Extracting Relations from Ecclesiastical Cultural Heritage Texts

Giulia Cruciani
Università degli Studi di Messina / Messina, Italy
giulia.cruciani@studenti.unime.it

Abstract

Motivated by the increasing volume of data and the necessity of getting valuable insights, this research describes the process of extracting entities and relations from Italian texts in the context of ecclesiastical cultural heritage data. Named Entity Recognition (NER) and Relation Extraction (RE) are paramount tasks in Natural Language Processing. This paper presents a traditional methodology based on a two-step procedure: firstly, a custom model for Named Entity Recognition extracts entities from data, and then, a multi-input neural network model is trained to perform Relation Classification as a multi-label classification problem. Data are provided by IDS&Unitelm (technological partner of the IT Services and National Office for Ecclesiastical Cultural Heritage and Religious Buildings of CEI, the Italian Episcopal Conference) and concerns biographical texts of 9,982 entities of type person, which can be accessed by the online portal BeWeb. This approach aims to enhance the organization and accessibility of ecclesiastical cultural heritage data, offering deeper insights into historical biographical records.

1 Introduction

In the current landscape, the abundance of data is unprecedented; technological advancements, the Internet of Things (IoT), increasing connectivity, and digitalization are some factors that led to today's scenario. Yet, whereas methods of data collection continue to expand, the true value of this phenomenon is not related to the mere accumulation of information, but to the acquisition of meaningful insights. For this reason, there is a major focus on applying innovative techniques to all possible domains; the cultural heritage environment is also experimenting with big attention to exploit new possibilities. In the post-COVID era, the National Recovery and Resilience Plan (NRRP) has funded numerous endeavors to underline the importance of preserving and exploring cultural heritage, signaling a great moment for the combination of innovative approaches with domains traditionally less associated with these topics. This study positions itself among those initiatives, aiming to bridge this gap by employing advanced methodologies within the cultural heritage landscape. At the core of this pursuit, a paramount role is played by Knowledge Graphs, tools based on the concept of Knowledge Bases. Knowledge Graphs enable us to navigate the complex and intricate depths of data; for this reason, this work aims at finding the basic components of a knowledge graph, by extracting entities and relations from texts. The proposed methodology uses a traditional approach structured as a two-step process: first Named Entity Recognition (NER), key for pinpointing names, locations, and other text elements, and later sentence-level Relation Extraction, implemented as a multi-class classification task. A multi-input neural network model is built to leverage a labeled dataset of sentences and entity types. In Natural Language Processing, extracting semantic relationships from text is a very crucial task. This process concerns converting unstructured data (text) into structured. Relation Extraction (RE) can be achieved in several ways; one of them consists of setting the problem as a Classification task: Relation Classification (RC) (Zeng et al., 2014), (Zhang et al., 2017). Relation Classification has been approached using pattern-based (Suchanek et al., 2006), (Kambhatla, 2004), or kernel-based methods (Zhou et al., 2016). Early approaches make use of pipelines that identify entities and then classify relations between pairs using CNNs or LSTMs to capture sentence-level semantics (Zeng et al., 2014), (Zhou et al., 2016). Moreover, current methods for sentence-level Relation Extraction employ Transformer models (Yamada et al., 2020) like BERT (Devlin et al., 2019). In the context of art, especially in the last years, increasing importance is reserved for applications of digital tools

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 41–50
November 16, 2024.

mainly focused on art discovery or recommendation mechanisms (Gonzalez and Andrew, 2014). In Santini et al. (2022) the authors describe several techniques (such as entity recognition and linking, coreference resolution, time extraction, and artwork extraction) applied to Vasari's most important piece "The Lives of the Artists". In Chen et al. (2022), instead, NLP techniques have been applied to biographical texts of artists to achieve sentence-pair binary classification, hence connections among artists, without considering the type of the relations, conversely to what is developed in this study. The paper is organized as follows: sections 2 and 3 respectively illustrate the data and the methodology that were employed, starting from the formulation of the problem and continuing with the construction of models, and their evaluation (section 4). The last part shows conclusions and potential future applications.

