Exploring Pre-training Methods for Verbal Autopsy

Karishma Mandyam

University of Washington

Verbal Autopsy

The Verbal Autopsy task involves gathering information about a deceased individual in order to determine their cause of death. Typically, field workers interview family and close friends of the recently deceased to learn more about their behavior, symptoms, and demographics. These surveys, consisting of categorical and text data, are then interpreted by models in order to predict cause of death (COD) without an official autopsy. This survey is typically done in regions where obtaining official COD reports are difficult or unavailable.

Objectives

- Utilize new advances in pre-training methods to improve performance on the Verbal Autopsy task
- Learn from and leverage strengths of statistical and neural methods
- Understand how pre-training methods translate to a low-resource task, with less data and more variance
- Develop a model to be realistically used by field workers performing Verbal Autopsy surveys

Motivation

- The Verbal Autopsy task has previously been explored with statistical methods, which do not take advantage of some textual aspects of VA surveys.
- Several pre-training methods in NLP typically require large amounts of text data (BERT, RoBERTa, etc.). Recent models, such as VAMPIRE, are not only suitable for smaller data, but also operate with fewer assumptions (a Bag-of-Words model) and allow for semi-supervised training.
- Our goal is to take advantage of the text data and other data from VA surveys by using a lightweight pre-training framework to improve performance on the task.

Data

Verbal Autopsy data consists of two important parts. The first is a list of answers to predetermined questions. These questions aim to identify symptoms and key demographics of the deceased individual. Some sample questions include:

- Last known age of the deceased?
- For how long was the decedent ill before s/he died?
- Did decedent have a cough?
- Did decedent have difficulty breathing?
- Did decedent have yellow discoloration of the eyes?

The second part of the data consists of a bag-of-words interpretation of the open narratives in the survey. For instance, if the word "abdomen" appeared anywhere in the written narrative for the survey, we associate the word "abdomen" with the data point. Other key words observed in this method include "chest", "shock", "upper", "inflame", "organ", "cardio", etc.
The label for each data point is one of 12 causes of death (External, Cardio, Cancer, Stroke, TB/AIDS, Other Non Communicable Diseases, Other Communicable Diseases, Pneumonia, Renal, Maternal, Diabetes, and Liver).

Our data comes from six sites around the world.

Site	Number of Examples
Andhra Pradesh	1554
Bohol	1259
Dar es Salaam	1726
Mexico City	1586
Pemba Island	297
Uttar Pradesh	1419

In the context of developing a model that will be useful for future sites with little labeled data, we separate our dataset as follows. We consider data from Andhra Pradesh, Bohol, Mexico City, Pemba, and Uttar Pradesh as labeled training data. We consider Dar es Salaam as the target site, from which we pull out a small set of labeled evaluation data. The rest of the Dar data is considered unlabeled.

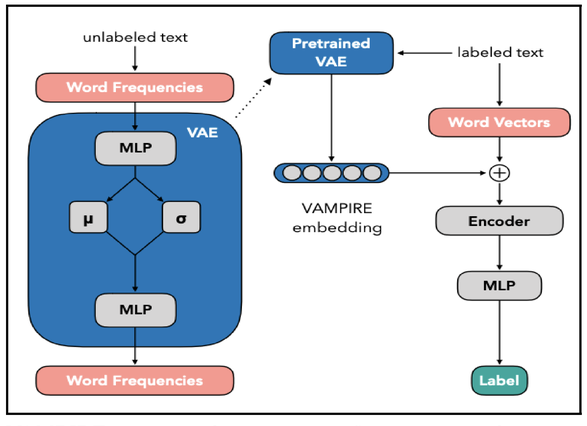
Baseline Models

Prior work from Clark et al. 2018 explored using four statistical methods on the Verbal Autopsy data.

- **Tariff**: Implemented in the OpenVA R package and based on the original algorithm from James et al. 2011
- **InterVA**: Uses physician provided conditional probabilities to determine COD
- **NBC**: Additional method implemented in OpenVA R package
- **InSilicoVA**: Uses same symptom cause information as InterVA

VAMPIRE

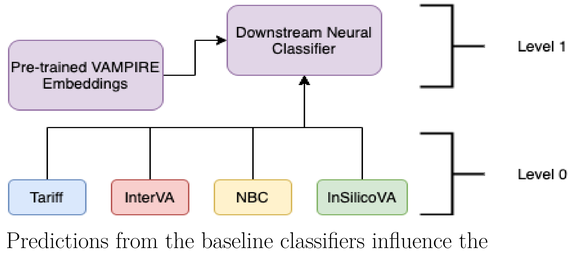
VAMPIRE allows us to utilize the unlabeled Dar data for pre-training.



VAMPIRE uses a similar pretraining-finetuning paradigm to BERT, but uses Variational Auto-Encoders and treats text as a Bag of Words. Figure taken from original VAMPIRE paper.

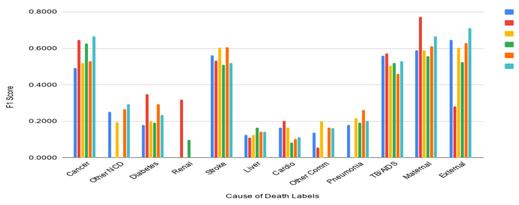
Stacked Learning

Our final model uses stacked learning, where we have two levels of classifiers, level 0 and level 1.



Predictions from the baseline classifiers influence the predictions and learning of the neural classifier, which already includes VAMPIRE embeddings from pre-training

Preliminary Results



The results from evaluating each model on the held-out Dar data. The first four bars refer to the baselines, the fifth bar represents VAMPIRE without stacking, and the final model is VAMPIRE with stacking.

While the InterVA baseline model outperforms all other classifiers in five of the 12 labels, it also has a higher level of variance (0 F1 score for Other NCD and Pneumonia). The stacked learning method outperforms all other models on three out of 12 labels and has a higher average F1 score (0.354) than any other model. This model is also able to take advantage of the unlabeled Dar data by using the semi-supervised pre-training approach in VAMPIRE.

Further Work

- Explore other stacked learning configurations
- Extensive cross validation tests to determine if our results are generalizable
- Evaluating other larger pre-trained models, such as BERT or RoBERTa

Acknowledgements

Suchin Gururangan, Noah Smith, Tyler H. McCormick, Brandon Stewart

Contact Information

krm28@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING