Entangled Relations: Leveraging NLI and Meta-analysis to Enhance Biomedical Relation Extraction

Anonymous EMNLP submission

Abstract

Recent research efforts have explored the potential of leveraging natural language inference (NLI) techniques to enhance relation extraction (RE). In this vein, we introduce METAENTAIL-RE, a novel adaptation method that harnesses NLI principles to enhance RE performance. Our approach follows past works by verbalizing relation classes into class-indicative hypotheses, aligning a traditionally multi-class classification task to one of textual entailment. We introduce three key enhancements: (1) Meta-class analysis which, instead of labeling non-entailed premise-hypothesis pairs with the less informative "neutral" entailment label, provides additional context by analyzing overarching meta relationships between classes; (2) Feasible hypothesis filtering, which removes unlikely hypotheses from consideration based on pairs of entity types; and (3) Group-based prediction selection, which further improves performance by selecting highly confident predictions. METAENTAIL-RE is conceptually simple and empirically powerful, yielding significant improvements over conventional relation extraction techniques and other NLI formulations. We observe F1 gains of 17.6 points on BioRED and 13.4 points on ReTACRED when compared to conventional methods, underscoring the versatility of METAENTAIL-RE across both biomedical and general domains.

1 Introduction

Relation extraction (RE) is an NLP task that distills factual information from text by identifying relationships between entities in the form of fact triplets (e.g.,⟨head, relation, tail⟩) (Califf and Mooney, 1997; Mintz et al., 2009; Soares et al., 2019; Wan et al., 2023). RE facilitates various downstream applications such as knowledge graph construction, question answering, and information retrieval (Yuan et al., 2022; He et al., 2023; Yamada et al., 2023); however, creating datasets for training RE models is costly and challenging, requiring annotators to identify entities and relations across large sections of text (Yao et al., 2019; Luo et al., 2022).

Recent efforts have explored adapting the RE task into an NLI task, enabling the use of relatively large NLI datasets to improve performance on an RE-adapted task (Sainz et al., 2021, 2022; Xu et al., 2023). RE-to-NLI works transform relation instances into premises paired with m class-indicative hypotheses where m is the number of relation classes in a dataset. A language model is trained to label premise-hypothesis pairs as *entailed*, *contradicted*, or *neutral*. We build on this work by introducing METAENTAIL-RE, a novel NLI adaptation method that improves RE performance by leveraging three key enhancements: meta-class analysis, automatic feasible hypothesis filtering, and group-based prediction selection.

Meta-class analysis: In past RE-to-NLI works, if a premise does not entail a hypothesis, the corresponding NLI label assigned is "neutral" (Sainz et al., 2021; Xu et al., 2023). However, this misses an implicit training signal we can gain by analyzing the semantics of a dataset's relation classes. When assigning NLI labels to adapted RE instances, we distinguish between *task-based* mutual exclusivity and *definition-based* mutual exclusivity. Task-based mutual exclusivity is an artifact of the single-class classification task inherent to a dataset. Each input instance is annotated with a single relation class, thereby arbitrarily making all classes mutually exclusive. In contrast, definition-based mutual exclusivity is derived from definitions of relation classes. For example, within the BioRED dataset (Luo et al., 2022), the "positive correlation" class is definitionally mutually exclusive and contradictory to the "negative correlation" class (Luo et al., 2022). We call this method *meta-class analysis* (MCA) and use it to determine the appropriate NLI labels for each premise-hypothesis pair. We show through ablation experiments that adding MCA leads to significant gains on the RE task.

Feasible hypothesis filter: We introduce a feasible hypothesis filter that automatically removes infeasible hypotheses based on the head and tail entity types within a relation instance. For instance, in the BioRED dataset (Luo et al., 2022), it is impossible for a gene to "bind" to a disease (i.e., the "bind" label is not applicable to gene-disease entity-type pairs). We therefore remove the "bind" hypothesis from all instances with gene-disease entity types. To develop this filter automatically, we approximate valid sets of entity-type pairs corresponding to each relation class by aggregating all relations in the training data. These approximated sets of valid type-pairs are then used to remove hypotheses that verbalize infeasible relationships. This filter improves training efficiency by reducing the number of NLI instances.

Group-based prediction selection: Group-based prediction selection exploits the feature of RE-to-NLI adaptation in that each relation instance is converted into a group of premise-hypothesis pairs where each hypothesis verbalizes a relation class in the dataset. When evaluating cases where the model predicts multiple "entail" labels within a single group, we can select the most confident "entail" prediction and ignore other predictions. Our results demonstrate that this group-based prediction selection method leads to additional gains.

METAENTAIL-RE as an RE-to-NLI adaptation method is technically domain agnostic; however, it is particularly well-suited for biomedical RE where associations often have opposing classes such as "positively correlated" and "negatively correlated" (Luo et al., 2022) or "agonist" and "antagonist" (Taboureau et al., 2010) enabling a rich MCA. We also find that associations in biomedical RE are often type-dependent compared to general domain RE, making the feasible hypothesis filter more effective at trimming infeasible hypotheses. Still, we extend our evaluations beyond the biomedical domain to determine how METAENTAIL-RE fares on general domain RE datasets. Notably, we observe improvements in both domains, reinforcing the effectiveness and versatility of METAENTAIL-RE. We summarize the main contributions of this work as the following:

- We introduce a novel RE-to-NLI adaptation method, METAENTAIL-RE, and showcase its robustness and versatility in RE datasets from general and biomedical domains.
- Through ablation experiments, we illustrate the effectiveness of components of METAENTAIL-RE.
- We openly provide all code, experimental settings, and datasets used to substantiate the claims made in this paper.[1]

2 Related Work

Traditionally, RE has been approached as a classification task, where input instances are classified as belonging to a relational class (Califf and Mooney, 1997; Mintz et al., 2009; Soares et al., 2019; Wan et al., 2023). These methods have several drawbacks: they tend to generalize poorly (Peng et al., 2020; Xu et al., 2023), and they heavily rely on relatively small and disjoint RE datasets. To account for these drawbacks, recent works have proposed clever adaptation methods to recast RE into adjacent NLP tasks, such as a question-answering (Levy et al., 2017) and NLI (Obamuyide and Vlachos, 2018; Sainz et al., 2021, 2022; Xu et al., 2023). Task adaptation presents an opportunity to leverage the relatively large datasets available for other tasks (e.g., SQuAD (Rajpurkar et al., 2016), MultiNLI (Williams et al., 2018), SNLI (Bowman et al., 2015a), etc.), which can be particularly advantageous in the context of biomedical RE where datasets are often limited.

Levy et al. (2017) recast RE into a question-answering task by associating relation instances with one or more natural-language questions, resulting in predicted spans denoting class indicative text. Obamuyide and Vlachos (2018) adapts general domain RE into an NLI task by using relation instances as premises where each premise is paired with a hypothesis generated by verbalizing a relation class. In doing so, they formulate a binary entailment task where they predict whether or not a premise entails the corresponding hypothesis.

