Advancing Community Directories: Leveraging LLMs for Automated Extraction in MARC Standard Venue Availability Notes

Mostafa Didar Mahdi
Thushari Atapattu
Menasha Thilakaratne
School of Computer and Mathematical Sciences
University of Adelaide
mostafadidar.mahdi@student.adelaide.edu.au
thushari.atapattu@adelaide.edu.au
menasha.thilakaratne@adelaide.edu.au

Abstract

This paper addresses the challenge of efficiently managing and accessing community service information, specifically focusing on venue hire details within the SAcommunity directory. By leveraging Large Language Models (LLMs), particularly the RoBERTa transformer model, we developed an automated system to extract and structure venue availability information according to MARC (Machine-Readable Cataloging) standards. Our approach involved fine-tuning the RoBERTa model on a dataset of community service descriptions, enabling it to identify and categorize key elements such as facility names, capacities, equipment availability, and accessibility features. The model was then applied to process unstructured text data from the SAcommunity database, automatically extracting relevant information and organizing it into standardized fields. The results demonstrate the effectiveness of this method in transforming free-text summaries into structured, MARC-compliant data. This automation not only significantly reduces the time and effort required for data entry and categorization but also enhances the accessibility and usability of community information.

1 Introduction

In the realm of digital information management, the seamless transition between unstructured text and structured data remains a case of efficiency and utility. Particularly within the context of event management where details range from facilities and capacities to rental fees and accommodations for the disabled, the need for sophisticated data extraction methods is paramount. This work proposes to enhance community directories by leveraging state-of-the-art deep learning models for automated data extraction.

Community directories are centralized databases or listings that provide information about local services, organizations, and resources available to residents within a specific community or region, in our case South Australia. The work focuses on converting open-field free-text summaries of community service information into structured, MARC (Machine-Readable Cataloging) standard-compliant data elements by the Library of Congress (Library of Congress, 2000), specifically targeting "venue availability" for meeting rooms and facilities. We have chosen this aspect due to its high demand, as indicated by significant searches in Google Analytics for "Venue Hire". Our strategy involves not only meeting the current demand but also laying the groundwork for creating truly closed fields in the future. We aim to address the gap in effectively utilizing unstructured text describing venue hire capabilities for SAcommunity, a free online community service established in 1981 and supported by the Government of South Australia. The work involves extracting information from open fields, specifically focusing on the Physical Description Fields (MARC21 3XX, 2000) section of MARC 21 Community Information library.

The primary challenges in extracting structured venue hire information from unstructured text include:

- Variability in Descriptions: Venue hire information is presented in diverse formats, with varying levels of detail and terminology.
- Complexity of Information: Details about venue hire encompass multiple dimensions—physical attributes, services, pricing, and policies, each requiring nuanced understanding.
- Need for Standardization: Extracting information that aligns with the MARC-21 format necessitates a methodological approach to categorize and structure data.

Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association, pages 118–129
December 2-4, 2024

- Lack of Labeled Data: There was no labeled data in the dataset that consisted of ground truth values, so we had to change our initial approach and label a subset of the data manually.

This study on automated extraction of venue availability information using a RoBERTa-based model demonstrated promising outcomes. The model achieved a peak accuracy of 0.78 on the test set, with balanced precision and recall scores of approximately 0.65 and 0.70, respectively. The F1 score reached 0.65, indicating a good balance between precision and recall. These results suggest that the model effectively learned to extract and classify venue availability information from unstructured text, potentially streamlining the process of updating and maintaining community information directories.

This research is critical because it tackles a prevalent issue in digital librarianship and information management: the efficient utilization of unstructured text. By developing a method to extract structured data from free-form text, the research supports better data management practices, improves accessibility, and enhances decision-making processes within community and event management sectors. It also contributes to the broader field of information science by integrating cutting-edge NLP technologies to solve real-world problems.

2 Related Works

Recent advancements in NLP, particularly in Named Entity Recognition (NER) and text classification, form the foundation of this research. Transformer models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have shown significant potential in understanding context and extracting relevant information from text.

For instance, Jehangir et al. (2023) provide a comprehensive survey across various domains, emphasizing the pivotal role of Deep Learning in enhancing NER capabilities. Lample et al. (2016) introduce innovative neural architectures that integrate character-based and distributional word representations, offering improved model sensitivity to both orthographic features and word context. Meanwhile, Dagdelen et al. (2024) propose a domain-specific approach to extracting relational information from scientific texts by fine-tuning GPT-3 models, thereby enabling non-NLP experts to generate structured datasets for specialized tasks. Shen et al. (2018) address large labeled data requirements in NER by combining deep learning with active learning, introducing a CNN-CNN-LSTM architecture for incremental training.

In medical NER, Cui et al. (2023) present the SoftLexicon-RoBERTa-BiLSTM-CRF model for Chinese electronic medical records, while Chuang et al. (2023) explore GPT-J for prompt generation in periodontal diagnosis extraction. Wu et al. (2021) propose the Ra-RC model for Chinese clinical NER, combining radical features with deep learning.

For legal NER, Zhang et al. (2023) introduce a method using RoBERTa and GlobalPointer for Chinese legal documents, fusing character-level and word-level features to identify nested entities.

