Few-Shot Document-Level Relation Extraction

Nicholas Popovic and **Michael Färber**
Karlsruhe Institute of Technology (KIT), Germany
{popovic, michael.faerber}@kit.edu

Abstract

We present FREDo, a few-shot document-level relation extraction (FSDLRE) benchmark. As opposed to existing benchmarks which are built on *sentence-level* relation extraction corpora, we argue that *document-level* corpora provide more realism, particularly regarding *none-of-the-above* (NOTA) distributions. Therefore, we propose a set of FSDLRE tasks and construct a benchmark based on two existing supervised learning data sets, DocRED and sciERC. We adapt the state-of-the-art sentence-level method MNAV to the document-level and develop it further for improved domain adaptation. We find FSDLRE to be a challenging setting with interesting new characteristics such as the ability to sample NOTA instances from the support set. The data, code, and trained models are available online[1].

1 Introduction

The goal of relation extraction is to detect and classify relations between entities in a text according to a predefined schema. The schema, defining which relation types are relevant is highly dependent on the specific application and domain. Supervised learning methods for relation extraction (Soares et al., 2019; Zhou et al., 2021; Zhang et al., 2021; Xu et al., 2021; Xiao et al., 2022), which have advanced rapidly since the introduction of pretrained language models such as BERT (Devlin et al., 2019), need large corpora of annotated relation instances to learn a schema. Since annotating data sets for relation extraction manually is expensive and time consuming, few-shot learning for relation extraction represents a promising solution for relation extraction at scale.

While the general N-way K-shot few-shot learning framework is relatively well defined and appears easy to apply to relation extraction, constructing realistic benchmark tasks has proven to be challenging. One of the core difficulties of establishing a realistic benchmark task for few-shot relation extraction is correctly modelling the most frequent situation a relation extraction will system encounter, *none-of-the-above* (NOTA) detection. NOTA refers to the case in which a candidate pair of entities does not hold any of the relations defined in the schema, a situation which is far more common than its reverse (for the document-level data set DocRED (Yao et al., 2019), 96.84% of candidate entity pairs are NOTA cases). While initial benchmarks (Han et al., 2018) ignored this scenario altogether, researchers working on few-shot relation extraction have pushed for more realistic NOTA modeling in tasks and developed methods that can better detect NOTA instances (Gao et al., 2019; Sabo et al., 2021).

Parallel to the outlined efforts towards realistic few-shot relation extraction benchmarks, research into *supervised* relation extraction has moved from *sentence-level* tasks, relation extraction within single sentences, to *document-level* relation extraction. The push towards document-level relation extraction is motivated by (1) extracting more complex, cross-sentence relations and (2) information extraction at scale. The latter is driven by an inherent challenge when increasing the scope from single sentences to multiple sentences: The number of entities involved increases and with that comes a quadratic increase in candidate entity pairs. While sentence-level approaches typically evaluate each candidate entity pair individually, this strategy is infeasible at the document-level (DocRED contains an average of 393.5 candidate entity pairs per document, compared to only 2 for many sentence level-tasks). In addition to the increased computation requirements, this results in a drastic increase in the amount of NOTA examples in a given query, demanding new methods of handling the imbalances that come with this change of distribution (Han and Wang, 2020; Zhou et al., 2021).

All current few-shot relation extraction bench-

[1] https://github.com/nicpopovic/FREDo

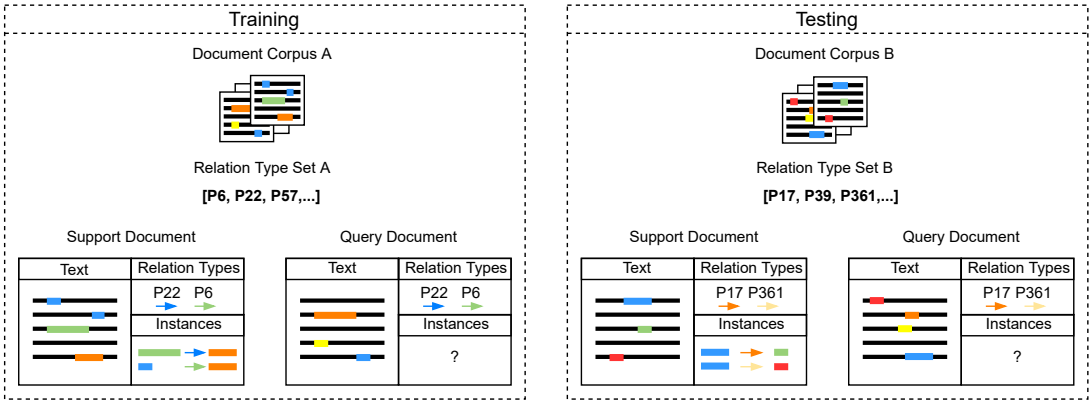

Figure 1: Illustration of the Few-Shot Document-Level Relation Extraction setting. Given a support document with annotated relation instances, the task is to return all instances of the same relation types for the query document. During testing a different corpus of documents, as well as a different set of relation types are used than during training.

marks are based on sentence-level tasks. We argue that moving few-shot relation extraction from the sentence-level to the document-level: (1) brings with it as an inherent characteristic the more realistic NOTA distribution which prior work has looked to emulate and (2) will make the resulting methods more suitable for large scale information extraction.