Sainz et al. (2021) expands on Obamuyide and Vlachos (2018) by incorporating a three-label classification objective where a model can predict *entail*, *contradict*, and *neutral* depending on the premise-hypothesis pair, bringing the task in line with a standard NLI formulation (Dagan et al., 2005). They manually generate hypothesis templates corresponding to each relation class in a dataset, and NLI labels are assigned based on the alignment of the premise-hypothesis pair. If the corresponding hypothesis is the verbalized version of the ground truth relation label, then "entail" is

[1] https://anonymous.4open.science/r/metaentail-re

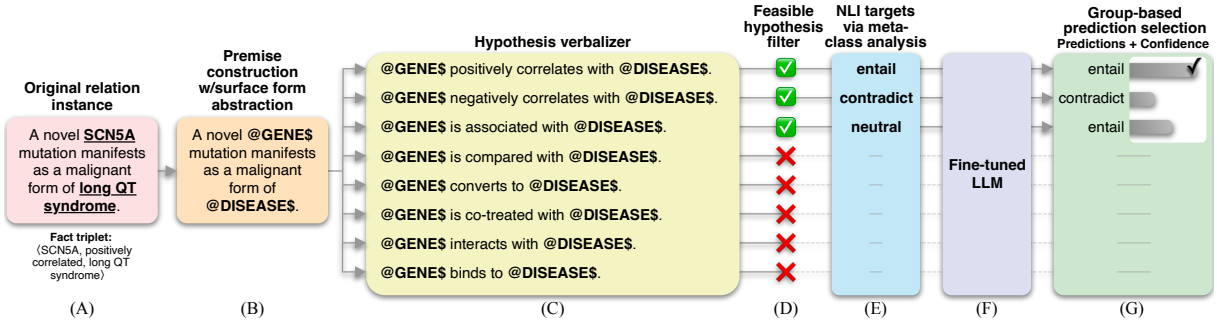

Figure 1: Data flow used for METAENTAIL-RE. The original RE input instance (A) is converted into a premise where surface forms are masked with corresponding entity types (B). Each relation class is verbalized into a hypothesis (C), and a feasible hypothesis filter (D) removes infeasible hypotheses based on the pair of entity types. NLI labels are generated via meta-class analysis (E), which are the labels used to fine-tune an LLM via cross-entropy (F). Finally, we use softmax probabilities as a proxy for the model's confidence and select the most confident "entail" prediction among the group of predictions (G). Note that the model makes three predictions in this example—one for each feasible hypothesis. The second "entail" prediction is incorrect but the group-based prediction module selects the first and correct "entail" prediction by assessing the model's confidence.

assigned as the NLI label for the instance. The "neutral" label is applied to positive class hypotheses which do not align with a given premise. The "contradict" label is applied in two cases: (1) if the premise is a positive relation instance (e.g., any class other than "no relation"), the "no-relation" hypothesis is labeled as "contradict," and (2) if the premise is a negative instance (e.g., "no relation"), then all other positive class hypotheses are labeled as "contradict." Sainz et al. (2021) fine-tune a language model pre-trained on the MultiNLI (Williams et al., 2018) dataset to predict generated NLI labels. They observe impressive results in zero- and few-shot scenarios on TACRED (Zhang et al., 2017), a general domain, sentence-level RE dataset.

Xu et al. (2023) explores cross-domain transfer learning, leveraging indirect supervision from general domain NLI datasets to improve biomedical RE-to-NLI adapted methods. Our work can be considered an extension of their proposed NBR method. However, we introduce a few key improvements: meta-class analysis, a feasible hypothesis filter, and group-based prediction selection. We also expand evaluations beyond sentence-level RE to include more challenging document-level RE (Li et al., 2016; Luo et al., 2022).

3 Problem Statement

Our problem is a hybridization of RE and NLI; as such, we describe both tasks, as well as the adapted RE-to-NLI task.

Relation Extraction (RE): RE takes inputs $\{x_1, x_2, \ldots, x_n\} \in X_{\text{RE}}$ where X_{RE} is a corpus of sentences, paragraphs, or documents of size n and x_i is a singular instance containing an entity pair e_{i_1} and e_{i_2}. Each input x_i has a corresponding label y_i. Labels $\{y_1, y_2, \ldots, y_n\} = Y_{\text{RE}}$ belong to a set of m relation classes $R = \{r_1, r_2, \ldots, r_m\}$. RE seeks to identify which class links the co-mentioned entities to form a fact triplet $\langle e_{i_1}, y_i, e_{i_2} \rangle$, or, semantically, $\langle \textit{head}, \textit{relation}, \textit{tail} \rangle$.

Natural Language Inference (NLI): NLI takes a premise $p_i \in P$ and a hypothesis $h_i \in H$, where P and H are the set of premises and hypotheses in a corpus, respectively, and seeks to determine whether the premise entails, contradicts, or is neutral to the respective hypothesis (Dagan et al., 2005; Bowman et al., 2015b). Using $\hat{y}_i$ to represent an NLI label applied to the i_{th} instance, $\hat{y}_i \in \{\textit{entail}, \textit{contradict}, \textit{neutral}\}$, and a single NLI example can be expressed as $\langle p_i, \hat{y}_i, h_i \rangle$.

RE-to-NLI Adaptation: RE-to-NLI adaptation converts RE inputs and labels into premise-hypothesis pairs such that each input instance maps to $|R|$ premise-hypotheses pairs: $(x_i, y_i) \rightarrow \{(p_i, \hat{y}_j, h_j)\}_{j=1}^{|R|}$. We decompose RE-to-NLI adaptation into the following sub-steps:

(a) Premise generation, $x_i \rightarrow p_i$: Input instances $x_i \in X_{\text{RE}}$ directly become premises $p_i \in P^{|X_{\text{RE}}|}$ where P is the collection of all premises generated from X_{RE}.

(b) Hypothesis generation, $H_i = \{h_j\}_{j=1}^{|R|}$: In the hypothesis generation step, a set of hypotheses H_i paired with each premise p_i. This is achieved by first verbalizing relation classes in R into a set of m hypothesis templates $T = \{t_1, t_2, \ldots, t_m\}$. Each hypothesis template contains head and tail entity placehold-

ers, which are replaced by the head and tail entities found in the corresponding premise p_i. The verbalizer function $f_{verbalizer}(\cdot)$ takes each hypothesis template and entity pair in premise p_i to produce the set of hypotheses $H_i = \{f_{verbalizer}(t_j, e_{i_1}, e_{i_2})\}_{j=1}^{|R|}$.

(c) NLI label generation, $\hat{Y} = \{\hat{y}_i\}_{i=1}^{|X_{\text{RE}}|\times|R|}$: The set of NLI labels $\hat{Y}$ is generated via a function which takes the original instance label y_i and the premise-hypothesis pair $f_{target}(y_i, p_i, h_j) \rightarrow \hat{y}_j$ where NLI label $\hat{y}_j = entail$ iff verbalized class-indicative hypothesis h_j aligns with the ground truth label y_i, and, depending on the adaptation method, $\hat{y}_j$ is assigned $neutral$ or $contradict$ for non-aligned hypotheses.