Addressing cross-lingual challenges, Chan et al. (2023) investigate task learning and data augmentation for NER in low-resource Filipino, highlighting transfer learning's importance.

Alshammari and Alanazi (2021) provide a comprehensive study of transformer-based models (BERT, ALBERT, XLM-RoBERTa) for NER using the CoNLL dataset, emphasizing preprocessing and fine-tuning.

In the realm of active learning, Chen et al. (2015) examine strategies for clinical NER, while Le et al. (2023) address train-test distribution misalignment using feature matching. Lastly, Tchoua et al. (2019) explore active learning for NER in scientific texts, developing the polyNER system to reduce dependency on large annotated datasets in polymer science.

These studies demonstrate the ongoing efforts to enhance NER performance across various domains and languages, often focusing on reducing annotation requirements and improving efficiency in specialized fields.

3 Methodology

3.1 Data Collection and Annotation

SAcommunity Database (CIVICRM-DB): The SAcommunity database provides a comprehensive report of all the listed organizations, their names, addresses, contact details, website urls, emails, services, offered, venue hire information, etc. The variables involved in our study are as follows: Organization Name, Organization ID, Subject ID (a unique identifier denoting the subject category of the organization), Venue Hire Information (an open text field containing venue hire details), Comments

(an open text field with additional information about the organization), Services (listing the services provided by the organization), and Subject (indicating the subject category under which the organization falls).

SAcommunity Subject (Subject-DB): The Subject-DB contains subjects with a subject ID. This study selected the subjects that has higher correlation to venue hiring capabilities (e.g. halls for hire, community facilities, community centers etc.). The full list of subjects used in this study is shown in table 3 in the appendix section. We performed an SQL inner join (figure 4 in appendix) to combine both the databases and consolidate a final dataset.

3.2 Handling Unlabeled Data and Data Annotation

An innovative solution to the challenge of limited labeled data for training our NER model is the integration of active learning strategies (Ren et al., 2021). This approach trains our baseline NER model on a small labeled set, uses it to predict on unlabeled data, and then has humans label the most uncertain predictions, repeating the cycle to iteratively enhance model performance. We use Doccano, an open-source tool, for manual annotation, supporting active learning by labeling key samples. Doccano uses JSON Lines format for their data types, We log the entire study, including runs, using Weights & Biases, a platform for tracking and visualizing machine learning experiments.

3.3 Pre-processing

An effective NER system requires well-prepared data that helps the model learn to recognize and categorize entities accurately. The proposed preprocessing steps are designed to enhance the dataset's quality, ensuring optimal model performance.

Custom Entity Patterns Recognition: Regular expressions are employed to identify and pre-tag recurring patterns such as phone numbers and venue capacities. This initial structuring facilitates the model's ability to learn from consistent entity representations.

Text Normalization: Text normalization involves converting all text data to a standardized format. It is essential to consider the NER task's sensitivity to proper nouns and maintain the original case where necessary, as it may carry significant meaning for entity recognition.

Preprocessing Text Data: The preprocessing stage addresses several key challenges in the dataset. Special characters within entities (e.g., "Hall/Clubrooms") are handled through established rules that guide the tokenizer to treat such instances as single entities. URLs are removed from the dataset, unless they are integral to entity information, such as when specifically mentioned in a venue's contact details. Numeric data, including phone numbers and capacity figures, are preserved during tokenization to maintain their entity status, as NER often requires the identification of numeric entities.

Entity Consolidation: To address variations in referring to the same concept, such as "Hall for hire" versus "Hall/Clubrooms for hire," we advise consolidating these variations into a singular representation. This consolidation enhances the model's ability to recognize and classify entities consistently (Phan et al., 2023).

IOB Tagging: RoBERTa, like other transformer models, processes text at the token level. IOB Tagging allows us to assign a label to each token, enabling the model to perform fine-grained classification at the token level. It's like giving RoBERTa a special pair of glasses that help it see the structure of information in text. By marking each word as the Beginning, Inside, or Outside of an entity, we're essentially teaching RoBERTa to recognize patterns in how venue information is described. This approach is particularly useful for our work because venue details often span multiple words. For example, "can seat 100 people" might all be part of the "capacity" entity. IOB tagging helps RoBERTa understand where each piece of information starts and ends, making it much more accurate in extracting the specific details we need about venues. The process can be likened to equipping the model with the ability to differentiate and categorize various types of information, similar to how one might assign distinct colors to different data categories. This approach enhances the precision and reliability of information extraction, enabling more accurate identification and classification of relevant entities.

We have used advanced NLP libraries like spaCy to streamline various pre-processing tasks, including tokenization and initial entity tagging, which have proven essential in creating accurately labeled datasets for model training.

We conducted a manual review during preprocessing to ensure entities were accurately la-

beled, safeguarding data integrity and preventing errors that could impact model training.

The 14,000 entries were divided into training (80%), validation (10%), and test (10%) sets, with each set undergoing the same pre-processing and review steps to ensure compatibility with the RoBERTa model.

