Taxonomy Induction Using Hierarchical Random Graphs

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

This paper presents a novel approach for inducing lexical taxonomies automatically from text. We recast the learning problem as that of inferring a hierarchy from a graph whose nodes represent taxonomic terms and edges their degree of relatedness. Our model takes this graph representation as input and fits a taxonomy to it via combination of a maximum likelihood approach with a Monte Carlo Sampling algorithm. Essentially, the method works by sampling hierarchical structures with probability proportional to the likelihood with which they produce the input graph. We use our model to infer a taxonomy over 541 nouns and show that it outperforms popular flat and hierarchical clustering algorithms.

1 Introduction

The semantic knowledge encoded in lexical resources such as WordNet (Fellbaum, 1998) has been proven beneficial for several applications including question answering (Harabgiu et al., 2003), document classification (Hung et al., 2004), and textual entailment (Geffet and Dagan, 2005). As the effort involved in creating such resources manually is prohibitive (cost, consistency and coverage are often cited problems) and has to be repeated for new languages or domains, recent years have seen increased interest in automatic taxonomy induction. The task has assumed several guises, such as term extraction — finding the concepts of the taxonomy (Kozareva et al., 2008;Navigli et al., 2011), term relation discovery — learning whether any two terms stand in an semantic relation such as IS-A, or PART-OF (Hearst, 1992; Berland and Charniak, 1999), and taxonomy construction — creating the taxonomy proper by organizing its terms hierarchically (Kozareva and Hovy, 2010; Navigli et al., 2011). Previous work has also focused on the complementary task of augmenting an existing taxonomy with missing information (Snow et al., 2006; Yang and Callan, 2009).

In this paper we propose an unsupervised approach to taxonomy induction. Given a corpus and a set of terms, our algorithm jointly induces their relations and their taxonomic organization. We view taxonomy learning as an instance of the problem of inferring a hierarchy from a network or graph. We create this graph from unstructured text simply by drawing an edge between distributionally similar terms. Next, we fit a Hierarchical Random Graph model (HRG; Clauset et al. (2008)) to the observed graph data based on maximum likelihood methods and Markov chain Monte Carlo sampling. The model essentially works by sampling hierarchical structures with probability proportional to the likelihood with which they produce the input graph. This is advantageous as it allows us to consider the ensemble of random graphs that are statistically similar to the original graph, and through this to derive a consensus hierarchical structure from the ensemble of sampled models. The approach differs crucially from hierarchical clustering in that it explicitly acknowledges that most real-world networks have many plausible hierarchical representations of roughly equal likelihood and does not seek a single hierarchical representation for a given network. This feature also bodes well with the nature of lexical taxonomies: there is no uniquely correct taxonomy for a set of terms, rather different taxonomies
are likely to be appropriate for different tasks and different taxonomization criteria.

Our contributions in this paper are three-fold: we adapt the HRG model to the taxonomy induction task and show that its performance is superior to alternative methods based on either flat or hierarchical clustering; we analyze the requirements of the algorithm with respect to the input graph and the semantic representation of its nodes; and introduce new ways of evaluating the fit of an automatically induced taxonomy against a gold-standard. In the following section we provide an overview of related work. Next, we describe our HRG model in more detail (Section 3) and present the resources and evaluation methodology used in our experiments (Section 4). We conclude the paper by presenting and discussing our results (Sections 4.1–4.4).

2 Related Work

The bulk of previous work has focused on term relation discovery following essentially two methodological paradigms, pattern-based bootstrapping and clustering. The former approach (Hearst, 1992; Roark and Charniak, 1998; Berland and Charniak, 1999; Girju et al., 2003; Etzioni et al., 2005; Kozareva et al., 2008) utilizes a few hand-crafted seed patterns representative of taxonomic relations (e.g., IS-A, PART-OF, SIBLING) to extract instances from corpora. These instances are then used to extract new patterns which are in turn used to find new instances and so on. Clustering-based approaches have been mostly employed to discover IS-A and SIBLING relations (Lin, 1998; Caraballo, 1999; Pan tel and Ravichandran, 2004). A common assumption is that words are related if they occur in similar contexts and thus clustering algorithms group words together if they share contextual features. Most of these algorithms aim at inducing flat clusters rather than taxonomies, with the exception of Brown et al. (1992) whose method induces binary trees.

