DARE: Data Augmented Relation Extraction with GPT-2

Yannis Papanikolaou and Andrea Pierleoni
Healx, Cambridge, UK
{yannis.papanikolaou, andrea.pierleoni}@healx.io

Abstract

Real-world Relation Extraction (RE) tasks are challenging to deal with, either due to limited training data or class imbalance issues. In this work, we present *Data Augmented Relation Extraction* (DARE), a simple method to augment training data by properly fine-tuning GPT-2 to generate examples for specific relation types. The generated training data is then used in combination with the gold dataset to train a BERT-based RE classifier. In a series of experiments we show the advantages of our method, which leads in improvements of up to 11 F1 score points against a strong baseline. Also, DARE achieves new state of the art in three widely used biomedical RE datasets surpassing the previous best results by 4.7 F1 points on average.

1 Introduction

Relation Extraction (RE) is the task of identifying semantic relations from text, for given entity mentions within it. This task, along with Named Entity Recognition, has recently become increasingly important due to the advent of knowledge graphs and their applications. In this work, we focus on supervised RE (Zeng et al., 2014; Lin et al., 2016; Wu et al., 2017; Verga et al., 2018), where relation types come from a set of predefined categories, as opposed to Open Information Extraction approaches that represent relations among entities using their surface forms (Banko et al., 2007; Fader et al., 2011).

RE is inherently linked to Natural Language Understanding in the sense that a successful RE model should manage to adequately capture language structure and meaning. So, almost inevitably, the latest advances in language modelling with Transformer-based architectures (Radford et al., 2018a; Devlin et al., 2018; Radford et al., 2018b) have been quickly employed to also deal with RE tasks (Soares et al., 2019; Lin et al., 2019; Shi and Lin, 2019; Papanikolaou et al., 2019).

These recent works have mainly leveraged the discriminative power of BERT-based models to improve upon the state of the art (SOTA). In this work we take a step further and try to assess whether the text generating capabilities of another language model, GPT-2 (Radford et al., 2018b), can be applied to augment training data and successfully deal with class imbalance and small-sized training sets.

Specifically, given a RE task we fine-tune one pretrained GPT-2 model per relation type and then use the resulting fine-tuned models to generate new training samples. We then combine the generated data with the gold dataset and fine-tune a pretrained BERT model (Devlin et al., 2018) on the resulting dataset to perform RE.

We conduct extensive experiments, studying different configurations for our approach and compare DARE against two strong baselines and the SOTA on three well established biomedical RE benchmark datasets. The results show that our approach yields significant improvements against the rest of the approaches. To the best of our knowledge, this is the first work augmenting training data with GPT-2 for RE. In Table 1 we show some generated examples with GPT-2 models fine-tuned on the datasets that are used in the experiments (refer to Section 4).

In the following, we provide a brief overview of related works in Section 2, we then describe our approach in Section 3, followed by our experimental results (Section 4) and the conclusions (Section 5).

2 Related Work

Relation Extraction is usually modelled as a text classification task. Therefore most methods to deal with class imbalance or limited data in RE follow the respective methods from text classification. In

Dataset(relation type)	Generated sentences
CDR(Induce)	DISEASE was the most common adverse reaction (21 %) reported for DRUG, and occurred in approximately 50 % of patients .
DDI2013(Effect)	DRUGA may enhance the effects of alcohol, barbiturates, DRUGB, and other cns depressants.
DDI2013(Advise)	caution should be observed when DRUGA and DRUGB are coadministered.
DDI2013(Mechanism)	co-administration of DRUGA decreased the oral bioavailability (48%) of DRUGB, a substrate for cyp2d6.
ChemProt(Activate)	DRUG enhances PROTEIN sensitivity via activation of the pi3k / akt signaling pathway.
ChemProt(Inhibit)	DRUG, a novel orally bioavailable xanthine PROTEIN inhibitor,
ChemProt(Product)	the enzyme PROTEIN catalyzes the two-electron reduction of DRUG to produce acetyl groups.

Table 1: Examples of generated sentences with fine-tuned GPT-2 models. Each model is fine-tuned on examples from the specific relation type.

the following, we describe the different approaches that have been followed in the literature.