3.4 Custom NER Model (Finetuning RoBERTa)

RoBERTa (Robustly Optimized BERT Approach) enhances the BERT language model while maintaining its core transformer-based architecture. Key modifications include dynamic masking, removal of Next Sentence Prediction, larger mini-batches and learning rates, and processing of longer sequences. It uses byte-level Byte-Pair Encoding with a 50,000 subword vocabulary. RoBERTa's training is more extensive, utilizing more data and computational resources. It offers both base (12 layers, 768 hidden size) and large (24 layers, 1024 hidden size) configurations. These enhancements result in a more robust model with state-of-the-art performance in various natural language understanding tasks. Detailed chart of the hyperparameters of our model is shown in table 1. The way our custom NER model works is as follows:

- The input text is fed into the tokenizer.
- Each sequence starts with a [CLS] token, representing the special classification token.
- The input is transformed into numerical representations called vector embeddings.
- The final hidden vector of the model begins with the final special [CLS] token.
- This token outputs the prediction after normalization by the softmax layer.
- This architecture, also visualized in figure 15 in appendix, allows RoBERTa to capture complex contextual relationships in the text, making it well-suited for our NER task.

Inference: After training and validating the RoBERTa model, we proceeded to the inference stage, where we applied the model to extract venue availability information from previously unseen community service directory entries. This phase was crucial in demonstrating the practical applicability of my approach. More details are provided in figure 1.

To begin the inference process, we first preprocessed the new text entries using the same pipeline developed during the training phase. This ensured consistency in how the data was presented to the model. Each entry was tokenized and encoded using the RobertaTokenizerFast, maintaining the format the model was trained on.

We then passed these preprocessed entries through the trained model. The model output predictions for each token, classifying them according to the IOB tagging scheme we had established. These predictions corresponded to various aspects of venue availability such as capacity, equipment available, and rental fees.

Post-processing: This was a critical step in making the model's output useful. We developed a script to convert the IOB-tagged output back into meaningful chunks of information. For example, consecutive tokens tagged as "B-CAPACITY" and "I-CAPACITY" were combined to form complete capacity descriptions.

One of the most challenging and rewarding aspects of this stage was aligning the extracted information with MARC standards. We mapped the extracted entities to corresponding MARC fields, ensuring that the output could be easily integrated into existing library and information management systems. For instance, information about equipment availability was mapped to the relevant MARC field for facility information.

To evaluate the model's performance on this unseen data, we calculated accuracy, precision, recall, and F1 scores, comparing the model's extractions against a small set of manually annotated entries. This gave me a realistic picture of how well the model would perform in a real-world setting.

The inference stage not only validated the effectiveness of my approach but also highlighted areas for future improvement. It demonstrated the potential of using advanced NLP techniques to automate the extraction of structured information from community service directories, paving the way for more efficient and standardized data management practices in this domain.

3.5 Integration of Active Learning

The research incorporates an active learning loop to iteratively enhance the NER model's performance. Starting with a manually annotated subset, the model predicts entities on unlabeled data, identifying instances of uncertainty. These uncertain predictions, determined by evaluating the model's

Hyperparameter Category	Details
Model Configuration	RoBERTa (base model)
Hyperparameters	Batch size: 16 Epochs: 50 Learning rate: 0.00012 (dynamic)
Training Configuration	Optimizer: AdamW Learning rate scheduler: Cosine with warmup TrainingArguments: Set up
Training Process	Framework: Hugging Face's Trainer Custom Metrics: Precision, Recall, F1, Accuracy Training Duration: 50 Epochs Logging: Weights and Biases

Table 1: Hyperparameters and Training Configuration

confidence, are then selected for manual annotation using Doccano. The model is subsequently retrained with the newly labeled data, refining its performance through iterative cycles. Key considerations in this process include defining appropriate stopping criteria, ensuring diversity in sample selection to avoid bias, and utilizing efficient annotation tools. This approach significantly improves model accuracy and efficiency by focusing annotation efforts on the most informative samples.

4 Results and Discussion

The results of the model training and evaluation are presented across three sets: Training, Validation, and Test.

4.1 Training Set Results

Loss: Started around 2.5-3.0 and decreased to near 0 as shown in figure 6 in appendix. Showed a smooth downward trend, indicating good learning progress.

Learning Rate: Followed a typical warmup and decay pattern. Peaked at approximately 0.00012 and gradually decreased.

Gradient Normalization: Showed some fluctuation, with extreme spikes indicating potential instability in the training process as shown in figure 2. This suggests room for improvement in the training process, possibly through implementing gradient clipping, adjusting the learning rate, or using more advanced optimization techniques.

4.2 Validation Set Results

Accuracy: Highest value: approximately 0.75 as shown in figure 9 in appendix. Demonstrated consistent improvement across epochs.

Loss: Started high (around 4.5) and decreased to approximately 1.2 as shown in figure 10. Indicated good learning progress.

Precision and Recall: Both metrics peaked around 0.6. Recall showed more stability compared to precision (figures 7 and 8 in appendix).

F1 Score: Peak performance at around approximately 0.59 shown on figure 3. Showed fluctuations but maintained an overall upward trend.

4.3 Test Set Results

Accuracy: Best performance at approximately 0.78 as shown in figure 13 in appendix. The graph demonstrated a steady improvement trend.