In this work, we therefore define a new set of few-shot learning tasks for document-level relation extraction and design a strategy for creating realistic benchmarks from annotated document corpora. Applying the above to the data sets DocRED (Yao et al., 2019) and sciERC (Luan et al., 2018), we construct a few-shot document-level relation extraction (FSDLRE) benchmark, FREDo, consisting of 2 main tasks, an in-domain and a cross-domain task requiring domain adaptation. Finally, building on the state-of-the-art few-shot relation extraction approach MNAV (Sabo et al., 2021) and document-level relation extraction concepts (Zhou et al., 2021), we develop 2 approaches for tackling the above tasks.

We begin by outlining key related work in section 2. In section 3 we give a description of the proposed tasks. Next, in section 4 we explain the construction of the benchmark, FREDo, followed by an overview of the proposed methods (section 5), an analysis and discussion of the observed results (section 6), and the conclusion (section 7).

2 Related Work

To the best of our knowledge, all current few-shot relation extraction benchmarks (Han et al., 2018; Gao et al., 2019; Sabo et al., 2021) focus on the extracting relations from single sentences. FewRel (Han et al., 2018) introduces a relation extraction benchmark in the N-way K-shot setting (Vinyals et al., 2016; Snell et al., 2017) in which a relation instance is to be assigned to one of N classes given only K examples for each of the classes. In this setting human performance was quickly surpassed (Soares et al., 2019), leading Gao et al. (Gao et al., 2019) to create FewRel 2.0 in an effort to increase the difficulty by adding a domain adaptation task, as well as a NOTA detection task. Sabo et al. (Sabo et al., 2021) argue that the way FewRel 2.0 models NOTA cases is not realistic due to the way NOTA instances are sampled, develop a framework for creating more realistic benchmarks and propose building such a benchmark using the sentence-level data set TACRED (Zhang et al., 2017). Tran et al. (Tran et al., 2021) forego labeled training data altogether and focus on the one-shot and weakly-supervised classification setting without NOTA cases.

While multiple relation extraction data sets based on annotated documents, rather than single sentences, are available in the form of CDR (Li et al., 2016), sciERC (Luan et al., 2018), SciREX (Jain et al., 2020), DialogRE (Yu et al., 2020), and GDA (Wu et al., 2019), the introduction of the large scale data set DocRED (Yao et al., 2019) seems to have significantly increased research interest into supervised relation extraction at the document-level more recently (Zhou et al., 2021; Zhang et al., 2021; Xu et al., 2021; Xiao et al., 2022).

Since documents contain considerably more entities than individual sentences and the amount of candidate entity pairs increases quadratically

with the amount of entities, applying sentence-level methods to document-level tasks is not feasible. Document-level relation extraction approaches therefore use a different architecture (Wang et al., 2019) than sentence-level approaches. Another challenge is the large imbalance in the amount of positive and negative examples of relations encountered during training. Some researchers approach the problem by resampling training examples to counteract the imbalance (Han and Wang, 2020), while others use more specialized solutions, such as modified loss functions (Zhou et al., 2021).

3 Task Description

In document-level relation extraction the task is to return a set S containing all valid triples of the format (e_h, r_i, e_t) for a document D. Here, e_h and e_t are the head- and tail-entity of a relation instance, respectively, and $r_i \in R$ is a relation type, with R being the set of relation types for which instances are to be extracted. The positions of any entity mentions, as well as any co-reference clusters are provided as part of the input[2]. In both supervised learning and few-shot learning the documents used at test time are sampled from a different corpus than those used at training time. The added complexity in few-shot learning is caused (1) by a change in the set of relation types R between training and test time, and (2) by a much smaller amount of annotated examples given for each relation type.

3.1 Document-Level Few-Shot Relation Extraction

In figure 1 we give an illustration of the proposed task setting. We define as *few-shot* document-level relation extraction (FSDLRE) the following: Given a set of *support documents* $\{D_{X,1}, ..., D_{X,k}\}$, the corresponding sets containing *all* valid triples $\{S_{X,1}, ..., S_{X,k}\}$, and a *query document* D_Q, the task is to return the set S_Q, containing all valid triples in the query document. The sets $\{S_{X,1}, ..., S_{X,k}\}$ and S_Q consist of triples for the relation types R_X. During training, annotations are based on a set R_{train}, while during testing they are based on a set R_{test}. The two sets are disjoint. The annotations of the support documents are complete, meaning that any candidate entity pair for which no relation type has been assigned can be considered NOTA.