One approach is to deal with imbalance at the classifier level, by penalizing misclassification errors differently for each class, depending on the class frequency (Lewis et al., 2004; Zhou and Liu, 2005) or by explicitly adjusting prior class probabilities (Lawrence et al., 1998).

Another popular approach relies on either undersampling the majority class(es) or oversampling the minority one(s), transforming the training data with the aim of balancing it. One of the simplest approaches, random majority undersampling, simply removes a random portion of examples from majority classes so that per class training examples are roughly equal (Japkowicz and Stephen, 2002). An improved version of the previous method, balanced bagging (Hido et al., 2009), employs an ensemble of classifiers that have been trained with random majority undersampling.

Oversampling approaches for textual data have been somehow limited as opposed to those for image data (Wong et al., 2016; Fawzi et al., 2016; Wang and Perez, 2017; Frid-Adar et al., 2018), since text semantics depend inherently on the exact order or structure of word tokens.

A simple approach is to replace words or phrases with their synonyms (Zhang et al., 2015). Chen et al. (2011) employed topic models to generate additional training examples by sampling from the topic-word and document-topic distributions. Ratner et al. (2016) proposed a data augmentation framework that employs transformation operations provided by domain experts, such as a word swap, to learn a sequence generation model. Kafle et al. (2017) used both a template-based method and an LSTM-based approach to generate new samples for visual question answering.

A similar method to our approach was proposed by Sun et al. (2019a) who presented a framework to successfully deal with catastrophic forgetting in language lifelong learning (LLL). Specifically and given a set of tasks in the framework of LLL, they fine-tune GPT-2 to simultaneously learn to solve a task while generating training samples for it. When dealing with a new task, the model is trained on the generated training samples from previous tasks alongside the data of the new task, therefore avoiding catastrophic forgetting.

Our work falls into the oversampling techniques for text, but our focus is RE. Importantly, we do not need any domain expertise, templates, synonym thesaurus or to train a model from scratch, which makes our approach easily adaptable to any domain, with relatively low requirements in resources.

3 Methods

In this section we present briefly the GPT-2 model and before giving a detailed introduction to our approach.

3.1 GPT-2

GPT-2 (Radford et al., 2018b) is a successor of the GPT language model (Radford et al., 2018a). Both models are deep neural network architectures using the Transformer (Vaswani et al., 2017), pre-trained on vast amounts of textual data. Both models are pre-trained with a standard language modelling objective, which is to predict the next word token given k previously seen word tokens. This is achieved by maximizing the following likelihood:

$$L(U) = \sum_i logP(u_i|u_{i-1}, ..., u_{i-k}; \Theta) \quad (1)$$

where Θ are the neural network parameters. The authors have gradually provided publicly four different flavours of GPT-2, with 124M, 355M, 774M and 1558M parameters respectively. In our experiments we use the second largest model (774M), since it seems to represent a good compromise between accuracy and hardware requirements[1].

3.2 Data Augmented Relation Extraction

Let $D = [s_0, ...s_d]$ be a RE dataset containing d sequences. Furthermore, we assume that each sequence $s = [w_0, ...w_n]$ will be a sequence of n word tokens and that $e_1 = [w_{e1i}, ...w_{e1j}]$ and $e_2 = [w_{e2k}, ...w_{e2l}]$ will represent a pair of entity mentions in s. Furthermore, let $L = [l_1, ..., l_c]$ be a set of c relation types. Then, RE is the task of learning a function that maps each triple $(s_i, e1, e2)$ to L, i.e.,

$$h = f_{\Theta}(s_i, e1, e2), h \in L \quad (2)$$

where Θ are the parameters of the model.

In this work we employ a RE classifier based on a pretrained BERT language model. This classifier follows the same principle followed by Devlin et al. (2018), using a special token (CLS) for classification. The only modification is that we mask entity mentions with generic entity types, i.e., *$ENTITY_A$* or *$ENTITY_B$*. It should be noted that the method that we introduce here is not classifier specific, so any other classifier can be used instead.