Loss: Lowest loss: approximately 1.2. Showed a decreasing trend across runs, indicating better model fit.

More information on precision, recall and F1 score is described in table 2.

4.4 Interpretation of Results

The application of LLMs, specifically the RoBERTa transformer, for automated extraction of venue availability information in MARC standard format represents a significant advancement in community information management. This discussion will delve into the implications of our results, limitations of our work and propose future directions for research and application.

Model Performance: The RoBERTa-based model demonstrated promising results in identifying and categorizing relevant information from unstructured text. The best performance achieved an accuracy of approximately 0.78 on the test set, with F1 scores around 0.65. These results indicate that the model has learned to extract and classify

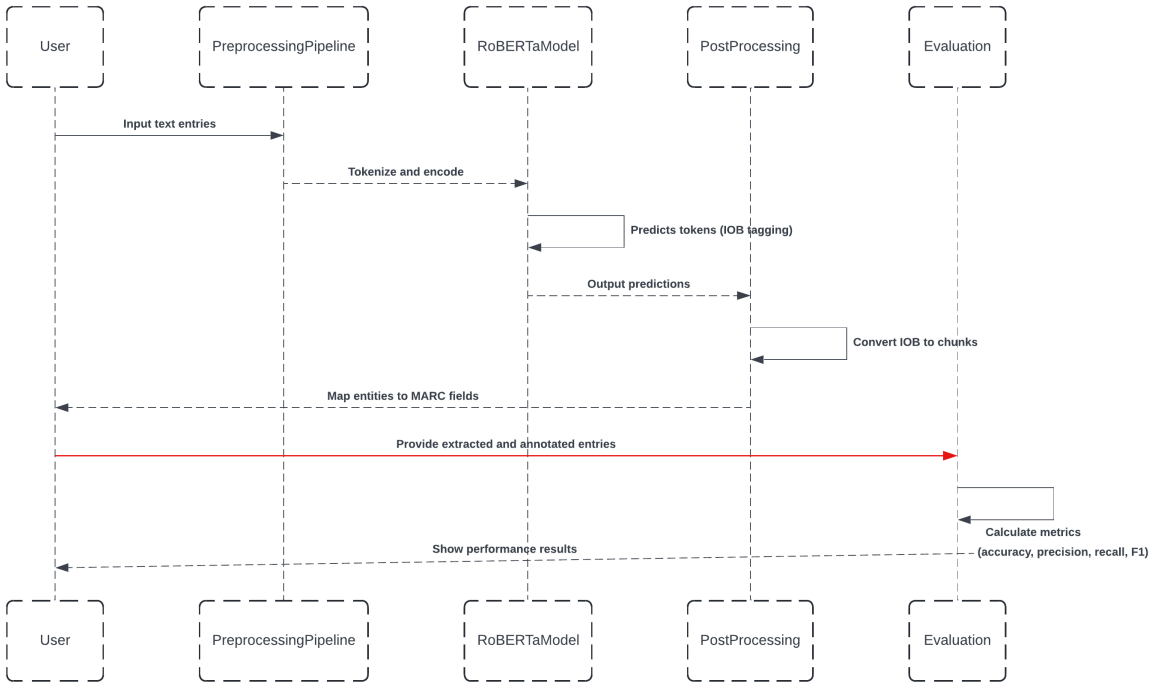

Figure 1: Shows the sequence diagram of how the system operates.

Metric	Best Performance	Date	Figure
Accuracy	0.78	21-07-2024	Figure 13 in Appendix
Recall	0.70	21-07-2024	Figure 11 in Appendix
Precision	0.65	21-07-2024	Figure 12 in Appendix
F1 Score	0.65	21-07-2024	Figure 14 in Appendix

Table 2: Best performance metrics for the NER model for Test Set Data. All metrics showed gradual improvement across runs, with the best performance achieved on 21-07-2024.

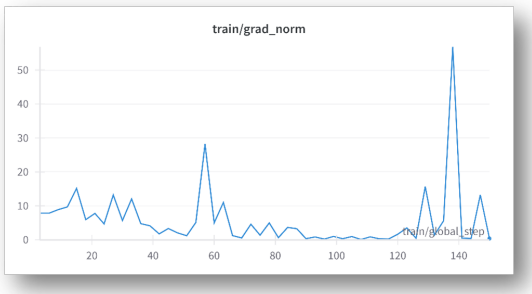

Figure 2: Shows a graph of gradient normalization on the training set.

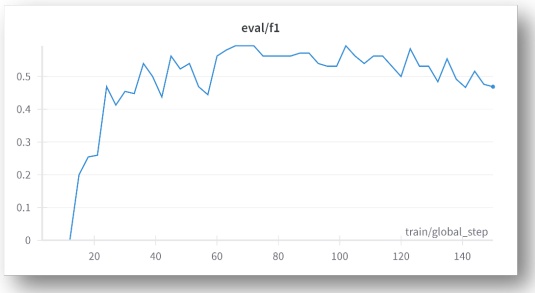

Figure 3: Shows an evaluation set f1 curve.

venue availability information with a reasonable degree of reliability.

The consistent improvement in performance metrics across training runs suggests that our iterative approach to model development was effective. The gradual increase in accuracy, precision, and recall indicates that the model's ability to identify relevant information improved over time, likely due to refinements in the training process and data preparation.

