Appendix A3: Experimental Procedure and Parameter Choices

In the training phase, each sentence in the training set is represented in the corresponding feature representation. In the Surface level feature (SuLF) Classifier, each sentence is represented as a vector of words. Similarity between the two sentences is computed as a weighted dot product between their vector representations. Weights (w_l,w_m,w_r) assigned to the left, middle and right vectors were set at ($w_l = 0.3, w_m = 0.6, w_r = 0.1$). The intuition behind the values is that domain experts found that the middle context contained most indicators of connectivity followed by left and then the right context. The same intuition for the middle context has been adopted by Agichtein et al [1] also. Similar vectors are clustered together and each cluster center, computed as the mean of all vectors in the cluster, forms a pattern.
In the Syntax level feature (SyLF) Classifier, each sentence in the training set is parsed by the link parser and represented in the form of its Bridge using the Iterative Least Cost Parsing (ILCP) approach. The weights associated with the words and links in the bridge representation are calculated to derive the weight of quadruples making up the bridges. In this weight computation, *weight(c)* for left, middle and right contexts are set at 0.3, 0.6 and 0.1 respectively with the same domain intuition about the importance of contexts. Pattern generation from the training set bridges is achieved by grouping similar bridges together using the weighted edit distance measure. The procedure *Substitute* in the Weighted Bridge Edit Distance algorithm computes the penalty values based on different similarity checks between the quadruple pair being compared, where the corresponding left words, right words and links in the quadruple pair are matched and penalties assigned as given in Table 1 .

Penalty value	Condition
0	Links and corresponding words (left and right) match
0.3	Links match, but only one of the corresponding words match
0.6	Links match, but none of the corresponding words match
1	Links do not match, but only one of the corresponding words match
2	Neither links, nor, any of the corresponding words match
2	Context of the links do not match

Table 1: Substitute penalty values used in Weighted Bridge Edit Distance algorithm

The generated patterns are ranked by their confidence score and the highly

confident patterns form the pattern bank, representative of the connectivity statements in the training corpus. In the testing phase, a test sentence is classified as *'Connected'* if its confidence is more than the threshold τ_c. The confidence value is a function of the similarity of test sentence and the most similar, top k confident patterns generated during the training phase.

10-fold cross validation has been performed on the training data to optimize the parameters used by the algorithm. In our experiments, values of the parameters were chosen by cross validation. In the *WhiteText* dataset, for the SuLF-CW classifier, the similarity threshold for clustering $\tau_{sim} = 0.11$ and confidence threshold $\tau_c = 0.24$ was arrived upon by 10-fold cross validation. Similarly, in case of SyLF-LPBridge classifier, value of parameters k=7 and $\tau_c = 0.4$ were derived using 10-fold cross validation. The models trained on the *WhiteText* dataset are applied with the corresponding parameter setting to the *10NeuroPubMed* dataset and results are reported.
The remaining parameters used by the algorithm, though set to specific values, the user of the system can configure these values to suit the needs of expected precision and recall.

References

[1] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections. In Proceedings of the fifth ACM conference on Digital libraries 2000 Jun 1 (pp. 85-94). ACM.