[2]The setting in which no such annotations are given is typically referred to as *joint entity and relation extraction* and is out of scope of this paper.

3.1.1 In-Domain vs. Cross-Domain

For *in-domain* FSDLRE training and test documents are taken from the same domain. For *cross-domain* FSDLRE the test documents are taken from a different domain. Consequently, text style, text content, entity types, and relation types will all differ from those seen in the training documents. While this increases the difficulty of the challenge, this also resembles a more realistic application of few-shot relation extraction methods: A key motivation for few-shot learning is to develop methods which can be applied to new data without the need for large-scale manual annotation. Restricting the applicability of a method to a specific domain and annotation procedure does not fit this idea.

3.2 Differences to Existing Benchmarks

The tasks described above differ from existing few-shot relation extracting benchmarks in multiple ways. (1) Operating at document-level means the data now includes instances of relations expressed across multiple sentences and that models need to be able to evaluate candidate entity pairs more efficiently. (2) Like for FS-TACRED (Sabo et al., 2021), the amount of candidate entity pairs for which no relation is to be extracted is significantly larger than in other benchmarks (96.4% compared to 15%/50% for FewRel 2.0 (Gao et al., 2019)) and the distribution from which NOTA instances are sampled, is more realistic than in FewRel 2.0, where NOTA instances are always instances of other, valid relation types. (3) By requiring support annotations to be complete we have access to a support NOTA distribution, which is not the case for any existing benchmarks. (4) Our tasks do not follow the fixed N-way K-shot format that related work has followed. Instead, N and K are variable between documents and therefore between individual episodes.

4 FREDo: Few-Shot Document-Level Relation Extraction Benchmark

4.1 Selected Data Sets

In order to construct a benchmark based on the tasks described in section 3 we require fully annotated data sets from 2 seperate domains. For the training set and the in-domain test set we use DocRED (Yao et al., 2019) due to it being, to the best of our knowledge, the largest and most widely used document-level relation extraction data set. For the cross-domain test set we use sciERC (Luan et al.,

Data Set	# Docs	# RT	# CP/Doc	# Words/Doc	# Sents/Doc	Domain
DocRED	4051	96	394	172	8	Non-specific
sciERC	500	7	187	118	5.4	Scientific Publications

Table 1: A comparison of DocRED (Yao et al., 2019) and sciERC (Luan et al., 2018), the datasets selected for the FREDo benchmark.

2018) due to its domain (abstracts of scientific publications), which differs from DocRED (Wikipedia abstracts), and the fact that it contains annotations for 7 relation types. In table 1 we show a comparison of the selected datasets. Additional document-level relation extraction data sets, SciREX (Jain et al., 2020), DialogRE (Yu et al., 2020), GDA (Wu et al., 2019), CDR (Li et al., 2016), were considered but ultimately not used for the cross-domain set, due to the amount of relation types annotated (too few), missing coreference links, or differing relation format (SciREX annotations are based on N-ary relations, while the other datasets annotate only binary relations).

4.2 Training and Test Data

4.2.1 Document Corpora

We begin by building 3 separate corpora of documents, 1 for training and development and 1 for testing each task (in-/cross-domain). Since the annotated test corpus for DocRED is not publicly available we use the documents in the development corpus as the test corpus for our in-domain task (meta-test). The DocRED training corpus is therefore used as the basis for both our training, and development set (meta-train). For the cross-domain task we require only a test set. This is because the training and development set for this task are identical to that of the in-domain task. We therefore use all documents in sciERC as our cross-domain test set (meta-test).

4.2.2 Assigning Relation Types

For preprocessing, we begin by comparing the relation types annotated in sciERC to those in the DocRED corpus[3]. We find 2 relation types (P279, P361), which are annotated in both DocRED and sciERC and remove these from the DocRED corpus in order to prevent data leakage between train and test sets.

For DocRED, we split the remaining 94 relation types into 4 disjoint sets, a training set (62) R_{train}, development set (16) R_{dev}, and in-domain test set (16) R_{test}. For the cross-domain test set we use all 7 relation types in the sciERC corpus. An overview of the relation types assigned to each set can be found in appendix A and B.

4.3 Test Episode Sampling

In few-shot learning, a each training/testing step consisting of support documents and query documents is called an episode. Since evaluating every possible combination of support and query documents would result in too many episodes (approx. 1 million episodes for the in-domain and 250k episodes for the cross-domain test set) we need to sample a smaller amount of episodes from our corpora. We chose our sampling procedure with the goal of producing a representative measurement of the macro F_1 score.