Contrary to the plethora of algorithms developed for relation discovery, methods dedicated to taxonomy learning have been few and far between. Caraballo (1999) was the first to induce a taxonomy from a corpus using a combination of clustering and pattern-based methods. Specifically, nouns are organized into a tree using a bottom-up clustering algorithm and internal nodes of the resulting tree are labeled with hypernyms from the nouns clustered underneath using patterns such as “B is a kind of A”.

Kozareva et al. (2008) and Navigli et al. (2011) both develop systems that create taxonomies end-to-end, i.e., discover the terms, their relations, and how these are hierarchically organized. The two approaches are conceptually similar: they both use the web and pattern-based methods for finding domain-specific terms. Additionally, in both approaches the acquired knowledge is represented as a graph from which a taxonomy is induced using task-specific algorithms such as graph pruning, edge weighting, and so on.

Our work also addresses taxonomy learning, however, without the term discovery step — we assume we are given the terms for which to create a taxonomy. Similarly to Kozareva et al. (2008) and Navigli et al. (2011), our model operates over a graph whose nodes represent terms and edges their relationships. We construct this graph from a corpus simply by taking account of the distributional similarity of the terms in question. Our taxonomy induction algorithm is conceptually simpler and more general; it fits a taxonomy to the observed network data using the tools of statistical inference, combining a maximum likelihood approach with a Monte Carlo Sampling algorithm. The technique allows us to sample hierarchical random graphs with probability proportional to the likelihood that they generate the observed network. The induction algorithm can operate over any kind of (undirected) graph, and thus does not have to be tuned specifically for different inputs. We should also point out that our formulation of the inference problem utilizes very little corpus external knowledge other than the set of input terms, and could thus be easily applied to domains or languages where lexical resources are scarce.

The Hierarchical Random Graph model (Clauset et al., 2008) has been applied to construct hierarchical decompositions from three sets of network data: a bacterial metabolic network; a food-web among grassland species; and the network of associations among terrorist cells. The only language-related application we are aware of concerns word sense induction. Klapaftis and Manandhar (2010) create a graph of contexts for a polysemous target word and use the HRG to organize them hierarchically, under the assumption that different tree heights correspond to different levels of sense granularity.
3 The Hierarchical Random Graph Model

A HRG consists of a binary tree and a set of likelihood parameters, and operates on input organized into a semantic network, an undirected graph in which nodes represent terms and edges between nodes indicate a relationship between pairs of terms (Figure 1a). From this representation, the model constructs a binary tree whose leaves correspond to nodes in the semantic network (Figure 1b); the model then employs a simple Markov chain Monte Carlo (MCMC) process in order to explore the space of possible binary trees and derive a consensus hierarchical structure from the ensemble of sampled models (Figure 1c).

3.1 Representing a Hierarchical Structure

Formally, we denote a semantic network \( S = (V, E) \), where \( V = \{v_1, v_2, \ldots, v_n\} \) is the set of vertices, one per term, and \( E \) is the set of edges between terms in which \( E_{a,b} \) indicates the presence of an edge between \( v_a \) and \( v_b \).

Given a network \( S \), we construct a binary tree \( D \) whose \( n \) leaves correspond to \( V \) and whose \( n - 1 \) internal nodes denote a hierarchy over \( V \). Because the leaves remain constant for a given \( S \), we define \( D \) as the set of internal nodes \( D = \{D_1, D_2, \ldots, D_n\} \) and associate each edge \( E_{a,b} \in E \) with an internal node \( D_i \) being the lowest common parent of \( a, b \in V \). The core assumption underlying the HRG model is that edges in \( S \) have a non-uniform and independent probability of existing. Each possible edge \( E_{a,b} \in E \) exists with a probability \( \theta_i \), where \( \theta_i \) is associated with the corresponding internal node \( D_i \).

