CONTRASTIVE FILTERING AND DUAL-OBJECTIVE SUPERVISED LEARNING FOR NOVEL CLASS DISCOVERY IN DOCUMENT-LEVEL RELATION EXTRACTION

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Nicholas Hansen

June 2024
COMMITTEE MEMBERSHIP

TITLE: Contrastive Filtering and Dual-Objective Supervised Learning for Novel Class Discovery in Document-Level Relation Extraction

AUTHOR: Nicholas Hansen

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Lubomir Stanchev, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Foaad Khosmood, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Jonathan Ventura, Ph.D.
Associate Professor of Computer Science
ABSTRACT

Contrastive Filtering and Dual-Objective Supervised Learning for Novel Class Discovery in Document-Level Relation Extraction

Nicholas Hansen

Relation extraction (RE) is a task within natural language processing focused on the classification of relationships between entities in a given text. Primary applications of RE can be seen in various contexts such as knowledge graph construction and question answering systems. Traditional approaches to RE tend towards the prediction of relationships between exactly two entity mentions in small text snippets. However, with the introduction of datasets such as DocRED [36], research in this niche has progressed into examining RE at the document-level. Document-level relation extraction (DocRE) disrupts conventional approaches as it inherently introduces the possibility of multiple mentions of each unique entity throughout the document along with a significantly higher probability of multiple relationships between entity pairs.

There have been many effective approaches to document-level RE in recent years utilizing various architectures, such as transformers and graph neural networks. However, all of these approaches focus on the classification of a fixed number of known relationships. As a result of the large quantity of possible unique relationships in a given corpus, it is unlikely that all interesting and valuable relationship types are labeled before hand. Furthermore, traditional naive approaches to clustering on unlabeled data to discover novel classes are not effective as a result of the unique problem of large true negative presence. Therefore, in this work we propose a multi-step filter and train approach leveraging the notion of contrastive representation learning to discover novel relationships at the document level. Additionally, we propose the use of an alternative pretrained encoder in an existing DocRE solution architecture to
improve F1 performance in base multi-label classification on the DocRED dataset by 0.46.

To the best of our knowledge, this is the first exploration of novel class discovery applied to the document-level RE task. Based upon our holdout evaluation method, we increase novel class instance representation in the clustering solution by 5.5 times compared to the naive approach and increase the purity of novel class clusters by nearly 4 times. We then further enable the retrieval of both novel and known classes at test time provided human labeling of cluster propositions achieving a macro F1 score of 0.292 for novel classes. Finally, we note only a slight macro F1 decrease on previously known classes from 0.402 with fully supervised training to 0.391 with our novel class discovery training approach.
ACKNOWLEDGMENTS

• Advisor: Dr. Lubomir Stanchev, for his valuable guidance, expertise, and support throughout the completion of this thesis.

• Committee member: Dr. Foaad Khosmood, for his participation on my committee and dedication to offering amazing NLP courses at Cal Poly which originally sparked my interest in NLP related work.

• Committee member: Dr. Jonathan Ventura, for his participation on my committee and expert instruction of a wide variety of ML courses at Cal Poly which has allowed me to gain more well-rounded understanding of this domain.

• All Cal Poly faculty for their commitment to my educational success and for fostering a welcoming learning environment.

• Friends and family for all of their love and support throughout the completion of my degree.
# Table of Contents

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
</tr>
</tbody>
</table>

## CHAPTER

1. Introduction .............................................. 1

2. Related Research ........................................ 5
   2.1 Document-Level Relation Extraction ...................... 5
      2.1.1 Graph-Based Models .................................. 6
      2.1.2 Transformer-Based Models ............................ 7
   2.2 Transformer Models (RoBERTa and LUKE) .................. 8
   2.3 Contrastive Learning .................................... 10
      2.3.1 Self-Supervised Contrastive Learning ............... 11
      2.3.2 Supervised Contrastive Learning .................... 12
      2.3.3 Contrastive Learning in NLP ......................... 13
   2.4 Novel Class Discovery ................................... 15
      2.4.1 Novel Class Discovery ................................ 15
      2.4.2 Generalized Category Discovery/Open World Semi-Supervised Learning .................................. 16
      2.4.3 Novel Class Discovery in Relation Extraction ....... 18

3. Problem Description and Proposed Solution ................. 19
   3.1 Terminology ............................................. 19
   3.2 Problem Definitions ..................................... 21
      3.2.1 Modified Novel Class Discovery ...................... 21
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1.1</td>
<td>The Purity Problem</td>
<td>26</td>
</tr>
<tr>
<td>3.2.1.2</td>
<td>The Semantically Meaningful Cluster Problem</td>
<td>28</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Novel Class Informed Document-Level Relation Extraction</td>
<td>29</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Final Combined Objective</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>Motivation</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td>Proposed Solution</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Dataset (DocRED)</td>
<td>33</td>
</tr>
<tr>
<td>4.1</td>
<td>Data Format</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Size and Label Distribution</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Label Correlations and Co-Occurences</td>
<td>38</td>
</tr>
<tr>
<td>4.4</td>
<td>Document Preprocessing</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>System Overview and Design</td>
<td>40</td>
</tr>
<tr>
<td>5.1</td>
<td>ATLOP</td>
<td>40</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Pretrained Encoder</td>
<td>40</td>
</tr>
<tr>
<td>5.1.2</td>
<td>LUKE Pooling Modification</td>
<td>41</td>
</tr>
<tr>
<td>5.1.3</td>
<td>LUKE-TACRED</td>
<td>42</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Localized Context Pooling</td>
<td>43</td>
</tr>
<tr>
<td>5.1.5</td>
<td>Adaptive Threshold Loss</td>
<td>45</td>
</tr>
<tr>
<td>5.2</td>
<td>Holdout Set</td>
<td>47</td>
</tr>
<tr>
<td>5.3</td>
<td>Self-Supervised Contrastive Candidate Filtering</td>
<td>48</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Self-Supervised Contrastive Learning</td>
<td>50</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Filtering</td>
<td>53</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Temperature</td>
<td>54</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Early Stopping</td>
<td>55</td>
</tr>
<tr>
<td>5.4</td>
<td>Dual-Objective Supervised Learning</td>
<td>59</td>
</tr>
</tbody>
</table>
5.4.1 Adaptive Threshold Multi-Label Learning .......................... 60
5.4.2 Binary Positive Logit Learning ................................. 60
5.4.3 Dual-Objective Loss Function .................................. 61
5.4.4 Binary Positive Logit Thresholding ............................... 62
5.5 Clustering and Novel Class Proposition ............................. 62
5.5.1 UMAP Dimensionality Reduction ................................. 62
5.5.1.1 The Curse of Dimensionality ............................... 63
5.5.1.2 UMAP - Preserving Global and Local Distances .......... 64
5.5.2 HDBSCAN Clustering ........................................... 65
5.5.2.1 Ranked Cluster Proposition ................................. 66
5.6 Human Labeling and Final Training ............................... 66
6 Experiments and Evaluation ........................................... 69
6.1 Base Encoder Selection ............................................ 69
6.2 Filtering Evaluations .............................................. 70
6.2.1 Metrics .......................................................... 70
6.2.1.1 Holdout Retention and True Negative Retention ........ 70
6.2.2 Results .......................................................... 71
6.3 Clustering Evaluation .............................................. 73
6.3.1 Metrics .......................................................... 74
6.3.1.1 Silhouette Coefficient .................................... 74
6.3.1.2 Average Weighted Holdout Cluster Purity ............... 74
6.3.1.3 Average Percent Holdout in Max Cluster ............... 74
6.3.1.4 Holdout Noise Percentage ................................ 75
6.3.2 Results .......................................................... 77
6.4 Final Model Evaluation ............................................ 77
6.4.1 Metrics ........................................ 78

6.4.1.1 Novel Holdout F1s .......................... 78

6.4.1.2 Known Class Macro F1 ................. 78

6.4.1.3 Overall F1 ................................. 78

6.4.2 Results .......................................... 79

7 Conclusion and Future Research ............... 83

7.1 Summary ........................................... 83

7.2 Limitations ........................................ 83

7.3 Future Work ....................................... 84

7.4 Threats to Validity ............................... 85

BIBLIOGRAPHY ....................................... 87
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>General statistics of instances in DocRED. Note the large difference between the number of negatives and number of positives along with the large difference between the maximum relation counts and median relationship counts.</td>
<td>36</td>
</tr>
<tr>
<td>5.1</td>
<td>Measuring only the amount of times these relationships appear alone in each set.</td>
<td>48</td>
</tr>
<tr>
<td>6.1</td>
<td>Performance comparison of the encoder variations tested for ATLOP [39] architecture.</td>
<td>69</td>
</tr>
<tr>
<td>6.2</td>
<td>We observe temperature of 0.01 performs the best in the holdout percentage of our candidate set and retains the highest amount of holdouts in the candidate set aside from no contrastive filtering.</td>
<td>71</td>
</tr>
<tr>
<td>6.3</td>
<td>Number of each holdout relation preserved in the candidate set after contrastive filtering.</td>
<td>71</td>
</tr>
<tr>
<td>6.4</td>
<td>Candidate set make up after binary logit filtering.</td>
<td>72</td>
</tr>
<tr>
<td>6.5</td>
<td>Number of each holdout relation preserved in the candidate set to be clustered after binary logit filtering. We observe that even with the bias against specific relationships such as ‘P569’ and ‘P570’ the retention of these relationships is still superior to random sampling.</td>
<td>72</td>
</tr>
<tr>
<td>6.6</td>
<td>Clustering statistics</td>
<td>76</td>
</tr>
<tr>
<td>6.7</td>
<td>Overall F1 Scores</td>
<td>79</td>
</tr>
<tr>
<td>6.8</td>
<td>Holdout Specific F1s</td>
<td>79</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Example of the document level relation extraction problem for entities “Albert Einstein” and “Switzerland”. There are two relationships between the entities: “moved to” and “studied in” in addition to multiple mentions of both entities.</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>LUKE’s Pretraining Objective [35].</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison of self-supervised contrastive learning and supervised contrastive learning applied to image data. Note that in self-supervised contrastive learning only the augmentation of an image is used as a positive example where in supervised contrastive learning pairs of elements of the same class and their augmentations are used as positive pair examples [16].</td>
<td>12</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison of text augmentation strategies tested for SimCSE. ‘None’ represents SimCSE pretrained dropout augmentation [11].</td>
<td>13</td>
</tr>
<tr>
<td>2.5</td>
<td>Self-Supervised SimCSE Augmentation Procedure: Involves leveraging the different dropout masks of the pretrained encoder to produce augmentations [11].</td>
<td>15</td>
</tr>
<tr>
<td>2.6</td>
<td>Image-Based NCD, the model must only identify novel classes at test time. Assumed testing set only composed of novel samples [27].</td>
<td>16</td>
</tr>
<tr>
<td>2.7</td>
<td>Image-Based GCD, the model must identify both known and novel classes at test time [5].</td>
<td>17</td>
</tr>
<tr>
<td>3.1</td>
<td>Definition of “instances” in the context of DocRE. Note that each entity-pair receives two instances based upon the subject to object relationship organization. Ideally, bidirectional relationships such as “married” mean both entity-pair permutations are positive instances.</td>
<td>20</td>
</tr>
<tr>
<td>3.2</td>
<td>Process of novel class discovery in our training data. We must discover novel classes in the unlabeled training set and then include these novel classes in our labeled training data for a supervised classifier evaluated on the test set. The test set includes known classes ($C_k$), novel classes ($C_n$), and true negatives.</td>
<td>21</td>
</tr>
</tbody>
</table>
3.3 Image-Based version of our modified NCD problem, the model must identify only novel classes with a presence of many true negative examples in the unlabeled set.  

3.4 Text version of our modified NCD problem for a single document with the assumption that only the relationship class “moved to” is known and labeled beforehand. Both relationship classes “studied in” and “worked in” are novel classes that must be discovered.  

3.5 Comparison of different NCD problems versus our modified NCD problem. We do not assume all instances in the unlabeled set belong to a class.  

3.6 Embedding space when just clustering on random subset of 150,000 unlabeled examples. We observe low cluster purity with relations ‘P69’ and ‘P577’. However, we do observe these instances clustering together indicating some semantic grouping behavior.  

3.7 Clustering on naive subset of 300,000 unlabeled examples. The purity problem further increases due to continuous introduction of true negative examples.  

3.8 Difference between candidate set sampling methods. Informed sampling strives to only retain novel class instances to be clustered.  

3.9 An example of the meaningful cluster problem applied to images. Images are clustered unintentionally by color rather than the objects in the image. This is similar to the meaningful cluster problem in DocRE when naively clustering unlabeled examples [27].  

3.10 Overall flow chart of our NCD approach in DocRE.  

4.1 Example of a multi-hop relationship. The relational fact “John lives in United States” can only be obtained by reasoning through the intermediary entity “Arizona” and corresponding intermediary relationships.  

4.2 Label distribution of relationships in the DocRED training dataset. Note that the y-axis represents a log scale of instances and the leftmost relation is the “NA”/“no relation” class.  