For few-shot learning tasks the episode sampling process can be split into 2 steps, the first step being the sampling of the support examples and the second step being the sampling of the query examples. Unlike the sentence-level scenario where each example contains exactly one relation instance, each document we sample contains multiple instances of different relation types. In order to balance the amount of times each relation type is seen as a support example during testing we use the following procedures for the first sampling step: We begin by selecting from the set R_T the relation type r_s which is currently least represented in the testing corpus. If there are multiple such relation types we randomly choose one. For this relation type we sample support documents which contain at least one instance of r_s each. Since the selected support documents may contain instances of other relation types in R_{test} we add all of the relation types contained in the support document[4] to the episodes annotation schema. Following Sabo et al. (Sabo et al., 2021), we randomly sample query documents from the training corpus[5] to realistically represent the NOTA distribution of the entire corpus.

[3]The mapping can be found in appendix A.

[4]The first, if there are multiple support documents.

[5]Note that we exclude the previously sampled support documents.

Task	N	K (micro)	K (macro)
in-domain			
1-Doc	2.18	2.36	2.24
3-Doc	3.47	4.30	4.31
cross-domain			
1-Doc	4.26	2.73	2.40
3-Doc	6.08	5.55	5.27

Table 2: Average values for N and K across test episodes. K (micro) denotes the average across all episodes, K (macro) denotes the weighted average of mean K for each relation type.

4.3.1 Choosing Test Set Sizes

In order to chose a sufficiently large amount of test episodes for a representative F_1 score we evaluate a trained model for 50k episodes, logging the macro f_1 score at intervals of 100 episodes. We repeat this for 5 different random seeds. Using the variance between the 5 measurements as a guide, we choose a number of episodes which we deem to satisfy a good balance between low variance and manageable test set size. For robustness we sample episodes with 3 different random seeds for the final test sets. The resulting test set sizes are: 15k episodes for the in-domain tasks and 3k episodes for the cross-domain tasks.

4.4 Characteristics of Resulting Tasks

Existing few-shot benchmarks typically set 2 tasks, a single-shot and a K-shot (3/5/10-shot) challenge, in order to determine the way performance may scale when adding annotated training data. Due to the nature of our tasks, N and K are variable from episode to episode, depending on the specific support documents and relation types. We measure the scalability of approaches by defining a 1-Doc and 3-Doc challenge.

Therefore, the proposed benchmark, FREDo, consists of 2 main tasks with a 1-Doc and a 3-Doc subtask each:

- The in-domain tasks for which an approach which has been trained on documents sampled from DocRED is evaluated on 15k episodes generated using documents from DocRED.

- The cross-domain tasks for which an approach which has been trained on documents sampled from DocRED is evaluated on 3k episodes generated using documents from sciERC.

In order to better characterize our tasks in relation to the common N-way K-shot format we measure the distribution of N and K across our test sets. All the average values for K and N are shown in table 2. We find that the mean values of N are (2.18/3.47) for the in-domain tasks (1-/3-doc) and (4.26/6.08) for the cross-domain task. For K we calculate both the mean values across all episodes (micro), as well as the mean across the different relation types (macro).

5 Experiments

5.1 Models

A common approach to relation extraction in general is to compute the similarity between embeddings produced by a fine-tuned language model such as BERT (Devlin et al., 2019). In order to produce a relation embedding for a given pair of entities, most approaches concatenate embeddings corresponding to each entity. One way to generate an entity embedding from the output of a language model is to average the embeddings of all tokens belonging to an entity. Another way is the use of so called entity markers, introduced by Soares et al. (Soares et al., 2019), which are tokens placed at the beginning and end of an entity mention within the input text. The embeddings of the tokens placed at the start of each entity mention are then used as the entity embeddings. In few-shot learning, a common way to use embedding similarity are prototypical networks (Snell et al., 2017). Here, the embeddings of all K support examples are averaged into a so-called prototype. Given a query embedding, the similarity to the N class-prototypes is then used for classification.

In order to assess the difficulty of our challenges we measure the performance of 3 approaches. We set an initial baseline using the pretrained language model $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) without fine-tuning. Next, we adapt the state-of-the-art sentence-level few-shot relation extraction method MNAV (Sabo et al., 2021) to the document-level (DL-MNAV). Finally, we make 2 modifications to DL-MNAV (SIE and SBN) to improve cross-domain performance. In figure 2 we show a comparison of how the different models handle decision boundaries with respect to support and query instances.

