Crowdsourcing for Building Knowledge Graphs at Scale from the Vatican Archives (Discussion Paper)

Donatella Firmani, Paolo Merialdo, Elena Nieddu, Andrea Rossi, and Riccardo Torlone

Roma Tre University, Rome, Italy

Abstract. Project *In Codice Ratio* is developing tools to extract the knowledge contained in the ancient manuscripts of the Vatican Archives. The scarcity of datasets suitable for our setting has led us to rely on crowdsourcing in all phases of our project. In this paper we discuss our approaches for leveraging inexpensive non-expert workers to fruitfully perform labelling operations on challenging manuscripts. We describe the range of different tasks we are devising, as well as the corresponding priority and redundancy policies we are employing. We describe the datasets collected thus far and the corresponding results.

Keywords: Crowdsourcing · Knowledge Graphs · Digital Humanities

1 Introduction

Recent advancements in Knowledge Extraction from unstructured data have opened up new, exciting possibilities in digital humanities. These novel methodologies typically require large amounts of labelled data to reach satisfactory performances; nonetheless, when dealing with specific tasks in vertical domains, the lack of suitable datasets makes it unfeasible to use such approaches. As a matter of fact, in many scenarios the only available resources are raw and unlabeled. This has led several projects to rely on crowdsourcing techniques, having human operators, or *workers*, manually process and label part of the available data. By gathering the outputs yielded by the workers it is possible to build new, extensive datasets to train automatic systems on.

Our Digital Humanities project (ICR) *In Codice Ratio* [2, 6] falls within this scenario. ICR aims at extracting and harnessing the knowledge contained in the manuscripts of the Vatican Apostolic Archive (VAA). The VAA is one of the largest historical libraries in the world, with more than 85 linear kilometres of shelving. We are currently focusing on its "Vatican Registers" corpus, consisting of 43 parchment registers for a total of 18650 pages; these documents date to the XIII century, under the papacy of Honorius III, and contain the official correspondence of the Roman Curia, including legal or political letters from and to kings, sovereigns and institutions throughout Europe. The manuscripts are written in Chancery script (also called *Cancelleresca*) and they have been recently digitized in high definition images, but despite their historical relevance they have never been integrally transcribed so far.

ICR aims at *(i)* obtaining a textual transcription of the manuscripts in the Archive, and *(ii)* extracting entities, relations and facts from the obtained text to construct an extensive Knowledge Graph (KG). Performing automatic transcription from the high definition images of the document pages amounts to perform *Handwritten Text Recognition* (HTR), while extracting structured knowledge from the resulting transcription is a well known task of *Text Mining*.

Machine Learning (ML) approaches have been shown to reach state-of-the-art performances in both tasks. Since our documents are written in Medieval scripts and languages, no comprehensive datasets could be found for the adequately training ML models. In this paper we discuss how we are leveraging crowdsourcing methodologies to process, label and enrich data, in order to collect suitable datasets, we describe the crowdsourcing tasks we have devised and the assignment policies we have implemented.

Our pool of workers counts over 700 members, and it is entirely constituted by high-school students in the city of Rome, that have joined the project for free as a part of their work-related learning program. In addition to performing tasks, students also attended frontal lessons in a variety of topics, from paleography to machine learning, managed by both the engineering and humanities departments of Roma Tre University. As a consequence, they are provided with an opportunity of personal growth, and they receive guidance for their future studies and careers.

So far we have focused on the transcription phase: with the data labelled by our workers, we were able to successfully generate datasets large enough for training and developing very promising HTR models. These results allow us to start assigning to our workers the first batches of Text Mining tasks, with the goal of extracting knowledge from the transcribed text.

The rest of this paper is structured as follows. In Section 2 we report the HTR-oriented tasks we have employed so far, and we describe the obtained datasets and the corresponding transcription results. In Section 3 we discuss the typologies of tasks we have devised for text mining and knowledge extraction. We discuss related works in Section 4, and provide concluding remarks in Section 5.

2 Crowdsourcing for automatic transcription

The extraction of textual transcriptions from images is nowadays generally performed with automatic HTR models. As already mentioned in the Introduction, the lack of representative labelled datasets for the Papal Registers (i.e., similar script, abbreviations, and layout) makes it difficult to train a full-fledged HTR system without dealing with a very expensive data preparation effort.

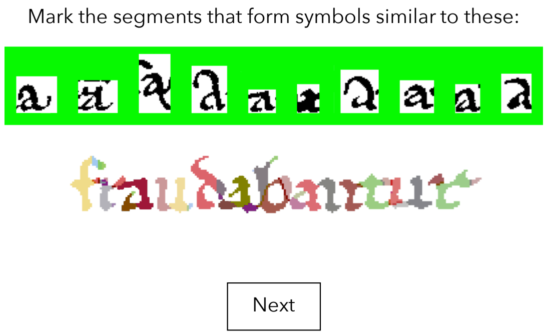
Mark the segments that form symbols similar to these:

Fig. 1. Example of labeling task: given the set of segments shown in different colors, the worker is asked to select the sets of segments corresponding to the required symbol.