The RE-to-NLI task is to correctly predict entailed premise-hypothesis pairs where each entailed pair has a 1-to-1 mapping to the original RE label.

4 Methods

This section sequentially discusses the modules used in METAENTAIL-RE (see Figure 1).

Premise Construction: Following Xu et al. (2023), a relation instance x_i is transformed into a premise by replacing surface forms of the subject and object entities, e_1 and e_2, respectively, with their corresponding entity types, $e_{1_{type}}$ and $e_{2_{type}}$. Abstracting entity surface forms into entity types helps alleviate the long-tail nature of biomedical entities and encourages language models to learn from context instead of shallow heuristics (Peng et al., 2020). The start and end spans of entity types are denoted with "@" and "$," respectively.

Hypothesis Verbalizer: Past works have manually generated hypothesis templates for each relation class in a dataset which are then used, in turn, to generate hypotheses to pair with a given premise. A secondary contribution of METAENTAIL-RE is that we reduce this human effort by leveraging LLMs to automatically generate the set of hypothesis templates $\{t_1, t_2, \ldots, t_m\} \in T$, where m corresponds to the number of relation classes in a dataset. We prompt an LLM[2] to verbalize each relation class using natural language and placeholders for subject and object entities (see Appendix A.1 for more details). The placeholders within the hypothesis templates are replaced by the entity types, $e_{1_{type}}$ and $e_{2_{type}}$, found in the corresponding premise.

[2]We use ChatGPT (GPT 3.5) via OpenAI's web interface.

Feasible Hypothesis Filter: There is an implicit multiplicative effect of adapting RE into an NLI task where each relationship instance produces m class-indicative hypotheses resulting in $|X_{\text{RE}}| \times m$ premise-hypothesis pairs. To mitigate this effect, we develop a feasible hypothesis filter which automatically filters out improbable hypotheses by aggregating valid sets of entity-type pairs by relationship classes across all training data: $E_{valid} = \{r_1 \mapsto S_1, r_2 \mapsto S_2, \ldots, r_m \mapsto S_m\}$ where $r_j \in R$ for $j = 1, 2, \ldots, m$ and each S_j is the set of tuples of entity-type pairs associated with all instances of relationship class r_j.

Using this filter, we assess the feasibility of hypotheses given a pair of entity types: $\hat{H}_i = \{h_j | (e_{1_{type}}, e_{2_{type}}) \in E_{valid}(r_j)\}_{j=1}^{|R|}$ where $\hat{H}_i$ is a set of feasible hypotheses given the entity-type pair found in instance i, and $\hat{H}_i \subset H$ where H is the set of all possible hypotheses.

Since sets of feasible hypotheses are approximated using the training data's relationships and entity-type pairs, the filter may remove valid hypotheses based on an entity-type pair and corresponding relation that exists only in the test set. For these instances, the entailed premise-hypothesis pair will not be presented to the model, leading to false negatives. However, in practice, we observe that this does not occur with the datasets we use for evaluation and should not occur as long as training data is sufficiently representative of the test data (i.e., the training data contains at least one relationship with a specific entity-type pair for every relation and entity-type pair found in the test set).

Meta-class Analysis (MCA): After applying the aforementioned feasible hypothesis filter, we leverage MCA to assign NLI labels, namely *entail*, *neutral*, and *contradict*, to the resultant premise-hypothesis pairs. To do this, we first construct definition-based mutually exclusive meta-relationships between relation classes. For example, in the ChemProt dataset, the "up regulator" class is, by definition, mutually exclusive to the "down regulator" class. For datasets with a negative class (e.g., "no relation"), the negative class is mutually exclusive to all positive classes and vice-versa. With this analysis, we construct NLI labels in the following way:

(a) *Entail*: Premise-hypothesis pairs are labeled "entail" when the hypothesis h_j aligns with the verbalized ground truth label y_i.

(b) *Neutral*: If the original instance expresses a

positive class (i.e., any class other than the "no relation" class), then all non-exclusive class hypotheses are labeled as "neutral."

(c) *Contradict*: The "contradict" label is assigned to hypotheses that verbalize definitionally exclusive classes.

See Appendix A.4 for tables showing how original relation labels map to NLI labels using MCA for each dataset.

LLM Fine-tuning: With generated premise-hypothesis pairs, we train a discriminative language model, namely BioLinkBERT$_{\text{large}}$ (Yasunaga et al., 2022), to predict NLI labels. We concatenate premise-hypothesis pairs as the input to the language model and send the resultant representation of the special [CLS] token through a fully connected layer, which is trained using cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{m} y_{o,i} \cdot \log\left(p\left(y_{o,i}\right)\right) \quad (1)$$

where y is a binary indicator that is 1 if and only if i is the correct classification for observation o, $p(y_{o,i})$ is the softmax probability that observation o is of class i, and m is the number of classes.

Group-based Prediction Selection: Given the multiplicative effect of adapting RE-to-NLI where one relation instance results in a group of up to m premise-hypothesis pairs, we can employ a group selection method to select the most confident *entail* prediction. If the model predicts two or more entailed instances within a group of premise-hypothesis pairs, we use the softmax probability from Equation 1 as a proxy for model confidence (Hendrycks and Gimpel, 2017) and select the prediction with the highest confidence. We allow the model to naturally abstain from making a prediction by predicting "neutral" for all premise-hypothesis pairs in a group.

5 Experiments

5.1 Datasets

We include a spread of experiments on various biomedical RE datasets. BioRED is a document-level RE dataset featuring eight relation classes (Luo et al., 2022). BioRED also provides an orthogonal and binary "Novel" class, which annotates whether an instance expresses a novel finding. BC5CDR is a document-level RE dataset featuring binary relations between chemical and disease entities (Li et al., 2016). DDI13 is a drug-drug interaction dataset with four relation classes (Herrero-Zazo et al., 2013), and ChemProt is a chemical-protein dataset featuring five relation classes (Taboureau et al., 2010). GAD is a gene-disease dataset with binary relations (Bravo et al., 2014). We only include GAD in our main experiment for comparative purposes to past works. We believe that the GAD dataset should be retired from future works due to significant label accuracy issues, which the authors acknowledge.[3]

As mentioned in Section 1, our method is designed to leverage features of biomedical domain RE, namely the prevalence of definitionally exclusive classes and the importance of entity types vis-à-vis feasible relationships. However, we also seek to assess our method beyond the biomedical domain and extend our experiments to general domain datasets ReTACRED (Stoica et al., 2021) and SemEval-2010 Task 8 (Hendrickx et al., 2010). ReTACRED is a re-annotated version of TACRED (Zhang et al., 2017) and features 40 relation classes—significantly more classes than any of the biomedical datasets we tested. SemEval-2010 is a sentence-level RE dataset with ten relation classes.

