Appendix

Classifying Cause of Revision on Basis of Postoperative Surgical Notes

Because no CPT or ICD codes are available to specify whether a knee revision is due to infection or a non-infectious cause, the full text of the postoperative note was used to classify revision as due or not due to infection. The VA's corporate data warehouse was queried for all of the postoperative surgical text notes (written between the day of surgery and a week after it) of the patients who had had a knee revision in our cohort. The first 1,000 characters from a randomly chosen set of 294 notes were used to manually curate a gold standard for classification, by one of the authors specifying either infection or not infection as the cause of the revision in each note. Numeric characters, punctuation, and stop words in the text were removed, followed by normalization using the Porter Stemmer[37]. Finally, a tf-idf matrix was created for the whole corpus. Using the curated labeled set, a Support Vector Machine (SVM) classifier[38] was trained and tuned to classify the surgical note as specifying either infection or not infection as the cause of the revision. A random train test size equal to 0.75 of the gold standard was used for training, and the reminder was used for testing. An optimal parameter and accuracy score were obtained. The classifier was than applied to the rest of the unseen notes. All analysis was done using Python 2.6, nltk 2.0[39], and Scikit-learn[40]. The classifier had an accuracy of 0.94 with an area under the curve of 0.99. A surgical note was not available for 223 revisions, leaving 1,128 notes that were classified successfully. The SVM parameters were $C = 16681.0$ and $\gamma = 0.00021$ using a radial basis function (RBF) kernel.