2 Data

Data for this research were provided by the partner company [1] and can be accessed through an online website: BeWeb (https://beweb.chiesacattolica.it/). The portal allows seamless exploration of diverse databases acquired through the joint efforts of UNBCE [2] and the Italian dioceses (Russo, 2014). This collaborative initiative began in 1996 and has resulted in a census of diocesan and ecclesiastical assets, including archives, libraries, and museums. Throughout the years they were able to assemble a database that comprises over 5 million records, including 4 million historical and artistic assets, 66,000 places of worship, 1.5 million library assets, 6,800 archival collections, and 1,588 cultural institutions (Weston et al., 2017). Thanks to the help of approximately 3,000 experts, the projects are constantly reviewed and updated. Upon validation, the data is integrated into the national database; BeWeb, therefore, shows data validated from several inventories and facilitates cross-domain navigation of databases making use of the specific descriptive standards for each sector. The project's development involves two main elements: a dataset derived from sector-specific descriptive standards and a clustering system that interconnects terms referring to the same entity across various catalog databases, resulting in clusters identified by cross-domain aggregates: CEI [3] Authority File. Hence, the CEI Authority File (AF-CEI) can be seen as a centralized repository that integrates authority records from diverse cataloging domains through a clustering mechanism. Each domain interacts with the AF-CEI to establish, associate, or revise clusters. These resulting clusters undergo enrichment by designated reviewers, incorporating additional elements such as alternative nomenclature, biographical and historical annotations, images representing the entity, where present, interrelationships with other AF-CEI, and potential references to external web resources (Weston et al., 2017). This research focuses on the biographical notes of 9,982 CEI Authority Files categorized as Person.

3 Methodology

3.1 Problem Definition

The primary goal of this study is to extract structured triplets from Italian biographical texts. The study concerning entities and relations can be formulated as the exploration of triplets, such as $\{e_1, rel, e_2\}$, where e_1 and e_2 are respectively the first and second entity in the sentence, and rel is the relationship existing between the two. The proposed work, therefore, starts with an initial assumption: given that the text analysis involves biographical notes of entities categorized as "Person", and considering a sentence-level relation classification, the e_1 in the triplet is always the entity whose texts is being analyzed (namely the CEI Authority File), while e_2 changes each time a new entity is extracted from the sentences. For this reason, these triplets consist of the implicit entity that is the subject of the biographical text (referred to as the Authority File entity), the second entity, which is explicitly mentioned in the sentence and extracted using Named Entity Recognition (NER), and the relationship between the two, which is classified based on predefined relationship categories. The final output of this process will eventually look like: $\{Authority\_File, relationship, entity\}$.

[1] This project is the result of a PhD program financed by an EU scholarship initiative designed for Italian public school students, aimed at fostering innovation, and is a collaboration between University of Messina (Italy) and the local company IDS&Unitelm.

[2] UNBCE stands for Ufficio Nazionale Beni Culturali Ecclesiastici, namely National Office of Ecclesiastical Cultural Assets

[3] CEI stands for "Conferenza Episcopale Italiana", namely the assembly of bishops of the Catholic Church, responsible for coordinating and promoting the Church's activities and policies in Italy.

Authority File	Text Note
Franco Margari	Painter, graphic designer, video artist. He trained at the Accademia di Belle Arti of Rome. He began his artistic experience in the graphic field in the 1980s and specialized in engraving techniques; since the early 1990s he has also simultaneously dedicated himself to painting: in 1993 he began his exhibition activity in 1993. In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence. There are numerous presences in public and private collections. He lives and works in Florence.

Table 1: Biographical Text Note of Franco Margari

3.2 Data Segmentation and Entity Type

Named Entity Recognition (NER) plays a crucial role in this research; however, its standalone application is insufficient and some adjustments to the data are imperative to facilitate the sentence-level extraction of relationships. As mentioned above, the aim is to identify triplets where the Authority File itself denotes the first component, the relation is intrinsic in the semantics of the sentence and the third unit encompasses every other entity identified by the model. Consider Table 1 as an illustrative example.