4.3 Co-Occurrence matrix of different relationships in the training set of DocRED. We note quite sparse co-occurrence rates among each relationship type with only a few relationship types being strongly correlated with one another.  

4.4 Association rules of relation co-occurrences in DocRED dataset.
4.5 Different preprocessing strategies for entities between base RoBERTa model and LUKE model. ................................................................. 39

5.1 LUKE modification of attention mechanism, introduces unique query matrices for each paired type (entity, word) [35]. ........................................ 43

5.2 Simplified flowchart of ATLOP localized context pooling mechanism. ................................................................. 44

5.3 Adaptive thresholding loss. Threshold logit is learned with respect to the output of the positive and negative classes. Positive logit values are pushed above this learned threshold, and negative class logits are pushed below this threshold [39]. ................................................................. 46

5.4 Our contrastive learning architecture flowchart. ................................................................. 49

5.5 Intuition behind contrastive learning applied to DocRE. ................................................................. 50

5.6 Visualization of contrastive filtering metric. Unlabeled instances with a higher cosine similarity to the positive mean than the average unlabeled cosine similarity to the positive mean are selected as candidate instances. ................................................................. 52

5.7 Higher temperatures in contrastive learning tend to be associated with more uniform distributions whereas lower temperatures are associated with more rigid distributions [33]. ................................................................. 54

5.8 The ratio of the holdouts and true negatives that are retained in the candidate set depending on the temperature used for self-supervised contrastive learning. ................................................................. 56

5.9 Ratio of holdouts with respect to quantity of samples in the candidate set over contrastive training epochs. ................................................................. 57

5.10 Ratio of holdouts with respect to all holdouts in dataset in the candidate set and cosine similarity between the average positive and average unlabeled example over contrastive training epochs. ................................................................. 58

5.11 Dual-Objective architecture learning on multi-label classification head and binary classification logit. ................................................................. 60

5.12 Binary positive logit filtering mechanism. Candidates from contrastive filtering are predicted after training with labeled data and unlabeled non-candidates. Candidates that are not predicted as “positive” relationships are merged back into the true negative set for final training. Candidates that are predicted as “positive” relationships are clustered and proposed to human labeler. ................................................................. 61
5.13 Human-in-the-loop labeling method. Clusters are inspected by random sample propositions from clusters and accepted (pseudolabeled and used for training) or rejected (used as true negatives). In this example, cluster 1 is accepted and all instances in its cluster are used for training the final model as a new class, whereas cluster 2 is rejected and its instances are relegated to true negatives for final model training.

6.1 UMAP clustering results on naive random subset of unlabeled data.

6.2 UMAP clustering results on our informed subset of unlabeled data. Red outlines indicate a top 30 cluster proposal.

6.3 Relationship between cluster purity and precision with our NCD (Whole Cluster) model.

6.4 Relationship between representation of holdout relationship type in its cluster vs. and recall with our NCD (Whole Cluster) model.
Chapter 1

INTRODUCTION

Document-level relation extraction (DocRE) is a relatively recent expansion of traditional relation extraction (RE) that aims to identify relationships between named entities that span an entire document. Extracting relationships at the document level serves a variety of applications in the construction of knowledge graphs, question-answering systems, and genealogical graphs [24], along with enhancing information retrieval in domains such as healthcare, business intelligence, legal document analysis, and scientific research.

Existing methods in DocRE primarily focus on multi-label classification of predefined relationship categories. While various techniques such as graph-based and transformer-based approaches have achieved significant advancements, there has not been much work examining the ability of these models to discover entirely new relationship types within a corpus. A major hurdle in this pursuit lies in the vast search space for clustering algorithms. In DocRE, most entity pairs represent a true negative example where there is no relationship between them. Clustering to find novel categories on this massive and noisy dataset with these irrelevant true negative examples leads to computationally expensive and ineffective exploration. For this work, we look towards contrastive learning and novel class discovery methods which are techniques popular within the image domain, to assist in reducing the search space for clustering techniques.
This thesis builds upon the foundation of the ATLOP architecture [39], a prominent DocRE model, and proposes a novel methodology for discovering new relationships within a corpus using the DocRED dataset.

We propose an improved base encoder for ATLOP by leveraging LUKE [35], a RoBERTa-based transformer architecture geared towards entity-related tasks. This enhancement slightly improves the multi-label F1 score of ATLOP, paving the way for a more robust foundation for novel class discovery.

Additionally, we propose a multi-step approach that tackles the challenge of novel class discovery in DocRE by addressing the large clustering search space filled with entity-pair examples with no concrete relationship, which we define as “true negative” examples. First, we leverage self-supervised contrastive learning to identify promising candidate entity pairs from the unlabeled data and reserve them for future prediction and clustering. Secondly, we introduce a dual-objective supervised model that learns both fine-grained and general positive/negative relationship representations and use the positive/negative binary prediction to further filter our candidate set. We propose a ranking system for clustered novel class candidates to facilitate human-in-the-loop evaluation of proposed clusters where human labelers can accept or reject clusters of samples for pseudolabeling and further training.

Our proposed approach departs from the naive approach of clustering on a subset of unlabeled relation embeddings, which risks overfitting novel classes as negatives and clustering on non-informative features. Our contributions are formalized as:

- **Improved ATLOP Base Encoder:** We demonstrate a 0.46 point increase in multi-label F1 score by replacing ATLOP’s base encoder with LUKE leveraging entity-aware attention and traditional relation extraction pretraining on the TACRED [38] dataset.
- **Self-Supervised Contrastive Filtering:** We introduce a self-supervised contrastive learning method to filter potential novel class candidates and set them aside during training, significantly reducing search space for clustering. We find this approach improves novel class sample representation in the candidate set by 1.65 times compared to the naive candidate set solution.

- **Dual-Objective Supervised Training:** We propose a dual training objective that learns both the fine-grained multi-label relation representations and general “positive” and “negative” relation representations in order to further filter our candidates and cluster only the most likely novel class candidates. We observe this further increases novel class representation in the candidate set by 3 times.

- **Ranked Novel Cluster Proposition:** We propose a ranked cluster proposition approach to guide human labelers to the most important clusters under the intuitive assumption that novel class clusters will be uniquely situated in the embedding space. We find that 4/6 of our holdout novel classes are retained in the top 30 cluster propositions and improve cluster purity for novel classes by nearly 4 times compared to the naive solution.

- **Final Combined Training:** We pseudolabel instances from clusters accepted by human labelers and use them as training data for novel classes in a final fully supervised training regimen. In doing so we achieve a macro F1 score of 0.292 for novel classes and 0.391 for known classes.

By bridging the gap in novel class discovery for DocRE, we hope to inspire more research into applications of novel class discovery within the text domain and contribute a meaningful pipeline for managing this task in DocRE.

In Chapter 2, we will examine a variety of related works and previous approaches applicable to this thesis. In Chapter 3, we will define our motivations for tackling
this problem in addition to defining the various sub-problems relevant to our final system. In Chapter 4, we examine the DocRED dataset’s structure, class distribution, and co-occurrence distribution. In Chapter 5, we explain our system architecture and approach to the NCD problem in DocRE in more detail. In Chapter 6, we evaluate the components of our system through a variety of experiments. Finally in Chapter 7, we summarize our contributions and additionally address limitations and future work.
2.1 Document-Level Relation Extraction

Document-level relation extraction (DocRE) is as a natural extension of the well-established task of sentence-level relation extraction (RE). Sentence-level RE focuses on identifying relationships between entities within a single sentence, whereas DocRE tackles the more challenging task of extracting relations that can span across multiple sentences or even the entire document (Figure 2.1). This generalization is motivated by the limitations of sentence-level approaches, which often fail to capture the rich context necessary to understand complex relationships.

The popularity of sentence-level RE is evident in the widespread use of datasets such as TACRED [38]. These datasets have been instrumental in fine-tuning various text transformer models for relation classification. However, the introduction of DocRED [36] has marked a shift in relation extraction tasks, prompting the exploration of DocRE as a new research area.

Figure 2.1: Example of the document level relation extraction problem for entities “Albert Einstein” and “Switzerland”. There are two relationships between the entities: “moved to” and “studied in” in addition to multiple mentions of both entities.
Unlike sentence-level RE, which typically deals with single relations per entity pair and one instance per entity, DocRE allows for the existence of multiple relations between entities along with multiple mentions of any given entity within a document. DocRE additionally necessitates reasoning across multi-hop relational paths between entities that may not be directly connected within a single sentence. These “multi-hop” relations require the model to trace connections and infer relationships through intermediate entities (referenced further in Chapter 4).

Two major architectural approaches have gained popularity in DocRE, graph-based and transformer-based [8].

2.1.1 Graph-Based Models

Graph-based network approaches to document-level relation extraction have emerged as a powerful methodology for understanding complex relationships between entities within long-form text documents. These approaches leverage the structure of graphs to model and infer relationships that span sentences, capturing both local and global contextual information.

One of the primary advantages of graph-based methods in DocRE is their ability to incorporate various types of dependencies and contextual information. By representing documents as graphs where nodes correspond to entities or spans of text and edges represent potential relationships or syntactic dependencies, these methods can effectively capture interactions that are not immediately adjacent in the text. This is particularly useful for document-level tasks where relevant entities and their relations may be dispersed across multiple sentences or paragraphs. Approaches such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) have been
applied successfully in this domain [8]. However we will not be exploring applications of graph-based approaches in this thesis.

2.1.2 Transformer-Based Models

Transformer-based approaches have also gained prominence in document-level relation extraction (DocRE) due to their ability to model long-range dependencies and contextual information effectively. The self-attention mechanism of transformers is particularly well-suited for capturing intricate patterns and relationships across entire documents, making them ideal for tasks that require understanding global context.

A foundational work in this domain is the paper “Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling” (ATLOP) [39]. This work introduces a novel approach that combines adaptive thresholding with localized context pooling to enhance relation extraction. Localized context pooling introduces a method for representing a pair of entities based upon the interactions between their shared context. Furthermore, the adaptive thresholding mechanism dynamically adjusts the threshold for relation classification, improving the model’s ability to differentiate between relevant and irrelevant relationships dependent on learned entity pair representations. The paper has laid the groundwork for subsequent advancements in transformer-based DocRE by demonstrating the effectiveness of combining adaptive techniques with robust context modeling.

Building on this foundation, the paper “DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction” [19] modified ATLOP to guide the transformer’s attention mechanism using evidence as a supervisory signal. DREEAM leverages an evidence-guided attention mechanism that highlights critical parts of the text, thereby directing the model’s focus towards the most relevant
information for relation extraction. This approach addresses one of the main challenges in DocRE of identifying and concentrating on the portions of text that contain essential relational information while ignoring irrelevant details. By incorporating evidence-based guidance, DREEAM improves the accuracy and reliability of relation extraction, especially in complex scenarios where relevant information is dispersed across various sections of the document.

Another significant work is presented in the paper “Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation” [26]. This research introduces adaptive focal loss, knowledge distillation and the usage of axial attention. Adaptive focal loss is designed to address the imbalance in relation extraction tasks, where some relations are more frequent than others. By adaptively focusing on harder to classify instances, the model learns more effectively from underrepresented relations. Knowledge distillation involves transferring knowledge from a larger, pre-trained model to a smaller, more efficient model, thereby enhancing its performance without significantly increasing computational complexity. Finally, the authors propose the use of axial attention in order to better tackle the multi-hop relationship problem.

This thesis will, similar to the works mentioned above, leverage ATLOP as its base architecture.

2.2 Transformer Models (RoBERTa and LUKE)

The transformer model introduced by Vaswani et al. [29] was an important step for natural language processing (NLP) tasks by departing from recurrent architectures. It leverages an encoder-decoder structure, where the encoder generates contextualized
representations for each element in an input sequence, which is achieved through stacked encoder layers, each containing self-attention and feed-forward sub-layers.

Transformer-based models, such as BERT [9] and RoBERTa [18], achieve state-of-the-art performance in many NLP tasks and are pre-trained on massive text corpora to learn general representations of language. Both BERT and RoBERTa use a masked language modeling objective and a next sentence prediction objective during pre-training. However, they differ in some key aspects. BERT uses a fixed vocabulary size and static masking, where tokens to be masked are determined beforehand, Conversely, RoBERTa utilizes a larger vocabulary and dynamic masking, where the masking pattern changes during training, potentially leading to a more robust understanding of language.

Traditional transformer models, including RoBERTa, process all words within a sequence identically. While powerful for various NLP tasks, this approach can be limiting for DocRE where specific attention to entities can be useful.

Building upon RoBERTa, LUKE [35] introduces entity embeddings as dense vector representations that encapsulate the meaning and context associated with named entities. By incorporating these embeddings, LUKE injects a deeper understanding of the crucial elements likely to be involved in relationships within the document. This enriched representation allows the model to prioritize information relevant to DocRE tasks, offering an advantage over RoBERTa’s more general-purpose approach.

The core innovation of LUKE lies in its entity-aware self-attention mechanism. The standard self-attention mechanism in transformers, employed by RoBERTa and BERT, enables each element within the sequence to attend to all others. LUKE modifies this process by additionally considering the type of tokens (word or entity) in the attention calculation. By introducing special considerations for interactions between
entities and non-entities, LUKE can achieve a more nuanced understanding of the relationships they participate in.

