George Mason University

Professor: Dr. Lindi Liao
AIT 526
Intro to Natural Language Processing

Universal Document-level Information Extraction

Checkpoint 3

Prepared by:

Aakiff Panjwani
Hamaad Zuberi
Vansh Setpal

Team 12

1 Abstract

This paper presents a universal approach to document-level information extraction, addressing the challenges of cross-domain generalization in Named Entity Recognition (NER) and Relation Extraction (RE). Leveraging the DocIE dataset spanning 34 diverse domains, we first establish a zero-shot baseline using GPT-4o and Llama3-8B, revealing significant performance gaps (15–25 F1-point drops) when models are applied to held-out domains. We then propose a graph neural network (GNN) framework that encodes each document as a token-level graph enriched with BERT embeddings and structured edges. To overcome data scarcity and domain shifts, we integrate Model-Agnostic Meta-Learning (MAML) into the GNN, enabling rapid adaptation to new contexts with few examples. Experimental results show that our GNN+MAML model achieves consistent improvements—raising NER and RE F1 scores into the 25–30% range across unseen domains—thus demonstrating the effectiveness of meta-learning for robust, low-resource document extraction. Future work will explore advanced graph architectures and alternative meta-learning strategies to further enhance cross-domain performance.

Keywords: Document-level information extraction, Named Entity Recognition, Relation Extraction, Graph Neural Networks, Model-Agnostic Meta-Learning, Cross-domain generalization, Few-shot adaptation

2 Introduction

Information Extraction is the process of automated extraction of structured information from unstructured text. This process involves processing natural language texts to identify and structuring relevant information using Natural Language Processing (NLP). **Document level Information Extraction** extends the scope of information extraction beyond individual sentences, enabling the capture of more complex and comprehensive information from documents while presenting unique challenges in processing and understanding long-range textual relationships.

It is crucial since most real-world information exists in unstructured documents, such as articles and reports. By extracting structured information from documents, computers are able process and analyze vast amounts of textual data efficiently.

However, documents vary greatly in format, structure, and content, making it difficult to design a single solution applicable in all contexts and domains. Understanding the context in which references occur is also crucial, as the meaning may change based on the surrounding information. Thus, understanding the context and relationships between different data elements in unstructured data is crucial. Converting and integrating unstructured data into structured formats without losing context or meaning requires sophisticated processing capabilities.

Named Entity Recognition (NER) is an NLP task focused on identifying and classifying specific entities such as people, organizations, locations, and dates within unstructured textual date. It plays a crucial role in extracting structured information, enabling enhanced data retrieval, analysis, and comprehension across various applications. [5]

Relation Extraction (RE) is the task of automatically identifying semantic relationships between entities mentioned within a text. It involves detecting entity mentions (such as people, organizations, locations) and classifying the type of relationships (like "employed at," "author of," or "located in") connecting these entities. RE facilitates structuring unorganized text data into structured knowledge, benefiting various applications including information retrieval, question answering, and knowledge management systems. [38]

3 Objectives

The DocIE initiative [1] aims to improve cross-domain generalization in document-level information extraction by developing robust and state-of-the-art NLP models capable of handling diverse textual data. The primary goal of DocIE is to create a benchmark framework that challenges and enhances adaptability in Named Entity Recognition (NER) and Relation Extraction (RE) when compared to conventional models. By leveraging a dataset spanning 34 domains, the competition ensures that models are tested against real-world complexities such as coreference resolution, implicit relationships, and variations in domain-specific schemas.

Additionally, DocIE helps introduce scalable and modular methodologies to information extraction, enabling the development of models that are more interpretable and adaptable to unstructured datasets.

DocIE serves as a stepping stone toward practical, deployable real-world NLP solutions for industries like engineering, law, finance, and healthcare, where accurate document-level information extraction is crucial.

Cross-domain generalization remains DocIE's most persistent hurdle. Medical texts demand recognition of nested entities (e.g., "Stage IIIB non-small cell lung cancer"), while legal documents require precise identification of referential phrases across hundreds of pages. The DocIE competition structure—with 5 training domains, 2 validation domains, and 27 held-out test domains—directly confronts this challenge by forcing models to transfer knowledge across disparate contexts. Recent benchmarks show performance drops of 15-25 F1 points when models trained on news articles are applied to clinical notes, underscoring the competition's practical relevance. [3]

4 Literature Review

4.1 Approaches in Document level Named Entity Recognition

Most of the information in this section is drawn from the Bibliometric analysis of NER trends by Yang et al. [2], which provides an extensive overview of the field.

Rule-based approaches for document-level NER employ manually crafted patterns, lexicons, and linguistic rules to identify named entities throughout a document. These systems typically incorporate domain knowledge and contextual cues that span multiple sentences. For example, the LASIE-II system [45] uses cascaded finite-state transducers to recognize nested entities across document sections. Similarly, FASTUS [47] employs multiple phases of pattern matching to build increasingly complex entity representations by consolidating information from different parts of a document. These approaches excel in domains with highly structured text and well-defined entity patterns but require substantial expert knowledge to develop and maintain.

Traditional machine learning methods like Conditional Random Fields (CRFs) [43], Support Vector Machines (SVMs) [44], and Hidden Markov Models (HMMs) [46] have been extended to document-level NER by incorporating document-wide features. The MENE (Maximum Entropy Named Entity) system [48], for example, uses document-level statistics and entity frequency information to improve recognition. These approaches typically employ feature engineering to capture cross-sentence dependencies, such as entity co-occurrence patterns, document section information, and positional features. While effective for many applications, these methods require extensive feature engineering and may struggle with complex entity relationships that span large portions of text.