The initial column denotes the name of the Authority File, specifically "Franco Margari", while the subsequent column contains the corresponding text data.[4] Firstly, texts undergo a segmentation into individual sentences. This segmentation is initiated at every punctuation dot occurrence, marking a new sentence's beginning. The NER model is employed to analyze each sentence within the data and as well as identifying the entities, it is asked to retain only sentences containing named entities, as depicted in Table 2.

Sentences such as "Painter, graphic designer, video artist", "He began his artistic experience in the graphic field in the 1980s and specialized in engraving techniques; since the early 1990s he has also simultaneously dedicated himself to painting: in 1993 he began his exhibition activity in 1993", "There are numerous presences in public and private collections" are excluded as the model did not detect any entities within them. Moreover, in cases where a single sentence contains multiple entities, the model generates an equivalent number of triplets. Table 2 shows an example: the second sentence ("In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence") encompasses two separate entities, namely "Fiorino d'Argento" and "Palazzo Vecchio in Florence". The model extracts both entities, resulting in the creation of not only three triplets - as suggested by the number of sentences - but rather four, accounting for the total number of named entities identified and extracted by the model. For this reason, the total number of potential triplets, hence the total number of relations, is obtained after applying the NER model to the dataset. In addition, while extracting the entities, the model is asked to create another field to account for the type of each entity. This augmentation is motivated by the fact that sentences may include multiple entities, some of which may correspond to distinct types of relations. Notably, the differentiation between entities often underlines the nature of these relations. Hence the augmentation with entity types can offer enhanced insights into the diverse nature of relations, facilitating a more comprehensive understanding of the associations and enabling a more granular analysis of the relationships within the data.

3.3 Identification of relations

The dataset used in this study is highly domain-oriented, consequently, the choice of multi-label classification for relation extraction relies on the fact that recognizing the topic of a sentence will likely reflect the nature of the relationship among the entities mentioned within that sentence. For this reason, four distinct categories were identified to represent all the potential relationships in the texts: Work/Study, Birth/Death/Travel, Kinship, and Ecclesiastical Titles. The classes were constructed intentionally to be broad and able to encompass similar relations. Table 3 shows the categories and relative explanations. After identifying the categories, relation classification is achieved with a multi-label classification model, part of supervised

[4]Texts are originally in Italian, but for the sake of understanding examples will show an English-translated version.

Authority File	Sentence-Level Note	Extracted Named Entity	Entity Type
Franco Margari	He trained at the Accademia di Belle Arti of Rome	Accademia di Belle Arti of Rome	Organization
Franco Margari	In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence	Fiorino d'Argento	Miscellaneous
Franco Margari	In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence	Palazzo Vecchio in Florence	Organization
Franco Margari	He lives and works in Florence	Florence	Location

Table 2: Named Entity Recognition on Texts

Relationship	Explanation	Examples
Work/Study	Connections expressing work or study relations.	"Franco Margari trained at the Accademia di Belle Arti of Rome."
Birth/Death/Travel	A relation between a Person and a Location, that represents the place where the Person was born, found dead, or traveled to.	"Cassiano Carpaneto died in 1998 and was buried in the Langasco cemetery."
Kinship	Relations showing a familiar bond, such as: "is son of", "is married to".	"Giovan Battista Del Tasso, son of Marco di Domenico."
Ecclesiastical Titles	Religious relationship.	"Piero Novati was a priest of the Diocese of Lodi."

Table 3: Classes for Relationships and Examples

machine learning. This branch of machine learning involves presenting the algorithm with input data along with the corresponding correct output so that the model can learn patterns and relationships between inputs and outputs. For this purpose, a sample of 1,000 sentences, was randomly selected and used for manual annotation. Table 4 shows an example of one annotated sentence.

