Can Classification of Publications into Translational Categories be Automated?

Sarah A. Wiegreffe[1], Paul E. Anderson, PhD[1], and Jihad S. Obeid, MD[2]
1 Department of Computer Science, College of Charleston, Charleston, SC
2 Medical University of South Carolina, Charleston, SC

INTRODUCTION

There is a significant lag, an average of 17 years, for moving biomedical discoveries from basic science research into clinical research and eventual treatments in practice. As a result, in 2007 there was significant investment in federal funding by the NIH to reduce this gap. A key measure of this funding can be assessed by evaluating the translational trend in resulting publications. Manual classification of each publication is impractical given the massive amount of scientific literature produced. There have been previous works in this area but generating training data for machine learning remains a major obstacle. We used *a priori* knowledge of existing study types in PubMed to facilitate this process.

METHODS

We pulled records from the Medline/PubMed database based on criteria for each translational category found in the Harvard Catalyst pathfinder[1]. To simplify our classifier into those categories most distinguishable from one another, we following a method used in existing work[2] and grouped the five classes into three (see Figure 1). We selected a subset of all the abstracts collected for a total of 97,049, approximately evenly distributed between the three classes.

To investigate the classification of translational categories, we compared traditional bags of words (BOWs) with inverse document frequency using a random forest (RF) classifier against variations of the Word2Vec (W2V) inversion technique[3], which acts as both a preprocessing step and a classifier in one (see Figure 2). Different Word2Vec-based classifiers were derived using statistical representations of sentence probabilities (e.g., mean, standard deviation, position of review). These features were then provided to random forest and decision tree (DT) classifiers to arrive at the final prediction (see Figure 3). This is in contrast to the original Word2Vec inversion which averages sentence probabilities.

T0 — **Basic Science Discovery**

Abstracts from one of the following journals:
- *Cell*
- *Methods in Molecular Biology*
- *Molecular* and *Cellular Biology*

T1/T2 — **Translation to Humans & Translation to Patients**

Publication Types:
- Phase 1, Phase 2, or Phase 3 clinical trials

MeSH headings/terms:
- "humans/physiology"

Keywords:
- "First-in-human"
- "proof of concept"

T3/T4 — **Translation to Practice & Translation to Population Health**

Publication Types:
- Phase 4 clinical trials & comparative studies

MeSH terms:
- "social determinants of health"
- "outcome assessment (health care)"
- "health services research" with subheadings "dissemination", "communication", or "implementation"

Figure 1: Makeup of dataset subsets..

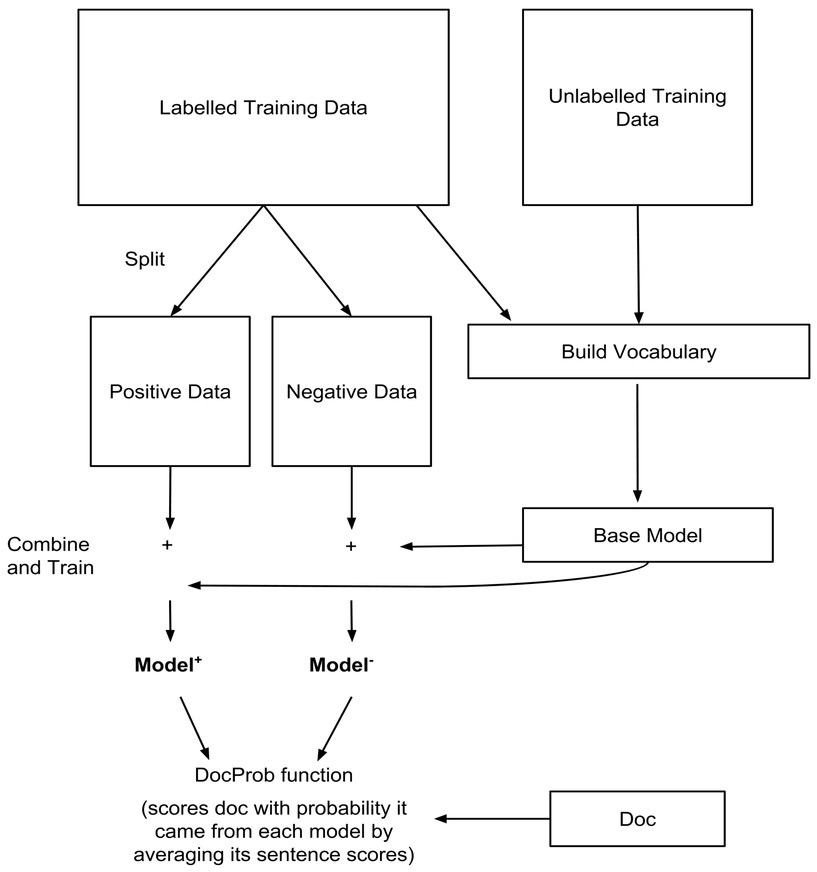

Figure 2: High-level overview of the Word2Vec Inversion[3] algorithm.

RESULTS

We used 5-fold cross validation to calculate the average area under the receiver operating curve score for each model, which served as our metric of performance.