We have thus employed a crowd-sourcing approach to build our own dataset with workers manually identifying portions from the original images. The obtained labelled segments are then gathered into datasets large enough to enable the training of HTR models. The registers are strikingly hard to decipher, and the extensive use of abbreviations makes them even less understandable. Therefore, unlike other transcription projects [5], in our case a non-expert worker is generally unable to directly transcribe entire paragraphs, lines or even words. Therefore we ask our workers to just recognize specific symbols inside words: the resulting task is more akin to pattern matching rather than to text transcription.

In a preparatory step, we apply to the images of the manuscripts pages (i.e. written facades) a pipeline of computer vision operations. We first run a custom binarization algorithm to transform them into black and white images. We then use the alternation between black and white sections to perform line segmentation and word segmentation. In each of the obtained word images, we analyse the upper and lower profile of the writing, and cut the word wherever the stroke is absent or exceedingly thin. This is an over-segmentation, because the obtained segments are finer-grained than the actual characters in the word.

We then use this setting to generate our crowdsourcing tasks for character recognition. In each task, we use as an input the oversegmented image of a word, and we ask our workers if a specific symbol is present within the word (e.g. symbol "⁊", that is a Tironian note meaning "et"). If it is, the worker should highlight the set of consecutive segments that belong to that character. In case the symbol occurs multiple times in the word, the worker can also highlight multiple sets of segments. In order to facilitate the selection, in our UI we display each segment in a different color, as shown in Figure 1, in which the worker is asked to identify the symbols for character 'a'.

Our approach to design the task was inspired by [11], in which the authors address the problem of extracting a number of items satisfying certain properties from a larger set, and design optimal algorithms for various settings in terms of cost and time. In our tasks, if the worker has identified the character in the word,

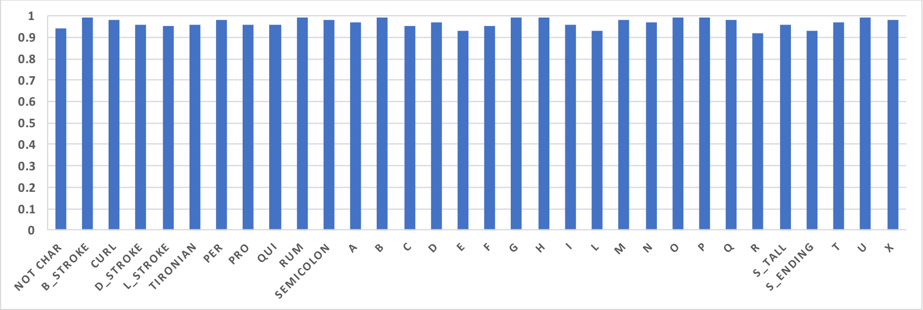

Fig. 2. Character-level predictive accuracy achieved by our HTR model on the crowd-based dataset. Class "NOT_CHAR" identifies inputs that do not correspond to any character in the alphabet. Classes from "B_STROKE" to "SEMICOLON" identify abbreviation symbols. Classes from "A" to "X" identify symbols that correspond to actual characters in the alphabet. note that there are two variants for character "S".

this correspond to a positive vote; otherwise it is a negative one. We assume that workers can make mistakes and employ redundancy to address this issue.

Specifically, we use the *rectangular* policy in [11] to aggregate votes and the *cost-optimal* paradigm in same paper to schedule tasks: for each class, we propose the same task to different workers until either positive or negative votes exceed respective thresholds N and M.[1] As soon as the crowd finds a pre-defined target amount of items for a class, we move to the next one sequentially. The same over-segmented word image can be re-used in multiple tasks, asking to recognize different symbols. At this regard, we prioritize the symbols that we lack the most in our datasets. The extracted sets of segments are used to generate labelled images, that are finally put in our transcription datasets. So far, with the contributions gathered from our workers, we were able to build various datasets, the latest - and largest - of which includes around 50k samples. We currently cover 32 classes; among these, 21 correspond to character symbols, 10 correspond to abbreviation symbols, and the remaining class identifies "not-character" elements, i.e. selections not corresponding to any valid item in the alphabet [2]. All our datasets are publicly available at our repository. [3]

We have then used the produced datasets to train our HTR model, based on Deep Convolutional Neural Networks. Our model is described in greater detail in [8]. As shown in Figure 2, character-level transcription results very promising: in each character class our prediction accuracy exceeds 90% - even in the "not-character" class. The average accuracy is 97%.

[1] We tried numerous configurations for these thresholds, and found that in our scenario using $N = M = 3$ is the best option.