For a given internal node \( D_i \), let \( \ell_i \) and \( r_i \) be the number of leaves in \( D_i \)’s left and right subtrees, respectively; let \( E_i \) be the number of edges in \( E \) associated with \( D_i \) (colloquially, the number of edges in \( S \) between leaves in \( D_i \)’s left and right subtrees). For each \( D_i \in D \), we can estimate the maximum likelihood for the corresponding \( \theta_i \) as \( \hat{\theta}_i = \frac{E_i}{\ell_i r_i} \). The likelihood \( L(D, \theta|S) \) of a HRG over a given semantic network \( S \) is then given by:

\[
L(D, \theta|S) = \prod_{i=1}^{n-1} \left( \hat{\theta}_i \right)^{E_i} (1 - \hat{\theta}_i)^{\ell_i r_i - E_i}
\]  

3.2 Markov Chain Monte Carlo Sampling

Given a representation for a HRG \( \mathcal{H}(D, \theta) \) and a method for estimating the likelihood of a given \( D \) and \( \theta \), we can focus on obtaining the binary tree \( D \) which best fits (or most plausibly explains) a given semantic network. Because the space of possible binary trees over \( V \) is super-exponential with respect to \( |V| \), we employ a MCMC process to sample from the space of binary trees. During each iteration of
Algorithm 1: MCMC Sampling

1. Compute the likelihood $\hat{L}(D, \theta)$ of the current binary tree.
2. Pick a random internal node $D_i \in D$.
3. Randomly permute $D_i$ according to Figure 2.
4. Compute the likelihood $\hat{L}(D, \theta)$ of the modified binary tree.
5. if $\hat{L}(D, \theta) > L(D, \theta)$ then
   accept the transition;
else
   accept with probability $\hat{L}(D, \theta)/L(D, \theta)$ (i.e., standard Metropolis acceptance).
end

Repeat;

Algorithm 2: Flat Clusters

1. Let $D_k$ be the root node of $D$.
2. if $\theta_k > \hat{\theta}$ then
   output the leaves of the subtree rooted at $D_k$ as a cluster
else
   repeat 2 with left and right children of $D_k$.
end

this process we randomly select a node within the tree and permute it according to Figure 2. If this permutation improves the overall likelihood of the dendrogram we accept it as a transition, otherwise it is accepted with a probability proportional to the degree to which it decreases the overall likelihood (i.e., standard Metropolis acceptance). This procedure is described in more detail in Algorithm 1.

3.3 Consensus Hierarchy

Once the MCMC process has converged, the model is left with a binary tree over the terms from the input semantic network. As in standard hierarchical clustering, however, this imposes an arbitrary structure which may or may not correspond to the observed data — the tree at convergence will be similar to an ideal tree given the graph, but may not be the most plausible structure. Indeed, for taxonomy induction it is quite unlikely that a binary tree will provide the most appropriate categorization.

To avoid encoding such bias we employ a model averaging technique to produce a consensus hierarchy. For a set of binary trees sampled after convergence, we first identify the set of possible clusters encoded in the tree, e.g., the binary tree in Figure 1b encodes the clusters $\{AB, ABC, EF, D, DEF, ABCDEF\}$. As in Clauset et al. (2008), each cluster instance is then weighted according to the likelihood of the originating HRG (Equation 1); we then sum the weights for each distinct cluster across all resampled trees and discard those whose aggregate weight is lower than 50% of the total observed weight. The remaining clusters are then used to reconstruct a hierarchy in which each subtree appears in the majority of trees observed after the sampling process has reached convergence, hence the term consensus hierarchy.

