'teachers' have a distance of one) and, in some languages, gender (in Portuguese, both 'professor' and 'professora' mean 'teacher', but they have a distance of one). Nevertheless, even a value of zero can be a mismatch, as words may have multiple meanings, such as 'cook', which can be a verb or a person who prepares food. Because of this, it is paramount to go beyond string matches.

Another important factor to be taken into account, which can prove to be a valuable ally in determining the string matches, is context. Following the previous example, it is rather hard to uncover the meaning of the word 'cook' by itself, however if taken into account the context ('I cook every other day' and 'He is a cook'), it becomes much easier to determine it. With that in mind, the next step in the pipeline will be to prompt the user not only with the string match between the historical text's value and the chosen HISCO occupation, but also with contextual data from the text, which will come in the form of the sentence that the word was taken from. Through this, it is hoped that the user will have enough information to make the decision of whether the comparison is sound or not (Figure 1).

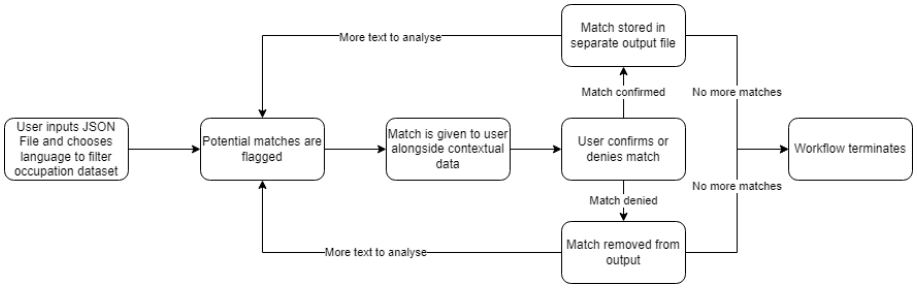

Fig. 1. Diagram representation of current workflow.

5 Result and Discussion

Finally, after the researcher goes through all matches found by the tool, the correct ones, alongside their context, will be stored in a separate file. This is for the user's future reference, and also to assist in whichever need they may have.

5.1 Evaluation setup

To evaluate the accuracy of the pipeline, it will be tested it on a dataset of structured biographical texts. The group will also have a domain expert from

the International Institute of Social History (IISG), Dr. Richard Zijdeman, use the workflow to provide feedback and get an understanding of how researchers would use it and what else is needed to make it a worthwhile research companion. The evaluation itself will be mostly centered around the user's experience with the tool, how it performs with the dataset, and how useful it is in non-testing scenarios.

As this project is mainly centered on developing a workflow, the final result will be an overview of both the performance of the workflow with relation to the aforementioned dataset as well as Dr. Zijdeman's remarks. Through this, it is hoped that the approach will be validated as a useful tool for researchers, and that it will also receive the necessary feedback to increment it for real-case scenarios.

After running the biographical dataset through the workflow, it was noticed that, overall, it was working as intended. Match candidates would be flagged by the system and prompted to the user's screen, providing useful information alongside it, such as: the word in question, the potential HISCO match and its code, the in-text contextualization of the word, the item's Levenshtein distance and the subject of the biography currently being analysed (Figure 2). This is an important early success, as it proves that the workflow can indeed perform its primary intended function.

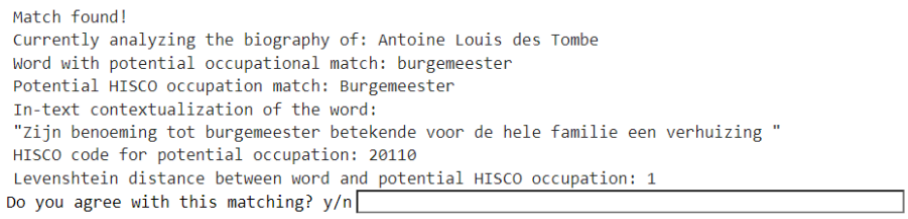

Fig. 2. Current workflow's output from any flagged potential match.