[2] Labelled samples for "not-character" class are generated by combining or removing segments from samples belonging to the other classes

[3] http://www.inf.uniroma3.it/db/icr

The whole process that generates tasks, assigns them to the workers, and gathers the corresponding results, has been implemented as a Web Application hosted in the engineering department at Roma Tre University. We used a standard stack of technologies, using Java and the Spring Boot Framework, and storing data in a centralized relational DB.

3 Crowdsourcing for KG generation

After the correct transcription of a manuscript has been obtained, it is possible to extract its entities and relations, and to use them to populate a KG. In modern languages, these tasks are usually tackled by automatic classifiers based on machine learning and NLP approaches. However, similarly to the HTR step, when it comes to Medieval Latin texts the scarcity of labelled datasets makes it hard to train automatic models. Therefore, we are going to approach this phase too with crowd-based techniques. We are currently engineering two main categories of tasks for our workers: *Entity Classification* and *Relation Classification*. Each category may contain multiple types of tasks, with different requirements.

Entity Classification. We have verified that, in our manuscripts, it is possible to automatically select almost all named entities by just applying simple heuristics, e.g. searching for capitalized words. Nonetheless, human intervention is still required to classify the selected entities into semantic categories, such as "people", "locations", "organizations", "temporal data" etc.

In a basic Entity Classification task, each worker may receive the text of an entire transcribed page, with the named entities to classify already highlighted; the task would consist in labelling each entity with the corresponding class. Note that, once the correct class for an entity has been identified, it is possible to propagate it to all the other occurrences of the same entity across the manuscript. For this task, we assume that workers can make mistakes, and manage redundancy with a policy similar to the *Rectangular* strategy employed for the transcription task, extending it to the case in which multiple classes are available. The order in which tasks will be handed to the workers will depend on their priority value, that is, the number of entities it contains that have not been classified yet.

Entity Classification can be further optimized by devising automatic labeling rules. For instance, if a named entity lies in close proximity to special keywords, such as the title "episcopus" (meaning bishop), one can infer that it belongs to the "people" category. Modeling these rules requires knowledge of the specific syntactic patterns that occur most often in the manuscripts under analysis. Our workers, after they have become accustomed to the basic Entity Classification task, should ideally possess this kind of knowledge. Therefore, we plan to design advanced Entity Classification tasks based on rule building. We will provide students with the basics of rule-based logic, and divide them into groups. Each group will be receive, as an input, sets of several already labelled pages, from which they will be asked to build entity-classifying rules. We will evaluate the resulting clauses on standard metrics such as precision, recall and f-measure, and automatically select the combinations that maximize performances. We will finally apply the best performing mix of rules to pages in which the entities have not been classified yet.

Relation Classification. In order to build a KG, after extracting entities and their classes it is also necessary to identify the relations linking them. Once again, we will model basic Relation Classification tasks as tasks of manual labelling. In each task, the worker will receive as an input a single sentence containing at least two distinct entities, and will be asked to identify the relations occurring among such entities. The relation class can be chosen from a fixed vocabulary, that can be manually extended, if necessary. Once again, we will assume that workers can make mistakes, and make the crowd results more robust through redundancy, employing policies similar to those described for the basic Entity Classification task.

The *Entity Classification* and *Relation Classification* tasks will yield large amounts of labelled data that we will use to train automatic classifiers, similarly to the HTR phase. Once the KG has been populated, further facts can be inferred by leveraging recent Link Prediction techniques based on KG embedding (see [10] for a comparative analysis). In terms of general prerequisites, our workers in this phase will just require a very basic knowledge of Latin language - which is generally imparted in Italian high schools. Furthermore, in case of doubts or unclear transcriptions, they will have the possibility to interact with domain experts and paleographers. In order to assign KG extraction tasks to our workers, as well as to gather the corresponding results, we are planning to expand the web application employed for the transcription phase (see Section 2).

4 Related Works

Works related to ours can be roughly divided into Text Recognition and KG Generation works. We summarize these works in the following paragraphs, focusing on those based on crowdsourcing approaches, and refer the reader to the recent works [1, 7] for further discussion on crowdsourcing in the Text Recognition area.

Text Recognition Projects. As a matter of fact, relying on crowd-based solutions is not unusual in text recognition projects. In this area general purpose crowdsourcing platforms may not be flexible enough, therefore the most common approach involves building specialized applications following the entire lifecycle of each task. Project Transcribe Bentham [5], for instance, aims at crowdsourcing the entire transcription of Jeremy Bentham's unpublished works. The writing is in modern English and relatively easy to read, so they ask their workers to transcribe whole paragraphs or even pages. Project Read [9], on the other hand, have developed a mobile application in order to expand the set of potential volunteers; in each task the worker is asked to read aloud a handwritten text line, thus relying on speech dictation. Project Monk [14], finally, focuses on word search in handwritten manuscripts; they typically ask their workers to transcribe entire words, and use the resulting labeled images to train automatic systems for word spotting and other word-level operations.