3.4 Obtaining Flat Clusters

For evaluation purposes we may want to compare the groupings created by the HRG to a simpler non-hierarchical clustering algorithm (see Section 4 for details). We thus defined a method of converting the tree produced by the HRG into a flat (hard) clustering. This can be done in a relatively straightforward, principled fashion using the HRG’s $\theta$ parameters. For a given $\mathcal{H}(D, \theta)$ we identify internal nodes whose $\theta_k$ likelihood is greater than the mean likelihood and who possess no parent node whose $\theta_k$ likelihood is also greater than the mean. Each such node is the root of a densely-connected subtree; each such subtree is then assumed to represent a single discrete cluster of related items, where $\hat{\theta} = mean(\theta)$ (illustrated in Figure 1c). This procedure is explained in greater detail in Algorithm 2.

4 Evaluation

Data

We evaluated our taxonomy induction algorithm using McRae et al.’s (2005) dataset which consists of for 541 basic level nouns (e.g., DOG and TABLE). Each noun is associated with features (e.g., has-legs, is-flat, and made-of-wood for TABLE) collected from human participants in multiple studies over several years. The original norming study does not include class labels for these nouns, however, we were able to exploit a clustering provided by Fountain and Lapata (2010), in which a set of online participants annotated each of the McRae et al. nouns with basic category labels.

The nouns and their class labels were further taxonomized using WordNet (Fellbaum, 1998). Specifically, we first identified the full hypernym path in WordNet for each noun in McRae et al.’s (2005) dataset, e.g., APPLE > PLANT STRUCTURE > NAT-
URAL OBJECT > PHYSICAL OBJECT > ENTITY (a total of 493 concepts appear in both). These hypernym paths were then combined to yield a full taxonomy over McRae et al.’s nouns; internal nodes having only a single child were recursively removed to produce a final, compact taxonomy containing 186 semantic classes (e.g., ANIMALS, WEAPONS, FRUITS) organized into varying levels of granularity (e.g., SONGBIRDS > BIRDS > ANIMALS).

Evaluation measures Evaluation of taxonomically organized information is notoriously hard (see Hovy (2002) for an extensive discussion on this topic). This is due to the nature of the task which is inherently subjective and application specific (e.g., a dolphin can be a Mammal to a biologist, but a Fish to a fisherman or someone visiting an aquarium). Nevertheless, we assessed the taxonomies produced by the HRG against the WordNet-like taxonomy described above using two measures, one that simply evaluates the grouping of the nouns into classes without taking account of their position in the taxonomy and one which evaluates the taxonomy directly.

To evaluate a flat clustering into classes we use the F-score measure introduced in the SemEval 2007 task (Agirre and Soroa, 2007); it is the harmonic mean of precision and recall defined as the number of correct members of a cluster divided by the number of items in the cluster and the number of items in the gold-standard class, respectively. Although informative, evaluation based solely on F-score puts the HRG model at a comparative disadvantage as the task of taxonomy induction is significantly more difficult than simple clustering. To overcome this disadvantage we propose an automatic method of evaluating taxonomies directly by first computing the walk distance between pairs of terms that share a gold-standard category label within a gold-standard and a candidate taxonomy, and then computing the pairwise correlation between distances in each tree (Lapointe, 1995). This captures the intuition that a ‘good’ hierarchy is one in which items appearing near one another in the gold taxonomy also appear near one another in the induced one. It is also conceptually similar to the task-based IS-A evaluation (Snow et al., 2006) which has been traditionally used to evaluate taxonomies.

Formally, let \( G = \{g_0, g_2, \ldots, g_{n-1}\} \), where \( g_{ab} \) indicates the walk distance between terms \( a \) and \( b \) in the gold standard hierarchy. Similarly, let \( C = \{c_{0,1}, c_{0,2}, \ldots, c_{n-1}\} \), where \( c_{ab} \) is the distance between \( a \) and \( b \) in the candidate hierarchy. The tree-height correlation between \( G \) and \( C \) is then given by Spearman’s \( p \) correlation coefficient between the two sets. All tree-height correlations reported in our experiments were computed using the WordNet-based gold-standard taxonomy over McRae et al.’s (2005) nouns.