Deep learning methods have revolutionized document-level NER by effectively capturing long-range dependencies without extensive feature engineering. BiLSTM-CRF models [42] enhanced with document-level attention mechanisms, as proposed by Xu et al., use global document contexts to inform local entity recognition decisions. Transformer-based architectures [41] like BERT and its variants have proven particularly effective for document-level NER. Models such as BioBERT [32] and SciBERT [28] have been fine-tuned for domain-specific document-level entity recognition in biomedical and scientific literature, respectively. These models utilize self-attention mechanisms to capture relationships between tokens across the entire document, significantly improving recognition of entities with complex contextual dependencies.

Hierarchical approaches decompose document-level NER into multiple levels of processing. These methods typically begin with sentence or paragraph-level entity recognition and then integrate these results at the document level using higher-order models. For example, Xu et al. [34] proposed an attention-based neural network architecture that relieves context dependency by using document-level global information obtained from documents represented by a pre-trained bidirectional language model with neural attention. Similarly, pyramid hierarchical models with multi-head adjacent attention mechanisms fuse information from adjacent inputs to better model dependency relationships between entity spans [4]. These hierarchical structures effectively balance computational efficiency with the need to capture document-wide contexts.

Graph-based methods represent documents as interconnected networks where nodes typically represent tokens, sentences, or candidate entities, while edges capture their relationships. For instance, the heterogeneous graph neural network approach by Zhao et al. [19] represents both entities and their relationships as nodes in a graph, using iterative fusion to capture dependencies in long texts. Wang et al. employed hypergraph neural networks to model complex entity relationships, particularly for nested entities that span multiple sentences [33]. These approaches excel at capturing non-sequential dependencies that may exist across distant parts of a document, making them particularly suitable for complex documents with intricate entity relationships.

Cross-document NER extends beyond single-document boundaries to leverage information across multiple related documents. These methods are particularly valuable for entity disambiguation and linking. For example, federated learning approaches, as demonstrated by Wu et al., use knowledge distillation techniques to share information across different documents and datasets while maintaining privacy [8]. These models typically involve entity co-reference resolution across documents and often incorporate external knowledge bases. Cross-document approaches are especially useful for analyzing document collections like news article sets, scientific paper corpora, or medical record databases where the same entities appear across multiple documents in different contexts.

Multimodal document-level NER integrates textual content with other data modalities such as images, tables, or document layout information. Yu et al. proposed a unified multi-modal transformer that combines text and visual information with an auxiliary entity span detection module to improve entity recognition in social media posts [25]. Similarly, Zhang et al. introduced a multi-modal graph fusion method that creates a graph structure merging textual and visual objects to facilitate deep semantic interaction [17]. These approaches are particularly effective for documents where visual elements provide critical context for entity recognition, such as in scientific papers with figures and tables, social media posts with images, or technical documentation with diagrams.

To address the challenge of limited annotated data for document-level NER, researchers have developed **weakly-supervised and distant supervision approaches**. These methods generate training data automatically by leveraging existing knowledge bases or heuristic rules. Zhou et al. proposed a distant supervision method that generates labels by matching entity mentions in text with entity types in knowledge bases, employing a reliability-based learning strategy to reduce false negatives from incomplete labels[11]. Meng et al. developed a noise-robust learning scheme with a modified loss function to handle the noisy labels that often result from distant supervision [13]. These approaches are particularly valuable for domain-specific document-level NER where manual annotation would be prohibitively expensive.

4.2 Approaches in Document Relation Extraction

Most of the information in this section is drawn from the survey conducted by Zheng et al. [3], which provides an extensive overview of the field.

Multi-granularity models. Jia et al. [31] first introduced this approach using multiscale representation learning, aggregating mention-level information to ensemble sub-relations. Tang et al. (2020) proposed the Hierarchical Inference Network (HIN) [23], employing Bi-LSTMs at token, sentence, and document levels combined with attention mechanisms to capture both local and global features.

Graph-based models represent documents as graphs, using nodes (words, entities, mentions) and edges to capture relationships. Early models such as DISCREX [40] and Graph-LSTM [39] used graph structures with dependencies. Later, AGGCNs [30] leveraged Graph Convolutional Networks (GCNs) enhanced by multi-head self-attention to extract global context. EoG [29] significantly influenced subsequent methods by introducing entities as nodes connected by meaningful paths. Approaches derived from EoG include homogeneous graphs like LSR [22], dynamically refined via structured attention, and heterogeneous graphs like GLRE [24], HeterGSAN [16], POR [6], and dual-graph methods such as GAIN [26], GEDA [21], DHG [27], and DRN [15], capturing hierarchical and structural information with specialized node types and paths.

Task-specific models use tailored neural network structures or customized loss functions to address document-level relation extraction (doc-RE). SSAN [14] integrates structural dependencies directly into network layers. ATLOP ([20] utilizes localized context pooling with adaptive thresholding loss to balance label predictions. DocuNet [18] applies a U-shaped semantic segmentation structure for refined entity feature extraction. KD [7] uses axial self-attention on paired entity tables with adaptive focal loss to address class imbalance.

Path-based models emphasize constructing evidence paths, focusing specifically on critical information rather than processing entire documents. THREE [12] identifies relation-supporting sentences through consecutive, multi-hop, or default paths. EIDER [10] defines minimal evidence sentences required for accurate relation extraction. SAIS [9] adopts a two-stage process to distinguish and retrieve pooled and fine-grained evidence, aligning closely with human intuition. These approaches consistently show strong performance by leveraging evidence that directly informs relation decisions.

5 Dataset

The DocIE dataset is a carefully curated resource that supports document-level information extraction, bringing together a broad variety of topics like Entertainment, Government, Education, Communication, Energy, Internet, and Human Behavior, among many others. It is designed to develop more accurate and context-aware NLP models for Information Extraction. Whether it is recognizing named entities, identifying relationships between those entities, or handling co-reference resolution, DocIE's structured annotations provide a rich playground for understanding how language is used in diverse real-world settings. By drawing from 34 different domains, the dataset ensures that models learn from an expansive range of writing styles and subject matter, boosting their ability to handle new or unexpected data in practical applications [1].