To generate new training data, we split the D dataset into c subsets where each D_c subset contains only examples from relation type c. Subsequently, we fine-tune GPT-2 on each D_c for five epochs and then prompt each resulting fine-tuned model to generate new sentences, filtering out sentences that do not contain the special entity masks or that are too small (less than 8 tokens). The generated sequences are combined for all relation types into a dataset $Dsynth$.

Subsequently, we build an ensemble of RE classifiers, each of them being fine-tuned on a subset of $Dsynth$ and the whole D, such that the per-relation type generated instances are equal to the number of gold instances for that relation, multiplied by $ratio$, i.e., $|Dsynth'_c| = |D_c| * r$. In our experiments we have set $r = 1.0$ (refer to Section 4.6 for a short study of its influence). Algorithm **??** illustrates our method. We would like to note that in early experiments, we also experimented with fine-tuning over the whole D, by adding a special token to the beginning of each sentence that encoded the relation type, e.g., <0>: or <1>:. Then during generation, we would prompt the model with the different special tokens and let it generate a training instance from the respective relation type. However, this approach did not prove effective leading to worse results than just using gold data, primarily because frequent classes ”influenced” more GPT-2 and the model was generating many incorrectly labeled samples.

4 Experimental Evaluation

In this section we present the empirical evaluation of our method. We first describe the experimental setup, the datasets used, the baselines against which we evaluate DARE and subsequently present the experiments and report the relevant results.

4.1 Setup

In all experiments we used the second-largest GPT-2 model (774M parameters). All experiments were carried out on a machine equipped with a GPU V100-16GB. For the implementation, we have used HuggingFace’s Transformers library (Wolf et al., 2019).

To fine-tune GPT-2 we employed Adam as the optimizer, a sequence length of 128, a batch size of 4 with gradient accumulation over 2 batches (being equivalent to a batch size of 8) and a learning rate of $3e-5$. In all datasets and for all relation types we fine-tuned for 5 epochs. For generation we used a temperature of 1.0, fixed the top-k parameter to 5 and generated sequences of up to 100 word tokens. An extensive search for the above optimal hyper-parameter values is left to future work.

Since all of our datasets are from the biomedical domain, we found out empirically (see Section 4.4 for the relevant experiment) that it was beneficial to first fine-tune a GPT-2 model on 500k PubMed abstracts, followed by a second round of fine-tuning per dataset, per relation type.

In all cases, we used a pre-trained BERT model

[1] https://openai.com/blog/gpt-2-1-5b-release/

Dataset	$\|L\|$	Training	Development	Test
CDR	1	3,597(1,453)	3,876	3,806
DDI2013	4	22,501(153 658 1,083 1,353)	4,401	5,689
ChemProt	5	14,266(173 229 726 754 2,221)	8,937	12,132

Table 2: Statistics for the datasets used in the experiments. For the training data we provide in parentheses the number of positives across each class. We do not include in $|L|$ the *null* class which signifies a non-existing relation.

(the largest uncased model) as a RE classifier, which we fine-tuned on either the gold or the gold+generated datasets. We used the AdamW optimizer (Loshchilov and Hutter, 2017), a sequence length of 128, a batch size of 32 and a learning rate of $2e-5$, We fine-tuned for 5 epochs, keeping the best model with respect to the validation set loss. Also, we used a softmax layer to output predictions and we assigned a relation type to each instance si as follows:

$$rel\_type(s_i) = \begin{cases} argmax(p_c) & \text{if } max(p_c) \geq t \\ null & \text{if } max(p_c) < t \end{cases} \tag{3}$$

where $c \in L$ and $0 < t < 1$ is a threshold that maximizes the micro-F score on the validation set.

For DARE, in all experiments we train an ensemble of twenty classifiers, where each classifier has been trained on the full gold set and a sub-sample of the generated data. In this way, we manage to alleviate the effect of potential noisy generated instances.

4.2 Datasets

To evaluate DARE, we employ three RE datasets from the biomedical domain, their statistics being provided in Table 2.

