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 $D_{synth}$. Subsequently, we build an ensemble of RE classifiers, each of them being fine-tuned on a subset of $D_{synth}$ 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., $|D_{synth}^c| = |D_c| \times r$.

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-tune on either the gold or the gold+generated datasets.

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