To support the full development life-cycle of an NLP model, DocIE is divided into training, validation, and testing sets. The training sets include extensive labeled examples, allowing a model to learn fundamental patterns and rules within the text. The validation sets, which cover domains like Human Behavior and Internet, help evaluate whether the model is truly grasping these patterns or merely memorizing examples. Finally, the testing sets, spanning areas such as Government, Entertainment, and Education, provide a final unbiased check to confirm whether the model can accurately extract entities and relationships from entirely unseen data.

6 Baseline Solution

The organizers of the DocIE challenge established a baseline performance benchmark to provide a reference point for evaluating submitted systems. This baseline utilizes two prominent Large Language Models (LLMs): OpenAI's GPT-4o (specifically version gpt-4o-2024-08-06 accessed via the Azure OpenAI API) and Meta's Llama3-8B-Instruct (referred to as llama3-8b-all in the challenge documentation) [1]. These models represent the current state-of-the-art in large-scale (GPT-4o) and smaller-scale, highly capable (Llama3-8B) language modeling.

The baseline implementation methodology for GPT-4o, detailed in the provided script (`gpt4o.py`), employs a zero-shot prompting strategy. For each document instance provided in the input file (`args.input_path`), the script extracts the task-specific instruction (prompt), the document's text content (input), and the target `schema_list` from the input data structure. The GPT-4o model is invoked with two textual inputs via the chat completions API: (1) a system message containing the prompt, and (2) a user message containing the document text.

Evaluation of the baseline and participant submissions is conducted using a standardized script (`scoring.py`) provided by the organizers. This script computes performance metrics by comparing the system's structured output predictions (`results.json`) against the official ground truth annotations (`reference.json`). The key evaluation metrics reported are Precision (P), Recall (R), and the F1-score, calculated independently for various predefined document categories.

The official baseline results (summarized in Table 1) reveal the inherent difficulty of the DocIE task for general-purpose LLMs, even when employing zero-shot prompting with structured output enforcement:

- **Significant Room for Improvement:** Overall low scores indicate the baseline performance is far from perfect.
- **Task/Domain Specificity:** Performance varies greatly depending on the document category, suggesting that generic approaches struggle with diversity.
- **Recall Often a Bottleneck:** Models frequently fail to find/extract existing information.
- **Precision Challenges:** Models also often extract incorrect or wrongly formatted information.

Category		**gpt4o**			**llama3-8b-all**	
	F1	P	R	F1	P	R
Academic_disciplines	3.25	2.78	3.9	5.07	7.23	3.9
Business	1.41	1.71	1.2	4.08	13.33	2.41
Communication	10.26	13.73	8.19	2.76	6.25	1.75
Culture	7.31	9.4	5.98	5.38	8.91	3.85
Economy	2.6	2.87	2.37	7.17	14.71	4.74
Education	1.21	1.28	1.15	2.59	2.96	2.3
Energy	3.31	3.33	3.3	3.37	3.23	4.4
Engineering	3.11	2.55	3.97	4.79	6.03	3.97
Entertainment	2.89	2.51	3.4	8.07	9.04	7.28
Food_and_drink	0.81	0.72	0.93	3.7	5.5	2.79
Geography	2.81	4.92	1.97	8.41	14.52	5.92
Government	3.55	3.39	3.72	4.51	3.63	5.95
Health	2.73	3.23	2.36	5.69	5.19	6.3
History	5.24	5.79	4.78	10.28	8.52	12.97
Human_behavior	2.34	2.37	2.31	6.27	8.15	5.09
Humanities	2.76	3.09	2.49	1.62	4.44	0.99
Information	6	12	4	4.45	5.51	3.73
Internet	9.2	12.09	7.43	6.71	6.11	7.43
Knowledge	2.15	2.26	2.05	1.63	3.85	1.03
Language	7.53	8.02	7.1	4.18	8.93	2.73
Law	3.66	4.8	2.96	1.68	2.22	1.35
Life	2.72	4.55	1.94	5.84	20.59	3.4
Mathematics	9.82	11.27	8.7	10.3	15.91	7.61
Military	4.21	4.15	4.27	6.64	5.43	8.55
Nature	11.25	15.45	8.85	5.22	5.88	4.69
People	0.69	0.84	0.58	3.26	2.72	4.07
Philosophy	9.28	8.8	9.82	13.71	19.05	10.71
Politics	4.3	4.39	4.21	1.64	1.26	2.34
Religion	4.13	4.62	3.73	4.51	5.71	3.73
Science	10.33	12.09	9.07	1.49	8	0.82
Society	0.95	0.92	0.98	2.26	2.7	1.95
Sports	1.17	1.29	1.07	6.76	9.17	5.35
Technology	4.85	5.19	4.55	7.72	12.05	5.68
Universe	1.6	1.65	1.55	2.61	8.33	1.55

Table 1: Performance comparison of GPT-4o and Llama3-8B-ALL across various categories

- **Model Size vs. Performance:** The smaller Llama-8B model competes well with or even outperforms the much larger GPT-4o in several domains, highlighting that factors beyond scale (like training data or architecture nuances) matter.

This established baseline provides a crucial quantitative benchmark against which the advancements proposed in this work are measured.

7 Proposed solution

The solution proposed in this project revolves around using a Graph Neural Network (GNN) to effectively understand and analyze documents by extracting entities and identifying relationships between them. First, each document is converted into a structured graph where each word or token becomes a node. Instead of assigning random vectors, the model uses meaningful, context-rich embeddings obtained from a pre-trained BERT model, giving each node a 768-dimensional representation that captures linguistic nuances. To define relationships between tokens, edges in the graph are established using three practical strategies: connecting consecutive words to maintain sentence structure; forming connections based on annotated relationships within the document by comparing token embeddings to special embeddings of known relation elements; and linking words that represent parts of the same named entity. When explicit connections are not provided, the system ensures each token influences its own feature update by creating self-loops, a common practice in GNNs[37].