Baselines We compared the HRG output against three baselines. The first is Chinese Whispers (CW; Biemann (2006)), a randomized graph-clustering algorithm which like the HRG also takes as input a graph with weighted edges. It produces a hard (flat) clustering over the nodes in the graph, where the number of clusters is determined automatically. Our second baseline is Brown et al.’s (1992) agglomerative clustering algorithm that induces a mapping from word types to classes. It starts with \( K \) classes for the \( K \) most frequent word types and then proceeds by alternately adding the next most frequent word to the class set and merging the two classes which result in the least decrease in the mutual information between class bigrams. The result is a class hierarchy with word types at the leaves. Additionally, we compare against standard agglomerative clustering (Sokal and Michener, 1958) which produces a binary dendrogram in a bottom-up fashion by recursively identifying concepts or clusters with the highest pairwise similarity.

In the following, we present our taxonomy induction experiments (Sections 4.1–4.3). Since HRGs provide a means of inducing a hierarchy over a graph-based representation, which may be constructed in an arbitrary fashion, our experiments were designed to investigate how the topology and quality of the input graph influences the algorithm’s performance. We thus report results when the semantic network is created from data sources of varying quality and granularity.

4.1 Experiment 1: Taxonomy Induction from Feature Norms

Method We first considered the case where the input graph is of high semantic quality and constructed a semantic network from the feature norms collected by McRae et al. (2005). Each noun was represented as a vector with dimensions corresponding to the possible features generated by participants of the norming study; the value of a term along a dimen-

---

1The taxonomy and flat cluster labels are available from http://homepages.inf.ed.ac.uk/s0897549/data.
Table 1: Cluster F-score and tree-height correlation evaluation; a semantic network constructed over McRae et al.’s (2005) nouns and features is given as input to the algorithms.

<table>
<thead>
<tr>
<th>Method</th>
<th>F-score</th>
<th>Tree Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRG</td>
<td>0.507</td>
<td>0.168</td>
</tr>
<tr>
<td>CW</td>
<td>0.464</td>
<td>—</td>
</tr>
<tr>
<td>Agglo</td>
<td>0.352</td>
<td>0.137</td>
</tr>
</tbody>
</table>

Results

Our results are summarized in Table 1. We only give the tree correlation for the HRG and agglomerative methods (Agglo) as CW does not induce a hierarchical clustering. In addition, we do not compare against Brown et al. (1992) as the input to this algorithm is not vector-based. When evaluated using F-score, the HRG algorithm produces better quality clusters compared to CW, in addition to being able to organize them hierarchically. It also outperforms agglomerative clustering by a large margin. A similar pattern emerges when the HRG and Agglo are evaluated on tree correlation. The taxonomies produced by the HRG are a better fit against the WordNet-based gold standard; the difference in performance is statistically significant \( p < 0.01 \) using a \( t \)-test (Cohen and Cohen., 1983).

4.2 Experiment 2: Taxonomy Induction from the British National Corpus

Method

The results of Experiment 1 can be considered as an upper bound of what can be achieved by the HRG when the input graph is constructed from highly accurate semantic information. Feature norms provide detailed knowledge about meaning which would be very difficult if not close to impossible to obtain from a corpus. Nevertheless, it is interesting to explore how well we can induce taxonomies using a lower quality semantic network. We therefore constructed a network based on co-occurrence statistics computed from the British National Corpus (BNC, 2007) and provided the resulting semantic network as input to the HRG, CW, and Agglo models; additionally, we employed the algorithm of Brown et al. (1992) to induce a hierarchy over the target terms directly from the corpus. Unfortunately, this algorithm requires the number of desired output clusters to be specified in advance; in all trials this parameter was set to the number of clusters in the gold-standard clustering (41), thus providing the Brown-induced clusterings with a slight oracle advantage.

Again, nouns were represented as vectors in semantic space. We used a context window of five words on either side of the target word and 5,000 vector components corresponding to the most frequent non-stopwords in the BNC. Raw frequency counts were transformed using pointwise mutual information (PMI). An edge was added to the semantic network between a pair of nouns if their similarity exceeded a predefined threshold (the same as in Experiment 1). The similarity of two nouns was defined as the cosine distance between their corresponding vectors.