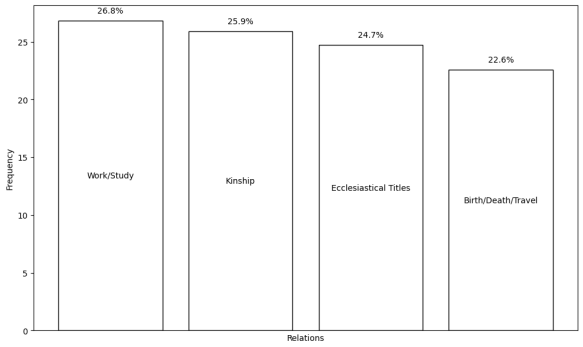

Figure 1: Distribution of relations in the first annotated dataset

Fig. 1 displays the distribution of the categories within the annotated dataset and shows a quite balanced division across all classes. As will be discussed later, the model is trained twice, once on the dataset containing 1,000 labeled sentences, and again on an augmented dataset that comprises 2,000 newly annotated sentences. Hence, the augmented dataset (composed of 3,000 labeled sentences) shows a different distribution for the categories, as depicted in Fig. 2, and a significative imbalance among classes. This problem is addressed later during the compiling of the model, when the optimizer, loss function, and custom metrics are defined.

3.4 Named Entity Recognition

Named Entity Recognition (NER) techniques aim at identifying significant elements such as names, locations, companies, or events within texts. While various methods exist for NER, this study focused on developing a custom pipeline using the publicly available SpaCy library. SpaCy is an open-source tool that employs machine learning models trained on extensive corpora to detect entities. It offers pre-trained models for different languages and domains. Specifically, the Italian version of

Authority File	Sentence	Extracted Named Entity	Entity Type	Label
Franco Margari	He trained at the Accademia di Belle Arti of Rome	Accademia di Belle Arti of Rome	Organization	Work/Study

Table 4: Annotation Example

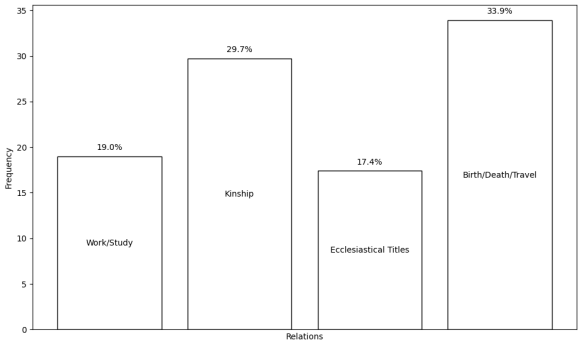

Figure 2: Distribution of relations in the second annotated dataset

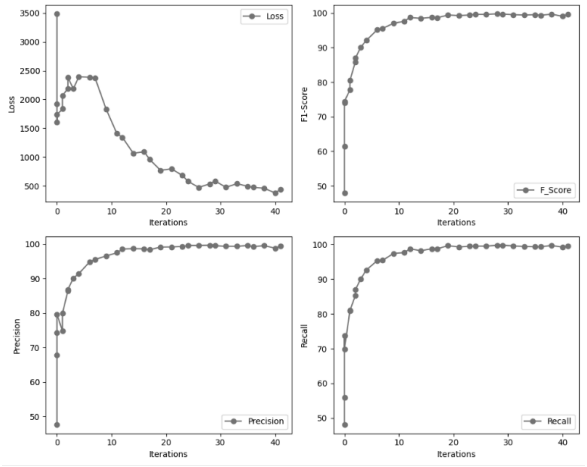

Figure 3: Evaluation Metrics for Named Entity Recognition

SpaCy classifies entities into four categories: person, location, organizations, and miscellaneous; the same notation was adopted in this application, with the only difference being that they were translated into Italian. However, given the domain-specific nature of our dataset, creating a custom model seemed the best choice. Training a custom pipeline using SpaCy is a straightforward process. Detailed guidelines and configuration resources can be found on SpaCy's training documentation (https://spacy.io/usage/training). For the model's training, a random sample was drawn from the original dataset, consisting of 1,000 biographic annotations, and was split into the standard 70% for training and 30% for testing. Using an online platform (https://tecoholic.github.io/ner-annotator/), the subset was manually annotated, considering four entity types, mirroring SpaCy's pre-trained NER model for Italian: Person, Organization, Location, and Miscellaneous. The training phase involved utilizing SpaCy's provided configuration files; following training, the model's performance was tested. Fig. 3 depicts evaluation metrics such as loss, precision, recall, and F1-score.