5.2 Baselines

5.2.1 Traditional Multi-Class Classification

We select leading biomedical language models and train them using a traditional RE multi-class approach where models directly predict relation classes. **BioM-ALBERT$_{\text{xxlarge}}$**, **BioM-BERT$_{\text{large}}$**, and **BioM-ELECTRA$_{\text{large}}$** (Alrowili and Shanker, 2021) are transformer architectures adapted into the biomedical domain by using a custom biomedical vocabulary and pre-training on PubMed abstracts (National Library of Medicine (US), 1946) and PubMed Central articles (National Library of Medicine (US), National Center for Biotechnology Information, 2000). **BioMed RoBERTa$_{\text{base}}$** (Gururangan et al., 2020) features the RoBERTa architecture (Liu et al., 2019) adapted to the biomedical domain via continued pre-training on papers from the S2OR Corpus (Lo et al., 2020). **PubMedBERT$_{\text{base}}$** (Gu et al., 2020) and **BioLinkBERT$_{\text{large}}$** (Yasunaga et al., 2022) are BERT (Devlin et al., 2019) variants. The former is trained on PubMed abstracts with a custom biomedical vocabulary. The latter is trained with two self-supervised objectives: masked language modeling and document relation prediction.

5.2.2 NLI Adapted Models

NBR is a biomedical domain RE-to-NLI method that leverages BioLinkBERT$_{\text{large}}$ as a backbone

[3] https://github.com/dmis-lab/biobert/issues/162

language model. Like our method, NBR converts relation instances and labels into premise-hypothesis pairs. Key differences between NBR and our method are that NBR does not use MCA or feasible hypothesis filtering, and they leverage a ranking loss training objective to rank entailed premise-hypothesis pairs over non-entailed pairs.

The RE-to-NLI adaptation method used in METAENTAIL-RE is architecture-agnostic, so we also experiment with auto-regressive architectures. We conduct the following experiment using identical data and methods to those discussed in Section 4; the only difference is the final training step.

We fine-tune **Phi-2** (2.7B) and **Phi-3** (3.8B).[4] For Phi-2 and Phi-3, we construct a seq-to-seq task and fine-tune the models to generate an NLI label for each premise-hypothesis pair. For more information about training Phi-2 and Phi-3, see Appendix A.2.2.

We also seek to assess the performance of large, frontier auto-regressive language models, **GPT 3.5** (OpenAI, 2024) and **GPT 4** (OpenAI et al., 2024),[5] leveraging few-shot, in-context learning. For more on the prompts we use to solicit predictions from GPT 3.5 and GPT 4, see Appendix A.2.3.

For all NLI-adapted models, only entailed premise-hypothesis pairs map directly to the original RE training instance. Thus, we only keep NLI instances labeled or predicted as entailed when mapping instances back into the original RE labels for evaluation. This ensures a fair comparison across adapted and non-adapted methods.

5.3 General Domain Experiments

For our general domain experiments, we use **DeBERTaV3$_{\text{large}}$** (He et al., 2021) and **RoBERTa-MNLI$_{\text{large}}$** (Liu et al., 2019). DeBERTaV3 is an improved version of BERT that uses replaced token detection, a more sample-efficient pre-training objective. RoBERTa-MNLI$_{\text{large}}$ is the RoBERTa architecture fine-tuned on the MNLI corpus (Williams et al., 2018).[6]

We make slight modifications to the general domain version of METAENTAIL-RE. We use RoBERTa-MNLI$_{\text{large}}$ as the backbone language model, and we do not leverage surface-form abstraction for entity types (i.e., we leave the original entities as they appear in the text and do not replace them with their corresponding types). Entity surface form abstraction is a method developed for the long-tail nature of biomedical entities (Peng et al., 2020). Also, some general domain RE datasets, such as SemEval-2010 Task 8, do not provide annotated entity type information.

6 Results

We observe an interesting comparison between the BioLinkBERT$_{\text{large}}$ model and METAENTAIL-RE. Both experiments share the same backbone language model, yet the performance of our METAENTAIL-RE method is significantly higher providing evidence of the effectiveness of adapting the RE task into one of textual entailment. We hypothesize that the boost in performance primarily comes from the additional data abstraction RE-to-NLI introduces by training the model to recognize entailed premise-hypothesis pairs instead of directly predicting suppositional classes. By combining RE-to-NLI adaptation with surface-form entity abstraction, the model is less prone to memorizing entities and shallow heuristics of relation classes; instead, it must understand the context and the natural language interplay between a premise and hypothesis. Furthermore, the boost in performance between the NBR model and METAENTAIL-RE highlights the effectiveness of leveraging MCA, feasible hypothesis filtering, and group-based prediction selection.

Within the biomedical domain experiments, the NLI-adapted auto-regressive models generally underperform compared to the discriminative models. Predictably, the larger Phi-3 outperforms Phi-2 and fine-tuning smaller auto-regressive models outperforms larger models, GPT 3.5 and GPT 4, leveraging few-shot in-context learning. This aligns with findings from Peng et al. (2024) that LLMs using in-context learning underperform relative to smaller, fine-tuned language models on information extraction tasks.

In the general domain, we observe better overall performance from auto-regressive architectures. The performance from Phi-3 approaches that of METAENTAIL-RE on both ReTACRED and SemEval-2010 Task 8 datasets. These results are promising for auto-regressive models, in general. We leave fine-tuning larger auto-regressive models to future work but expect additional gains to be made, potentially overtaking the smaller discriminative models.

[4] We use the *microsoft/phi-2* and *microsoft/Phi-3-mini-4k-instruct* checkpoints from Hugging Face.

[5] Specifically, we use *gpt-3.5-turbo-0125* and *gpt-4-turbo-2024-04-09* via OpenAI's API.

[6] We use the *FacebookAI/roberta-large-mnli* checkpoint from Hugging Face.

Model	BC5CDR	BioRED	BioRED (novel)	ChemProt	DDI13	GAD
TRADITIONAL MULTI-CLASS CLASSIFICATION						
BioM-ALBERT$_{xxlarge}$ (Alrowili and Shanker, 2021)	0.679	0.668	0.863	0.940	0.911	0.815
BioM-BERT$_{large}$ (Alrowili and Shanker, 2021)	0.681	0.709	0.904	0.934	0.917	0.795
BioM-ELECTRA$_{large}$ (Alrowili and Shanker, 2021)	0.687	0.657	0.903	0.925	0.885	0.830
BioMed RoBERTa$_{base}$ (Gururangan et al., 2020)	0.664	0.714	0.897	0.919	0.911	0.803
PubMedBERT$_{base}$ (Gu et al., 2020)	0.651	0.715	0.891	0.923	0.916	0.803
BioLinkBERT$_{large}$ (Yasunaga et al., 2022)	0.682	0.699	0.899	0.931	0.917	0.806
NLI ADAPTED MODELS						
NBR (Xu et al., 2023)	0.679	0.543	0.664	0.883	0.846	0.831
Phi-2 (Li et al., 2023)	0.653	0.715	0.824	0.852	0.873	0.729
Phi-3 (Abdin et al., 2024)	0.749	0.688	0.840	0.930	0.915	0.721
GPT 3.5† (OpenAI, 2024)	0.282	0.470	0.594	0.494	0.386	0.548
GPT 4† (OpenAI et al., 2024)	0.418	0.532	0.680	0.626	0.492	0.660
METAENTAIL-RE	**0.757**	**0.891**	**0.917**	**0.968**	**0.957**	**0.878**

Table 1: Micro F1 scores for traditional RE and NLI adapted methods. †Results from GPT 3.5 and GPT 4 are via in-context learning (see Appendix A.2.3 for details), whereas other models were fine-tuned directly on the task from our own implementations. Results show averages over five runs.