Our work is fundamentally different from these projects as the tasks we propose to workers do not require any transcription skills: each task only requires to identify specific symbols or characters inside the image of a word and thus it can be solved by non-experts such as high-school students.

KG Generation Projects. Many open KGs nowadays are built in a collaborative fashion, meaning that they rely on contributions to add missing information or correct wrong pieces of the current contents. This collaborative effort can be seen as a very broad form of crowdsourcing, in which any user can temporarily become a volunteer worker. In this setting each task has a very loose formulation, with the worker herself choosing the entities, relations or facts to add or update. Examples of KGs built this way include FreeBase [3] and Wikidata [13].

While these approaches are viable for general purpose KGs, they do not work well when handling vertical topics that the workers may not have knowledge of, or when facts must actually be extracted from text. This has led researchers to devise more structured approaches for these scenarios.

In a completely different domain (drugs and their side effects), the work in [4] has a similar objective to ours, as they aim at building a KG from a collection of articles (from the PubMed archive) by leveraging a crowsourced dataset of annotations: in each task workers receive a sentence and a relation, and are asked whether the sentence actually conveys that relation or not. A similar approach is followed by [12], whose goal is to build a scientific KG by extracting entities, relations and facts from unstructured texts of research publications. Their approach is to perform fact extraction, integration, and analysis in a semi-superivised way. On the one hand, automatic tools are employed in each step; on the other, users are involved in all activities through visual interfaces that allow them to perform quality control, data enrichment and discovery.

5 Conclusions

In this paper we have described how crowdsourcing methodologies are allowing us to tackle the extraction of knowledge from Medieval handwritten manuscripts in our In Codice Ratio project.

We have shown that the lack of datasets suitable for our scenario affects both the transcription step and the text mining step. We have thus discussed how a pool of not expert workers can be put in condition to perform very simple tasks, yielding enough labeled data to train effective automatic systems. We have described in detail our strategies to assign tasks to specific workers, including the policies we employ to increase robustness to potential mistakes made by workers. We have reported our very promising results in character-level transcription, as well as our plans for the oncoming KG construction phase.

Among future work, we plan to inject data provenance methods in the whole process of knowledge extraction from ancient manuscripts, with the goal of improving the understanding of the results and simplifying the ability to trace errors back to the root cause.

Acknowledgments

Work funded by Regione Lazio LR 13/08 Project "In Codice Ratio" (14832).

References

1. Ammirati, S., Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E.: In codice ratio: Machine transcription of medieval manuscripts. In: Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings. pp. 185–192 (2019)
2. Ammirati, S., Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E., Rossi, A.: In codice ratio: Scalable transcription of historical handwritten documents. In: SEBD (2017)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data (2008)
4. Bravo, À., Li, T.S., Su, A.I., Good, B.M., Furlong, L.I.: Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text. Database (2016)
5. Causer, T., Tonra, J., Wallace, V.: Transcription maximized; expense minimized? crowdsourcing and editing the collected works of jeremy bentham. Literary and Linguistic Computing (2012)
6. Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E.: Towards knowledge discovery from the vatican secret archives. in codice ratio - episode 1: Machine transcription of the manuscripts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, ACM (2018)
7. Firmani, D., Merialdo, P., Maiorino, M.: In codice ratio: Scalable transcription of vatican registers. ERCIM News (111) (2017)
8. Firmani, D., Merialdo, P., Nieddu, E., Scardapane, S.: In codice ratio: Ocr of handwritten latin documents using deep convolutional networks. In: AI* CH@ AI* IA. pp. 9–16 (2017)
9. Granell, E., Martínez-Hinarejos, C.: Multimodal crowdsourcing for transcribing handwritten documents. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(2), 409–419 (Feb 2017). https://doi.org/10.1109/TASLP.2016.2634123
10. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., Barbosa, D.: Knowledge graph embedding for link prediction: A comparative analysis. CoRR (2020), https://arxiv.org/abs/2002.00819
11. Sarma, A.D., Parameswaran, A., Garcia-Molina, H., Halevy, A.: Crowd-powered find algorithms. In: 2014 IEEE 30th International Conference on Data Engineering. pp. 964–975. IEEE (2014)
12. Seifert, C., Granitzer, M., Höfler, P., Mutlu, B., Sabol, V., Schlegel, K., Bayerl, S., Stegmaier, F., Zwicklbauer, S., Kern, R.: Crowdsourcing fact extraction from scientific literature (2013)
13. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM (2014)
14. Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M., Schomaker, L.: Towards a digital infrastructure for illustrated handwritten archives. In: Digital Cultural Heritage, pp. 155–166. Springer (2018)