DARE: Data Augmented Relation Extraction with GPT-2

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

Introduction Relation Extraction (RE) is the task of identifying semantic relations from text, for given entity mentions within it. 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 [2] to generate examples for specific relation types. We sample subsets of the generated data and combine them with the gold data to fine-tune an ensemble of RE classifiers that are based on BERT [1].

Data Augmented Relation Extraction To generate new training data, we split the gold dataset D into c subsets where each D_c 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 r, i.e., $|Dsynth'_c| = |D_c| * r$.

Results Since all of our datasets are from the biomedical domain, we found out empirically 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 (the largest uncased model) as a RE classifier, which we fine-tuned on either the gold or the gold+generated datasets.

Results Comparing DARE against the SOTA, we observe a steady advantage of our method across all datasets, ranging from 3 to 8 F1 points. DARE is better from 2 to 4 F1 points against the baselines, an improvement that is smaller than that against the SOTA, but still statistically significant in all cases. Overall, we observe that DARE manages to leverage the GPT-2 automatically generated data, to steadily improve upon the SOTA and two competitive baselines. 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.

Dataset	Configuration	Precision	Recall	F1
CDR	SOTA papanikolaou2019deep	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 peng2018extracting	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 sun2019drug	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 1: 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.

References

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

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