However, the gap between training and evaluation loss suggests some degree of overfitting. While not severe, this indicates that there's room for im-

provement in the model's ability to generalize to new, unseen data. This challenge is common in NLP tasks, especially when dealing with domain-specific information like venue availability.

Balanced Precision and Recall: Similar values for precision and recall (both around 0.65-0.70) indicate a balanced model performance. This balance is crucial for the practical application of the model. High recall (0.70) suggests that the model is effective at identifying relevant information about venue availability. This is important for ensuring that critical details about facilities are not missed. The precision of 0.65 indicates that when the model identifies information as relevant, it is correct about 65% of the time. While there is room for improvement, this level of precision is promising for an initial implementation.

The balanced performance suggests that the model is equally capable of identifying relevant information (recall) and avoiding false positives (precision). This balance is particularly important in the context of community information management, where both completeness and accuracy of information are crucial.

4.5 Implications for Community Information Management

Improved Data Standardization: By automating the extraction and structuring of venue availability information according to MARC standards, this research contributes significantly to data standardization efforts in community information management. Standardization has several important implications:

- **Interoperability:** MARC-compliant data can be easily shared and integrated across different systems and organizations, potentially leading to more comprehensive and accessible community information networks.
- **Improved Search and Retrieval:** Standardized data structures enable more efficient and accurate information retrieval, benefiting both information managers and end-users seeking venue information.
- **Data Quality:** Automated extraction can help maintain consistency in how venue information is recorded, potentially reducing errors and inconsistencies that can occur with manual data entry.

Efficiency Gains: The automation of information extraction has the potential to significantly streamline the process of updating and maintaining community information directories:

- **Time Savings:** Manual extraction and categorization of venue information from free-text descriptions is time-consuming. Automation can dramatically reduce the time required for these tasks.
- **Resource Allocation:** By reducing the manual effort required for data entry and categorization, organizations can reallocate human resources to higher-value tasks such as community engagement and service improvement.
- **Scalability:** As the volume of community information grows, automated systems can handle increased data loads more efficiently than manual processes.

Enhanced Accessibility and User Experience: Structuring venue availability information in a standardized format has the potential to greatly enhance the accessibility and usability of this information:

- **Improved Search Functionality:** Structured data enables more advanced search capabilities, allowing users to filter and find venues based on specific criteria (e.g., capacity, equipment available, accessibility features).
- **Consistency Across Platforms:** Standardized data can be presented consistently across different platforms and interfaces, improving the user experience for those seeking venue information.
- **Integration with Other Services:** Structured venue data could be more easily integrated with other services, such as event planning tools or community calendars, providing added value to users.

4.6 Limitations

Our work has several factors that limit the full potential of the models developed. The model's performance heavily relies on the quality and balance of the training data. One key challenge is data imbalance, where certain categories of venue information are underrepresented, potentially leading to biased outcomes. Additionally, annotation consistency posed a challenge, as maintaining uniformity in manual annotations, especially for nuanced categories, proved difficult and may have introduced

noise into the dataset. The limited dataset size from the SAcommunity database, while substantial, could benefit from further expansion and diversity to improve model generalization and performance.

Another limitation of our work is the use of a complex transformer model like RoBERTa, which, while effective, introduces challenges in interpretability. The "black box" nature of deep learning models makes it difficult to fully understand or explain their decision-making processes, which raises concerns in contexts where transparency and accountability are critical, such as community information. Additionally, the model's heavy reliance on the training data increases the risk of perpetuating any existing biases or inconsistencies, potentially affecting the fairness of the output.

Additionally, our work stems from the domain-specific focus on venue availability information, which affects the model's ability to generalize across different contexts. The highly specific vocabulary used to describe venues and facilities may limit the model's effectiveness when encountering new or unseen descriptions. Additionally, regional variations in terminology and the way venues are characterized introduce challenges, as the model may not fully capture these differences, potentially reducing its applicability to broader datasets or other geographical areas.

5 Conclusion

This paper demonstrates the feasibility and potential of using LLMs for automated extraction of venue availability information in MARC standard format. The RoBERTa-based model showed promising results in identifying and categorizing relevant information from unstructured text, with consistent improvements observed throughout the training process. This research enhances data management by automating the extraction and structuring of venue availability information, improving accessibility through MARC standards for better usability across stakeholders. The scalability of the transformer-based RoBERTa model allows for adaptation to larger datasets and other community service types, while also representing an innovative use of advanced NLP techniques to address real-world challenges in community information management.

Further experimentation with model architectures, training regimes, and hyperparameters could enhance performance, while exploring ensemble methods may improve robustness by leveraging the strengths of different models. Additionally, investigating few-shot learning techniques might enable the model to adapt to new types of venue information or regional variations with minimal training. Moreover, data enhancement can be achieved through several strategies: employing data augmentation techniques like back-translation or synonym replacement to artificially expand the training dataset may enhance model generalization; increasing experimentation with active learning, where the model identifies informative samples for human annotation, could more efficiently improve the training dataset; and incorporating venue information from various geographic regions could better equip the model to manage regional variations in terminology and venue descriptions.