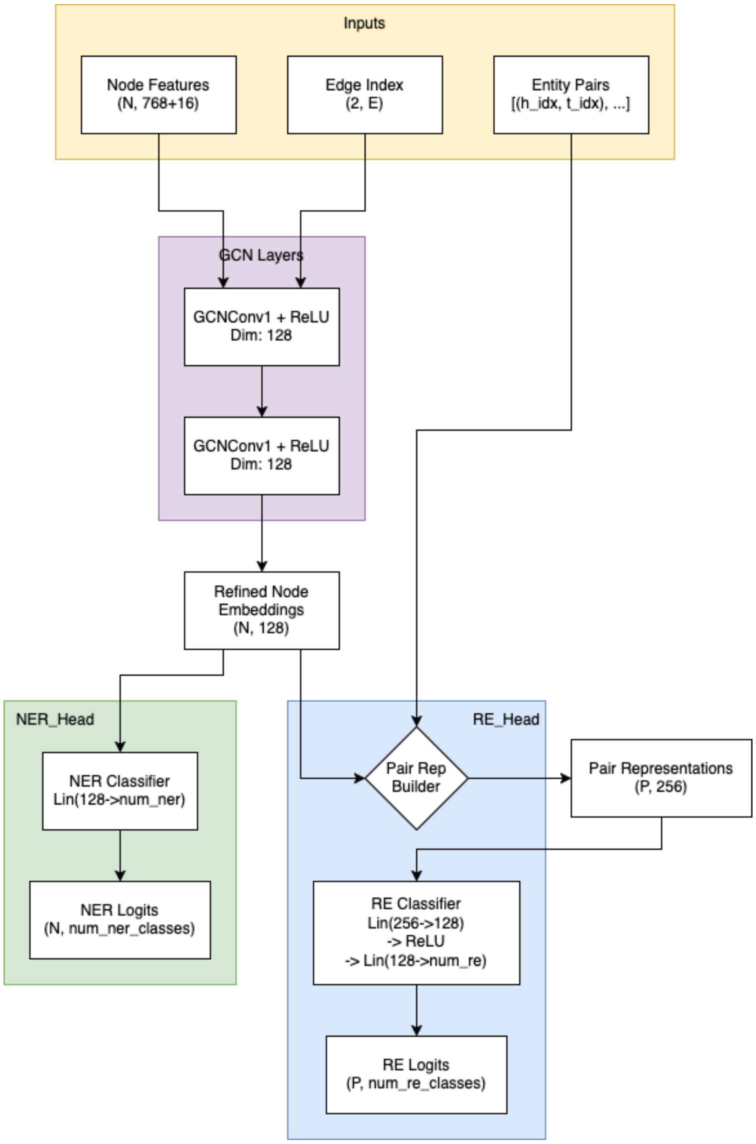
Inputs
Node Features (N, 768+16)
Edge Index (2, E)
Entity Pairs [(h_idx, t_idx), ...]
GCN Layers
GCNConv1 + ReLU Dim: 128
GCNConv1 + ReLU Dim: 128
Refined Node Embeddings (N, 128)
NER_Head
NER Classifier Lin(128->num_ner)
NER Logits (N, num_ner_classes)
RE_Head
Pair Rep Builder
Pair Representations (P, 256)
RE Classifier Lin(256->128) -> ReLU -> Lin(128->num_re)
RE Logits (P, num_re_classes)

Figure 1: GNN Model

After building this detailed graph representation, the model reduces the complexity of token embeddings through a dimensionality reduction step, typically from 768 to 64 dimensions. These simplified features are then processed using a multi-head attention-like mechanism, helping the model capture diverse, local contexts and patterns within the text. After aggregating insights from these multiple perspectives, the information is further refined, enabling the model to grasp broader contextual cues. The refined node representations are then directed toward two key tasks. First, for Named Entity Recognition (NER), the model classifies each token into categories such as PERSON, ORGANIZATION, DATE, and so on, calculating loss based on how well these predictions align with annotated entities[35]. Simultaneously, the model tackles Relation Extraction by identifying promising entity candidates based on their predicted classes, pairing them, and evaluating their combined features to predict potential relationships.

The input to the system comes as a series of JSON files, each containing one or more documents. Each document provides the raw text under a “doc” field and may also include a pre-tokenized list of words or “tokens”. Additionally, the input may come with extra details like edges to describe connections between words, entity labels for each token to help the system learn what kind of entity each token might be, and relation triples that indicate how entities relate to one another. There’s also metadata like a “document id” or “title” that helps uniquely identify each document.

On the output side, the system produces a structured JSON array where each element corresponds to a processed document. For each document, the output includes an “id” and “domain” to identify the source and context. The output is divided into two main sections: “entities output” for Named Entity Recognition and “triples output” for Relation Extraction. In “entities output”, each entity is represented by a list of its “mentions” and its “type” (such as PERSON, ORGANIZATION, etc.). In “triples output”, each relationship is clearly broken down into “head” (subject), “relation” (the type of connection), and “tail” (object). This format makes it easy to see both the individual pieces of information in the text and how they are connected.

We have, at present, achieved a precision, recall, and F1 score of approximately 0.5 for NER, but have been unable to obtain any results for RE as of now.

To overcome the challenges observed in the initial system, a new solution was proposed that incorporates Model-Agnostic Meta-Learning (MAML)[36] into the Graph Neural Network architecture. Rather than training the GNN in a conventional supervised manner, the meta-learning approach allows the model to adapt rapidly to new domains and scarce relation patterns by simulating a series of learning tasks during training. Each training document is split into a support set and a query set: the model first adapts to the support set with a few gradient steps (inner loop) and then updates its global parameters based on query set performance (outer loop). This enables the GNN to learn parameters that generalize better across different domains and documents, even when only limited examples are available.

