Brown Clustering: Overview and Applications

Zacharia Arthur Rupp

Text clustering is a text mining technique that allows grouping of unlabeled text data into clusters such that each element within a single cluster is similar in such a way that their syntagmatic or paradigmatic relationships are captured. This is an immensely useful tool, allowing not only a good general overview of the corpus to which the method has been applied, but also the automation of tasks such as offensive language detection. This review will explore Brown Clustering, providing a high-level overview of the algorithm and some potential applications as explored in recent research.

Brown Clustering is a hierarchical clustering method, meaning that it produces from an unlabeled corpus input an output that has a root cluster which spreads out like a binary tree. The tree is structured such that taking “horizontal cuts” of it can yield flat clusters of semantically related terms, and small subtrees can yield near synonymous pairs. The logic behind Brown Clustering is to determine the ability of adjacent words to predict words in the currently considered cluster using a key quality function which measures mutual information.

The algorithm works as follows: first, an initial value, k, representing the initial number of clusters is chosen. After having chosen k, the top k most frequent words are placed into their own individual clusters, c1, c2, …, ck. Having established k clusters of the top k most frequent elements in the corpus, we iterate over the remaining unique terms in the corpus and do the following: create a new cluster representing the next most frequent word in the corpus; choose two clusters to be merged, settling on the merge which maximizes our quality function, which is to say the merge that yields the highest mutual information. Finally, we create k-1 merges to construct a full hierarchy. This is an agglomerative method, beginning with single-term clusters which are unified through each iteration. Brown Clustering’s strength is that it’s a fairly simple algorithm which nonetheless allows the creation of a cluster hierarchy from a reasonably large corpus of unlabeled text data from which we can extract clusters with different levels of granularity (Brown et al., 1992, 473).

One application of Brown Clustering explored by Tian and Kubler is to use Brown Clusters in the task of offensive language detection. As noted by Tian and Kübler, research in this direction has typically found that Brown Clusters are inadequate for sentiment analysis because terms for opposing sentiments tend to be semantically related and are thus clustered together (Tian & Kübler, 2020, 5080). In their experiment, Tian and Kübler challenge this idea by suggesting that the reason for the noted inadequacy in previous research stems from having trained the clusters on a general genre; in their work, Tian and Kübler train separate clusters on the general genre and specialized datasets, one comprised of negative sentiment tweets and another

comprised of positive sentiment tweets, and then combine these clusters into a single feature to determine whether or not the resulting feature improves prediction accuracy (Tian & Kübler, 2020, 5081). The results from their experiment showed improved performance not only when combining the individually trained clusters, but also when using each cluster trained on positive and negative datasets as a feature, respectively. Furthermore, Tian and Kübler found the greatest performance when combining the clusters trained on the negative and positive sentiment data with a cluster trained on the general genre to create a single feature, suggesting that each cluster provides varying levels of information to aid in the task. These findings upend previously-held assumptions about the inadequacy of using Brown Clusters for sentiment analysis. Indeed, when including the feature resulting from the combined clusters, Tian and Kübler report improvement in nearly all currently used models for the task of sentiment analysis (Tian & Kübler, 2020, 5085).

One shortcoming of Brown Clustering is that in most current implementations it only considers bigram contexts, which is to say that it only takes into account shared context of immediately adjacent terms. Because of this, Brown clusters can miss out on important relational data which might be more easily captured in other clustering methods. Further complicating things is that Brown clusters may overemphasize relationships between certain bigram pairs when one somewhat contextually meaningless word is frequently paired with another (for a somewhat convoluted example, consider the pair, “throws a,” in a document about pitching). Aiming to improve on this shortcoming of Brown Clustering, Šuster and van Noord explore a modified version of Brown Clustering that uses dependency relationships in place of bigram relationships, thus eliminating the aforementioned disadvantages of most current implementations of Brown Clustering (Šuster & van Noord, 2014, 1382).

In the original Brown Clustering formulation, the clustering function is one that essentially seeks to maximize mutual information between bigram pairs, which is represented by the product of the product of two conditional probabilities,

$$L(w; C) = \prod_i p(C(w_i)|C(w_{i-1}))p(w_i|C(w_i)),$$

where w represents words tokens, and C represents the clustering function which maps words in the vocabulary of the corpus to a cluster as described above. Šuster & van Noord make a slight modification to this formulation by altering the clustering transition probability so that it is conditioned on parent cluster dependency rather than dependency on the preceding word,

$$L'(w; C) = \prod_i p(C(w_i)|C(w_{(i)}))p(w_i|C(w_i)).$$

In this modified formulation, π represents a function from children to their unique parents in the tree (Šuster & van Noord, 2014, 1384).

The small modification to the original Brown Cluster formulation yields impressive results, with Šuster and van Noord reporting significantly better cluster formation when using noun-only dependencies when compared to the original formulation. Furthermore, because dependency relationship data yields more meaningful information than simple bigram relationship data, Šuster and van Noord were able to achieve equivalent evaluation scores to the traditional formulation while using only a fraction of the data.

Though not suitable for all tasks, Brown Clustering provides an excellent clustering strategy; with a conceptually simple algorithm we're able to construct a cluster hierarchy from reasonably large corpora which we can then sample from in meaningful ways. As the above two studies have shown, it provides flexibility such that tasks not typically associated with Brown Clustering can benefit from it with simple tweaks to its application. Furthermore, it proves extensible to such a degree that two of its greatest shortcomings, specifically its use of bigram contexts and its relatively long running time, can be mitigated with simple alterations of the original formulation.

Bibliography

Brown, P. F., deSouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992, December). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467-480.

Šuster, S., & van Noord, G. (2014). From neighborhood to parenthood: the advantages of dependency representation over bigrams in Brown clustering. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, (August 23-29), 1382–1391.

Tian, Z., & Kübler, S. (2020, May). Offensive Language Detection Using Brown Clustering. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 5079–5087.