6 Ethical Considerations

We have carefully considered the ethical implications of working with community service information and leveraging AI technologies, ensuring that data privacy, transparency, and fairness are maintained throughout the process. We adhered to strict ethical guidelines throughout the project by fully anonymizing all data, ensuring no personally identifiable information was included. The data usage remained aligned with its original sharing intent, and the training data was carefully examined for potential biases. Regular bias checks were implemented during model development to mitigate risks, while safeguards were established to prevent the aggregation of sensitive information. Additionally, guidelines emphasizing human oversight were developed to promote responsible system use.

Acknowledgments

We thank Catherine McIntyre from SAcommunity and Connecting Up (an Infoxchange service) for her valuable practical insights. We're grateful to SAcommunity for being our industry partner, enabling us to work on a meaningful real-world application that addresses community needs. This project stands as a testament to the power of academic-industry collaboration, and we are deeply thankful for their guidance and partnership.

References

Norah Alshammari and Saad Alanazi. 2021. The impact of using different annotation schemes on named

entity recognition. *Egyptian Informatics Journal*, 22(3):295–302.

Kurt Chan, Kristian Alfonso Delas Alas, Carmina Orcena, Don Justin Velasco, Queni John San Juan, and Charibeth Cheng. 2023. Practical approaches for low-resource named entity recognition of filipino telecommunications domain. In *Proceedings of the 2023 American Medical Informatics Association (AMIA) Annual Symposium*.

Yukun Chen, Thomas A. Lasko, Qiaozhu Mei, Joshua C. Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18.

Yu-Shiang Chuang, Xiao Jiang, Chao-Te Lee, Riddhiman Brandon, Dat Tran, Oluwabunmi Tokede, and Muhammad F. Walji. 2023. Use gpt-j prompt generation with roberta for ner models on diagnosis extraction of periodontal diagnosis from electronic dental records. In *Proceedings of the 2023 American Medical Informatics Association (AMIA) Annual Symposium*.

Xiuying Cui, Yongmin Yang, Dongsheng Li, Xiaolong Qu, Lingling Yao, Shuai Luo, and Chuanqi Song. 2023. Fusion of softlexicon and roberta for purpose-driven electronic medical record named entity recognition. *Applied Sciences*, 13(24):13296.

John Dagdelen, Amalie Trewartha, Sanghoon Lee, Alexander Dunn, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

Doccano. Doccano: Open source text annotation tool for machine learning practitioner. Accessed: 2024-03-15.

IOB Tagging. Nlp | iob tags. Accessed: 2024-03-15.

B. Jehangir, S. Radhakrishnan, and R. Agarwal. 2023. A survey of ner. *Natural Language Processing Journal*, 3:100017.

JSON Lines. Json lines: Text sequence format. Accessed: 2024-03-15.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Le Le, Gianluca Demartini, Guido Zuccon, Guihua Zhao, and Xin Zhang. 2023. Active learning with feature matching for clinical named entity recognition. *Natural Language Processing Journal*, 4:100015.

Library of Congress. 2000. Marc 21 format for community information.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

MARC21 3XX. 2000. Marc 21 format for community information: Physical description fields (3xx). Accessed: 2024-03-15.

Doan Thai Binh Phan, Phuoc Vinh Linh Le, Ngoc Hoang Luong, Tahar Kechadi, and Hung Q. Ngo. 2023. Domain adaptation in nested named entity recognition from scientific articles in agriculture. In *Proceedings of the 12th International Symposium on Information and Communication Technology (SOICT '23)*, pages 48–55, New York, NY, USA. Association for Computing Machinery.

Traian-Radu Ploscă, Christian-Daniel Curiac, and Daniel-Ioan Curiac. 2024. Investigating semantic differences in user-generated content by cross-domain sentiment analysis means. *Applied Sciences*, 14:2421.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys*, 54(9):180.

SAcommunity. Sacommunity - south australia's community information directory. Accessed: 2024-03-15.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Roselyne B. Tchoua, Aswathy Ajith, Zhuozhao Hong, Logan T. Ward, Kyle Chard, Daniel J. Audus, Shrayesh N. Patel, Juan J. de Pablo, and Ian T. Foster. 2019. Active learning yields better training data for scientific named entity recognition. In *2019 15th International Conference on eScience (eScience)*. IEEE.

Weights & Biases. Weights and biases: Developer tools for ml. Accessed: 2024-03-15.

Yuhang Wu, Jing Huang, Chao Xu, Hongbo Zheng, Luoxin Zhang, and Jie Wan. 2021. Research on named entity recognition of electronic medical records based on roberta and radical-level feature. *Wireless Communications and Mobile Computing*.

Xuan Zhang, Xiaojun Luo, and Jinqiu Wu. 2023. A roberta-globalpointer-based method for named entity recognition of legal documents. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

A Appendix

A.1 What is MARC?

MARC (Machine-Readable Cataloging) standards are a set of digital formats for the description of items cataloged by libraries, such as books and articles. Developed by the Library of Congress, these standards are designed to be comprehensive and allow for the encoding of various types of bibliographic materials across different types of content and media. In this project, the MARC-21 format for community information is utilized to structure data related to venue hires, ensuring that the extracted data aligns with widely recognized library and information science standards.

