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]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: T0 (basic science discovery), T1/T2 (translation to humans/translation to patients), and T3/T4 (translation to practice/translation to population health). Our criteria for the T0 class were records from basic science journals, including Cell, Methods in Molecular Biology, and Molecular and Cellular Biology. For the T1/T2 class, we selected records with publication types of clinical trials in phase 1, 2, or 3, those with the MeSH heading and term "humans/physiology" and those with the keywords "first-in-human" or "proof of concept". For the final class, T3/T4, we included records with publication types of phase 4 clinical trials or comparative studies, as well as those with the MeSH terms "social determinants of health", "outcome assessment (health care)", or "health services research" (subheadings "dissemination", "communication", and "implementation"). 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 classifier against variations of Word2Vec inversion technique[3], which acts as both a preprocessing step and a classifier in one. Different Word2Vec-based classifiers were derived using statistical representations of sentence probabilities (e.g., mean, standard deviation, quartile, bins). These features were then provided to random forest (RF) and decision tree (DT) classifiers to arrive at the final prediction. This is in contrast to the original Word2Vec inversion which averages sentence probabilities. 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.

Results:

Table 1. Performance of various classifiers on the dataset.

Technique	BOWs & RF	Word2Vec 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
Avg. accuracy	90.70	87.51	88.32	87.18

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 do not appear helpful relative to a traditional BOWs & RF approach, but the slight improvement seen for T1/T2 over the standard Word2Vec inversion scores indicates that there is promise in this approach. 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.

Conclusion: Overall, the Word2Vec inversion technique and variations did not improve on the accuracy of the traditional classification method. However, considering the fact that no preprocessing must be done, the Word2Vec inversion technique has similar performance to bags of words and random forest, especially for the T0 and T3/T4 categories, and is simpler in implementation. The resulting accuracies of all of the methods confirm that it is possible to automate the classification of publications into translational categories as defined above. Specifically, basic science (T0) records appear highly distinguishable.

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.

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.