Additionally, LUKE inherits the transformer’s ability to capture long-range dependencies, a capability also present in RoBERTa. This enables the model to consider entities even if they are far apart in the document, as long as they are relevant to the potential relationship being analyzed. This capability is essential for DocRE tasks, where relationships can sometimes be established between entities spread across the document.

2.3 Contrastive Learning

Contrastive learning is a powerful technique for learning informative representations from unlabeled data. Unlike traditional supervised learning that relies on large amounts of labeled data, contrastive learning leverages the inherent structure within unlabeled samples for training. This approach hinges on the principle of contrasting similar and dissimilar data points, by bringing similar instances closer together and pushing dissimilar ones apart in a latent embedding space. The model learns meaningful representations that capture essential features and relationships within the data.
This ability to extract knowledge from unlabeled data has served as a valuable pre-training task for various learning domains, especially for computer vision. Original contrastive learning techniques serve as a type of self-supervised learning, where the model itself generates supervisory signals from unlabeled data via artificial positive pairs created through sample augmentation.

2.3.1 Self-Supervised Contrastive Learning

Self-supervised contrastive learning has primarily been applied for learning image representations from unlabeled data. Unlike traditional supervised learning approaches that rely on explicit labels, self-supervised contrastive learning focuses on maximizing the agreement between different augmented views of the same image. This is achieved by passing these views through an encoder network that projects them into a latent space. A contrastive loss function then pulls together the representations of the augmented views, forcing the model to learn features that capture the underlying structure of the image, independent of its specific identity.

One influential self-supervised contrastive approach is SimCLR [6]. SimCLR utilizes a stochastic data augmentation pipeline to generate multiple augmented versions of a single image. These can involve random crops, color jittering, or random rotations. They then use these random augmentations of the image as positive representation pairs and the rest of the images in the batch as negative representation pairs and apply the proposed contrastive loss function to learn meaningful representations of images.
Figure 2.3: Comparison of self-supervised contrastive learning and supervised contrastive learning applied to image data. Note that in self-supervised contrastive learning only the augmentation of an image is used as a positive example where in supervised contrastive learning pairs of elements of the same class and their augmentations are used as positive pair examples [16].

2.3.2 Supervised Contrastive Learning

Building upon the success of self-supervised contrastive learning, recent work such as SupCon [16] has explored leveraging class labels to further enhance representation learning. Supervised contrastive learning incorporates class information into the contrastive loss function, aiming to create an embedding space where data points from the same class are clustered even closer, while maintaining the separation between dissimilar classes. This is achieved by contrasting positive pairs formed by all samples from the same class against negative pairs drawn from the rest of the batch. Compared to unsupervised methods that rely solely on data similarity, SupCon explicitly exploits class labels to guide the model towards learning class-discriminative representations.

Experimental results have shown that models pretrained with SupCon can outperform traditional supervised learning approaches that rely solely on cross-entropy loss in benchmark datasets like ImageNet. This suggests that supervised contrastive learning
Figure 2.4: Comparison of text augmentation strategies tested for SimCSE. 'None' represents SimCSE pretrained dropout augmentation [11].

can effectively leverage both labeled and unlabeled data, potentially leading to better generalization capabilities, particularly in scenarios with limited labeled data.

2.3.3 Contrastive Learning in NLP

While contrastive learning has been primarily popular in the image domain, this thesis is interested in examining its applications to NLP and document-level relation extraction. One of the core obstacles in utilizing contrastive learning in NLP is the difficulty of preserving meaning in augmentation. Augmentation is much easier to perform on image data (flipping, cropping, etc), while augmentation in NLP often results in some sort of disruption of semantic meaning. Although there have been major advancements for more complicated augmentation strategies in NLP [3], a very simple practical application of text augmentation comes from SimCSE [11].

The SimCSE framework [11] provides an approach to self-supervised contrastive learning in NLP to learn sentence embeddings. SimCSE utilizes a standard pre-trained language model (e.g., BERT) as the encoder and treats internal model dropout, a common regularization technique during training, as a form of data augmentation rather than the traditional, more complicated, text augmentation schemes [3]. By
feeding the original document through the pretrained encoder twice (Figure 2.5), SimCSE creates a positive pair. The model is then trained to minimize the distance between the embeddings of these positive pairs in the representational space. This approach allows SimCSE to exploit the vast amount of unlabeled text data to learn sentence embeddings that capture semantic similarity effectively. This augmentation strategy was shown in their experiments to be more effective than traditional text augmentation techniques for self-supervised contrastive learning (Figure 2.4).

The success of SimCSE highlights the potential of unsupervised contrastive learning for learning meaningful representations in NLP tasks. This approach offers an alternative to traditional supervised techniques that require large amounts of labeled data for sentence embedding generation.

Building on these advancements, a recent work “An Effective Deployment of Contrastive Learning in Multi-label Text Classification” [17] proposes a series of novel contrastive loss functions specifically designed for multi-label text classification. Their approach moves beyond simple label overlap and explores various aspects of the label relationships. This includes losses that focus on intra-label similarity, Jaccard similarity-based losses, and even a stepwise contrastive loss that progressively builds upon the relationships between labels. Their experiments demonstrate that these tailored contrastive losses can outperform previous methods, highlighting the importance of considering the intricacies of multi-label data when designing supervised contrastive learning strategies. Such investigations of contrastive learning in various NLP objectives lay the groundwork for our motivations for leveraging this method in DocRE.
2.4 Novel Class Discovery

Representation learning methods such as contrastive learning have found prominent usage in the task of novel class discovery. Novel class discovery (NCD) and its variations aim to discover novel classes in an unlabeled set, that can be used for further downstream tasks. There exists two major versions of NCD, traditional, and generalized category discovery (GCD). However, this thesis must make a modification to some of the assumptions of the NCD problem variations which will be described in Chapter 3.

2.4.1 Novel Class Discovery

Traditional novel class discovery takes the form of a closed-world scenario, where given a labeled set of known classes and unlabeled set of unknown classes, the model is tasked with the goal of clustering unlabeled data into unknown novel classes (Figure 2.6).
Figure 2.6: Image-Based NCD, the model must only identify novel classes at test time. Assumed testing set only composed of novel samples [27].

Prominent approaches of NCD are usually found in the domain of image classification in either one-stage or two-stage methods. Works such as [14] act as two-stage methods that first learn a similarity function on the labeled data which is then leveraged to pseudolabel pairs of similar unlabeled examples and cluster to learn novel classes. Additionally, AutoNovel [12] exists as a single stage method where a clustering objective is learned on the unlabeled data, while labeled data is learned with a classification objective. Approaches to NCD in two-stage methods are primarily concerned with learning domain knowledge on the labeled set first and then applying this knowledge to the unlabeled set in order to cluster novel classes effectively. One-stage approaches tend to leverage labeled data and unlabeled data simultaneously in a joint learning objective. Traditional NCD does not concern itself with the correct classification of labeled data, rather it only focuses on the detection of novel classes.

2.4.2 Generalized Category Discovery/Open World Semi-Supervised Learning

Generalized Category Discovery (GCD), also known as Open World Semi-Supervised Learning (Open SSL), is a generalization of the novel class discovery paradigm to concern itself with both the correct classification of known classes along with the
Figure 2.7: Image-Based GCD, the model must identify both known and novel classes at test time [5].

discovery of novel classes in the unlabeled set (Figure 2.7). GCD removes the restriction in NCD that instances in the unlabeled set will only be members of novel classes. Therefore, the task objective in GCD then becomes to leverage the labeled and unlabeled data to correctly classify instances of known classes and cluster novel instances. A popular approach to GCD [30] utilizes a mix of self-supervised contrastive learning and supervised contrastive learning on the training data and then uses a semi-supervised k-means++ [2] to cluster all instances (labeled and unlabeled). Furthermore, an approach called ORCA [5] introduces a semi-supervised approach to GCD where bias towards known classes during training is addressed utilizing an uncertainty adaptive margin mechanism combined with a pairwise similarity objective. This mechanism controls the speed of learning on known classes in order to allow the network to learn a variable amount of novel classes.

Most approaches to both NCD and GCD have been related to image-based tasks. However there has been emerging research of NCD/GCD applied to text and, more particularly, relation extraction.
2.4.3 Novel Class Discovery in Relation Extraction

OpenRE is the application of NCD to the relation extraction task. There has been limited but interesting research in this area, including [34] where the authors utilize relational twin networks (RTNs/RSNs) and approach the NCD problem in RE as a knowledge transfer problem to learn a similarity metric from the labeled set applied to the unlabeled set (similar to work in [14]). However, this method does not consider the existence of true negative examples which are a significant problem in the RE domain. Furthermore, work in [13] approach NCD in RE while considering the inclusion of true negatives. The authors utilize language model prompt templates for entity pairs and average representations of both constrained and unconstrained vocabulary predictions for relationships. Clustering is then performed and positive class instances act as votes to bifurcate the clustering space between novel and known classes. This method considers the existence of negative examples, however the quantity of the true negative examples is somewhat unrealistic. Additionally, neither of these approaches focus on RE at the document level.

One of the key problems with many of the NCD and GCD approaches is that true negatives in the unlabeled set are either not considered altogether, or not considered in realistic volume. Additionally, these approaches to NCD applied to RE are not expanded to DocRE. We augment the problem statement of NCD in order to apply the realistic scenario of DocRE in Chapter 3.
Chapter 3

PROBLEM DESCRIPTION AND PROPOSED SOLUTION

There are two main goals of this thesis: to both improve the classification accuracy of a base multi-label document-level relation classifier on known relationships, in addition to the discovery and labeling of novel classes within the unlabeled training set.

3.1 Terminology

Throughout this work we will refer to representations of individual direction dependent (subject to object) entity-pairs from a document as “instances” or “examples”. Additionally, we will often refer to instances in the context of being either a “positive” instance, where there is some relationship that exists between the two entities in the pair, or a “negative” instance, where no tangible directional relationship exists between the entities in the pair (Figure 3.1).

Furthermore, when referring to our unlabeled set, we introduce the notions of “true negative” instances and “novel class” instances. True negative instances refer to the “negative” examples in the unlabeled set whereas “novel class” instances refer to undiscovered “positive” examples that express a relationship class that is not present in the labeled set (Figure 3.4).
Figure 3.1: Definition of “instances” in the context of DocRE. Note that each entity-pair receives two instances based upon the subject to object relationship organization. Ideally, bidirectional relationships such as “married” mean both entity-pair permutations are positive instances.
Figure 3.2: Process of novel class discovery in our training data. We must discover novel classes in the unlabeled training set and then include these novel classes in our labeled training data for a supervised classifier evaluated on the test set. The test set includes known classes ($C_k$), novel classes ($C_n$), and true negatives.

3.2 Problem Definitions

We formalize three distinct problems that must be addressed in our pipeline for both discovering novel relationships and classifying these novel relationships in tandem with known positive relationships.

3.2.1 Modified Novel Class Discovery

The problem statement of NCD/GCD applied to DocRE must be slightly modified to account for the presence of true negative instances (Figure 3.3, 3.4). We do not assume that all sample instances in our unlabeled set must be assigned to a novel class. However, we do assume that none of the samples in the unlabeled set belong to any of our already known classes. We additionally do not assume the number of novel
Figure 3.3: Image-Based version of our modified NCD problem, the model must identify only novel classes with a presence of many true negative examples in the unlabeled set.

classes in our unlabeled set as opposed to other novel class discovery methods, and introduce the notion of novel class proposals that will be either confirmed or rejected with human evaluation. Finally, we restrict the problem so that the distributions of our training and test sets will be identical, meaning no new novel classes will be present in the test distribution that are not already in the unlabeled training distribution. This restriction allows us to use the instances of discovered novel classes in our unlabeled training set to train a supervised model for later downstream inference (Figure 3.2). Therefore, our approach can be seen as a mix of modified novel class discovery including true negatives amongst training data and generalized category discovery including true negatives at test time. It should be noted that labeled instances are considered fully labeled, meaning that novel classes are not considered to be applicable to entity-pair instances in the labeled set.

We discover novel classes disjointly from our known classes in the unlabeled training data, pseudolabel members of these novel classes, and then perform traditional supervised training and inference utilizing these pseudolabeled instances along with the originally known class instances.
We divide this problem into two steps of discovery and inference. The end goal of this pipeline is for human labelers to either accept or reject a few instances from each proposed novel cluster, and then all samples within each accepted cluster are used as training data for the novel class. Inference is then done on the test set, which is composed of a mix of both known class instances, novel class instances, and true negative instances to evaluate the quality of novel class representation captured in the analysis of the training data (Figure 3.2).

Formally, we define the steps of our modified novel class discovery objective as follows:

1. Given a labeled training set of instances $X_l$ with known relationship classes $C_k$ and unlabeled set of instances $X_u$, discover an arbitrary number of novel classes $C_n$ in $X_u$ with the assumptions $C_k \cap C_n = \emptyset$, not every $x \in X_u$ has a label in $C_n$, and no $x \in X_u$ has a label in $C_k$. 