A.2 Stakeholders of the Research

- Event and Community Service Managers: These professionals will benefit from easier access to standardized information, improving their ability to plan and manage venues.
- Government Entities: Local and state governments, especially those supporting community services like SAcommunity, rely on structured data to better serve their constituents and manage community resources.
- Librarians and Information Scientists: Professionals in these fields are key users of MARC standards and will benefit from enhanced methods of cataloging and accessing information.
- Technology Developers and Researchers: Individuals and teams developing NLP and data extraction technologies have a vested interest in the methodologies and outcomes of this research.
- End Users: General public users of community directories who will experience improved usability and access to information regarding venue hires.

A.3 Performance Metrics Calculation

Calculate accuracy, precision, recall, and F1 scores to assess the NER model's performance on the evaluation dataset.

Subject
Halls for Hire
Community Facilities
Convention Facilities
Community Centers
Conference Venues
Conference Venues (Residential)
Reception Facilities
Recreation Facilities
Recreation Centers
Sports Clubs & Centers
Clubs/Groups
Meeting Rooms

Table 3: Subjects Covered in the Database

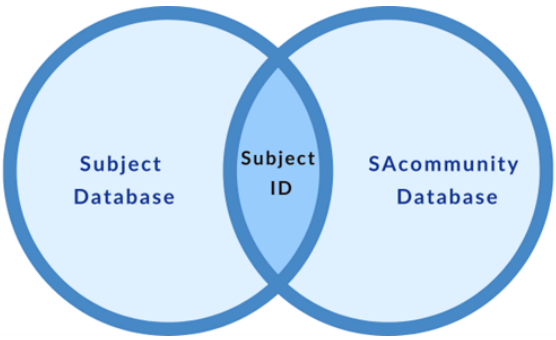

Figure 4: SQL Inner Join of both databases: A visualization.

- Accuracy: Measures the overall correctness of the model.

  Accuracy = (TP + TN) / (TP + TN + FP + FN)

- Precision: Measures the accuracy of positive predictions.

  Precision = TP / (TP + FP)

- Recall: Measures the proportion of actual positives correctly identified.

  Recall = TP / (TP + FN)

- F1 Score: The harmonic mean of precision and recall, providing a balanced measure.

  F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

A.4 MARC 21 Format for Physical Description Notes for Venue Hire

- $a - General description of facilities

Figure 5: Wordcloud exploratory data analysis for the "Venue Hire" feature from our dataset.

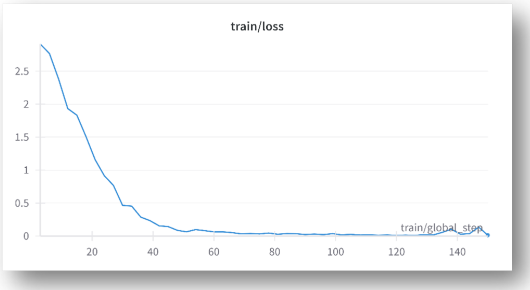

Figure 6: Shows a graph of train/loss over global steps.

- \$b - Name and location
- \$c - Physical description
- \$d - Capacity
- \$e - Equipment available
- \$f - Rental fee
- \$g - Special restrictions
- \$h - Accommodations for the disabled
- \$m - Miscellaneous information
- \$p - Contact person
- \$6 - Linkage
- \$8 - Field link and sequence number

A.5 Critical Reflection

Reflecting on these ethical considerations, we recognize that our project exists in a complex ethical landscape. While we have taken steps to address key ethical issues, we acknowledge that ethical challenges in AI and data management are evolving. One area for future consideration is the long-term impact of automating information extraction on human roles in community information management. While our project aims to enhance efficiency, it's crucial to balance this with the value of human expertise and judgment. Additionally, as AI technologies advance, the ethical framework for projects like mine will need continuous reassessment. We're committed to ongoing ethical evaluation and adjustment of our approach as new insights and standards emerge in the field. In conclusion, ethical considerations have been integral to our research process, shaping decisions from data handling to model development and deployment strategies. By maintaining this ethical focus, we aim to ensure that my project contributes positively to community information management while respecting individual privacy and promoting fairness and accessibility.

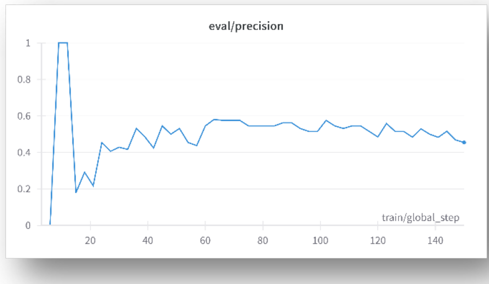

Figure 7: Shows a graph for precision on the evaluation set.

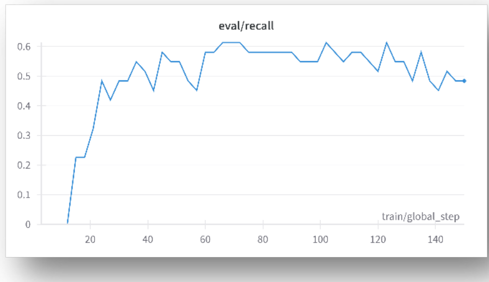

Figure 8: Shows a graph for recall on the evaluation set.