The BioCreative V CDR corpus (Li et al., 2016) contains chemical-disease relations. The dataset is a binary classification task with one relation type, *chemical induces disease*, and annotations are at the document level, having already been split into train, development and test splits. For simplicity, we followed the work of Papanikolaou et al. (2019) and considered only intra-sentence relations. We have included the dataset in our GitHub repository to ease replication. In the following, we dub this dataset as *CDR*.

The DDIExtraction 2013 corpus (Segura Bedmar et al., 2013) contains MedLine abstracts and DrugBank documents describing drug-drug interactions. The dataset has four relation types and annotations are at the sentence level. The dataset is provided with a train and test split for both MedLine and DrugBank instances. Following previous works, we concatenated the two training sets into one. Also, we randomly sampled 10% as a development set. In the following this dataset will be referred to as *DDI2013*.

GPT-2	Precision	Recall	F1
Vanilla	**0.71**	0.69	0.70
fine-tuned	0.68	**0.75**	**0.73**

Table 3: DARE results on CDR when using a vanilla GPT-2 model or a model that has first been fine-tuned on 500k abstracts from PubMed. In either case the resulting model is then fine-tuned per relation type to generate new examples.

The BioCreative VI-ChemProt corpus (Krallinger et al., 2017) covers chemical-protein interactions, containing five relation types, the vast majority of them being at the sentence level. The dataset comes with a train-development-test split. In the following we will refer to it as *ChemProt*.

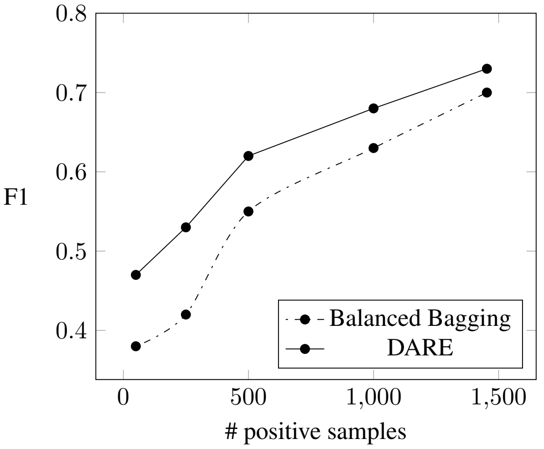
4.3 Baselines

The above datasets suffer both from class imbalance and a limited number of positives. For example the rarest relation type in DDI2013 has only 153 instances in the training set, while the respective one in ChemProt has only 173 data points. Therefore, we consider two suitable baselines for such scenarios, the balanced bagging approach and the class weighting method, both described in Section 2. Both baselines use the base classifier described in Section 4.1. Also, in both cases we consider an ensemble of ten models[2]. Finally, for the class weighting approach we set each class's weight as

$$weight_c = \frac{freq_{min}}{freq_c} \tag{4}$$

with min being the rarest class.

[2] We considered up to 20 models in initial experiments, but there is hardly any improvement after even five models, since the data are repeated.



Dataset	BB	DARE
50	0.38	**0.47**
250	0.42	**0.53**
500	0.55	**0.62**
1000	0.63	**0.68**
all(1453)	0.70	**0.73**

Figure 1: DARE vs balanced bagging(BB) for different sizes of positive samples on CDR dataset. Both methods employ ensembles of BERT RE classifiers.

4.4 Fine-tuning GPT-2 on In-domain Data

Since all our datasets come from the biomedical domain, we hypothesized that a first round of fine-tuning GPT-2 on in-domain data could be beneficial instead of directly employing the vanilla GPT-2 model. We designed a short experiment using the CDR dataset to test this hypothesis. To clarify, any of the two models (i.e, the vanilla and the one finetuned in in-domain data) would then be fine-tuned per relation type to come up with the final GPT-2 models that would generate the new training examples.

Table 3 illustrates the results of this experiment. As we expect, this first round of fine-tuning proves significantly favourable. We note that when inspecting the generated examples from the vanilla GPT-2 model, generated sentences often contained a peculiar mix of news stories with the compound-disease relations.

4.5 DARE on Imbalanced Datasets

In this experiment, we wanted to evaluate the effect of our method when dealing with great imbalance, i.e., datasets with very few positive samples. To that end, we considered the CDR dataset and sampled different numbers of positive examples from the dataset (50, 250, 500, 1000 and all positives) and combined them with all the negative instances. The resulting five datasets were used to train either a balanced bagging ensemble or DARE.