Model	ReTACRED	SemEval
TRADITIONAL MULTI-CLASS CLASSIFICATION		
DeBERTaV3$_{large}$ (He et al., 2021)	0.809	0.807
RoBERTa-MNLI$_{large}$ (Liu et al., 2019)	0.800	0.828
NLI ADAPTED MODELS		
NBR (Xu et al., 2023)	0.875	0.826
Phi-2 (Li et al., 2023)	0.862	0.855
Phi-3 (Abdin et al., 2024)	0.880	0.871
GPT 3.5† (OpenAI, 2024)	0.306	0.340
GPT 4† (OpenAI et al., 2024)	0.565	0.616
METAENTAIL-RE	**0.943**	**0.902**

Table 2: Micro F1 scores from general domain RE experiments.

Model	BioRED	ChemProt	ReTACRED
METAENTAIL-RE	0.891	0.968	0.943
(w/o Feasible Hypothesis Filter)	0.876	N/A	DNC
(w/o Meta-class Analysis)	0.853	0.911	0.916
(w/o Grouped Selection)	0.805	0.950	0.875

Table 3: Micro F1 scores from ablation experiments which remove each proposed module within METAENTAIL-RE. Each module has a significant impact on performance. ChemProt is monolithic in its entity types (*chemicals* and *diseases*), which prevents the use of the feasible hypothesis filter. On ReTACRED, we observe that without applying the feasible hypothesis filter, the model does not converge (DNC).

6.1 Ablation Experiments

We conduct ablation experiments to better understand METAENTAIL-RE's performance gains by removing modules and reporting the performance. Note that the performance of BioLinkBERT$_{large}$ in Table 1 can be considered a type of ablation of METAENTAIL-RE that does not leverage NLI adaptation or any additional modules since METAENTAIL-RE uses BioLinkBERT$_{large}$ as its backbone language model. For our ablations, we choose to examine the BioRED, ChemProt, and ReTACRED datasets because they feature more than two relation classes and contain one or more definition-based mutually exclusive relations as determined by MCA.

(a) *w/o Feasible Hypothesis Filter*: We remove the feasible hypothesis filter, and, in doing so, each original relation instance is converted into m premise-hypothesis pairs, with m being the number of classes in a dataset. Removing the feasible hypothesis filter produced a moderate drop in performance on BioRED ($m = 8$). Since the feasible hypothesis filter is based on entity type pairs, it is not available (N/A) for datasets such as ChemProt, which only feature a single entity type pair (namely, *chemical* and *gene*) associated with every relation class. However, the feasible hypothesis filter is essential in model convergence when a dataset consists of many relation classes, such as ReTACRED ($m = 40$). In our experiments on ReTACRED without the feasible hypothesis filter, the model did not converge (DNC), likely due to the overwhelming number of non-informative "neutral" premise-hypothesis pairs used in training.

(b) *w/o Meta-class Analysis*: Removing MCA and using "neutral" as the NLI label for all non-entailed premise-hypothesis pairs led to a considerable drop in performance, indicating the benefit of training the model with the additional training signal obtained via MCA. Note that in this ablation experiment, we maintain mutual exclusive NLI labels between positive and negative (i.e., "no relation") classes.

(c) *w/o Group Prediction Selection*: To remove this module, we select all entailed predictions

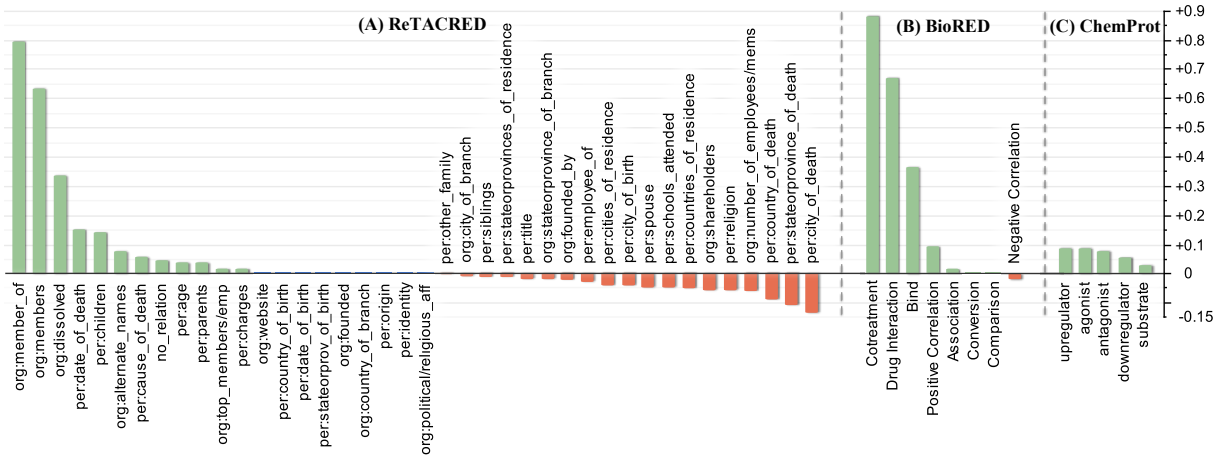

Figure 2: ΔF1 per relation class when leveraging meta-class analysis to assign NLI labels.

regardless of how many *entail* predictions are made within a group of premise-hypothesis pairs. Doing this removes the constraint imposed by the single-class classification task and allows the model to freely predict multiple classes for a single relation instance. This ablation experiment led to a drop in performance across all datasets but most significantly on BioRED, which we suspect results from the closeness in BioRED's "positively correlated" and "associated" relation classes. These classes were typically confused since "associated" can sometimes be considered a hypernym of "positively correlated," leading the model to predict *entail* for both of the corresponding hypotheses.

6.2 Meta-class Analysis Case Study

To further explore the impacts of leveraging MCA, we decompose results from ReTACRED, BioRED, and ChemProt by evaluating the change in Micro F1 scores (ΔF1) for each class. We isolate the effect of MCA by training identical models with and without MCA-informed NLI labels and report the results in Figure 2.

MCA results in a net benefit in performance across classes and datasets, but the specific nature of these benefits varies. In ReTACRED, we observe notable improvements in the "member of" and "members" classes, which are definitionally exclusive. Conversely, some classes experience minor decreases in performance. For BioRED, we observe a slight drop in predictive performance for the "negative correlation" class, while all other classes get a significant boost. The largest performance gains are seen in classes that are not mutually exclusive, suggesting that the additional training signal from MCA aids the model in disentangling adjacent relation class representations. In ChemProt, we observe near-uniform, albeit relatively small, boosts in performance across all classes.