Figure 3.4: Text version of our modified NCD problem for a single document with the assumption that only the relationship class “moved to” is known and labeled beforehand. Both relationship classes “studied in” and “worked in” are novel classes that must be discovered.
Figure 3.5: Comparison of different NCD problems versus our modified NCD problem. We do not assume all instances in the unlabeled set belong to a class.

2. Given the discovered novel classes $C_n$, create a new labeled set $X_z = X_l \cup X_d$ where $X_d \in X_u$ and $x$ has a single label in $C_n$ for all $x \in X_d$.

3. Perform the subsequent DocRE multi-label classification problem with training set $X_z \cup X_u - X_d$ and unlabeled test set $X_i$, where $x \in X_i$ can have either no label, a single label in $C_n$, or multiple labels in $C_k$.

It is important to note that we do not consider multi-label instances for $C_n$. We assume discovered novel instances $x \in X_d$ have a single label corresponding to a single class in $C_n$. We leave it to future work to handle the multi-label NCD problem.
Figure 3.6: Embedding space when just clustering on random subset of 150,000 unlabeled examples. We observe low cluster purity with relations ‘P69’ and ‘P577’. However, we do observe these instances clustering together indicating some semantic grouping behavior.
3.2.1.1 The Purity Problem

A naive solution to the NCD problem would be to simply train a supervised model on both the labeled and unlabeled training set with a portion of the unlabeled set removed and reserved for inference. One could cluster on the inferred embeddings of the reserved samples (also called the candidate set) and use these clusters for novel class proposal. Theoretically, this approach is effective in situations where the distribution of the novel classes in the unlabeled set is somewhat uniform and free from true negative examples. However, in reality, with regards to NCD in DocRE, positive class instances are extremely sparse and in a long-tailed distribution where there are many instances of a few classes and few instances of many classes. Therefore, a subset of the unlabeled examples will not contain sufficient relative evidence for novel classes with respect to the true negatives for clustering to be effective. Furthermore, part of the DocRE problem is distinguishing positive from “no relation” instances, which
Figure 3.8: Difference between candidate set sampling methods. Informed sampling strives to only retain novel class instances to be clustered.

requires training on negative instances. Random sampling is likely to select the majority of the novel class examples and overfit to them as negative representations. We observe this problem when attempting cluster in the manner described above, we see that true novel class instances are enveloped in clusters polluted with many true negative examples (Figure 3.6). Additionally, simply increasing the subset of unlabeled data to use for clustering exacerbates the purity problem (Figure 3.7).

It should be noted that this method is still quite effective in bringing together unlabeled examples in a (presumed) semantically meaningful way. However, in practical use cases where clusters are proposed to a human labeler, the purity of the proposed clusters pose an issue for selecting or rejecting them based upon a subsample of the cluster instances. For example, in the human labeling stage for a cluster, if only 5% of the proposed cluster instances are comprised of true examples of the novel class, then that cluster is unlikely to be deemed meaningful as opposed to if the cluster was comprised of 50% true novel class instances. Additionally, the majority of the training
data constructed from that cluster will be falsely labeled. Since our method relies on human labeling and supervised training based upon cluster samples, we strive for cluster proposals that are as pure as possible with respect to the novel class instances.

3.2.1.2 The Semantically Meaningful Cluster Problem

Clustering directly on a subset of the unlabeled examples does tend to group instances together in some sort of semantically meaningful way. However, there is no indication or guidance that this clustering behavior is based upon the unlabeled examples’ similarity with respect to the relationships they express. Samples could be grouped based upon some other irrelevant semantic features in their respective documents (Figure 3.9). We can infer that similar relationship expression is part of the contribution to the clustering behavior as we observe that members of the holdout relation set that we describe later in Chapter 5 tend to cluster together dependent on their class. Randomly selecting a candidate set for clustering is not informative as to whether or not these instances share a common relational “positiveness”. However, our approach is motivated by this innate clustering behavior as something to be leveraged. We examine the possibility to refine this behavior to only apply to the features relevant for relationship extraction. We aim to retain the more granular separation ability of the

Figure 3.9: An example of the meaningful cluster problem applied to images. Images are clustered unintentionally by color rather than the objects in the image. This is similar to the meaningful cluster problem in DocRE when naively clustering unlabeled examples [27].
supervised approach on the unlabeled set, but with a subset of the unlabeled examples that express a similar relational “positiveness”. Our hope is that our approach will cluster only on examples that express some sort of relationship, and that clustering on this informed subset will allow for better relationship separation and purity. This will essentially reduce our problem to the traditional NCD problem where instances in our candidate set are assumed to only correspond to novel classes.

3.2.2 Novel Class Informed Document-Level Relation Extraction

Our detection of novel classes at test time will be dependent on a traditional DocRE method utilizing a training a set of labeled examples (comprising of known relationship categories along with novel relation categories) and unlabeled examples (treated as “no relation” examples) to infer relations of instances at test time. Document-level relation extraction is a difficult task due to the possibility of multiple entity instances and multiple relationships for each unique entity pair. Furthermore, a minor contribution of this thesis is an enhanced approach to an existing pretrained encoder of a popular DocRE approach by replacing the base encoder, RoBERTa, with an encoder more suitable for document-level relation extraction, LUKE-TACRED.

Formally, we define our document-level relation extraction task as:

- Given a document $D$ and entities within that document $e_1, e_2, e_3, ... e_n$ with mentions for each entity $m_{e_1}^1, m_{e_1}^2, ... m_{e_n}^k$ correctly classify each entity pair within the document $(e_1, e_2), (e_1, e_3)...(e_n, e_m)$ with zero or many relationship classes $c \in C_n \cup C_k$. 
3.2.3 Final Combined Objective

Our final combined objective is to utilize the samples in accepted novel class cluster propositions to train a normal DocRE classifier to detect the new and previously known relationships in the test set. We aim to avoid the need for additional manual labeling of novel classes in the training set beyond the samples contained in accepted cluster propositions. Our hope is that novel cluster propositions contain examples that are sufficiently reflective of the novel classes such that they can be effectively utilized as training examples.

3.3 Motivation

DocRE is an interesting and more difficult generalization of traditional relation extraction that is useful for identifying more complicated and meaningful relationships between entities. There has been little, if any, exploration of NCD applied to DocRE, and effective application of NCD in this domain could be immensely useful for discovering interesting and relevant relationships in a given corpus/domain. NCD in DocRE poses a novel challenge resultant of the necessary consideration of a large quantity of true negative examples in the unlabeled set. There has been little research for NCD in explicitly addressing the true negative pollution of the unlabeled set, and research that does consider this problem does not consider true negatives in a quantity realistic for DocRE.

3.4 Proposed Solution

Our approach motivation stems from the idea that naively running inference on a randomly selected candidate set of examples from the unlabeled training set seems to be...
a promising approach in segmenting instances in some sort of semantically meaningful way. However, we aim to address both the purity problem and meaningful cluster problem by informatively instantiating the candidate set based upon the targeted retention of examples in the unlabeled set that are the most likely to be instances of positive relationships.

Our NCD approach can be broken into many steps (Figure 3.10):

1. Propose a candidate set of samples that are the most similar to the known positive examples in the unlabeled set utilizing self-supervised contrastive learning.

2. Train a dual-objective supervised model on both the unlabeled and labeled data (excluding the candidate set) that learns both on the specific multi-label relationships and a binary presence or lack of a relationship.
3. Run inference on the informed candidate set from Step 1, utilizing the output of the binary classifier to filter down the set to only the most likely positive candidates for clustering.

4. Cluster on the latent representations of the remaining instances learned in Step 3.

5. Rank cluster propositions based upon their isolation to other clusters and allow for human rejection/acceptance of novel class clusters.

6. Create a unique pseudolabel for the instances in each cluster and use these cluster samples as training data for the novel class.

7. Evaluate a final model on its ability to detect all classes (both known and novel) and reject true negatives in the test set.

Additionally, we note a minor contribution of an improvement to a popular DocRE architecture known as ATLOP through the replacement of its base encoder, RoBERTa, with a pretrained encoder that specifically targets entity related tasks, LUKE.

We begin our approach description in Chapter 4 with an analysis of the dataset used in all of our experiments, DocRED.
For all of our methods we utilize is one of the first and most widely adopted datasets for DocRE known as DocRED [36]. DocRED was explicitly created to address the lack of comprehensive and quality document-level relation extraction datasets. Previous work in relation extraction primarily focused on the sentence/phrasal level. Therefore the introduction of DocRED sparked research interest into expanding relation extraction methods to the document-level. DocRED contains both hand-annotated and distantly supervised samples from Wikipedia documents. However for the purposes of this thesis we will only be focusing on the hand-annotated examples.

DocRED caters to both intra-sentence and inter-sentence relations allowing for more comprehensive coverage of different realistic relationship scenarios. One of the most differentiating relation types that exist in DocRE, as opposed to traditional RE, is the prevalence of multi-hop relationships. Multi-hop relationships refer to connections between entities that go beyond a single direct link. Instead of just identifying relationships explicitly stated between two entities in a sentence, multi-hop extraction

![Diagram of a multi-hop relationship]

Figure 4.1: Example of a multi-hop relationship. The relational fact “John lives in United States” can only be obtained by reasoning through the intermediary entity “Arizona” and corresponding intermediary relationships.
aims to capture implicit connections by leveraging intervening entities and relationships (Figure 4.1). Additionally in DocRE, it is possible for a given entity to have multiple mentions of itself along with the possibility of there being from zero to multiple relationships between any two particular entities.

4.1 Data Format

The DocRED dataset utilizes a JSON-based format to represent documents, entities, relations, and supporting evidence. A breakdown of the data format is described below.

- **title**: The title of the Wikipedia document as a string.

- **sents**: A list of sentences within the document. Each sentence is itself a list of strings representing the individual words.

- **vertexSet**: A list of entities mentioned in the document. Each entity is represented as a list of dictionaries each corresponding to a single mention of that entity in the document.
  - **name**: The name of the entity mention.
  - **sent-id**: Sentence ID indicating the sentence the entity mention appears in.
  - **pos**: Word positions in the sentence that the entity mention appears in.
  - **type**: The named entity recognition type (NER) (person, location, etc.). We do not utilize this information in this thesis.
• **labels**: A list of dictionaries representing relationship instances. Each dictionary references “head” and “tail” entities where “head” entities refer to the subject of the relation and “tail” entities refer to the object of the relation:

  - **h**: Index of the head entity in the vertex set.
  - **t**: Index of the tail entity in the vertex set.
  - **r**: Relationship ID
  - **evidence**: Evidence sentence IDs. We do not utilize this information in this thesis.

It should be noted that there is no explicit labeling for negative instances (entity pairs with no relationship). These negative/unlabeled instances are generated based upon the pairs of entities in the document that are not mentioned in the document’s “labels” dictionary list.

### 4.2 Size and Label Distribution

DocRE requires the handling of a very imbalanced distribution of negative instances to positive instances. Every unique entity pair in a particular document must be considered, and the large majority of entity pairs are not related in any given document. Additionally, when only examining the distribution of positive examples, it can be seen that the distribution is quite long-tailed where there exists many instances for the most frequent relationships and very few instances for the remainder of the relationships (Table 4.1, Figure 4.2)
Figure 4.2: Label distribution of relationships in the DocRED training dataset. Note that the y-axis represents a log scale of instances and the leftmost relation is the “NA”/“no relation” class.

Table 4.1: General statistics of instances in DocRED. Note the large difference between the number of negatives and number of positives along with the large difference between the maximum relation counts and median relationship counts.

<table>
<thead>
<tr>
<th>General DocRED Statistics</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. Documents</td>
<td>3,053</td>
<td>998</td>
</tr>
<tr>
<td>Num. Relation Types</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Num. Positives</td>
<td>35,615</td>
<td>11,470</td>
</tr>
<tr>
<td>Num. Negatives</td>
<td>1,163,035</td>
<td>384,102</td>
</tr>
<tr>
<td>Max Rel. Ct.</td>
<td>8921</td>
<td>2831</td>
</tr>
<tr>
<td>Min Rel. Ct.</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Median Rel. Ct.</td>
<td>163</td>
<td>56</td>
</tr>
</tbody>
</table>
Figure 4.3: Co-Occurrence matrix of different relationships in the training set of DocRED. We note quite sparse co-occurrence rates among each relationship type with only a few relationship types being strongly correlated with one another.
4.3 Label Correlations and Co-Occurences

We run an initial association rule experiment (Figure 4.4) to obtain a better sense of the co-occurrences of relationships in the dataset. We note that label co-occurrences tend to be very sparse (Figure 4.3) as well with the most common relationship co-occurrences happening resultant of an implicit hierarchy of the relationships as seen in ‘P17’ (country) and ‘P1376’ (capital of). The most frequent label co-occurrences tend to happen with regards to geographical relationships with the most frequent co-occurrence being ‘P17’ (country) and ‘P131’ (located in administrative territorial entity) with 1182 instances. We note that repeated unique label co-occurrences are also quite sparse with the second most frequent label co-occurrence being ‘P17’ (country) and ‘P1001’ (applies to jurisdiction) with only 149 instances.