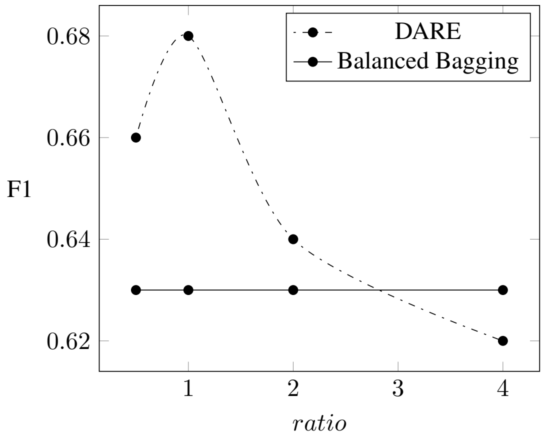


Figure 2: DARE performance for different generated dataset sizes in each base classifier. For each relation type we add *ratio*$*|D_c|$ examples.

In Figure 1, we show the results, averaging across five different runs. In all cases, our approach has a steady, significant advantage over the balanced bagging baseline, their difference reaching up to 11 F1 score points when only few positives ($\leq$ 250) are available. As we add more samples, the differences start to smooth out as expected. These results clearly illustrate that DARE can boost the predictive power of a classifier when dealing with few positive samples, by cheaply generating training data of arbitrary sizes.

4.6 Effect of Generated Data Size

Our next experiment focuses in studying the effect of different sizes of generated data on DARE's performance.

As explained, our method relies on fine-tuning GPT-2 to generate examples for each relation type that will, ideally, come from the same distribution as the ones from the gold training data. Nevertheless, we should expect that this procedure will not be perfect, generating also noisy samples. As mentioned previously, we try to alleviate this effect by training an ensemble of classifiers, each trained on the whole gold and a part of the generated dataset.

An important question that arises therefore, is how to determine the optimal ratio of generated examples to include in each classifier. If too few, the improvements will be insignificant, if too many we risk to have the model being influenced by the noise.

In order to gain empirical insight into the above question we design a short experiment using the CDR dataset, for different sizes of generated data. As gold set, we consider a random subset of 1,000 positive examples and all negatives, to make more prominent the effect of class imbalance.

In Figure 2 we show the results for five different generated data sizes. Interestingly, adding more data does not necessarily boost classifier performance, since the noisy patterns in the generated data seem to influence more the classifier than those in the gold data. In the following, we choose a $ratio = 1$, adding for each relation type a number of generated instances equal to the number of gold instances. It should be noted that we are not limited in the total generated data that we will use since we can fine-tune an arbitrary number of classifiers on combinations of the gold data and subsets of the generated data.

4.7 DARE against the SOTA and Baselines

Taking into account the previous observations, we proceed to compare DARE against the SOTA and the two previously described baselines. Table 4 describes the results. For the multi-class datasets we report the micro-F score in order to make our results comparable with previous works. Also, in the Supplementary Material we report the per class results for DARE against the SOTA and the class weighting baseline, for the two multi-class datasets in order to ease comparison with past or future works.

Comparing DARE against the SOTA, we observe a steady advantage of our method across all datasets, ranging from 3 to 8 F1 points. These results are somehow expected, since we employ BERT-large as our base classifier which has proven substantially better than Convolutional (CNN) or Recurrent neural networks (RNN) across a variety of tasks (Devlin et al., 2018). In CDR, Papanikolaou et al. (2019) have used BioBERT(Lee et al., 2019) which is a BERT base (cased) model pre-trained on PubMed, while we use BERT large (uncased), in ChemProt, Peng et al. (2018) use ensembles of SVM, CNN and RNN models while in DDI2013 Sun et al. (2019b) have used hybrid CNN-RNN models.

When observing results for the baselines, we notice that they perform roughly on par. DARE is better from 2 to 5 F1 points against the baselines, an improvement that is smaller than that against the SOTA, but still statistically significant in all cases.