Loss values initially start quite high but gradually decrease over epochs, while all evaluation metrics (F1-score, Precision, and Recall) showcase upward trends over the iterations. The precision and recall values in Fig. 3 reflect only the identification of named entities, not their types. These metrics were calculated based on the test set, ensuring the model's performance was properly evaluated without influencing subsequent training iterations. The number of extracted entities is the following: 4243 Location, 2471 Miscellaneous, 11731 Organization, and 10651 Person. However, it is pertinent to note that these figures are distinct values. A higher number of entities was extracted, but they have been manipulated to account for instances where a singular entity might be referenced multiple times, albeit with slight variations in notation or expression. For this reason, whereas the sum of all distinct entities is 29,098, the total number of triplets resulting from NER is, instead, 65,289.

3.5 Relation Classification

This section introduces a recurrent neural network multi-label classification model constructed using TensorFlow, an open-source framework created by Google specifically for machine learning applications (Abadi et al., 2016); its workflow comprises defining an architecture, compiling, training, and finally evaluating the model. Fig. 4 shows the architecture of the model.

As mentioned previously, the model is constructed by defining two input layers: the sentences

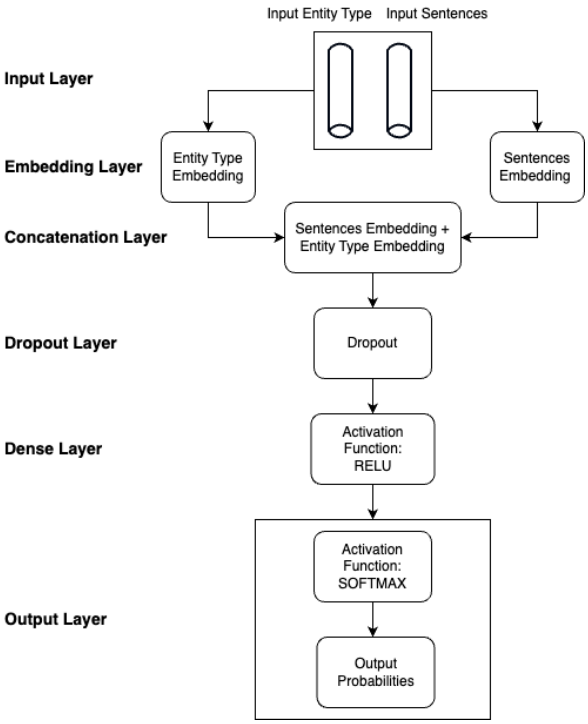

Figure 4: The model's architecture

and the type of the entities extracted in that specific sentence. The difference in the inputs defines a difference in the approach as well. Sentences, indeed, are passed through the Universal Sentence Encoder (Cer et al., 2018) which preprocesses and then encodes the sentences into embedded vectors. Since entity types, instead, are categorical data they are treated with the so-called one-hot encoding, namely a technique used to transform qualitative features into a format that can be easily comprehended by machine learning algorithms. In this specific case, there are four categories (Person, Location, Organization, Miscellaneous), and applying one-hot encoding results in a binary vector for each. For instance, the class "Person" might be represented as [1, 0, 0, 0] whereas "Location" might be represented as [0, 1, 0, 0] and so on. Each position in the vector corresponds to a category, and only one position has a value of 1, while the others are 0. After obtaining embedding vectors, the two inputs go through a concatenation layer, where they are merged to allow the model to learn from both sources of information simultaneously. This combined representation, then, is fed into a dropout layer, that behaves as a regularization process created to prevent overfitting in neural networks, by dropping a fraction of input units randomly. To that end, during the training a portion of inputs is randomly set to zero preventing the co-adaptation of neurons; this process encourages the network to learn from robust features that generalize better to unseen data. Given an input x (which in this case is represented by the output from the concatenation of sentence and entity type embeddings) the dropout layer randomly sets a fraction (p) of input units to zero, where p is the dropout rate. In this case, the dropout rate is set to 0.20. Next, a dense layer is introduced, establishing a fully connected hidden layer where the Rectified Linear Unit (ReLU) activation function is applied to the input tensor x. In this layer the input data is transformed, allowing the network to learn and extract higher-level representations and meaningful features from the merged inputs. Mathematically, a ReLU activation function can be defined as in (1); for a certain input x

$$f(x) = \max(0, x) \tag{1}$$

meaning that for any input x, the function will output the maximum of that input's value or zero. This essentially signifies that the function "activates" (returns a non-zero value) only when the input is positive, and otherwise, remains inactive (returns zero). ReLU is used in neural networks to introduce non-linearity, which enables the network to learn complex patterns and relationships in intricate data. The last dense layer in the model serves as the output layer; its purpose is to produce the final predictions, namely the classifications, based on the features learned from the previous steps. The activation function, in this case, is the SoftMax. Given a vector z of K real numbers (where K is the number of classes, four in this case), the SoftMax function computes the probability $p(y_i)$ for each class i as in (2).