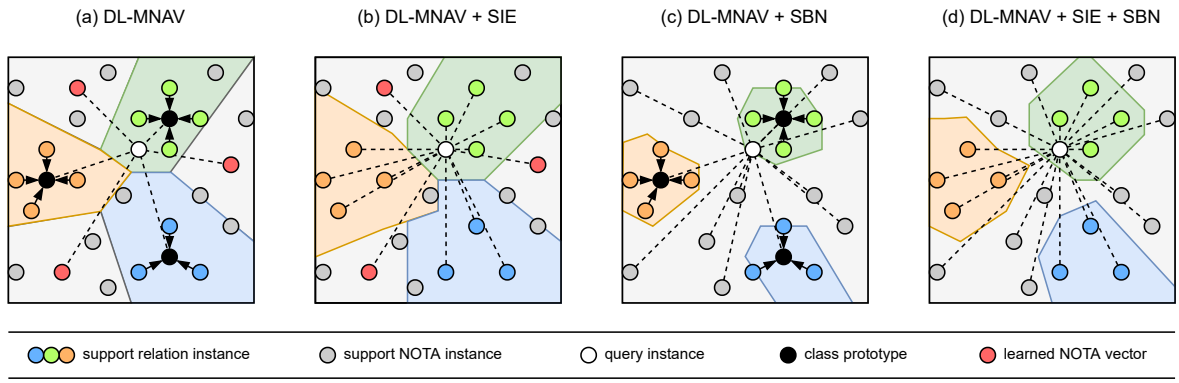

Figure 2: Illustration of the different models used at inference time. Relation prototypes are computed as mean of support relation instances. Learned NOTA vectors are not based on support NOTA instances but learned during training. Dotted lines indicate distances affecting the classification of the query instance. Background colors illustrate approximate classification boundaries. For the baseline model, decision boundaries are the same as in (d).

5.1.1 Baseline

We set an initial baseline using the pretrained language model $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) without fine-tuning in the following way: We encode each document using the language model and then average the output tokens of each entity mention. Following Han and Wang (Han and Wang, 2020), we then average the mention representations for each entity. The resulting entity mentions are then concatenated for each candidate pair of entities to form relation embeddings. The similarity between a relation embedding in a query document to a relation embedding in a support document is calculated via their dot product. The relation type of the support embedding producing the highest dot product is output as the predicted relation type.

5.1.2 Adapting MNAV to Document-Level (DLMNAV)

The current state-of-the-art few-shot sentence-level relation extraction method MNAV (Sabo et al., 2021) uses a combination of entity embeddings based on entity markers and prototypical networks. Furthermore, it introduces the idea of learning M prototypes to represent the NOTA class. In order to use MNAV at the document-level one key architectural change is required: Instead of only marking two entities, a head- and a tail-entity, with two different tokens we mark all spans with the same entity marker tokens. Furthermore, following other document-level approaches (Han and Wang, 2020; Zhou et al., 2021) we apply a pooling step[6] to create representations for entities which are mentioned more than once in a document.

One of the challenges in realistic relation extraction is the large imbalance in the amount of positive and negative examples of relations encountered during training. In document-level relation extraction this challenge is even more central to the task than in sentence-level relation extraction. Preliminary experiments showed that simply using cross-entropy loss, as is done for MNAV, yields sub-par results. To tackle this[7], we adopt the adaptive thresholding loss function used by Zhou et al. (Zhou et al., 2021) which is an adaptation of categorical cross entropy loss designed specifically for classifiers which treat NOTA as a relation type during classification, as is the case for MNAV.

Finally, we modify the initialization procedure of the NOTA vectors. While Sabo et al. (2021) initialize the vectors using an averaged value of relation representations, we sample NOTA representations from the support documents during the first training step.

5.1.3 Support Instance Evaluation (SIE)

MNAV being based on prototypical networks (Snell et al., 2017) means that the embeddings of all support instances of a relation type are averaged into a single prototype. While this has proven to be an effective strategy, we argue that it may not be ideal during inference in a cross-domain setting where the change in data distribution may break the assumption that the mean of support instances

[6]For the pooling step we use mean pooling as, during preliminary experiments, it performed better than the logsumexp pooling used by Zhou et al..

[7]We also examined the option of resampling the training samples such that the amount of negative examples in each training step roughly matches the amount of positive examples (Han and Wang, 2020), but found this to be less effective.

Model	1-Doc Precision [%]	1-Doc Recall [%]	1-Doc F_1 [%]	3-Doc Precision [%]	3-Doc Recall [%]	3-Doc F_1 [%]
Baseline	0.36	9.69	0.60	0.60	10.75	0.89
DL-MNAV	6.26 ± 0.22	21.08 ± 2.71	7.05 ± 0.18	7.71 ± 0.69	22.80 ± 3.82	$\mathbf{8.42 \pm 0.64}$
DL-MNAV$_{SIE}$	5.57 ± 0.04	23.12 ± 1.69	$\mathbf{7.06 \pm 0.15}$	5.16 ± 0.19	33.61 ± 2.68	6.77 ± 0.21
DL-MNAV$_{SIE+SBN}$	1.02 ± 0.05	22.94 ± 1.87	1.71 ± 0.04	1.75 ± 0.16	23.41 ± 0.76	2.79 ± 0.24

Table 3: Results for FREDo in-domain task. Reported results are macro averages across relation types.