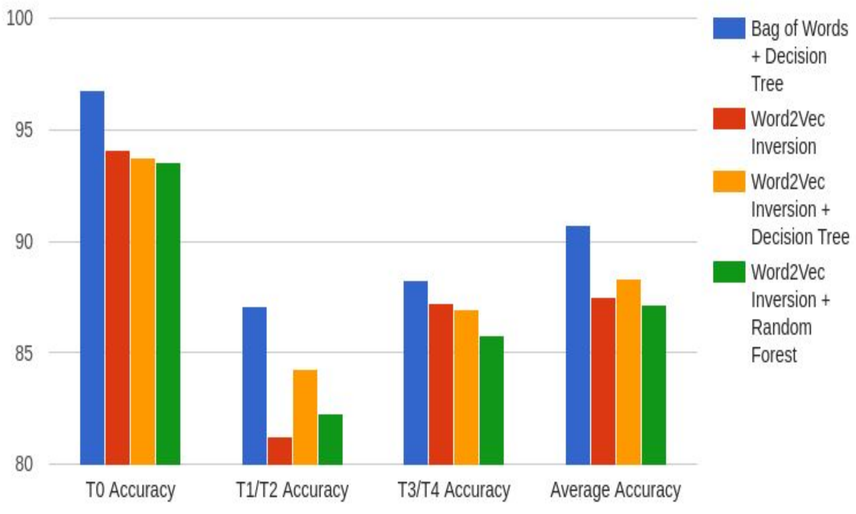

Figure 4: Classification performance of each method with 5-fold cross-validation on the PubMed abstracts dataset.

Technique	BOWs + DT	W2V Inversion	Inversion + DT	Inversion + RF
T0 Accuracy	96.75	94.11	93.74	93.55
T1/T2 Accuracy	87.09	81.22	84.27	82.25
T3/T4 Accuracy	88.25	87.19	86.95	85.75
Average Accuracy	90.70	87.51	88.32	87.18

Table 1: Associated performance values from Figure 4.

DISCUSSION

Performance varies across the categories, with models performing very well on the T0 classification and the worst on T1/T2. The variations of the Word2Vec inversion method did not increase accuracy over a traditional bag of words + random forest approach for this dataset, but the slight improvement seen for T1/T2 over the standard Word2Vec inversion scores indicates that there is promise in this approach, and the methods could work well on other datasets.

For future improvement, we will fine-tune the selection criteria used to build the datasets for each translational category. We will also reference outside expertise to determine what may be misclassification of records, and to validate our selection criteria for the dataset. Final models will be externally validated on another set of labeled publications, which will provide a further metric of how well our constructed datasets are capturing translational category. Our T0 and T1/T2 models are showing preliminary performance improvements of 2-3% over existing work that uses manual classification[2], but we will be able to better compare our results once we have validated on an external, manually labelled dataset. Future work also will include an effort to distinguish between the T1 and T2, and T3 and T4 categories.

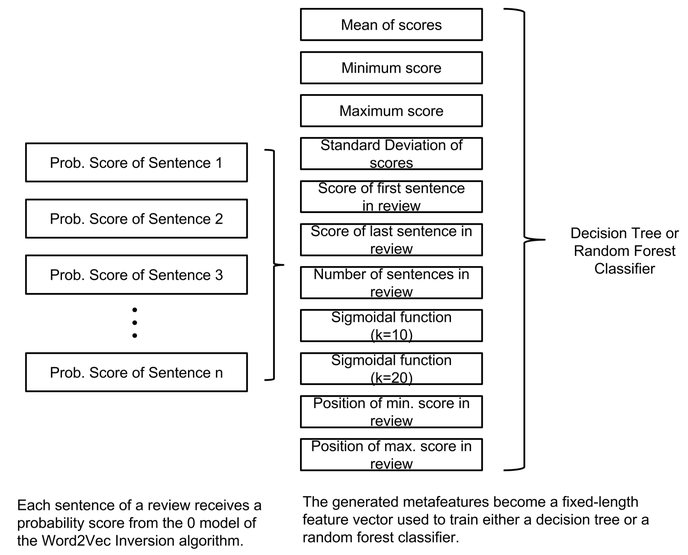

Figure 3: Training of variations of the Word2Vec-based classifier.

REFERENCES

[1] Harvard Catalyst Pathfinder. http://catalyst.harvard.edu/pathfinder/. Harvard Clinical and Translational Science Center, 2016.

[2] Surkis, A. et. al. "Classifying publications from the clinical and translational science award program along the translational research spectrum: a machine learning approach". Journal of Translational Medicine Vol. 14, Issue 235. 5 Apr. 2016.

[3] Taddy, M. "Document Classification by Inversion of Distributed Language Representations". arXiv.org, 2015.

ACKNOWLEDGEMENTS

This work is funded in part by the National Institutes of Health (NIH) Grant # UL1 TR001450, the Medical University of South Carolina, the College of Charleston and the SmartState Program in SC. Presentation of results is funded in part by the College of Charleston Office of Undergraduate Research and Creative Activities Grant # RP2017-024.