In this enhanced approach, each document is still converted into a graph where each node represents a token enriched by BERT embeddings. However, instead of relying solely on standard training, the node features are supplemented with learned entity-type embeddings, allowing the model to better differentiate between entity classes during both entity recognition and relation prediction. The GNN encoder uses two layers of Graph Convolutional Networks (GCN), with ReLU activations, to propagate contextual information. For Relation Extraction, entity pairs are explicitly sampled with a 1:1 positive-to-negative ratio, meaning that for every true relation, a corresponding unrelated entity pair is introduced. This careful balancing of relation examples, performed before meta-training, mitigates biases toward predicting ”no relation” and strengthens the model’s ability to recognize subtle relational patterns.

The output structure remains consistent with the earlier design: the system produces, for each document, a structured JSON object containing predicted entities and triples. However, during inference, a softmax confidence threshold is applied to relation predictions, ensuring that only confident triples are retained for evaluation and downstream tasks. Evaluation metrics are now computed separately for NER and RE at the type level, allowing a detailed analysis of how well the system generalizes across domains such as Education, Health, Technology, and Politics. Using this GNN+Meta-learning framework, we have been able to achieve significant improvements compared to the baseline. Both the NER and RE components now reach precision, recall, and F1 scores in the range of approximately 25% to 30%, with consistent improvements observed across multiple domains, validating the effectiveness of meta-learning in low-resource document extraction settings.

8 Result & Analysis

The results of Table 1 compare two well-known large language models, GPT-4o and Llama3-8b-all, across various domains. GPT-4o generally excels at recalling, meaning it is good at identifying relevant information, especially in fields like Mathematics, Philosophy, and Science. However, it struggles with precision, and often also identifies some irrelevant information. In contrast, Llama3 tends to pick fewer but more accurate results, showing higher precision in areas such as Economy, Geography, and Life. Despite these strengths, both models produce relatively low overall scores, revealing considerable challenges in accurately classifying information across diverse topics. These differences suggest that further tuned specifically to each domain

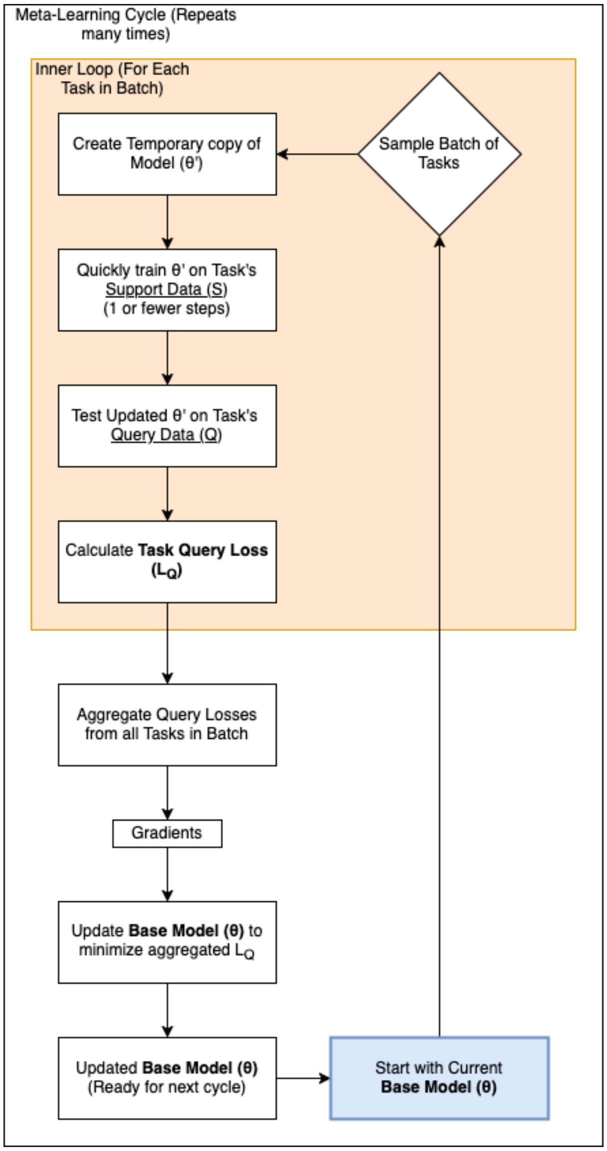
Meta-Learning Cycle (Repeats many times)
Inner Loop (For Each Task in Batch)
Create Temporary copy of Model (θ')
Sample Batch of Tasks
Quickly train θ' on Task's Support Data (S) (1 or fewer steps)
Test Updated θ' on Task's Query Data (Q)
Calculate Task Query Loss (L_Q)
Aggregate Query Losses from all Tasks in Batch
Gradients
Update Base Model (θ) to minimize aggregated L_Q
Updated Base Model (θ) (Ready for next cycle)
Start with Current Base Model (θ)

Figure 2: Model Agnostic Meta-Learning Approach

could greatly enhance their performance.

The result of Table 2 focuses on a basic Graph Neural Network (GNN) model. Unfortunately, this simpler GNN faces major hurdles, especially when it comes to recall, which means that it rarely recognizes enough relevant information to be effective. Although it occasionally shows flashes of precision—particularly in fields like Government and Politics—the extremely low recall makes the overall performance weak and practically limited. This shortcoming likely stems from insufficient data, inadequate representation of the context, or overly strict decision-making criteria within the model itself. Consequently, to make this basic GNN more useful, it would need substantial improvements in its ability to generalize and recognize relevant details.