6.3 Additional Experiments

Given that RE-to-NLI adaptation leads to models predicting the same *entail*, *neutral*, *contradict* labels across disparate datasets, we naturally sought to investigate the potential of combining the relatively small and disjoint biomedical RE datasets into a single, unified task. Unfortunately, these experiments failed to produce significant performance gains, indicating that these biomedical datasets have limited synergistic effects when adapted to the NLI task. We present experiment details and results in Appendix A.3.

7 Conclusion

The exploration of NLI techniques to enhance relation extraction has opened new avenues in natural language processing, and our study introduces METAENTAIL-RE as an advancement in this area. By adapting the RE task into an NLI framework and incorporating innovative strategies such as meta-class analysis, feasible hypothesis filtering, and group-based prediction selection, METAENTAIL-RE demonstrates remarkable improvements in RE performance. Our experiments, conducted across biomedical and general domain datasets, highlight the robustness and versatility of METAENTAIL-RE. By openly sharing our code, experimental settings, and datasets, we aim to facilitate further research and development in this promising intersection of NLI and RE, paving the way for more sophisticated and accurate information extraction systems in diverse domains.

Limitations

METAENTAIL-RE is not without its limitations. By verbalizing a hypothesis for each relation class, the training data is multiplied by the number of relation classes in the dataset, necessitating additional training resources. Our introduced module, the feasible hypothesis filter, relies heavily on accurate entity-type information. This information is crucial for the success of the adaptation process. However, the filtering process becomes ineffective if this information is unavailable or if numerous feasible hypotheses (e.g., 40+) exist for a given relation class and entity type pair. In these scenarios, the "entail" class becomes a minority class in a sea of "neutral" NLI instances, potentially causing the model to collapse to a trivial state of simply predicting "neutral" for every premise-hypothesis pair. Such a scenario would require the design of manually tuned sampling strategies or bespoke learning objectives to handle the overwhelming number of "neutral" premise-hypothesis pairs. We defer the exploration of such challenging settings to future research.

Additionally, in our study, meta-class analysis is performed manually, which introduces an extra layer of human effort. This manual effort involves reading annotation guidelines for a specific dataset to determine which relation classes are mutually exclusive based on their definitions. While this task is relatively quick and straightforward, it does require additional human involvement.

Ethics Statement

We do not anticipate any major ethical concerns; relation extraction is a fundamental problem in natural language processing. A minor consideration is the potential for introducing certain hidden biases into our results (i.e., performance regressions for some subset of the data despite overall performance gains). However, we did not observe any such issues in our experiments, and indeed these considerations seem low-risk for the specific datasets studied here because they are all published.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Àlex Bravo, Janet Piñero, Núria Queralt, Michael Rautschka, and Laura Inés Furlong. 2014. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16.

Mary Elaine Califf and Raymond J. Mooney. 1997. Relational learning of pattern-match rules for information extraction. In *CoNLL*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Yichen He, Xiaofeng Liu, Jinlong Hu, and Shoubin Dong. 2023. Entity relation aware graph neural ranking for biomedical information retrieval. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1118–1124.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *International Workshop on Semantic Evaluation*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of biomedical informatics*, 46 5:914–20.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **Phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia Noemi Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23.

Mike D. Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

National Library of Medicine (US). 1946. PubMed [Internet]. `https://www.ncbi.nlm.nih.gov/pubmed/`. [cited 2024 Apr 05].

National Library of Medicine (US), National Center for Biotechnology Information. 2000. PubMed Central (PMC) [Internet]. `https://www.ncbi.nlm.nih.gov/pmc/`. [cited 2024 Apr 05].

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2024. Chatgpt: A large language model trained on the gpt-3.5 architecture. `https://openai.com/blog/chatgpt`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane

Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. *ArXiv*, abs/2010.01923.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. Metaie: Distilling a meta model from llm for all kinds of information extraction tasks.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. *ArXiv*, abs/2109.03659.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. *ArXiv*, abs/2205.01376.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *ArXiv*, abs/1906.03158.

George Stoica, Emmanouil Antonios Platanios, and Barnab'as P'oczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *AAAI Conference on Artificial Intelligence*.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak, and Tudor I. Oprea. 2010. Chemprot: a disease chemical biology database. *Nucleic Acids Research*, 39:D367 – D372.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *ArXiv*, abs/2305.02105.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *ArXiv*, abs/2303.03846.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long Papers), pages 1112–1122. Association for Computational Linguistics.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2023. Can nli provide proper indirect supervision for low-resource biomedical relation extraction? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Koshi Yamada, Makoto Miwa, and Yutaka Sasaki. 2023. Biomedical relation extraction with entity type markers and relation-specific question answering. In *Workshop on Biomedical Natural Language Processing*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Li Yuan, Yi Cai, Jin Zhen Wang, and Qing Li. 2022. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *AAAI Conference on Artificial Intelligence*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Appendix

A.1 Automatic Generation of Hypothesis Templates

To reduce human effort in our methods, we turn to LLMs, specifically GPT 3.5 (OpenAI, 2024), to automatically generate hypothesis templates. Some datasets, such as BC5CDR, GAD, and BioRED Novel, feature two classes, making the template generation process relatively trivial. The benefits of automating the generation of hypothesis templates are more significant for datasets such as Re-TACRED, which feature 40 relation classes.

We use the following prompt where the ellipsis is replaced with the list of natural language relation classes (e.g., relation classes with underscores removed and spaces inserted) used in each dataset:

> Verbalize the following relation classes in the form "subj [verbalized relation] obj": [. . .].

A special case arose for the DDI13 dataset where each relation instance describes a relation between two drugs. We referenced the verbalized hypotheses proposed by Xu et al. (2023) and included instructions about describing two drug entities:

> Verbalize the following relation classes using the form "[verbalized relation] two drugs is described": [. . .].

Table 4 contains the generated hypothesis templates for each dataset.

A.2 Baselines

A.2.1 GPU Resources

All baselines were trained on a single NVIDIA A100, and training times ranged from 1 to 12 hours, changing based on the size of the dataset and the number of parameters in the model.

A.2.2 Phi-2 and Phi-3

Since responses from auto-regressive models may sometimes include additional text, all responses are aligned to ground truth labels using partial string matching. We do this by searching for the matches of the first three letters in each NLI label (e.g., "ent" → *entail*, "con" → *contradict*, "neu" → *neutral*). When a class cannot be matched, we assign "none," which, during evaluation, is equivalent to the NLI label *neutral*.

For Phi-2, we use the following prompt to finetune the model on our task:

> [INST]You are given a premise and a hypothesis below. If the premise entails the hypothesis, return "entail." If the premise contradicts the hypothesis, return "contradict." Otherwise, if the premise does neither, return "neutral."[/INST]
>
> ### Premise: [premise]
> ### Hypothesis: [hypothesis]
> ### Label: [nli_target]

Phi-3 uses a similar prompt that differs only in format:

> <|system|>
> You are given a premise and a hypothesis below. If the premise entails the hypothesis, return "entail." If the premise contradicts the hypothesis, return "contradict." Otherwise, if the premise does neither, return "neutral."
> <|end|>

<|user|>
Premise: [premise]
Hypothesis: [hypothesis]
Label:
<|end|>

<|assistant|>
[nli_target]
<|end|>

Both Phi-2 and Phi-3 were fine-tuned using the hyperparameters in Table 5.