Model	1-Doc Precision [%]	1-Doc Recall [%]	1-Doc F_1 [%]	3-Doc Precision [%]	3-Doc Recall [%]	3-Doc F_1 [%]
Baseline	1.34	3.04	1.76	1.84	2.47	1.98
DL-MNAV	2.30 ± 0.45	0.58 ± 0.12	0.84 ± 0.16	3.02 ± 2.38	0.29 ± 0.13	0.48 ± 0.21
DL-MNAV$_{SIE}$	1.77 ± 0.60	2.08 ± 0.34	1.77 ± 0.60	2.51 ± 0.66	2.52 ± 0.31	2.51 ± 0.66
DL-MNAV$_{SIE+SBN}$	2.26 ± 0.11	4.37 ± 0.13	$\mathbf{2.85 \pm 0.12}$	3.47 ± 0.14	4.24 ± 0.21	$\mathbf{3.72 \pm 0.14}$

Table 4: Results for FREDo cross-domain task. Reported results are macro averages across relation types.

provides a good prototype. In SIE we therefore use all individual support instances during inference, instead of their average.

5.1.4 Support Based NOTA Vectors (SBN)

In treating NOTA as a relation type and learning a persistent set of vectors during training, MNAV works on the assumption that the NOTA distribution during testing will match that seen during training. While this assumption is warranted and seems to work well for in-domain few-shot learning, we argue that this may not be the case for cross-domain settings. For this reason, we additionally add NOTA representations from the support documents to our set of NOTA vectors during training and inference.[8] Instead of randomly sampling NOTA vectors from the support documents we sample the most similar $k = 5$ NOTA instances measured via their dot product for each relation prototype[9]. During inference in a new domain, we use only the NOTA vectors sampled from the support document and ignore the learned vectors.

5.2 Sampling Training & Development Episodes

We compare 2 different ways of sampling episodes during training. First we sample training and development episodes in the same way as for the test sets. In order to get sufficient coverage to calculate representative macro F_1 scores on the development set, we sample 4k episodes. As an alternative we modify the query sampling by ensuring that for each episode at least one of the query documents contains an instance of the relation type r_s. This way we increase the amount of non-NOTA examples the model sees during training. Another effect is that we need fewer development episodes (we use 500) to calculate macro F_1 scores.

[8]The sampled NOTA representations do not persist across episodes.

[9]Since, with SIE, we do not use prototypes at inference time, we then perform this sampling step for each relation instance rather than for prototypes and increase k to 20.

6 Analysis of Results

6.1 Experimental Setup

All our models are based on BERT$_{\text{BASE}}$ (Devlin et al., 2019) implemented using Huggingface's Transformers (Wolf et al., 2020) and trained using mixed precision. We follow Zhou et al. (Zhou et al., 2021) in using AdamW (Loshchilov and Hutter, 2019) as optimizer (learning rates $\in [1\text{e}{-5}, 3\text{e}{-5}, 5\text{e}{-5}, 1\text{e}{-4}]$, of which $1\text{e}{-5}$ generally performs best) and training using linear warmup (1k/2k steps) (Goyal et al., 2017) followed by a linear learning rate decay. We use gradient clipping of 1.0. We train each model for 50k episodes and perform early stopping based on the macro F_1 score on the development set which we measure every 1k/2k steps (when random sampling/ensuring positive examples). Each 1-doc training episode consists of 1 support document and 3 query documents, 3-doc training episodes contain 3 support documents and 1 query document. We run each model 5 times using different random seeds and select the learning rate with the highest mean macro F_1 score on the development set for testing. For test scores we report the mean and standard deviation of macro F_1 scores for models trained using 5 different random seeds. For this model we report the macro F_1 score on the test set.

Model	macro F_1 [%] 1-Doc	3-Doc
Random Sampling	5.77	5.29
Ensure Positive	**7.26**	**9.37**

Table 5: Results for different query sampling strategies on the in-domain task

Results are shown in tables 5, 3, and 4. All models were trained on either NVIDIA V100 or NVIDIA 3090 GPUs.

6.2 Baseline Results

As expected for a baseline which is not fine-tuned to the task at hand, the resulting macro F_1 scores are very low. We argue, however, that the baseline is nevertheless relevant for 2 reasons. For the in-domain challenge, the baseline proves that the tasks are not trivially solvable by using a pre-trained language model out of the box. For the cross-domain challenge, our baseline lets us see whether models overfit on the training domain.

6.3 Comparing Sampling Strategies

In table 5 we compare the test macro F_1 scores of the best models (chosen according to development set) trained using the 2 sampling strategies described in section 5.2 on the 1-Doc challenge using the model DL-MNAV. We find that ensuring positive query documents during training and development helps increase the F_1 scores. Due to the better performance, as well as the reduced computation time needed for validation (thanks to the smaller development set), we use the latter strategy in all following experiments.