4.4 Document Preprocessing

Document preprocessing is done to format input in a way digestible for the transformer model that serves as the base of our methods. We first tokenize each document based upon the predefined implementation associated with the pretrained encoder.
We then associate each entity in the document with either an individual injected token (with RoBERTa) or a token span (LUKE) (Figure 4.5). Additionally, we expand the input limitations of the base models of 512 tokens to 1024 tokens by breaking the 1024 token input into two 512 token chunks and then subsequently overlapping the encoder outputs later in the pipeline.

We now continue on to describe our methods for generating entity pair representations in Chapter 5 and how these representations are leveraged to discover novel classes in DocRED.
Chapter 5
SYSTEM OVERVIEW AND DESIGN

5.1 ATLOP

Before we describe the methods utilized to perform novel category discovery in Do-
cRED, we must first define our strategy for creating representations of entity pairs
that emphasize their unique interactions between one another and therefore, their
relationship. We begin by detailing the transformer-based DocRE architecture used
in all of our experiments (Adaptive Thresholding and Localized Context Pooling (AT-
LOP) [39]) and how we have slightly modified this architecture to extract superior
performance.

5.1.1 Pretrained Encoder

ATLOP begins with the pretrained encoder RoBERTa to encode documents of length
up to 1024 tokens (through overlapping outputs). Each entity in a document is marked
with two “*” tokens prepended and appended around the mention spans (Figure 4.5),
with the first “*” token being utilized as the token that captures the entity mention’s
entire representation. The output representation for each token input to the encoder
for each document $D$ comprised of tokens $x_1, x_2, \ldots x_n$ can be expressed as:

$$H = h_1, h_2, \ldots h_n = RoBERTa(x_1, x_2, \ldots x_n)$$
Additionally, the output attention value between tokens $i$ and $j$ can be expressed with $A_{i,j}$. Each entity $e_1, \ldots, e_d$ is comprised of one or more mentions $m_{e_1}^1, m_{e_1}^2, \ldots, m_{e_d}^k$, and therefore multiple tokens in the output representation.

5.1.2 LUKE Pooling Modification

In order to consolidate the output representations for each entity from their entity mention output representation, the authors of ATLOP propose to use logsumexp pooling on the hidden representations of each entity mention \[15\] defined as:

$$h_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp(h_{m_{e_i}^j})$$

Where $N_{e_i}$ is the number of mentions for entity $e_i$ and $h_{m_{e_i}^j}$ is the hidden state generated by RoBERTa for mention $j$ of entity $e_i$. This results in an averaged representation of an entity $h_{e_i}$ in the document based upon its mentions.

However, this approach for entity mention consolidation using RoBERTa does not make any special consideration between ordinary words and entities. Therefore, we propose the usage of LUKE which adopts an entity-aware attention mechanism that results in superior hidden representations for entity-related tasks \[35\]. LUKE introduces four unique query matrices in its attention mechanism that consider different interactions to occur based upon what the tokens are representing, either a common word or entity (Figure 5.1). LUKE’s model output is then comprised of hidden states of the base sequence in addition to special hidden state outputs for the defined entity mention instances. We speculate that performing subsequent localized context pooling based upon these entity specific representations will allow for better performance in downstream inference. The output of LUKE providing an input document
\( D \) comprised of tokens \( x_1, x_2, \ldots x_n \) and token spans for the mentions of all entities \( s_{e1}^{m1}, s_{e2}^{m2}, \ldots s_{ed}^{mk} \):

\[
H = h_1, h_2, \ldots h_n = \text{LUKE}(x_1, x_2, \ldots x_n)
\]

\[
P = p_{e1}^{m1}, p_{e2}^{m2}, \ldots p_{ed}^{mk} = \text{LUKE}(s_{e1}^{m1}, s_{e2}^{m2}, \ldots s_{ed}^{mk})
\]

Where \( H \) are the hidden representations for every token in the sequence and \( P \) are entity specific hidden states for every entity mention in the document.

We then modify the logsumexp pooling mechanism to utilize the special entity mention representations in \( P \) rather than the ordinary sequence representations in \( H \):

\[
h_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp(p_{e_i}^{m_j})
\]

We call this modification entity-enriched logsumexp pooling.

5.1.3 LUKE-TACRED

We further select a version of LUKE named LUKE-TACRED that has been pretrained on the TACRED dataset for traditional relation extraction. We believe the pretrained knowledge from the TACRED RE task can be usefully transferred to the DocRE task.

TACRED [38], as mentioned earlier in this thesis, is a dataset for traditional relation extraction. TACRED is composed of text from a corpus disjoint from the Wikipedia documents in DocRED with 106,264 sentences from the newswire and TAC Knowledge Base Population corpus. There is some overlap in the relationship classes between TACRED and DocRED such as “date of birth” and “date of death” in
addition to the shared inclusion of the “no relationship” class. However, it is assumed that the examples for these overlapping relationship classes in TACRED are distinct from those in DocRED. Therefore, we believe that this pretraining on the TACRED dataset could provide transferable knowledge for document-level relation extraction with respect to distinguishing “no relationship” instances from positive instances in addition to offering more examples of diverse semantic structures for relationship expressions. It should be noted that we only utilize TACRED pretraining for LUKE in fully supervised multi-label classification and do not use it for our subsequent experiments in novel class discovery where we utilize the base LUKE architecture with no pretraining. We consider our improvements to ATLOP’s fully supervised DocRE architecture with respect to TACRED pretraining as a contribution separate from our novel class discovery pipeline.

5.1.4 Localized Context Pooling

Additionally, we are also provided the attention matrix output between each individual token and all other tokens $A$. The ATLOP authors leverage the attention outputs of the encoder to create a context vector representation for any entity pair $e_s$ and $e_o$ (subject and object respectively). This context vector $c$ is meant to capture context that is relevant to both entities in the sequence. The attention values for two entities

$$
e_{ij} = \begin{cases} 
Kx_j^TQx_i, & \text{if both } x_i \text{ and } x_j \text{ are words} \\
Kx_j^TQ_{\text{eq2w}}x_i, & \text{if } x_i \text{ is word and } x_j \text{ is entity}
Kx_j^TQ_{\text{eq2w}}x_i, & \text{if } x_i \text{ is entity and } x_j \text{ is word}
Kx_j^TQ_{\text{eq2w}}x_i, & \text{if both } x_i \text{ and } x_j \text{ are entities}
\end{cases}$$

Figure 5.1: LUKE modification of attention mechanism, introduces unique query matrices for each paired type (entity, word) [35].
Figure 5.2: Simplified flowchart of ATLOP localized context pooling mechanism.

\( e_s \) and \( e_o \) are denoted as \( A_s \) and \( A_o \) and are the result of averaging the attention matrices for each mention of each respective entity. This calculation of the context vector \( c \) is defined by:

\[
A_{(s,o)} = A_s \cdot A_o
\]

\[
q_{(s,o)} = \sum_{i=1}^{H} A_{i(\cdot,s,o)}
\]

\[
a_{(s,o)} = q_{(s,o)}/1^T q_{(s,o)}
\]

\[
c_{(s,o)} = H^T a_{(s,o)}
\]

The context representation \( c \) essentially constructs a representation for the relationship between \( e_s \) and \( e_o \) based upon the tokens important to both entities in the sequence (Figure 5.2).
Further representation of each entity in the relationship is learned based upon a learnable linear layer in the network with a tanh activation function as:

\[ z^s_{(s,o)} = \tanh \left( W_s ( h_{e_s} ; c_{(s,o)}) \right) \]

\[ z^o_{(s,o)} = \tanh \left( W_o ( h_{e_o} ; c_{(s,o)}) \right) \]

Where ; denotes a concatenation of the entity representation calculated from the logsumexp pooling on tokens in \( P \) and the relational context vector \( c \). Finally, the combined representation is learned for both entities with a bilinear layer and a multi-label classification head with an output node/logit for each relationship:

\[ e_{(s,o)} = z^T W_e z_o \]

\[ y_{(s,o)} = W_r e_{(s,o)} \]

\[ P(r | s, o) = \sigma(y^r_{(s,o)}) \]

Where \( e_{(s,o)} \) represents the bilinear output we utilize as the relationship embedding for the entity-pair in all further learning tasks, and \( y_{(s,o)} \) represents the output of the classification head for multi-label relationship classification.

### 5.1.5 Adaptive Threshold Loss

Multi-label classification problems traditionally rely on the tuning of a threshold parameter for individual class output probabilities. However, this assumption does not account for the fact that different entity pairs may necessitate different thresholds to express their relationships. ATLOP considers the threshold on which class labels
Figure 5.3: Adaptive thresholding loss. Threshold logit is learned with respect to the output of the positive and negative classes. Positive logit values are pushed above this learned threshold, and negative class logits are pushed below this threshold [39].

are activated as an additional learnable logit, therefore the threshold can change between any two entity pairs. By learning entity-dependent thresholds, the model can make more nuanced and accurate relation predictions within the context of a document. The core advantage of adaptive thresholding lies in its ability to account for the inherent variability in relation likelihood based on specific entity representations.

The loss function for adaptive thresholding is defined as:

\[ L_1 = -\sum_{r \in P_T} \log \left( \frac{\exp(\logit_r)}{\sum_{r' \in P_T \cup TH} \exp(\logit_{r'})} \right) \]

\[ L_2 = -\log \left( \frac{\exp(\logit_{TH})}{\sum_{r' \in N_T \cup TH} \exp(\logit_{r'})} \right) \]

\[ L_{AT} = L_1 + L_2 \]

Where \( P_T \) are relations that exist between entities in the label vector \( T \), \( N_T \) are the relations that do not exist between entities, and \( TH \) is the threshold logit. It should be noted that the only modification we propose that distinguishes our entity-pair representation approach from the original ATLOP [39] methods is the use of LUKE-TACRED as the base encoder rather than RoBERTa. We utilize the entity specific
hidden states unique to LUKE for entity representations in logsumexp pooling rather than base sequence hidden representations from RoBERTa.

5.2 Holdout Set

The majority of experimental approaches to evaluating the efficacy of a novel class discovery solution rely on the creation of a holdout set from the known labeled relationships [27]. Therefore, we select 6 different relation types (Table 5.1) from our labeled data and use these as our “novel classes” throughout our experiments. It should be noted that this work will not explore the discovery of multi-label relationships. Instead, we are primarily concerned with extracting individual novel classes and leave multi-label cases to future work. These relationships were chosen based upon a variety of different criteria, we chose some relations with high support in the training set and some relations with smaller supports. Additionally, we chose to include the very similar relationships ‘P570’ (date of death) and ‘P569’ (date of birth) to evaluate our method’s ability to distinguish between relationships that appear in similar contexts. We note that with our holdout selections, we are inherently defining that a relationship must appear at least once in every 3,900 entity pairs (based upon “P69”) to be considered significant enough for novel class discovery. We aim to select holdout relationships that are sufficiently variable based upon the types of entities (person, place, thing, date, etc.) and contexts involved in these relationships, however we acknowledge possible biases and the sensitivity of model performance to these holdout selections in “Threats to Validity” in Chapter 7.
Table 5.1: Measuring only the amount of times these relationships appear alone in each set.

<table>
<thead>
<tr>
<th>Rel. ID</th>
<th>Rel. Desc</th>
<th>Train Instances</th>
<th>Dev. Instances</th>
<th>Entity Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>P27</td>
<td>country of citizenship</td>
<td>2562</td>
<td>769</td>
<td>PER -&gt; LOC</td>
</tr>
<tr>
<td>P570</td>
<td>date of death</td>
<td>804</td>
<td>255</td>
<td>PER -&gt; TIME</td>
</tr>
<tr>
<td>P577</td>
<td>publication date</td>
<td>1124</td>
<td>404</td>
<td>OBJ -&gt; TIME</td>
</tr>
<tr>
<td>P569</td>
<td>date of birth</td>
<td>1044</td>
<td>343</td>
<td>PER -&gt; TIME</td>
</tr>
<tr>
<td>P175</td>
<td>cast member</td>
<td>1016</td>
<td>316</td>
<td>PER -&gt; OBJ</td>
</tr>
<tr>
<td>P69</td>
<td>educated at</td>
<td>307</td>
<td>86</td>
<td>PER -&gt; ORG</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>6857</td>
<td>2173</td>
<td>-</td>
</tr>
</tbody>
</table>

5.3 Self-Supervised Contrastive Candidate Filtering

Our approach to NCD in DocRE begins with the guided proposition of a candidate set that is held out from training. This is a set of examples that we will later cluster on in order to determine novel classes within the dataset. As mentioned previously in this thesis, a naive approach to creating this candidate set relies on a random subsampling of the unlabeled data. This approach is quite ineffective as a result of the large amount of true negatives (> 1,000,000) and relatively small amount of samples belonging to novel classes (6857). A random subsampling of the unlabeled set will lead to an absence of sufficient support for these examples in the embedding space, which prevents these novel instances from being clustered together effectively. Furthermore, it is likely when using random subsampling of the unlabeled set that during training of the supervised model, the majority of the instances of the novel relationships are included in the training set and their representations are learned as negatives. In addition, clustering on these unlabeled examples without any sort of learned filtering will lead to errant cluster propositions that are grouped together based upon semantic similarities other than that of a shared relationship expression.
However, we believe that this method of clustering based upon a subset of the unlabeled data offers potential based upon our exploration of the embedding space when just using random sampling. We see that our holdout examples tend to cluster together based upon their relationship type (Figure 6.1). It is simply that the sheer volume of true negative examples pollute the space, preventing the holdouts from forming their own unique clusters.