Overall, and in accordance with the results from the experiment in Section 4.5, we observe that DARE manages to leverage the GPT-2 automatically generated data, to steadily improve upon the SOTA and two competitive baselines.

5 Conclusions

We have presented DARE, a novel method to augment training data in Relation Extraction. Given a gold RE dataset, our approach proceeds by fine-tuning a pre-trained GPT-2 model per relation type and then uses the fine-tuned models to generate new training data. We sample subsets of the synthetic data with the gold dataset to fine-tune an ensemble of RE classifiers that are based on BERT. Through a series of experiments we show empirically that our method is particularly suited to deal with class imbalance or limited data settings, recording improvements up to 11 F1 score points over two strong baselines. We also report new SOTA performance on three biomedical RE benchmarks.

Our work can be extended with minor improvements on other Natural Language Understanding tasks, a direction that we would like to address in future work.

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, pages 2670–2676.

Enhong Chen, Yanggang Lin, Hui Xiong, Qiming Luo, and Haiping Ma. 2011. Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing & Management*, 47(2):202–214.

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

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. 2016. Adaptive data augmentation for image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3688–3692. Ieee.

Dataset	Configuration	Precision	Recall	F1
CDR	SOTA (Papanikolaou et al., 2019)	0.61	0.80	0.70
	BERT+class weighting	0.66	0.74	0.69
	BERT+balanced bagging	0.61	0.79	0.70
	DARE	0.68	0.75	**0.73**
ChemProt	SOTA (Peng et al., 2018)	0.72	0.58	0.65
	BERT+class weighting	0.75	0.67	0.70
	BERT+balanced bagging	0.69	**0.71**	0.70
	BERT+DARE	**0.79**	0.68	**0.73**
DDI2013	SOTA (Sun et al., 2019b)	0.77	0.74	0.75
	BERT+class weighting	0.81	0.71	0.76
	BERT+balanced bagging	0.74	0.72	0.73
	BERT+DARE	0.82	0.74	**0.78**

Table 4: Comparison of DARE vs the previous SOTA and two baselines suited for imbalanced datasets. Only statistically significant results to the second best model are marked in bold. Statistical significance is determined with a McNemar p-test at 0.05 significance level.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331.

Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. 2009. Roughly balanced bagging for imbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2(5-6):412–426.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*.

Steve Lawrence, Ian Burns, Andrew Back, Ah Chung Tsoi, and C Lee Giles. 1998. Neural network classification and prior class probabilities. In *Neural networks: tricks of the trade*, pages 299–313. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Yannis Papanikolaou, Ian Roberts, and Andrea Pierleoni. 2019. Deep bidirectional transformers for relation extraction without supervision. *EMNLP-IJCNLP 2019*, page 67.

Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Extracting chemical–protein relations with ensembles of svm and deep learning models. *Database*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training. Technical report, Technical report, OpenAi.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners. Technical report, Technical report, OpenAi.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

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

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019a. Lamal: Language modeling is all you need for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Xia Sun, Ke Dong, Long Ma, Richard Sutcliffe, Feijuan He, Sushing Chen, and Jun Feng. 2019b. Drug-drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy*, 21(1):37.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 872–884.

Jason Wang and Luis Perez. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit.*

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. 2016. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhi-Hua Zhou and Xu-Ying Liu. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77.

A Supplemental Material

In this section we present additionally the results per class for ChemProt and DDI2013, for DARE against the class weighting baseline and the SOTA.

relation type	SOTA	Class Weighting	DARE
CPR-3	-	0.66	**0.70**
CPR-4	-	0.75	**0.79**
CPR-5	-	0.73	**0.81**
CPR-6	-	0.69	**0.73**
CPR-9	-	0.57	**0.59**

Table 5: ChemProt results per relation type for DARE vs SOTA and best baseline in terms of F1.

relation type	SOTA	Class Weight	DARE
advise	**0.81**	**0.81**	0.80
effect	0.73	0.76	**0.78**
int	**0.59**	0.52	0.58
mechanism	0.78	0.78	**0.80**

Table 6: DDI2013 results per relation type for DARE vs state-of-the-art and best baseline in terms of F1.