A.2.3 GPT 3.5 and GPT 4

GPT 3.5 and GPT 4 often perform better on tasks with the help of in-context learning (Wei et al., 2023; Wang et al., 2023). We construct a prompt that lists the NLI labels and offers four examples of premise-hypothesis pairs expressing each NLI label.

The following is the prompt we used for soliciting predictions for our tests:

You are given a premise and a hypothesis below. If the premise entails the hypothesis, return "entail." If the premise contradicts the hypothesis, return "contradict." Otherwise, if the premise does neither, return "neutral." The following are some examples:

Premise: [premise]
Hypothesis: [hypothesis]
Label: [nli_target]
...
{*4x examples of each NLI class are provided*}
...

For responses from GPT 3.5 and GPT 4, we use the same partial string matching used for Phi-2 and Phi-3 (Appendix A.2.2) for evaluation.

A.2.4 Hyperparameters for METAENTAIL-RE

Table 6 contains the hyperparameters used to train METAENTAIL-RE.

A.3 Task Unification Results

We explore unifying biomedical relation extraction datasets in hopes of boosting performance on a target dataset. We investigate two task-unification training methodologies: single-stage training and double-stage training. Single-stage training can be viewed as multi-task learning, where the model is trained simultaneously on multiple datasets and tested on a target dataset. Double-stage training can be viewed as an initial pre-training stage on all data except the target dataset, followed by fine-tuning and evaluation on the target dataset.

Unfortunately, we did not observe a significant performance boost across our task-unification experiments (see Table 7), potentially indicating that these biomedical datasets do not provide complementary information when adapted into an NLI task. Generally, the two-stage training is more effective than the single-stage training, but both fail to realize significant performance gains on the target datasets. We leave investigating other task-unification methods for future works.

A.4 Meta-class analysis

We conduct a meta-class analysis for each dataset used in Section 5. We leverage class definitions to determine sets of mutually exclusive classes. The following tables show how meta-class analysis converts original RE labels (row headers) into NLI labels. The $h(class)$ column headers denote verbalized hypotheses using the corresponding class. For each table, we use the following denote NLI labels:

- 0 $\rightarrow$ *contradict*
- 1 $\rightarrow$ *neutral*
- 2 $\rightarrow$ *entail*

Dataset	Relation Classes	Hypothesis Templates
BC5CDR	Associated	“subj is associated with obj.”
	Not Associated	“subj is not associated with obj.”
BioRED	Positive Correlation	“subj positively correlates with obj.”
	Negative Correlation	“subj negatively correlates with obj.”
	Association	“subj is associated with obj.”
	Comparison	“subj is compared with obj.”
	Conversion	“subj converts to obj.”
	Cotreatment	“subj is co-treated with obj.”
	Drug Interaction	“subj interacts with obj (as drugs).”
	Bind	“subj binds to obj.”
BioRED Novel	Novel	“subj introduces a novel relationship to obj.”
	Not novel	“subj does not introduce a novel relation to obj.”
ChemProt	Upregulator	“subj upregulates obj.”
	Downregulator	“subj downregulates obj.”
	Agonist	“subj acts as an agonist for obj.”
	Antagonist	“subj acts as an antagonist for obj.”
	Substrate	“subj is a substrate for obj.”
DDI13	Advise	"Advice regarding two drugs is described.”
	Effect	"An effect between two drugs is described.”
	Interaction	"An interaction between two drugs is described.”
	Mechanism	"The mechanism involving two drugs is described.”
GAD	Associated	“subj is associated with obj.”
	Not Associated	“subj is not associated with obj.”
ReTACRED	No relation	“subj has no relation with obj.”
	Org:alternate names	“subj has alternate names as obj.”
	Org:city of branch	“subj’s branch is located in the city of obj.”
	Org:country of branch	“subj’s branch is located in the country of obj.”
	Org:dissolved	“subj has been dissolved.”
	Org:founded	“subj was founded on the date obj.”
	Org:founded by	“subj was founded by obj.”
	Org:member of	“subj is a member of obj.”
	Org:members	“subj has members including obj.”
	Org:number of employees/members	“subj has obj number of employees/members.”
	Org:political/religious affiliation	“subj has political/religious affiliation with obj.”
	Org:shareholders	“subj has shareholders including obj.”
	Org:state or province of branch	“subj’s branch is located in the state or province of obj.”
	Org:top members/employees	“subj’s top members/employees include obj.”
	Org:website	“subj’s website is obj.”
	Per:age	“subj’s age is obj.”
	Per:cause of death	“subj’s cause of death is obj.”
	Per:charges	“subj is charged with obj.”
	Per:children	“subj has obj as children.”
	Per:cities of residence	“subj resides in cities including obj.”
	Per:city of birth	“subj was born in the city of obj.”
	Per:city of death	“subj died in the city of obj.”
	Per:countries of residence	“subj resides in countries including obj.”
	Per:country of birth	“subj was born in the country of obj.”
	Per:country of death	“subj died in the country of obj.”
	Per:date of birth	“subj was born on the date obj.”
	Per:date of death	“subj died on the date obj.”
	Per:employee of	“subj is an employee of obj.”
	Per:identity	“subj’s identity is obj.”
	Per:origin	“subj’s origin is obj.”
	Per:other family	“subj has obj as other family members.”
	Per:parents	“subj’s parents include obj.”
	Per:religion	“subj’s religion is obj.”
	Per:schools attended	“subj attended schools including obj.”
	Per:siblings	“subj’s siblings include obj.”
	Per:spouse	“subj’s spouse is obj.”
	Per:state or province of birth	“subj was born in the state or province of obj.”
	Per:state or province of death	“subj died in the state or province of obj.”
	Per:state or provinces of residence	“subj resides in states or provinces including obj.”
	Per:title	“subj’s title is obj.”
SemEval 2010	Other	“subj and obj are related in some other way.”
	Component-Whole	“subj is a component of obj.”
	Instrument-Agency	“subj is used by obj.”
	Member-Collection	“subj is a member of obj.”
	Cause-Effect	“subj causes obj.”
	Entity-Destination	“subj is taken to obj.”
	Message-Topic	“subj is about obj.”
	Entity-Origin	“subj comes from obj.”
	Product-Producer	“subj is produced by obj.”
	Content-Container	“subj contains obj.”

Table 4: Auto-generated hypothesis templates for each relation class in each dataset. Hypotheses are generated using GPT 3.5 and the prompt described in Appendix A.1.

Parameter	Value
Epochs	3
Max seq. length	1,024
Batch size	3
Grad. accumulation steps	2
Max gradient norm	0.3
Learning rate	2e-4
Lr scheduler type	cosine
Weight decay	0.001
Warm-up ratio	0.03

Table 5: Hyperparameters used to fine-tune Phi-2 and Phi-3.