The result of Table 3 evaluates a more advanced GNN model, enhanced through meta-learning techniques, specifically Model-Agnostic Meta-Learning (MAML). This approach dramatically improves the performance of the model in all domains compared to basic GNN, achieving significantly better recall (around 20%) while consistently maintaining solid precision (approximately 50 to 65%). Particularly strong improvements are observed in areas like Academic Disciplines, Business, and People, indicating that meta-learning significantly boosts the model's capability to adapt across different contexts. Although these results are promising, there

Category	**P**	**R**	**F1**
Academic_disciplines	46.87	2.15	4.11
Business	20.00	1.58	2.86
Communication	30.00	2.11	3.90
Culture	35.00	1.86	3.49
Economy	36.90	2.18	4.09
Education	30.00	1.75	3.30
Energy	51.67	3.84	7.10
Engineering	42.31	2.37	4.48
Entertainment	53.47	3.66	6.79
Food_and_drink	30.00	1.57	2.98
Geography	42.71	4.54	8.20
Government	83.33	5.99	11.00
Health	37.50	2.30	4.32
History	54.17	4.02	7.43
Human_behavior	41.67	2.75	5.12
Humanities	36.36	2.77	5.11
Information	44.23	2.98	5.54
Internet	35.00	3.16	5.72
Knowledge	29.69	1.66	3.13
Language	30.00	2.37	4.32
Law	40.00	2.79	5.17
Life	51.52	3.66	6.77
Mathematics	35.00	2.30	4.29
Military	48.15	4.43	8.03
Nature	11.11	0.88	1.61
People	25.00	1.58	2.95
Philosophy	42.86	2.10	4.00
Politics	70.83	4.77	8.81
Religion	27.50	2.03	3.75
Science	50.00	3.16	5.89
Society	41.67	2.94	5.43
Sports	58.33	3.30	6.21
Technology	41.67	3.51	6.43
Universe	20.00	1.05	2.00

Table 2: Performance of GNN across various categories

is still room to further improve recall. Future steps might include integrating richer contextual features or experimenting with other sophisticated meta-learning strategies to achieve even better generalization.

9 Conclusion

In conclusion, this project highlights a thoughtful progression from a traditional GNN-based document understanding system to a more adaptable and robust GNN model enhanced with meta-learning. Initially, while the system was able to recognize named entities with reasonable accuracy, extracting relationships between those entities proved much more challenging, largely due to the scarcity and imbalance of relation examples across different domains. By integrating Model-Agnostic Meta-Learning (MAML), the project addressed this gap, enabling the model to quickly adapt to new domains and low-resource scenarios. As a result, both entity recognition and relation extraction performance saw noticeable improvements, making the system more effective and reliable across a wide range of document types.

Overall, the project demonstrates that combining rich contextual embeddings, careful graph construction, positive-to-negative balancing, and meta-learning strategies can significantly enhance information extraction

Category	**P**	**R**	**F1**
Academic_disciplines	66.25	19.86	29.23
Business	54.69	22.86	31.95
Communication	58.33	19.28	28.74
Culture	51.79	19.87	28.55
Economy	56.10	18.79	27.85
Education	54.00	18.86	27.71
Energy	63.46	17.64	26.70
Engineering	54.69	18.57	27.47
Entertainment	53.03	17.79	26.43
Food_and_drink	50.00	16.32	24.60
Geography	54.55	19.91	28.86
Government	69.23	18.02	27.53
Health	50.00	19.08	27.62
History	51.67	19.37	28.05
Human_behavior	56.10	20.10	29.31
Humanities	53.33	20.00	28.86
Information	53.12	17.69	26.35
Internet	62.96	18.76	28.15
Knowledge	51.19	19.82	28.46
Language	51.92	18.34	26.98
Law	51.92	20.95	29.67
Life	50.00	16.75	25.09
Mathematics	53.57	19.05	27.99
Military	60.42	18.41	27.74
Nature	50.00	15.79	24.00
People	58.06	21.45	30.91
Philosophy	52.94	17.59	26.24
Politics	53.70	17.16	25.79
Religion	51.92	18.87	27.54
Science	52.50	18.65	27.35
Society	53.33	20.65	29.48
Sports	60.87	16.90	25.65
Technology	50.00	19.30	27.85
Universe	55.88	20.38	29.78

Table 3: Performance of GNN+Meta-Learning across various categories

tasks in real-world, domain-diverse settings. It shows that even modest meta-learning enhancements can lead to a system that not only fits better on known domains but also generalizes to unseen ones without extensive retraining. This adaptability is a critical step forward, especially for applications where labeled data is limited or rapidly changing across fields.

Looking ahead, future work could focus on further strengthening the model's ability to capture complex relationships by exploring more advanced graph architectures, such as graph attention networks or heterogeneous GNNs. Domain-specific fine-tuning, smarter candidate pair selection for relation extraction, and experimenting with alternative meta-learning techniques like Reptile or Proto-MAML are also promising directions. Additionally, improving the confidence calibration of relation predictions could help refine outputs even further, making the system ready for deployment in real-world document analysis pipelines. Another exciting direction would be to integrate a reinforcement learning component that continuously adapts and improves its extraction policies based on feedback from new data, with the reinforcement agent itself being trained through meta-learning. This could enable the system to not only generalize quickly across domains but also progressively enhance its performance as more labeled or unlabeled documents are encountered over time.

10 Lessons Learned

Working with Graph Neural Networks (GNNs) for document understanding has provided several valuable insights throughout this project. One of the most important lessons was realizing that while GNNs are powerful for modeling complex structures, their performance heavily depends on the quality and design of the input graph. Early versions of the system, which relied on minimal graph connections, struggled because the model could not effectively capture the contextual relationships between tokens. Introducing dependency-based edges, self-loops and meaningful feature embeddings from BERT significantly improved the model's ability to reason over text, highlighting how critical graph construction strategies are when using GNNs for language tasks.