6.4 In-Domain Setting

Test scores for the in-domain challenge are shown in table 3. We observe large improvements in F_1 scores over the baseline, especially for DL-MNAV which reaches 7.05% in the 1-Doc and 8.42% in the 3-Doc task. SIE, does not seem to affect the accuracy of the model in the 1-Doc task, in the 3-Doc task, however, the F_1 score drops by 1.65 percentage points. SBN, on the other hand causes the F_1 scores to drop by 5.34% and 5.63%. This clearly illustrates the effectiveness of learned NOTA vectors for in-domain tasks. In table 6 we compare the best F_1 scores of different few-shot relation extraction benchmarks. Overall, compared to scores for benchmarks such as FewRel (Han et al., 2018) FewRel 2.0 (Gao et al., 2019), the F_1 scores are considerably lower, illustrating the difficulty of such a realistic challenge. When compared to the more realistic sentence-level benchmark FS-TACRED (Sabo et al., 2021) for which Sabo et al. report F_1[10] scores of 12.39% (1-shot) and 30.04% (5-shot) MNAV, these results are in-line with our expectations for an even more realistic (and thereby evidently more difficult) challenge. Notably, the scaling behavior seen in FS-TACRED between the 1-shot and the 5-shot setting is not as pronounced for FREDo. We hypothesize that this is due the fact that the change in K is not as large (see table 2), meaning that (1) our 1-Doc setting does not correlate perfectly to the 1-shot setting, and (2) due to the way that additional support documents are sampled, the 3-Doc setting does not *guarantee* additional support examples for infrequently occuring relation types.

Benchmark	input length	realistic NOTA	best F_1 [%]
FewRel	sentences	✗	97.85
FewRel 2.0	sentences	✗	89.81
FS-TACRED	sentences	✓	12.39
FREDo (ours)	documents	✓	7.06

Table 6: A comparison highlighting the levels of difficulty of different few-shot relation extraction benchmarks. For all sentence-level benchmarks, we report the highest F_1 scores (at the time of writing) in the 5-way 1-shot setting. For FREDo we report the 1-Doc setting. For FewRel and FewRel 2.0, we report the highest scores found at the respective CodaLab competition websites.

6.5 Cross-Domain Setting

Test scores for the cross-domain challenge are shown in table 4. For DL-MNAV we see a significant drop in F_1 scores over the baseline, illustrating the problem with learned NOTA vectors in a cross-domain setting. SIE brings the scores back to the baseline level, illustrating that the distribution of support examples is no longer well represented by their mean values. Switching to SBN (DL-MNAV$_{SIE+SBN}$), we find that our model exceeds the baseline scores, suggesting that the NOTA distribution on sciERC seems to be sufficiently different to cause an overfitting effect for learned NOTA vectors. While SBN improves the results over the naive baseline, even the improved F_1 scores are extremely low. This is, however, un-

[10]The reported results are micro F_1 scores

surprising given the increase in difficulty over the previous setting.

6.6 Scalability of DL-MNAV

Although our methods show improvements over the proposed baseline in both tasks the results are currently severely lacking, especially compared to the state-of-the-art *supervised* learning approaches on both data sets (65.92% for DocRED (Xu et al., 2021) and 52.0% for sciERC (Ye et al., 2022)). This performance gap raises the question of whether our models will achieve similar performance if given enough support documents. In order to assess the scalability of DL-MNAV when given amounts of annotated data resembling the supervised setting, we initialize a trained model with the full DocRED training corpus as support documents (96 classes, 3053 documents) and evaluate the performance on the full development set (96 classes, 998 documents). We measure an increase in recall to 45.75% combined with a drop in precision to 5.75%, resulting in a F_1 score of 8.86%. While a direct comparison of this score with the few-shot settings assessed in FREDo is not appropriate, due to the nature of how the task is posed (different relation types are examined), the score can be compared to results obtained from supervised learning. Here we see clearly that simply averaging support instance embeddings to prototypes does not scale well to the supervised setting. We do not experiment with SIE or SIE+SBN, as the amount of support instances would result in prohibitively large model sizes.

6.7 Limitations

Regarding the limitations of the proposed benchmark, FREDo, we believe that while it represents a good basis for model development, it will be beneficial to add other cross-domain data sets from a greater variety of domains in the future. With the current, low F_1 scores seen in our tests, overestimating the performance of approaches does not seem to be too critical a danger. We are, however, hopeful that new methods might achieve significantly better results. At that point we suggest a reassessment of how representative this benchmark is of cross-domain performance in general. For the time being, however, we are confident that our tasks represent a valuable contribution to advancing the field.