Parameter	Value
Epochs	3
Batch size	32
Grad. accumulation steps	1
Max seq. length	1,024
Learning rate	2e-5
Seeds	{41, 42, 43, 44, 45}
Optimizer	AdamW
LR Scheduler warm-up steps	0
LR Scheduler training steps	1,000

Table 6: Hyperparameters used to fine-tune BioLinkBERT$_{\text{large}}$ for the METAENTAIL-RE method.

MULTI-TASK LEARNING (SINGLE-STAGE)

ENSEMBLE TRAINING DATA →	TEST SET	Δ F1
BC5CDR + BioRED + ChemProt + DDI13	BC5CDR	-0.049
BioRED + ChemProt + DDI13 + BC5CDR	BioRED	-0.031
ChemProt + BioRED + DDI13 + BC5CDR	ChemProt	+0.012
DDI13 + BioRED + ChemProt + BC5CDR	DDI13	-0.007

CONTINUED PRE-TRAINING WITH SUPERVISED FINE-TUNING (DOUBLE-STAGE)

PRE-TRAINING CORPUS →	FINE-TUNING →	TEST SET	ΔF1
BioRED + ChemProt + DDI13	BC5CDR	BC5CDR	-0.005
ChemProt + DDI13 + BC5CDR	BioRED	BioRED	-0.014
BioRED + DDI13 + BC5CDR	ChemProt	ChemProt	-0.008
BioRED + ChemProt + BC5CDR	DDI13	DDI13	-0.011

Table 7: Results from single-stage and double-stage task unification experiments. ΔF1 scores are relative to METAENTAIL-RE scores from Table 1. We do not observe signification performance improvements from our task unification experiments and leave further experimentation to future work.

	h(Associated)	h(Not Associated)
Associated	2	0
Not Associated	0	2

Table 8: Meta-class analysis for BC5CDR. The "Associated" class is definitionally mutually exclusive to the "Not Associated" class.

	h(Positive Correlation)	h(Negative Correlation)	h(Association)	h(Comparison)	h(Conversion)	h(Co-treatment)	h(Drug Interaction)	h(Bind)
Positive Correlation	2	0	1	1	1	1	1	1
Negative Correlation	0	2	1	1	1	1	1	1
Association	1	1	2	1	1	1	1	1
Comparison	1	1	1	2	1	1	1	1
Conversion	1	1	1	1	2	1	1	1
Co-treatment	1	1	1	1	1	2	1	1
Drug Interaction	1	1	1	1	1	1	2	1
Bind	1	1	1	1	1	1	1	2

Table 9: Meta-class analysis for BioRED. The "Positive Correlation" class is mutually exclusive to the "Negative Correlation" class.

	h(Novel)	h(Not Novel)
Novel	2	0
Not Novel	0	2

Table 10: Meta-class analysis for BioRED Novel. The "Novel" class is mutually exclusive to the "Not Novel" class.

	h(Up regulator)	h(Down regulator)	h(Agonist)	h(Antagonist)	h(Substrate)
Up regulator	2	0	1	1	1
Down regulator	0	2	1	1	1
Agonist	1	1	2	0	1
Antagonist	1	1	0	2	1
Substrate	1	1	1	1	2

Table 11: Meta-class analysis for ChemProt. "Up regulator" is mutually exclusive to "down regulator" and "agonist" is mutually exclusive to "antagonist."

	h(Advise)	h(Effect)	h(Interact)	h(Mechanism)
Advise	2	1	1	1
Effect	1	2	1	1
Interact	1	1	2	1
Mechanism	1	1	1	2

Table 12: Meta-class analysis for DDI13. No classes in DDI13 are mutually exclusive based on class definitions.

	h(no relation)	h(org:alternate names)	h(org:city of branch)	h(org:country of branch)	h(org:dissolved)	h(org:founded)	h(org:founded by)	h(org:member of)	h(org:members)	h(org:number of employees/members)	h(org:political/religious affiliation)	h(org:shareholders)	h(org:state or province of branch)	h(org:top members/employees)	h(org:website)	h(per:age)	h(per:cause of death)	h(per:charges)	h(per:children)	h(per:cities of residence)	h(per:city of birth)	h(per:city of death)	h(per:countries of residence)	h(per:country of birth)	h(per:country of death)	h(per:date of birth)	h(per:date of death)	h(per:employee of)	h(per:identity)	h(per:origin)	h(per:other family)	h(per:parents)	h(per:religion)	h(per:schools attended)	h(per:siblings)	h(per:spouse)	h(per:state or province of birth)	h(per:state or province of death)	h(per:state or provinces of residence)	h(per:title)
No relation	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
org:alternate names	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:city of branch	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:country of branch	0	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:dissolved	0	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:founded	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:founded by	0	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:member of	0	1	1	1	1	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:members	0	1	1	1	1	1	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:number of employees/members	0	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:political/religious affiliation	0	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:shareholders	0	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:state or province of branch	0	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:top members/employees	0	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
org:website	0	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:age	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:cause of death	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:charges	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:children	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0	1	1	1	1	1	1
per:cities of residence	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:city of birth	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:city of death	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:countries of residence	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:country of birth	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:country of death	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:date of birth	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:date of death	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
per:employee of	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1
per:identity	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	2	1	0	0	1	1	0	0	1	1	1	1	1
per:origin	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1
per:other family	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	2	0	1	1	0	0	1	1	1	1	1
per:parents	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	2	1	1	0	0	1	1	1	1	1
per:religion	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
per:schools attended	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
per:siblings	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	2	0	1	1	1	1	1
per:spouse	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	2	1	1	1	1	1
per:state or province of birth	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
per:state or province of death	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
per:state or provinces of residence	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
per:title	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1

Table 13: Meta-class analysis for ReTACRED. Classes involving familial relations are all mutually exclusive to each other (e.g., "per:spouse," "per:parents," "per:other family," "per:siblings," "per:identity," "per:children"). Classes "org:members" and "org:member of" are mutually exclusive since each denotes an opposing directional relationship between a subject and an object.

	h(Other)	h(Component-Whole)	h(Instrument-Agency)	h(Member-Collection)	h(Cause-Effect)	h(Entity-Destination)	h(Message-Topic)	h(Entity-Origin)	h(Product-Producer)	h(Content-Container)
Other	2	1	1	1	1	1	1	1	1	1
Component-Whole	1	2	1	1	1	1	1	1	1	1
Instrument-Agency	1	1	2	1	1	1	1	1	1	1
Member-Collection	1	1	1	2	1	1	1	1	1	1
Cause-Effect	1	1	1	1	2	1	1	1	1	1
Entity-Destination	1	1	1	1	1	2	1	1	1	1
Message-Topic	1	1	1	1	1	1	2	1	1	1
Entity-Origin	1	1	1	1	1	1	1	2	1	1
Product-Producer	1	1	1	1	1	1	1	1	2	1
Content-Container	1	1	1	1	1	1	1	1	1	2

Table 14: Meta-class analysis for SemEval-2010 Task 8. No classes in SemEval-2010 Task 8 are mutually exclusive based on class definitions.

	h(Associated)	h(Not Associated)
Associated	2	0
Not Associated	0	2

Table 15: Meta-class analysis for GAD. The "Associated" class is definitionally mutually exclusive to the "Not Associated" class.