Another key takeaway was the difficulty GNNs face when dealing with extreme class imbalance, particularly in Relation Extraction. Even though GNNs are excellent at capturing node interactions, they tend to be biased toward majority classes if the dataset is not carefully balanced. The decision to enforce a 1:1 positive-to-negative relation pair ratio before training proved essential, dramatically stabilizing performance and making the model more sensitive to minority classes. This taught us that pre-training data preparation is just as crucial as model architecture when applying GNNs to real-world datasets.

Finally, the integration of meta-learning techniques revealed both the strengths and limitations of GNNs in few-shot or low-resource settings. While GNNs can generalize well within domains once trained, their adaptability to new domains without meta-learning was limited. Adding Model-Agnostic Meta-Learning allowed the GNN to fine-tune itself more effectively to unseen documents, showing that combining GNNs with adaptive learning strategies is a powerful way to extend their capabilities. Overall, this project reinforced that successful use of GNNs for information extraction requires not just careful modeling, but also thoughtful graph engineering, balanced supervision, and flexible learning frameworks.

11 Acknowledgment

We would like to sincerely thank Dr.Liao for her invaluable guidance, encouragement and feedback throughout the course of this project. Her insights into GNNs and real-world information extraction challenges provided essential direction at every stage of the work. The critical discussions during class and one-on-one meetings helped us not only to better understand the theoretical foundations of GNNs but also to apply them thoughtfully to complex and domain-diverse datasets.

We are also grateful for the emphasis on rigorous experimentation and careful model evaluation, which pushed us to go beyond standard approaches. The skills and mindset developed during this project, from careful graph construction to balancing training data and adapting models to low resource settings are lessons that will continue to shape our future work in machine learning. I deeply appreciate the time, support and mentorship provided throughout this project, without which this work would not have reached its current depth and quality.

We also appreciate the maintainers of open-source libraries such as PyTorch, PyTorch Geometric, Hugging Face Transformers and spaCy, whose tools formed the backbone of our development and experimentation. Finally, we are also thankful for the computational resources provided by Office of Research and Computing(ORC) at George Mason University which helped us to rigorously test and improve our models across multiple domains. This project would not have been possible without the collective knowledge, resources and encouragement from the academic and open-source communities.

References

[1] *XLLM ACL 2025 Shared Task-IV: Document-level Information Extraction — xllms.github.io*. `https://xllms.github.io/DocIE/`. [Accessed 23-03-2025].

[2] Jun Yang et al. "Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023". In: *Heliyon* (2024).

[3] Hanwen Zheng, Sijia Wang, and Lifu Huang. “A Comprehensive Survey on Document-Level Information Extraction”. In: *Proceedings of the Workshop on the Future of Event Detection (FuturED)*. Ed. by Joel Tetreault et al. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 58–72. DOI: 10.18653/v1/2024.futured-1.6. URL: https://aclanthology.org/2024.futured-1.6/.

[4] Shengmin Cui and Inwhee Joe. “A multi-head adjacent attention-based pyramid layered model for nested named entity recognition”. In: *Neural Computing and Applications* 35.3 (2023), pp. 2561–2574.

[5] Kalyani Pakhale. *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges*. 2023. arXiv: 2309.14084 [cs.CL]. URL: https://arxiv.org/abs/2309.14084.

[6] Wang Xu, Kehai Chen, and Tiejun Zhao. “Document-Level Relation Extraction with Path Reasoning”. In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22.4 (Mar. 2023). ISSN: 2375-4699. DOI: 10.1145/3572898. URL: https://doi.org/10.1145/3572898.

[7] Qingyu Tan et al. *Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation*. 2022. arXiv: 2203.10900 [cs.CL]. URL: https://arxiv.org/abs/2203.10900.

[8] Chuhan Wu et al. “Communication-efficient federated learning via knowledge distillation”. In: *Nature communications* 13.1 (2022), p. 2032.

[9] Yuxin Xiao et al. “SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2395–2409. DOI: 10.18653/v1/2022.naacl-main.171. URL: https://aclanthology.org/2022.naacl-main.171/.

[10] Yiqing Xie et al. “Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 257–268. DOI: 10.18653/v1/2022.findings-acl.23. URL: https://aclanthology.org/2022.findings-acl.23/.

[11] Kang Zhou, Yuepei Li, and Qi Li. “Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning”. In: *arXiv preprint arXiv:2204.09589* (2022).

[12] Quzhe Huang et al. “Three Sentences Are All You Need: Local Path Enhanced Document Relation Extraction”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 998–1004. DOI: 10.18653/v1/2021.acl-short.126. URL: https://aclanthology.org/2021.acl-short.126/.

[13] Yu Meng et al. “Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training”. In: *arXiv preprint arXiv:2109.05003* (2021).

[14] Benfeng Xu et al. “Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14149–14157. DOI: 10.1609/aaai.v35i16.17665. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17665.

[15] Wang Xu, Kehai Chen, and Tiejun Zhao. *Discriminative Reasoning for Document-level Relation Extraction*. 2021. arXiv: 2106.01562 [cs.CL]. URL: https://arxiv.org/abs/2106.01562.

[16] Wang Xu, Kehai Chen, and Tiejun Zhao. “Document-Level Relation Extraction with Reconstruction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14167–14175. DOI: 10.1609/aaai.v35i16.17667. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17667.

[17] Dong Zhang et al. “Multi-modal graph fusion for named entity recognition with targeted visual guidance”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 16. 2021, pp. 14347–14355.

[18] Ningyu Zhang et al. *Document-level Relation Extraction as Semantic Segmentation*. 2021. arXiv: 2106.03618 [cs.CL]. URL: https://arxiv.org/abs/2106.03618.

[19] Kang Zhao et al. “Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction”. In: *Knowledge-Based Systems* 219 (2021), p. 106888.

[20] Wenxuan Zhou et al. “Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.16 (May 2021), pp. 14612–14620. DOI: 10.1609/aaai.v35i16.17717. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17717.