$$p(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

The numerator in (2) calculates the exponential value of the ith element of z; this value is then normalized by dividing by the sum of all exponential values. This process forms a probability distribution through the function, that guarantees that the total sum of probabilities across all classes equals 1 and that allows the model to make predictions based on the class with the highest probability; the resulting values represent the likelihood of the input belonging to each class. In the compilation step, the model's configurations and op-

timizations are determined. For this architecture, the Adam optimizer (Kingma and Ba, 2014) was chosen, along with categorical cross-entropy as the loss function. This combination is commonly regarded as effective for training neural networks, particularly in multi-class classification problems. Adam optimizer is popular due to its effectiveness, adaptability, and efficiency in optimizing the training process for neural networks. The objective during training is to minimize the categorical cross-entropy loss by adjusting the model's weights and biases using the Adam optimizer. Additionally, specific metrics are defined to monitor both training and validation phases. In this case, the metrics (Precision, Recall, and F1-score) are designed for multi-class scenarios where class imbalance might exist, ensuring that the evaluation considers performance equally across all classes rather than being biased towards majority classes. The model is then trained following two separate phases. Initially, the labeled dataset comprises 1,000 manually annotated sentences, forming the initial training data for the model. The training duration is set to 20 epochs and the model undergoes iterative learning, gradually improving its ability to predict relations between entities within the provided sentences. Upon training the model, it is then applied to a new dataset containing 2,000 sentences for label prediction. Once the model classifies the relations, after a manual analysis where any incorrect predictions are rectified, the corrected ones are appended to the original annotated dataset. This process results in an augmented dataset useful to refine and improve the model. At this point, the model benefits from a larger corpus of labeled sentences and shows a more robust and accurate performance due to the increased diversity and quantity of annotated data. Table 5 illustrates the performance of the two models on the test set, namely a subset that the model does not see during the training phase and that consists of 30% of the annotated datasets.

As depicted in Table 5, the second model outperforms the first significantly in terms of loss, indicating lower errors in predictions, and demonstrates higher accuracy, achieving approximately 89% compared to the first model's 83%. There are also improvements concerning Balanced Recall, Precision, and F1-score implying enhanced ability in correctly identifying instances across different classes while minimizing false positives. Overall, the second model displays superior performance across all evaluation metrics.

Metrics	**First Model**	**Second Model**
Loss	0.47	0.33
Accuracy	0.83	0.89
Balanced Precision	0.85	0.88
Balanced Recall	0.80	0.85
Balanced F1-Score	0.82	0.86

Table 5: Test Sets Evaluation

4 Evaluation of Relationship Classification

To ensure the model grasps relationships effectively without merely memorizing the training data, the evaluation is carried out in two distinct ways. First, evaluation metrics are observed for both training and validation sets during the training process. The validation set is created using 30% of the training data and is used to assess how well the model generalizes to unseen examples. This step aids in identifying any signs of overfitting, where the model might excessively adapt to the training data and struggle to perform well on new data. By validating the model's performance on this subset, we ensure it can make accurate predictions beyond the examples it was trained on. Initially, the model starts with a loss of 0.9039 and an accuracy of approximately 61.8%, however, as training proceeds, the model shows notable improvements, with decreasing loss values and increasing accuracy, as well as balanced metrics such as Recall, Precision, and F1-score. These metrics exhibit positive trajectories across different classes. The model is further evaluated on a new subset composed of 250 manually annotated sentences. Out of these 250 sentences, 29 are wrongly predicted, representing an error rate of 11.6%. Table 6 provides the values for Accuracy, F1-score, Precision, and Recall. Finally, the complete dataset is fed into the model to classify the 65,289 relations. Figure 5 illustrates the final distribution of the relationships, where "Work/Study" is the most frequent class (40.3%), and "Kinship" is the least frequent (12.4%).