6 Result and Discussion

Nevertheless, a few challenges were also spotted during experimentation. Firstly, the group perceived a particularly slow run time to analyse entries, with potential matches taking up to 5.2 seconds to be declared. This by itself is already a fundamental drawback, given that any dataset inputted to the tool by a researcher may contain vast numbers of occupation matches, thus creating a dragged experience that may impair research. Another potential backlog is word repetition leading to unnecessary instances of the same token, as given that the tool does not currently have a system in place to handle repetitions, every single instance of a potential match will be flagged and displayed, which would lead to the

researcher having to read each of them. Moreover, the nature of the project's method of using Levenshtein distances also presents a problem of its own, as it will always flag tokens that are similar to an occupation even if it is known that such word means something else. An example would be the workflow tagging the surname 'Backer' because it has a distance of 1 to the Dutch occupation 'Bakker'.

Alongside this, a test demo was also presented to the domain expert, Dr. Richard Zijdeman, who provided some much-needed feedback. Firstly, he affirmed that the workflow was indeed both useful and important in his research field, as the automatic tagging of potential occupations is necessary for researchers planning to analyse a great quantity of texts. Dr. Zijdeman mentioned that while the use of Levenshtein distance was a good first step, future work would indeed benefit from the implementation of other NLP techniques that can further compare matches based on grammatical context and other cues prior to flagging them. Finally, the expert also suggested quality-of-life improvements for the output, to be added to a potential UI implementation, which included highlighting the matched token in the context, proving quicker input options (such as just pressing one button to move to the next match instead of the current two) and step-back and browse options, so that users may fix any potential mistakes they made to remove any errors from the generated results file. Dr. Zijdeman also suggested adding another layer to the comparison by also utilizing the HISCAM dataset, which includes social association comparisons to each occupation.

7 Conclusion

Given the four weeks that were provided for working on this project, the results have been satisfactory. The group has produced a program based on the workflow that effectively addresses the initial challenge of assigning HISCO occupations to unstructured texts. The program not only assigns the codes but also provides the user with context and the ability to independently analyse each instance. While some issues remain regarding the specifics of the workflow, these problems can be resolved by future work in the field and its customisable nature. For example, future users can choose to either use the Levenshtein distance as it is currently, change its specifics, or use a completely different NLP method. The demo presented to Dr. Zijdeman proved useful for the evaluation of the workflow and gave valuable insight into what researchers might expect from it, and what else they might need from the tool. Further work is required for the refinement of the workflow, but the one that has been created thus far is already useful for its intended purpose, and, although not as efficient as it could be, the end result will simplify the work that researchers have to do in order to study the past.

References

1. van Leeuwen, M.H.D., Maas, I. and Miles, A.G. (2004) "Creating a historical international standard classification of occupations an exercise in multinational interdisciplinary cooperation" Historical Methods: A Journal of Quantitative and Interdisciplinary History, 37(4), pp. 186–197. Available at: https://doi.org/10.3200/hmts.37.4.186-197.
2. Yujian, L. and Bo, L. (2007) "A normalized Levenshtein distance metric" IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), pp. 1091–1095. Available at: https://doi.org/10.1109/tpami.2007.1078.
3. Karsvall, Olof and Lars Borin. "SDHK meets NER: Linking Place Names with Medieval Charters and Historical Maps." Digital Humanities in the Nordic Countries Conference (2018).
4. Plu, J., Rizzo, G. and Troncy, R. (2015) "A hybrid approach for entity recognition and linking," Semantic Web Evaluation Challenges, pp. 28–39. Available at: https://doi.org/10.1007/978-3-319-25518-7_3.
5. NAPP (no date). What is Napp?, North Atlantic Population Project. Available at: https://www.nappdata.org/napp/intro.shtml. Last Accessed: February 2, 2023
6. Cambridge University (no date). The Cambridge Group for the history of Population and Social Structure, Cambridge, The Cambridge Group for the History of Population and Social Structure, Cambridge. Available at: https://www.campop.geog.cam.ac.uk/. Last Accessed: February 1, 2023.
7. Huysman, I., Kloek, E. (no date). Biografisch Woordenboek Gelderland. Het Biografisch Woordenboek Gelderland. Available at: https://www.biografischwoordenboekgelderland.nl/. Last Accessed: February 1, 2023.

Appendix

A Example of HISCO code

Majorgroup 2: Administrative and managerial workers
- 2.0 Legislative Officials and Governament Administrators
- 2.1 Managers
- 2.2 Supervisors, Foremen and Inspectors
- 2.11 General Managers
- 2.12 Production Managers
- 2.13 Sales Managers
- 2.12.10 Production Manager, Specialization Unknown
- 2.12.20 Production Manager (except Farm)
- 2.12.30 Farm Manager
- 2.12.40 Contractor