The primary intuition for our approach is that we can somehow create an informed candidate set that reduces or eliminates the amount of true negative examples in the clustered embedding space while retaining the novel class examples. This essentially transforms our problem from an NCD problem with true negatives to a traditional NCD problem where our unlabeled set is assumed to contain exclusively novel classes.

We speculate that true positive relationship examples share some general features that distinguish them from the true negatives, we therefore propose a self-supervised contrastive method to learn these generalized features unbiased to the presence of labels. We learn this model as a sort of filter independent of training any supervised models. The contrastive model simply serves as a preprocessing step to propose a candidate set to exclude from downstream supervised training.
5.3.1 Self-Supervised Contrastive Learning

Self-supervised contrastive learning is a type representational learning that is quite popular in image-based tasks. It has been primarily applied to image-based tasks as a result of its reliance on input augmentations to learn good representations [6].

Self-supervised contrastive learning essentially operates by utilizing augmentations of individual samples as positive-pair instances and all other examples in the batch as negative-pair instances. Contrastive learning attempts to push the representations of negative-pair examples away from one another while also bringing representations of positive-pair instances closer together. This method of learning has gained some traction in the natural language processing domain. However, its applications have been limited due to the difficulty of preserving semantic meaning while augmenting examples.

Figure 5.5: Intuition behind contrastive learning applied to DocRE.
Recent research for contrastive learning applied to NLP [11] has shown that the random drop out masks of pretrained transformers can serve as a type of data augmentation for input sequences. We therefore leverage such work in this thesis in order to apply self-supervised contrastive learning for NCD in DocRE. We believe learning representations of all examples (labeled and unlabeled) in this manner will lead to a separation between the representations known positive relationship examples and true negative examples. We therefore can select a candidate set from the unlabeled set based upon the similarity of representations with respect to the known positive examples (Figure 5.5).

The architecture of our self-supervised contrastive approach can be seen in Figure 5.4. Each document is passed through the base encoder twice with different drop out masks of probability $p = 0.1$. The entity pair representations created by the localized context pooling and bilinear output mechanism are passed to a projection head that compresses the embedding representation and then the self-supervised contrastive loss can be applied to this projection:

$$
L_{contr}^i = -\log \frac{\exp(sim(h_i, h'_i)/\tau)}{\sum_{j=1}^{N} \exp(sim(h_i, h'_j)/\tau)}
$$

In the above loss function, $\tau$ is a temperature hyperparameter and $h_i$ and $h'_i$ refer to two dropout mask augmentations of a particular entity pair representation $i$. The contrastive loss function aims to pull together representations of similar instances, while pushing apart dissimilar ones. The projection head acts as a bottleneck, compressing the high-dimensional encoder outputs into a lower-dimensional space. This compression process encourages the model to focus on the most informative aspects of the data, creating more discriminative and transferable representations. These improved representations ultimately benefit downstream tasks.
Figure 5.6: Visualization of contrastive filtering metric. Unlabeled instances with a higher cosine similarity to the positive mean than the average unlabeled cosine similarity to the positive mean are selected as candidate instances.

Another advantage of the projection head is the ability to reduce sensitivity to data augmentations. Contrastive learning relies on the previously mentioned data augmentation to create diverse representations from the same underlying data, however these augmentations can introduce irrelevant variations. The projection head can act as a filter, mitigating the impact of such variations and promoting learning of features robust to these transformations [16]. This leads to representations that are less sensitive to the specific augmentations used and generalize better to unseen data.

Similar to most implementations of contrastive learning [16], the projection head is discarded after training and we utilize the bilinear embedding outputs from our modified ATLOP implementation for our subsequent similarity calculations and filtering.
5.3.2 Filtering

After learning these representations, we must determine a metric for defining whether or not an unlabeled example should be included in the candidate set for clustering or if the unlabeled example should be included as a negative sample in the subsequent supervised training. In experimentation, we determined that unlabeled examples that had a larger than average (amongst all unlabeled examples) cosine similarity to the labeled (positive) mean were good candidates for retention (Figure 5.6). We therefore take these examples and hold them out from later training. We define this metric for the candidate set $C$ as:

$$sim(h_c, \bar{p}) > \frac{\sum_{i=1}^{U} sim(h_i, \bar{p})}{|U|}$$

$$\forall c \in C$$

Where $h_c$ is any given embedding for an entity pair/relationship instance the new candidate set $C$, $\bar{p}$ is the average embedding of all positive/labeled examples, and $U$ is the set of all unlabeled examples.

We note that utilizing this metric initially with only the embedding outputs of the pretrained transformer (without contrastive training) results in an initial holdout candidate ratio of 0.445 and negative candidate ratio of 0.503, which is far worse in terms of holdout candidate preservation when compared to our models trained with the contrastive loss (Figure 5.8) and even worse than random sampling (Figure 6.2). This indicates that contrastive training does learn representations that are useful for our purposes.
5.3.3 Temperature

We noted the presence of the hyperparameter $\tau$ for the temperature term in the contrastive loss. Temperature acts as a scaling factor for the logits within the contrastive loss function which impacts the model’s sensitivity to the distances between embedding vectors in the representation space. When temperature is increased, the loss function becomes less sensitive to the magnitude of the distance between embeddings. This leads to a softer contrastive loss, encouraging a more uniform distribution of embeddings in the representation space. While this promotes overall similarity preservation, it might limit the model’s ability to distinguish subtle differences between similar data points. Conversely, a low temperature makes the loss function highly sensitive to embedding distances, which leads to a harder contrastive loss, pushing the model to create a more separated distribution of embeddings [33] (Figure 5.7). We examined three different values for the temperature in our experimentation [0.01, 0.05, 0.1]. We follow the precedent set by [11] and began our contrastive loss experimentation with lower temperature values and evaluate the best temperature.
value based upon the number of our holdouts that are retained in the candidate set versus the number of true negatives that are retained in our candidate set. We aim to minimize the amount of true negatives that are included in the set while maximizing the amount of holdouts. We apply some priority on maximizing the amount of holdouts retained in order to avoid later negative overfit on the missed novel sample representations.

We observe in Figure 5.8 that the temperature 0.01 performed the best for our purposes. We suspect that based upon the nature of our filtering metric that softer distributions allow for more true negatives to be oriented close to the labeled positive mean increasing the average similarity of unlabeled instances to the positive mean. This restricts the amount of holdout examples that can be retained based upon the average distance metric. We therefore utilize contrastive filtering with a temperature of 0.01 for all future experiments.

5.3.4 Early Stopping

As the self-supervised contrastive learning objective is focused on grouping sample representations quite granularly, samples will be grouped into finer and finer clusters as training progresses. As our goal is to determine more general “positive” or “negative” relationship representations of the examples rather than well-defined clusters, we propose the early termination of the self-supervised contrastive model in order to ensure the model captures only general features. We discovered by monitoring the quantity of holdouts in our candidate set after two epochs the holdout inclusion in the candidate set began to decrease and true negative inclusion in the candidate set began to increase (Figure 5.9).
Figure 5.8: The ratio of the holdouts and true negatives that are retained in the candidate set depending on the temperature used for self-supervised contrastive learning.
Figure 5.9: Ratio of holdouts with respect to quantity of samples in the candidate set over contrastive training epochs.
Figure 5.10: Ratio of holdouts with respect to all holdouts in dataset in the candidate set and cosine similarity between the average positive and average unlabeled example over contrastive training epochs.
We speculate that this is the result of positive examples becoming more finely grouped together with unlabeled examples in various locations around the embedding space, leading the average calculation of the positive examples to lose meaning as a general representation of the positives in our set. We consider this behavior as a kind of model overfit for our purposes. To determine the best model checkpoint at training time without prior knowledge of the holdout set, we propose to use the model that minimizes the cosine similarity between the average positive example and average negative example (Figure 5.10). We observe an inverse relationship with regards to the separation between the positive and negative means and the amount of holdouts that are retained in our candidate set. We speculate that this occurs because the separation between these averages decreases as the samples are grouped together on more fine-grained features.

Future work is required to determine a superior metric that could consider this overfit via positive cluster segmentation more robustly.

5.4 Dual-Objective Supervised Learning

After our self-supervised contrastive filtering step, we are left with two sets of examples, the training set, which is comprised of the labeled examples in addition to unlabeled examples not selected for the candidate set, and the candidate set, which is comprised of the most “positive” samples of the unlabeled set. We therefore begin training utilizing the base ATLOP methodology with slight objective modifications to encourage learning of generalized positive/negative representations. It is important to note that the unlabeled examples that are left in the training set (not included in the candidate set) are treated as negative examples per the original DocRED training regimen.
5.4.1 Adaptive Threshold Multi-Label Learning

We retain the adaptive threshold classification head from the ATLOP architecture in order to retain the previously observed behavior of ATLOP’s ability to segment unlabeled data effectively (Figure 3.6).

5.4.2 Binary Positive Logit Learning

We include the addition of a binary logit utilizing a separate loss function from ATLOP’s Adaptive Threshold (AT) loss to allow the model to learn what an instance of a general positive relationship is and what an instance of a “no relationship” is. We opt to not include this as a logit thresholded in the AT loss function as the AT loss function learns based upon the presence of a particular relationship and the type of the relationships associated with any given entity pair representation [39]. We simply want to guide the model to learn to distinguish between the presence and lack of a relationship for a given entity pair representation.
Figure 5.12: Binary positive logit filtering mechanism. Candidates from contrastive filtering are predicted after training with labeled data and unlabeled non-candidates. Candidates that are not predicted as “positive” relationships are merged back into the true negative set for final training. Candidates that are predicted as “positive” relationships are clustered and proposed to human labeler.

5.4.3 Dual-Objective Loss Function

We define the loss function for the dual-objective supervised model as the mean of the adaptive thresholding loss and binary cross entropy loss on the binary positive logit. As this model is not directly concerned with classifying known relationships correctly, we allow the binary positive logit to contribute equally to the loss function:

$$\mathcal{L}_{tot} = 0.5 \cdot \mathcal{L}_{AT} + 0.5 \cdot \mathcal{L}_{BCELogit}$$

We borrow this method of weighted loss function optimization from various literature [30, 17]. The architecture diagram for training this intermediary model is seen in Figure 5.11.
5.4.4 Binary Positive Logit Thresholding

After model training on the dual-objective function, we then run inference upon the samples in our candidate set. Samples whose binary positive logit outputs (after sigmoid activation) are below or equal to 0.5 are removed from consideration for later clustering and will be merged back into the unlabeled set as negative examples later on (Figure 5.12). We believe this further effectively reduces the amount of true negative instances that will pollute the later embedding space. This threshold for “positiveness” is chosen based upon the notion that sigmoid outputs above 0.5 indicate above a 50% probability for a given sample to express a positive relationship.

5.5 Clustering and Novel Class Proposition

Finally, the remaining candidate examples after positive logit thresholding are used for clustering and cluster proposition. As a result of the high dimensional embeddings needed to properly represent complicated document-level relationships, in addition to the high quantity of embeddings remaining to be clustered. We propose a technique for clustering that reduces the dimensionality of the embeddings in our candidate set while mitigating the information loss related to dimensionality reduction.

5.5.1 UMAP Dimensionality Reduction

We propose the use of a relatively recent alternative to classic dimensionality reduction techniques like PCA and t-SNE in the form of UMAP.
5.5.1.1 The Curse of Dimensionality

The curse of dimensionality refers to various problems that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings. As the number of dimensions increases, the volume of the space increases exponentially, causing several challenges for data analysis [31]. The distance between any two points tends to become uniform, making it difficult to distinguish between close and distant points, which causes traditional distance-based algorithms to be less effective. This uniformity of distances undermines the assumption of many clustering algorithms that closer points are more similar, thereby reducing their effectiveness.

High-dimensional data tends to be sparse, because the large number of combinations of dimensions implies locations in the space are rarely populated, which complicates the process of identifying meaningful patterns. This sparsity makes it challenging for clustering algorithms to find dense regions of data. The computational resources required to process high-dimensional data also increase dramatically, making algorithms that perform efficiently in low dimensions computationally infeasible as dimensionality rises. The increased computational load can lead to longer processing times and higher memory usage, further complicating clustering tasks.

Furthermore, in high-dimensional spaces, models can become overly complex, capturing noise in the data rather than the underlying pattern, leading to overfitting and poor generalization to new data.

Addressing these challenges robustly is still an open problem. However, we decide to approach it simply through dimensionality reduction techniques, which aim to reduce the number of dimensions while preserving the essential characteristics of the data, thereby making clustering more feasible and effective. We note prior successful work
on clustering high dimensional word embeddings with UMAP [25, 23, 1] as precedent for this technique.