5 Results and Conclusion

The evaluation of the NER model demonstrates its growing capacity to accurately recognize entities within the provided text data, suggesting that the

Metrics	Values
Accuracy	0.884
Precision	0.90
Recall	0.884
F1-Score	0.888

Table 6: Evaluation of a manually annotated subset of 250 sentences

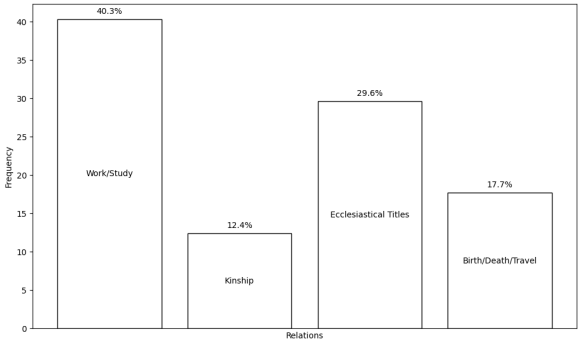

Figure 5: The distribution of all extracted relations across classes in the final and complete dataset

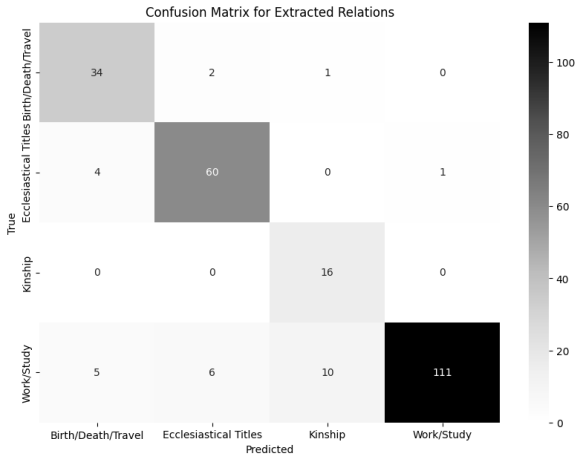

Figure 6: Confusion Matrix for the Extracted Relations

model effectively fine-tunes its parameters to better capture these entities, leading to significantly improved performance. Similarly, the results from the evaluation of the Relation Classification model reveal consistent and progressive enhancements throughout the training process. Beginning with a moderate initial accuracy of 61.8%, the model exhibits significant improvements over time. By the 20th epoch, it achieves a loss of 0.2037 and an accuracy of approximately 92.1%, indicating substantial progress from the initial stages. The declining loss values reflect a reduction in prediction errors, while the positive trends in Recall, Precision, and F1-score demonstrate the model's increasing ability to correctly identify relationships across various classes. The steady improvement in these metrics highlights the model's enhanced capability to capture true positives while minimizing false positives and false negatives. The final evaluation on a new subset of 250 manually annotated sentences, which resulted in an error rate of 11.6%, further supports the model's effectiveness in classification tasks. The high overall accuracy and balanced metrics indicate robust performance. Figure 6 shows that the model excelled in classifying "Work/Study" and "Ecclesiastical Title," achieving high true positive counts of 111 and 60, respectively. However, there are notable misclassifications, particularly in the "Birth/Death/Travel" and "Kinship" categories. These issues may be attributed to the class imbalances highlighted in Figure 5, suggesting areas for improvement in distinguishing these categories. Finally, the classification of 65,289 relations within the complete dataset reveals a meaningful distribution of relationships, with "Work/Study" as the most frequent and "Kinship" as the least common. These findings not only underline the model's ability to accurately recognize and categorize different relationships within the text data, but also align with the patterns that are evident from simply reading the texts, suggesting that the model is well-optimized for the task at hand.