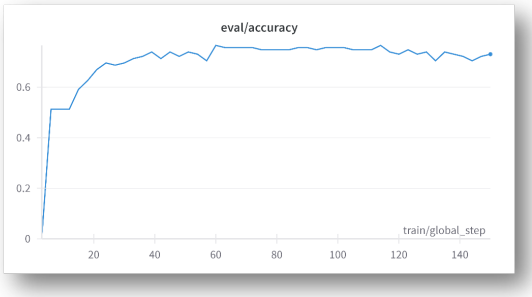

Figure 9: Shows a graph for accuracy on the evaluation set.

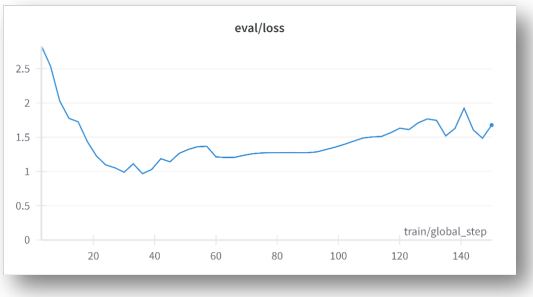

Figure 10: Shows the loss curve on the evaluation set.

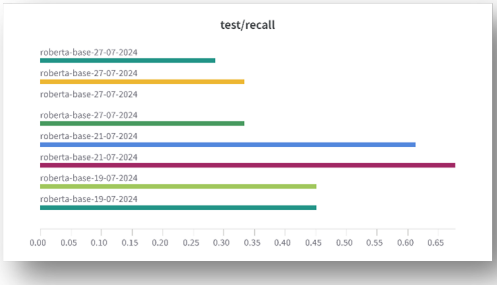

Figure 11: Test Set Recall over multiple experimentation.

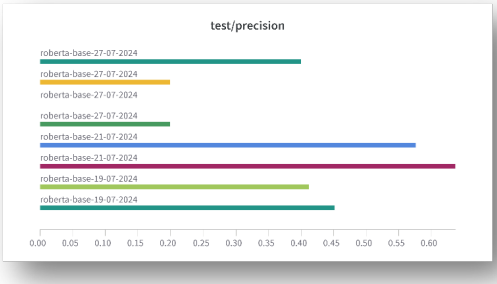

Figure 12: Test Set Precision over multiple experimentation.

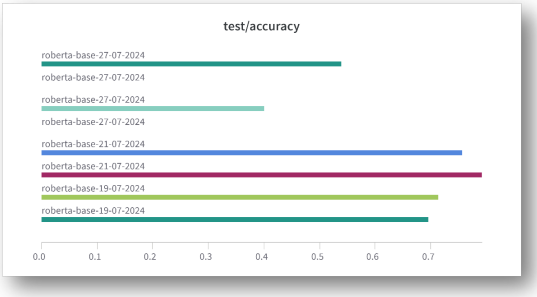

Figure 13: Shows our best test accuracy on 21-07-2024.

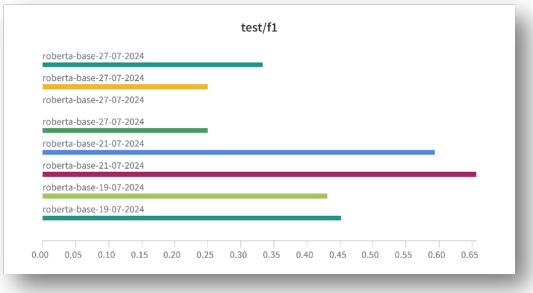

Figure 14: Shows F1 score graph on the test set across multiple iterations.

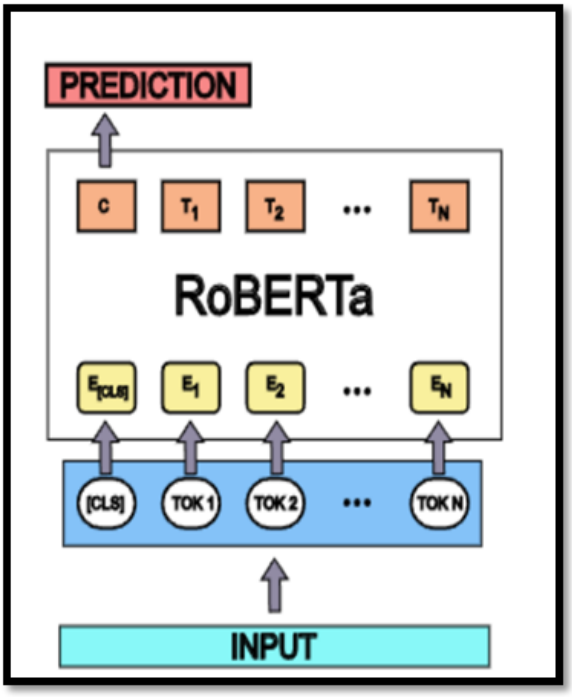

Figure 15: Roberta architecture adopted from (Ploscă et al., 2024)