[21] Bo Li et al. “Graph Enhanced Dual Attention Network for Document-Level Relation Extraction”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1551–1560. DOI: 10.18653/v1/2020.coling-main.136. URL: https://aclanthology.org/2020.coling-main.136/.

[22] Guoshun Nan et al. *Reasoning with Latent Structure Refinement for Document-Level Relation Extraction*. 2020. arXiv: 2005.06312 [cs.CL]. URL: https://arxiv.org/abs/2005.06312.

[23] Hengzhu Tang et al. “HIN: Hierarchical Inference Network for Document-Level Relation Extraction”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Hady W. Lauw et al. Cham: Springer International Publishing, 2020, pp. 197–209. ISBN: 978-3-030-47426-3.

[24] Difeng Wang et al. “Global-to-Local Neural Networks for Document-Level Relation Extraction”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 3711–3721. DOI: 10.18653/v1/2020.emnlp-main.303. URL: https://aclanthology.org/2020.emnlp-main.303/.

[25] Jianfei Yu et al. “Improving multimodal named entity recognition via entity span detection with unified multimodal transformer”. In: Association for Computational Linguistics. 2020.

[26] Shuang Zeng et al. *Double Graph Based Reasoning for Document-level Relation Extraction*. 2020. arXiv: 2009.13752 [cs.CL]. URL: https://arxiv.org/abs/2009.13752.

[27] Zhenyu Zhang et al. “Document-level Relation Extraction with Dual-tier Heterogeneous Graph”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1630–1641. DOI: 10.18653/v1/2020.coling-main.143. URL: https://aclanthology.org/2020.coling-main.143/.

[28] Iz Beltagy, Kyle Lo, and Arman Cohan. *SciBERT: A Pretrained Language Model for Scientific Text*. 2019. arXiv: 1903.10676 [cs.CL]. URL: https://arxiv.org/abs/1903.10676.

[29] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. *Connecting the Dots: Document-level Neural Relation Extraction with Edge-oriented Graphs*. 2019. arXiv: 1909.00228 [cs.CL]. URL: https://arxiv.org/abs/1909.00228.

[30] Zhijiang Guo, Yan Zhang, and Wei Lu. “Attention Guided Graph Convolutional Networks for Relation Extraction”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 241–251. DOI: 10.18653/v1/P19-1024. URL: https://aclanthology.org/P19-1024/.

[31] Robin Jia, Cliff Wong, and Hoifung Poon. "Document-Level N-ary Relation Extraction with Multiscale Representation Learning". 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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3693–3704. DOI: `10.18653/v1/N19-1370`. URL: `https://aclanthology.org/N19-1370/`.

[32] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (Sept. 2019), pp. 1234–1240. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btz682`. eprint: `https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/48983216/bioinformatics\_36\_4\_1234.pdf`. URL: `https://doi.org/10.1093/bioinformatics/btz682`.

[33] Bailin Wang et al. "A Neural Transition-based Model for Nested Mention Recognition". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1011–1017. DOI: `10.18653/v1/D18-1124`. URL: `https://aclanthology.org/D18-1124/`.

[34] Guohai Xu, Chengyu Wang, and Xiaofeng He. "Improving clinical named entity recognition with global neural attention". In: *Web and Big Data: Second International Joint Conference, APWeb-WAIM 2018, Macau, China, July 23-25, 2018, Proceedings, Part II 2*. Springer. 2018, pp. 264–279.

[35] Yuhao Zhang, Peng Qi, and Christopher D. Manning. "Graph Convolution over Pruned Dependency Trees Improves Relation Extraction". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2205–2215. DOI: `10.18653/v1/D18-1244`. URL: `https://aclanthology.org/D18-1244/`.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. 2017. arXiv: `1703.03400 [cs.LG]`. URL: `https://arxiv.org/abs/1703.03400`.

[37] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: `1609.02907 [cs.LG]`. URL: `https://arxiv.org/abs/1609.02907`.

[38] Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. *Relation Extraction : A Survey*. 2017. arXiv: `1712.05191 [cs.CL]`. URL: `https://arxiv.org/abs/1712.05191`.

[39] Nanyun Peng et al. "Cross-Sentence N-ary Relation Extraction with Graph LSTMs". In: *Transactions of the Association for Computational Linguistics* 5 (Apr. 2017), pp. 101–115. ISSN: 2307-387X. DOI: `10.1162/tacl_a_00049`. eprint: `https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl\_a\_00049/1567450/tacl\_a\_00049.pdf`. URL: `https://doi.org/10.1162/tacl%5C_a%5C_00049`.

[40] Chris Quirk and Hoifung Poon. *Distant Supervision for Relation Extraction beyond the Sentence Boundary*. 2017. arXiv: `1609.04873 [cs.CL]`. URL: `https://arxiv.org/abs/1609.04873`.

[41] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[42] Guillaume Lample et al. "Neural architectures for named entity recognition". In: *arXiv preprint arXiv:1603.01360* (2016).

[43] John Lafferty, Andrew McCallum, Fernando Pereira, et al. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: *Icml*. Vol. 1. 2. Williamstown, MA. 2001, p. 3.

[44] M.A. Hearst et al. "Support vector machines". In: *IEEE Intelligent Systems and their Applications* 13.4 (1998), pp. 18–28. DOI: `10.1109/5254.708428`.

[45] George Krupka and Kevin Hausman. "IsoQuest Inc.: description of the NetOwl™ extractor system as used for MUC-7". In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*. 1998.

[46] Sean R Eddy. "Hidden markov models". In: *Current opinion in structural biology* 6.3 (1996), pp. 361–365.

[47] Douglas Appelt et al. "Sri international fastus systemmuc-6 test results and analysis". In: *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. 1995.

[48] Jagat Narain Kapur. *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989.