5.5.1.2 UMAP - Preserving Global and Local Distances

UMAP [22] is a dimensionality reduction technique designed to maintain the topological structure of data when reducing its dimensions. It works by constructing a high-dimensional graph representation of the data, which captures both local and global relationships. This graph is then optimized to minimize the differences between the high-dimensional data and its low-dimensional projection. The primary steps involve creating a fuzzy topological representation of the data, optimizing the low-dimensional layout to match this representation, and using a force-directed graph layout to produce the final embedding. UMAP is highly efficient and scalable, making it suitable for large datasets.

UMAP differs from other popular dimensionality reduction techniques like t-SNE [28] and PCA (Principal Component Analysis) [21] in several ways. PCA is a linear technique that projects data onto the principal components that maximize variance, thus preserving global structure but often failing to capture the complex local structures of non-linear data. Whereas t-SNE focuses on preserving local similarities by converting the distances between data points into joint probabilities and minimizing the Kullback-Leibler divergence between these probabilities in high and low dimensions. While t-SNE is effective at preserving local structures, it can struggle with maintaining global structure and is computationally intensive for large datasets. UMAP bridges these gaps by preserving both local and global distances more effectively and expresses the structure of non-linear data.
We believe UMAP’s ability to maintain the integrity of both local and global distances makes it a robust choice for dimensionality reduction before clustering. By preserving the relationships between data points in a low-dimensional space, UMAP ensures that clusters are more distinct and representative of the true structure of the data.

5.5.2 HDBSCAN Clustering

HDBSCAN [4] is a clustering algorithm that extends DBSCAN [10] by converting it into a hierarchical clustering algorithm over a range of parameters and then extracting the most stable clusters. HDBSCAN works by first constructing a minimum spanning tree of the data points based on their mutual reachability distances, then creating a hierarchy of clusters by progressively removing edges with the largest mutual reachability distance. This hierarchy is used to extract a flat clustering by selecting clusters based on their stability, defined as the persistence of clusters across different scales. Unlike k-means [2], which requires the number of clusters to be specified beforehand and is sensitive to the initial placement of centroids, HDBSCAN does not require specifying the number of clusters and can find clusters of varying shapes and densities. This feature of density based clustering algorithms is useful for our NCD task where we do not assume prior knowledge of the quantity of novel classes. Compared to DBSCAN, HDBSCAN can handle varying density clusters better and is less sensitive to the choice of parameters, providing a more flexible approach to clustering complex data.

HDBScan only requires the definition of one parameter \textit{min-cluster-size} which defines the number of points in a dense region to be considered a cluster. For all of our experiments, we set \textit{min-cluster-size} as equal to 100 as we will assume that any significant relationship will have greater than or equal to 100 instances in our entire
corpus. HDBScan is further suggested by the UMAP authors as an ideal clustering algorithm to be utilized with dimensionality reduced data [23].

5.5.2.1 Ranked Cluster Proposition

We further define a method of ranking our cluster propositions, intuition infers that true negative example clusters that are arbitrarily created even after the filtering steps we have implemented will tend to form multiple small clusters tightly packed among one another due to these clusters being created from noise. We also infer that true novel class instances will likely cluster separately from the general noisy true negative distribution. Therefore, we define a ranked cluster proposition method based upon the most isolated clusters, which is defined as the clusters with their centroids the highest average distance away from other clusters proposed in the embedding space.

5.6 Human Labeling and Final Training

After the proposition of novel clusters we assume human intervention in evaluating the quality of the clusters proposed (Figure 5.13). The assumed technique for cluster propositions would be for random instances from each cluster being proposed to the human labeler to either reject or accept the cluster as a novel class. This particular section of the pipeline is where we find the utility in increasing the cluster purities through our filtering methods. For each accepted cluster the elements from the cluster are then given pseudolabels as novel classes and used as positive training examples for the particular novel class. To mimic the behavior of this human labeling, for every holdout class, we select the cluster with the highest amount of elements of that holdout class as the cluster that represents the novel relationship.
Figure 5.13: Human-in-the-loop labeling method. Clusters are inspected by random sample propositions from clusters and accepted (pseudolabeled and used for training) or rejected (used as true negatives). In this example, cluster 1 is accepted and all instances in its cluster are used for training the final model as a new class, whereas cluster 2 is rejected and its instances are relegated to true negatives for final model training.
It should be noted that the addition of the positive logit, weighted dual-objective loss function, and exclusion of the candidate set during training reduces the model’s ability to classify on the known labeled data. This is due to the fact we are removing hard negative examples from the training pool in the contrastive filtering step such that the model is not robust to negative examples that are difficult to distinguish from positive ones. We therefore complete our proposed system with a complete retrain of the ATLOP system with the injected novel pseudolabels and relegate all samples not members of novel clusters from the candidate set as negative examples. We then evaluate the model’s capacity to classify on the known labeled examples in addition to its ability to classify on the novel classes in the test set.

We now continue to Chapter 6 where we evaluate our filtering, clustering, and final training methodologies.
6.1 Base Encoder Selection

We evaluate our replacement of RoBERTa with LUKE-TACRED as the base encoder for ATLOP based upon the official evaluation metrics of the DocRED dataset [36]. The DocRED paper and subsequent research on this task only entails the evaluation of the micro F1 score for each relationship excluding the “no relationship” class. Additionally, the authors introduce two different F1 scores resultant of some discovered intersection between the relational facts in the training and test set. Therefore, there exists the normal micro F1 and the “Ignore” (Ign.) micro F1 where the latter excludes the intersecting relational facts in the test set.

We observe, after the 30 epochs of training prescribed by the ATLOP paper, a slight improvement to both the normal F1 and Ign. F1 scores (+0.39 and +0.46 above 1 standard deviation respectively) over the base RoBERTa model when using the LUKE architecture pretrained on the TACRED dataset. We believe this demonstrates that both entity-aware attention and pretraining on normal phrasal-level relation extraction is beneficial for the document-level relation extraction task (Table 6.1).

<table>
<thead>
<tr>
<th>Encoder</th>
<th>DocRED F1</th>
<th>DocRED Ign. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa-Large</td>
<td>63.18 ± 0.19</td>
<td>61.32 ± 0.14</td>
</tr>
<tr>
<td>LUKE-TACRED</td>
<td><strong>63.76</strong></td>
<td><strong>61.92</strong></td>
</tr>
</tbody>
</table>
6.2 Filtering Evaluations

As our next experiments are not primarily focused on the base multi-label classification of only known relationships, we reduce, for the sake of experimental efficiency, our model to the LUKE-BASE architecture (as opposed to LUKE-TACRED) trained from scratch on all DocRED data for 10 epochs per experiment. The pretrained LUKE-TACRED model is only available in conjunction with the LUKE-LARGE architecture which has a significantly higher number of trainable parameters than LUKE-BASE. Additionally, we do not want to introduce possible bias for or against certain relationships in the NCD pipeline resultant of TACRED’s class overlap with certain DocRED relationships. Therefore, our NCD methods only leverage LUKE’s entity-aware attention mechanism and not TACRED pretraining.

We evaluate both of our filtering methods, contrastive and binary logit filtering here.

6.2.1 Metrics

6.2.1.1 Holdout Retention and True Negative Retention

We evaluate the efficacy of our self-supervised contrastive filtering and binary logit filtering methods based upon the amount of holdouts retained in the candidate set compared to the amount of true negatives retained in the candidate set. We aim to maximize the number of holdout relations in the candidate set while minimizing the number of true negatives.

We additionally inspect the inclusions of each individual holdout relation to evaluate any bias towards or against certain relations. General inspection of the impact of varied temperature values in contrastive filtering can be seen in Table 6.2. However,
Table 6.2: We observe temperature of 0.01 performs the best in the holdout percentage of our candidate set and retains the highest amount of holdouts in the candidate set aside from no contrastive filtering.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Filtering (Random 600,000)</td>
<td>51.47%</td>
<td>51.6%</td>
<td>0.59%</td>
<td>600,000</td>
</tr>
<tr>
<td>0.1</td>
<td>74.7%</td>
<td>49.7%</td>
<td>0.88%</td>
<td>594,422</td>
</tr>
<tr>
<td>0.05</td>
<td>78.9%</td>
<td>49.6%</td>
<td>0.92%</td>
<td>582,635</td>
</tr>
<tr>
<td><strong>0.01</strong></td>
<td><strong>89.2%</strong></td>
<td><strong>50.6%</strong></td>
<td><strong>1.02%</strong></td>
<td><strong>583,444</strong></td>
</tr>
</tbody>
</table>

Table 6.3: Number of each holdout relation preserved in the candidate set after contrastive filtering.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P69</td>
<td>307</td>
<td>281</td>
<td>91.53%</td>
<td>0.047%</td>
</tr>
<tr>
<td>P570</td>
<td>804</td>
<td>728</td>
<td>90.55%</td>
<td>0.12%</td>
</tr>
<tr>
<td>P569</td>
<td>1044</td>
<td>948</td>
<td>90.81%</td>
<td>0.16%</td>
</tr>
<tr>
<td>P175</td>
<td>1016</td>
<td>823</td>
<td>81.00%</td>
<td>0.14%</td>
</tr>
<tr>
<td>P577</td>
<td>1124</td>
<td>867</td>
<td>77.14%</td>
<td>0.15%</td>
</tr>
<tr>
<td>P27</td>
<td>2562</td>
<td>2469</td>
<td>96.37%</td>
<td>0.42%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6857</strong></td>
<td><strong>6116</strong></td>
<td><strong>89.2%</strong></td>
<td><strong>1.02%</strong></td>
</tr>
</tbody>
</table>

for subsequent inspections (Tables 6.3, 6.4, 6.5), we evaluate with a temperature of 0.01 according to our previous analysis in Chapter 5.

### 6.2.2 Results

We see in Table 6.2 that the temperature value of 0.01 preserves most of the holdout examples and results in the highest ratio of holdout examples to true negative examples in the candidate set.
Table 6.4: Candidate set make up after binary logit filtering.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No Filtering (Random 150,000)</td>
<td>12.06%</td>
<td>12.89%</td>
<td>0.55%</td>
<td>150,000</td>
</tr>
<tr>
<td>Binary Logit Filtering</td>
<td>59.58%</td>
<td>11.1%</td>
<td>3.06%</td>
<td>133,229</td>
</tr>
</tbody>
</table>

Table 6.5: Number of each holdout relation preserved in the candidate set to be clustered after binary logit filtering. We observe that even with the bias against specific relationships such as ‘P569’ and ‘P570’ the retention of these relationships is still superior to random sampling.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P69</td>
<td>307</td>
<td>215</td>
<td>70.03%</td>
<td>0.1%</td>
</tr>
<tr>
<td>P570</td>
<td>804</td>
<td>270</td>
<td>33.58%</td>
<td>0.2%</td>
</tr>
<tr>
<td>P569</td>
<td>1044</td>
<td>229</td>
<td>21.93%</td>
<td>0.17%</td>
</tr>
<tr>
<td>P175</td>
<td>1016</td>
<td>588</td>
<td>57.87%</td>
<td>0.44%</td>
</tr>
<tr>
<td>P577</td>
<td>1124</td>
<td>564</td>
<td>50.18%</td>
<td>0.42%</td>
</tr>
<tr>
<td>P27</td>
<td>2562</td>
<td>2220</td>
<td>86.65%</td>
<td>1.67%</td>
</tr>
<tr>
<td>Total</td>
<td>6857</td>
<td>4086</td>
<td>59.58%</td>
<td>3.06%</td>
</tr>
</tbody>
</table>

After contrastive filtering we observe a near doubling of the ratio of holdout examples to true negative examples in the candidate set as opposed to random subsampling (Table 6.2). Additionally, the quantity of holdouts in general is well-preserved in our candidate set. The percent reduction of true negatives in the candidate set is significantly higher than the reduction in the number of holdouts. We can see after contrastive filtering we do not see any major bias against filtering out specific relations except for possibly ‘P175’ and ‘P577’ indicating that contrastive filtering is generally sufficient at selecting for relational ‘positiveness’.

Furthermore, after binary logit filtering we observe another increase in the ratio of holdouts compared to true negatives in the candidate set from 1.02% to 3.06% (Table
6.4). We observe a much more significant reduction in true negative examples in our candidate set than from contrastive filtering, which is expected as the threshold for being inferred as a positive relationship is much more selective. We do see significant bias for and against different samples that is irrespective of the total number of examples for that holdout relation. For example (Table 6.5), we see that most samples of ‘P569’ are not considered examples of a positive relationship after training on the unlabeled set, whereas most examples of ‘P69’ are considered positive examples. We can conclude that both contrastive filtering and binary logit filtering are effective in biasing towards preservation of true positive relation examples, however significant bias against certain relations is introduced in the binary logit filtering step likely resultant of the make up of the unlabeled/negative set used for supervised training. Examining the most two most pruned relationships ‘P569’ and ‘P570’ (‘date of birth’ and ‘date of death’ respectively) we can hypothesize these relationships may not have sufficiently unique contexts that distinguish them from true negative contexts. It is possible that because DocRED is a dataset of Wikipedia articles, the “date of birth” and “date of death” relationship likely appear in every document related to a person. Even with this bias, however, binary logit filtering preserves relations at a higher rate than random subsampling.