7 Conclusion

In order to encourage the development of few-shot relation extraction approaches which are useful in real world scenarios, we propose FREDo, a few-shot document-level relation extraction benchmark. By moving to the document-level, the settings become more realistic, a problem which existing benchmarks are struggling with. For both in-domain and cross-domain tasks we present an approach which performs better than a simple baseline. Our experiments confirm that, even though some existing benchmarks imply that impressive, even superhuman performance can already be achieved in few-shot relation extraction, realistic tasks are very difficult using current approaches and that significant advances are required for few-shot relation extraction approaches to be usable in a real word scenario. In providing a benchmark that reveals this performance gap, we look to pave the way towards new methods with a potentially high impact on domain-specific and cross-domain relation extraction at scale.

Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

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.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. pages 6251–6256.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour.

X. Han and L. Wang. 2020. A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information. *IEEE Access*, 8:96912–96919. Conference Name: IEEE Access.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A

Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809. Association for Computational Linguistics.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the 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*, 2016:baw068.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations 2019*, page 18.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting Few-shot Relation Classification: Evaluation Data and Classification Schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2021. One-shot to Weakly-Supervised Relation Classification using Language Models. In *Automated Knowledge Base Construction (2021)*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1371–1377, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. RENET: A Deep Learning Approach for Extracting Gene-Disease Associations from Literature. In *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 272–284, Cham. Springer International Publishing.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction. In *NAACL*.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14149–14157. AAAI Press.

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.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Pack Together: Entity and Relation Extraction with Levitated Marker. In *Proceedings of ACL 2022*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and

sciERC ID	Wikidata ID	DocRED
hyponym-of	P279	✓
part-of	P361	✓
used-for	P366	✗
compare	P2210	✗
evaluate-for	P5133	✗
feature-of	-	-
conjunction	-	-

Table 7: Overlap of relation types in sciERC and DocRED

Huajun Chen. 2021. Document-level Relation Extraction as Semantic Segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Overlap of DocRED and sciERC

In table 7 we show the mapping of sciERC relation types onto Wikidata (Vrandečić and Krötzsch, 2014) properties and whether these relation types are contained in DocRED.

B Relation Types in in-domain dataset

In tables 8, 9, 10, 11, we list the relation types in the different datasets based on DocRED.

Wikidata ID	Description	Number of instances
P131	located in the administrative territorial entity	4193
P577	publication date	1142
P175	performer	1052
P569	date of birth	1044
P570	date of death	805
P161	cast member	621
P264	record label	583
P527	has part	632
P19	place of birth	511
P54	member of sports team	379
P40	child	360
P30	continent	356
P69	educated at	316
P26	spouse	303
P607	conflict	275
P159	headquarters location	264
P22	father	273
P400	platform	304
P1344	participant of	223
P206	located in or next to body of water	194
P127	owned by	208
P170	creator	231
P178	developer	238
P20	place of death	203
P1412	languages spoken, written or signed	155
P155	follows	188
P710	participant	191
P6	head of government	210
P108	employer	196
P276	location	172
P156	followed by	192
P166	award received	173
P123	publisher	172
P800	notable work	150
P449	original network	152
P58	screenwriter	156
P706	located on terrain feature	137
P162	producer	119
P37	official language	119
P241	military branch	108
P31	instance of	103
P403	mouth of the watercourse	95
P580	start time	110
P585	point in time	96
P749	parent organization	92
P937	work location	104
P36	capital	85
P576	dissolved, abolished or demolished	79
P172	ethnic group	79
P205	basin country	85
P1376	capital of	76

Table 8: Relation types present in training data (continued on next page).

Wikidata ID	Description	Number of instances
P171	parent taxon	75
P740	location of formation	62
P840	narrative location	48
P676	lyrics by	36
P1336	territory claimed by	33
P551	residence	35
P1365	replaces	18
P737	influenced by	9
P190	sister city	4
P807	separated from	2
P1198	unemployment rate	2

Table 9: Relation types present in training data (continued).

Wikidata ID	Description	Number of instances
P27	country of citizenship	2689
P150	contains administrative territorial entity	2004
P571	inception	475
P50	author	320
P1441	present in work	299
P57	director	246
P179	series	144
P137	operator	95
P112	founded by	100
P86	composer	79
P176	manufacturer	83
P355	subsidiary	92
P136	genre	111
P488	chairperson	63
P1366	replaced by	36
P1056	product or material produced	36

Table 10: Relation types present in development/validation data.

Wikidata ID	Description	Number of instances
P17	country	2831
P361	part of	194
P495	country of origin	212
P102	member of political party	98
P463	member of	113
P3373	sibling	134
P1001	applies to jurisdiction	83
P118	league	56
P674	characters	74
P194	legislative body	56
P140	religion	82
P35	head of state	51
P364	original language of work	30
P272	production company	36
P279	subclass of	36
P25	mother	15
P582	end time	23
P39	position held	8

Table 11: Relation types present in test data.