Cultural heritage has a huge role in shaping national and cultural identity, promoting tourism, and attracting visitors; on the other hand, digital technology enables cultural institutions to offer global access to their collections, including items rarely exhibited due to space constraints or fragility (Sporleder, 2010). Therefore, in the era of rapid digital transformation where the volume of data has surged dramatically, the challenge is not merely increasing the number of information but transforming it into valuable insights. Whereas digitalization has initiated a new era, the true potential of data can only be unlocked through innovative methodologies that transcend conventional approaches and offer new opportunities for exploration. Within this context, the endeavor to employ new tools holds profound significance, because whereas traditional methodologies, reliant on manual studies of literature and documents by academics, have laid the groundwork, they may not have fully revealed all the potentiality of knowledge. This work wants to be an opportunity to bridge this gap, leveraging nat-

ural language methodologies to delve deeper into these big collections of data. The present study proposes a two-step procedure to extract entities and relations from text data concerning ecclesiastical cultural heritage. A custom pipeline for Named Entity Recognition and a multi-classification model are exploited with manually annotated data, to extract and classify 65,289 relations. A differentiated evaluation process is performed to assess the significance of the proposed methodology. For future work, we intend to train the model using a larger corpus of annotated data to enhance the classification performance. Finally, the ultimate goal of this study is to transform unstructured text into structured data to create a graph database, specifically a Knowledge Graph. This Knowledge Graph will serve as a valuable tool for operators studying extensive texts, enabling them to uncover new connections that may not have previously emerged. Additionally, the extracted information will be integrated into a recommender system designed to assist users on the BeWeb platform. BeWeb is a website that provides access to a vast amount of data concerning ecclesiastical cultural heritage. By offering personalized suggestions, the recommender system aims to enhance user engagement and foster the creation of communities around shared interests.

6 Limitations

This research presents some limitations that can be viewed as opportunities for future studies. One key aspect is the number of annotators involved in the annotation process. Increasing the number of annotators could lead to more standardized annotations and, ultimately, more accurate predictions, thereby enhancing the overall quality of the research. Moreover, as suggested by the reviewers, it would be interesting to explore the differences in predictions between sentences containing only one entity and those with multiple entities, possibly by incorporating entity name embeddings. This analysis could provide valuable insights into how the model handles different situations. This research direction has the potential to yield meaningful results, further increasing the accuracy and reliability of the described system. Finally, it is important to emphasize that this research is focused solely on the Italian language; however, a similar approach could be implemented for texts of other languages.

Acknowledgments

This work was sponsored by the Italian Research Program PON (Programma Operativo Nazionale) "Ricerca e Innovazione" 2014-2020 - Azione IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione" - Università degli Studi di Messina - Dottorato di Ricerca in "Economics, Management and Statistics" - XXXVII ciclo. This study was supported by IDS&Unitelm; a special thank goes to Angelo Cingari, Maria Teresa Rizzo, Massimo Currò, Nuccio Castorina, Silvia Tichetti, and finally the supervisor of the project Professor Edoardo Otranto. Gratitude is extended to the reviewers for their time and effort in reviewing the manuscript; their valuable comments and suggestions have significantly contributed to enhancing its quality.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, and X. ... Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *Preprint*, arXiv:1603.04467.

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Q. Chen, M. El-Mennaoui, A. Fosset, A. Rebei, H. Cao, C. O'Beirne, S. Shevchenko, and M. Rosenbaum. 2022. Towards mapping the contemporary art world with artlm: an art-specific nlp model.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

D. Gonzalez and L. Andrew. 2014. Rethinking recommendations: Digital tools for art discovery.

N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, page 22–es, USA. Association for Computational Linguistics.

D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

S. Russo. 2014. Beweb. the cross portal of cultural ecclesiastical heritage. *JLIS.it*, 5(2):147–157.

C. Santini, M. Tan, O. Bruns, T. Tietz, E. Posthumus, and H. Sack. 2022. Knowledge extraction for art history: the case of vasari's the lives of the artists (1568).

C. Sporleder. 2010. Natural language processing for cultural heritage domains. *Language and Linguistics Compass*, 4(9):750–768.

Fabian Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 712–717.

P. G. Weston, F. D'Agnelli, S. Tichetti, C. Guerrieri, and M. T. Rizzo. 2017. Authority data and cross-domain intersection within aggregation portals. the case of beweb. *JLIS.it*, 8(1):139–154.

I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.