6.3 Clustering Evaluation

We evaluate the results of our cluster proposals utilizing four metrics, we aim to balance between the cohesiveness of clusters for each holdout class in addition to the separation of holdout clusters compared to other clusters formed in the candidate set.
6.3.1 Metrics

6.3.1.1 Silhouette Coefficient

We define the first metric of our clustering evaluation as the silhouette score of all clusters proposed, both those created by the true negatives and holdout examples. This measures the separability of our proposed clusters and the strength of cohesion within the clusters. We assume it is ideal for the clusters we propose to have high separability and high cohesion.

6.3.1.2 Average Weighted Holdout Cluster Purity

For each relationship in our holdout set we calculate a weighted purity of the clusters the relation type appears in. This is especially relevant for the human labeling portion of our system as we want the highest likelihood of proposing examples from a cluster that are actually examples of a novel relationship. Therefore, we determine for a given holdout relation cluster proposition, what percent of that cluster is comprised of our novel holdout examples. For each holdout relation, the purity for each cluster the holdout relation appears in is weighted in the calculation by the amount of examples of that relation in the cluster.

6.3.1.3 Average Percent Holdout in Max Cluster

We calculate this metric to determine the amount of 'spread' the holdout relations have across clusters. We measure this as the percent of a particular holdout relation within the cluster it appears most frequently in.
6.3.1.4 Holdout Noise Percentage

Finally holdout noise percentage is simply the percentage of holdouts from the clustering solution that are classified as noise and not considered in clusters.
Figure 6.2: UMAP clustering results on our informed subset of unlabeled data. Red outlines indicate a top 30 cluster proposal.

Table 6.6: Clustering statistics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Candidate Set</td>
<td>0.4</td>
<td>0.086</td>
<td>0.86</td>
<td>0.023</td>
<td>80</td>
<td>N/A</td>
</tr>
<tr>
<td>Informed Candidate Set</td>
<td>0.52</td>
<td>0.342</td>
<td>0.80</td>
<td>0.02</td>
<td>102</td>
<td>4/6</td>
</tr>
</tbody>
</table>
6.3.2 Results

We observe that clustering upon our informed candidate set is superior in every metric excluding the average ratio in a single cluster. However, we believe this metric is skewed for the naive solution due to the fact that the naive solution results in very large impure clusters. Our informed candidate set results in a higher amount of dense, pure and separated clusters, however in some cases this results in the examples for a holdout relation being split into more than one small, but adjacent clusters. This works to our benefit in some cases for separating relationships that are very contextually similar such as ‘P569’ and ‘P570’. In the naive clustering solution we see that these relations are clustered together (Figure 6.1) whereas in our solution these relationships are separated into two clusters (Figure 6.2). Upon inspection of the clustered embedding space, we can see that clusters containing holdout relations tend to exist on the periphery of the distribution. Therefore, our approach to ranking novel clusters based upon their separation to other clusters is applicable. Furthermore, we note a significant amount of small clusters towards the center of the distribution surrounded by detected noise. We speculate this noisy center of the distribution accounts for many of the missed true negative examples with the exception of ‘P69’.

6.4 Final Model Evaluation

To evaluate the performance of the final model trained on the pseudolabeled novel class samples, we evaluate the performance of holdout and known classes separately with their macro F1 score, and then together with the overall micro and macro F1 score. We compare three versions of the model, the first being a model trained as normal ATLOP with no novel class discovery and with full access to all training data (Supervised), the second being a NCD model where all samples in a selected holdout
cluster are used for training data (NCD Whole Cluster [WC]), and the third being a model where only the actual holdout examples from the cluster proposition are used in the training data (NCD Only True Holdouts [OTH]). The final scenario is somewhat idealistic and would require the human labeler to inspect all instances of the cluster and determine if they belong or do not belong to the relationship. It should be noted that none of the models perform very well at the overall macro-F1 level due to the long-tailed distribution of relationships in DocRED in both the train and test sets, however we have selected holdout relationships where ATLOP performs sufficiently to perform a fair comparison of our NCD techniques.

6.4.1 Metrics

6.4.1.1 Novel Holdout F1s

This is F1 score calculated for each holdout relation type calculated individually and then averaged.

6.4.1.2 Known Class Macro F1

This is the F1 score calculated for each known relation type calculated individually and then averaged.

6.4.1.3 Overall F1

This is simply the calculation of the micro and macro F1 metrics on all data.
Table 6.7: Overall F1 Scores
General Final Model Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Macro F1 (Holdouts)</th>
<th>Macro F1 (Others)</th>
<th>Macro F1 (All)</th>
<th>Micro F1 (All)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>0.794</td>
<td>0.402</td>
<td>0.427</td>
<td>0.622</td>
</tr>
<tr>
<td>NCD (WC)</td>
<td>0.292</td>
<td>0.391</td>
<td>0.385</td>
<td>0.501</td>
</tr>
<tr>
<td>NCD (OTH)</td>
<td>0.472</td>
<td>0.391</td>
<td>0.396</td>
<td>0.572</td>
</tr>
</tbody>
</table>

Table 6.8: Holdout Specific F1s
Holdout Specific Model Results

<table>
<thead>
<tr>
<th>Rel. ID</th>
<th>F1</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>Prec.</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>P69</td>
<td>0.74</td>
<td>0.66</td>
<td>0.84</td>
<td>0.33</td>
<td>0.22</td>
<td>0.70</td>
<td>0.67</td>
<td>0.74</td>
<td>0.62</td>
</tr>
<tr>
<td>P570</td>
<td>0.89</td>
<td>0.86</td>
<td>0.93</td>
<td>0.14</td>
<td>0.09</td>
<td>0.33</td>
<td>0.48</td>
<td>0.88</td>
<td>0.33</td>
</tr>
<tr>
<td>P569</td>
<td>0.93</td>
<td>0.90</td>
<td>0.95</td>
<td>0.31</td>
<td>0.89</td>
<td>0.19</td>
<td>0.34</td>
<td>0.96</td>
<td>0.21</td>
</tr>
<tr>
<td>P175</td>
<td>0.77</td>
<td>0.76</td>
<td>0.77</td>
<td>0.43</td>
<td>0.61</td>
<td>0.34</td>
<td>0.45</td>
<td>0.83</td>
<td>0.31</td>
</tr>
<tr>
<td>P577</td>
<td>0.78</td>
<td>0.75</td>
<td>0.82</td>
<td>0.22</td>
<td>0.32</td>
<td>0.16</td>
<td>0.25</td>
<td>0.89</td>
<td>0.14</td>
</tr>
<tr>
<td>P27</td>
<td>0.65</td>
<td>0.67</td>
<td>0.64</td>
<td>0.31</td>
<td>0.19</td>
<td>0.82</td>
<td>0.64</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.79</td>
<td>0.77</td>
<td>0.83</td>
<td>0.29</td>
<td>0.39</td>
<td>0.42</td>
<td>0.47</td>
<td>0.83</td>
<td>0.36</td>
</tr>
</tbody>
</table>

6.4.2 Results

We observe a significant drop in the macro F1 score (Table 6.7) on the holdout relations compared to a purely supervised setting. However, this is expected as the purely supervised model treats the holdouts as known labeled classes and has access to the largest quantity of pure training examples. We observe slight drops on performance for both NCD approaches among the known classes as well. When comparing the two NCD approaches, training with only known holdouts from the proposed cluster results in higher performance when compared to training on all samples from the proposed cluster. This is a result of not labeling the true negatives within the proposed clusters as examples of the novel class, leading to more precise predictions.

We find a very strong relationship between the proposed cluster purity and precision on the test set for the holdout examples (Figure 6.3). In addition, we observe an
equally strong relationship between the percent retention of the holdout relations in the candidate training data with the relation’s ability to be recalled in the test set. Therefore, our approaches to creating an informed candidate set that retains a larger amount of holdout examples and results in higher cluster purities assists in better performance comparative to the naive candidate solution. However, we believe that more work may be needed to automatically select examples from accepted clusters that are the most useful to training in order to avoid the false labeling of true negatives in the cluster. Interestingly, we observe a universal increase in precision when moving from the fully supervised approach to the NCD (Only True Holdouts) (Table 6.8) approach which suggests the examples we retain for the holdout novel classes are some of the most distinguishing instances of the relationships of interest.
Figure 6.3: Relationship between cluster purity and precision with our NCD (Whole Cluster) model.
Figure 6.4: Relationship between representation of holdout relationship type in its cluster vs. and recall with our NCD (Whole Cluster) model.
Chapter 7

CONCLUSION AND FUTURE RESEARCH

7.1 Summary

We have offered an interesting first approach into exploring novel class discovery techniques in document-level relation extraction by explicitly addressing the problem of true negatives. Our main contribution exists as a two-stage filtering mechanism leveraging contrastive learning and dual-objective supervised learning to extract the most likely novel class candidate samples for clustering. We believe our system acts as a superior alternative to random subsampling of unlabeled data to detect novel classes and sufficiently considers the issue of true negative pollution of the unlabeled candidate set. Additionally, we have offered a re-implementation of a popular DocRE architecture leveraging the alternative base encoder, LUKE, to improve supervised multi-label relation extraction.

7.2 Limitations

There are a few limitations in this thesis that we believe can be resolved in future work. Firstly, our system relies on a human-in-the-loop assumption where novel cluster proposals must be confirmed or rejected by a human labeler. Secondly, utilizing all examples from an accepted novel cluster proposition is sub-optimal for training a supervised model as many of the examples in the clusters are still true negative examples. This negatively impacts model precision as the representations the model has learned for novel classes are too lenient for effective inference on the test set. We
believe this system, in its current state, is moreso useful as a novel class proposition technique rather than an end-to-end training methodology as a result of the reduced, but still prevalent, true negative examples in each cluster. It should be noted that even fully supervised DocRE with exclusively known classes struggle to deal with the presence of true negative examples in the training pool with current best systems achieving micro F1 scores below 0.70 as a result of low prediction precision. We believe there is still much room for improvement regarding methods used for proper document-level relation representation that can be leveraged within this system in the future. Finally, our clustering approach relies on the usage of dimensionality reduction which is a controversial preprocessing method for clustering, we believe it could be useful to leverage alternative clustering techniques or learned similarity functions such as those used in [34] to deal with the high dimensional embedding vectors necessary to represent complicated document-level relationships.

7.3 Future Work

We believe there is a variety of options for future work that could be explored in the context of our system. Firstly, we did not consider the discovery of novel classes in the context of already known relationships, we believe it could be interesting to explore clustering novel classes and known classes together to reveal interactions or hierarchical relationships between classes. Furthermore, we believe there could be methods for weighted clustering techniques leveraging the binary logit output to enforce tighter clusterings of more confidently positive examples. Additionally, we wish to address the problem of novel class cluster segmentation where novel classes end up in multiple small clusters and cite further research [20] for avoiding the micro-clustering of the noisy high-density regions. To address the problem of novel class representations being learned as negative when they are not captured in the candidate
set, we believe that soft labeling [32] unlabeled examples based upon some likelihood metric of novelty offers an interesting approach for targeting the discovery of novel classes. Finally, we did not handle multi-label novel class discovery in this work, but are curious to explore avenues for multi-label discovery via some sort of hierarchical method in the future.

7.4 Threats to Validity

Many of the assumptions and choices we make in our experiments along with the nature of the data we rely on from the DocRED dataset likely have a significant impact on model performance and the validity of NCD propositions. Firstly, although we aimed to select holdout classes that are both sufficiently distinct from one another and representative of true novel classes that may appear in a corpus, it is possible that selection bias has been introduced through holdout class selection. Our future work will include a comparison study on performance based upon different random holdout class selections. Furthermore, our clustering methods inherently assume that similar context representations among instances that express some sort of general relationship indicate that these instances share a relationship class. This is a superior assumption to the naive approach which does not consider any shared general relationship expression, however the meaningful cluster problem mentioned in Chapter 3 can still arise. It is possible that these instances that share a general relational "positiveness" can cluster together based upon context information such as the names of entities involved in their respective relationships rather than the relationship type itself.

Additionally, our detection of novel classes within the unlabeled set is subject to the nature and structure of the relationship types existent in our labeled set. It can be
argued that all entity pairs within a document are related in some way depending on the level of abstraction considered by a labeler. Therefore, our detection of novel relationships is likely restricted to instances that follow the general structure prescribed by instances in the labeled set. Furthermore, this notion of abstraction for what can be considered a relationship between two entities can disrupt the idea of what a “true negative” is. This ambiguity makes it difficult to concretely state that a model has the ability to distinguish between entity pairs that express a relationship and those that do not due to the fact that all instances are likely to express a relationship in some capacity. It is therefore more accurate to state that a model has developed the ability to identify general relationships of a contextual structure similar to those in the labeled set.
BIBLIOGRAPHY