The HRG algorithm was used to produce a taxonomy from this network and was also compared against Brown et al. (1992). The latter induces a hierarchy from a corpus directly, without the intermediate graph representation. All resulting taxonomies were evaluated against gold standard flat and hierarchical clusterings, again as in Experiment 1.

Results

Results are shown in Table 2. With regard to flat clustering (the F-score column in the table), the HRG has a slight advantage against CW, and Brown et al.’s (1992) algorithm (Brown). However, differences in performance are not statistically significant. Agglomerative clustering is the worst performing method leading to a decrease in F-score of approximately 1.5. With regard to tree correlation, the output of the HGRG is comparable to Brown (the difference between the two is not statistically significant). Both algorithms are significantly better \( p < 0.01 \) than Agglo.
Figure 3: The original semantic network as derived from the BNC (a) and the same network re-weighted using a flat clustering produced by CW (b). As \( s \) approaches 1.0 the network exhibits an increasingly strong small-world property, eventually reconstructing the input clustering only (c).

<table>
<thead>
<tr>
<th>Method</th>
<th>F-score</th>
<th>Tree Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRG</td>
<td>0.276</td>
<td>0.104</td>
</tr>
<tr>
<td>CW</td>
<td>0.274</td>
<td>—</td>
</tr>
<tr>
<td>Brown</td>
<td>0.258</td>
<td>0.124</td>
</tr>
<tr>
<td>Agglo</td>
<td>0.122</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Table 2: Cluster F-score and tree-height correlation evaluation for taxonomies inferred over McRae et al.’s (2005) nouns; all algorithms are run on the BNC.

Performance of the HRG is better when the semantic network is based on feature norms (compare Tables 1 and 2), both in terms of tree-height correlation and F-score. This suggests that the algorithm is highly dependent on the quality of the semantic network used as input. In particular, HRGs are known to be more appropriate for so-called small-world networks, graphs composed of densely-connected subgraphs with relatively sparse connections between (Klapaftis and Manandhar, 2010). Indeed, inspection of the semantic network produced from the BNC (see Figure 3a) shows that our corpus-derived graph is emphatically not a small-world graph, yet the HRG is able to recover some taxonomic information from such a densely-connected network.

In the following experiments we first assess the difficulty of the taxonomy induction task to get a feel of how well the algorithms are performing in comparison to humans and then investigate ways of rendering the BNC-based graph more similar to a small-world network.

4.3 Experiment 3: Human Upper Bound

Method The previous experiments evaluated the performance of the HRG against a gold-standard hierarchy derived from WordNet. For any set of concepts there will exist multiple valid taxonomies, each representing an accurate if differing organization of identical concepts using different criteria; for the set of concepts used in Experiments 1–2 the WordNet hierarchy represents merely one of many valid hierarchies. Noting this, it is interesting to explore how well the hierarchies output by the model fit within the set of possible, valid taxonomies over a given set of concepts.

We thus conducted an experiment in which human participants were asked to organize words into arbitrary hierarchies. To render the task feasible, they were given a small subset of 12 words rather than the full set of 541 nouns over which the HRG operates. We first selected a sub-hierarchy of the WordNet tree (‘living things’) along with its subtrees (e.g., ‘animals’, ‘plants’), and chose target concepts from within these trees in order to produce a taxonomy in which some items were differentiated at a high level (e.g., ‘python’ vs. ‘dog’) and others at a fine-grained level (e.g., ‘lion’ vs ‘tiger’). The experiment was conducted using Amazon Mechanical Turk\(^2\), and involved 41 participants, all self-reported native English speakers. No guidelines as to what features participants were to use when organizing these concepts were provided. Participants were presented with a web-based, graphical, mouse-driven interface for constructing a taxonomy over the cho-

\(^2\)http://mturk.com


### Results

As shown in Table 3, participants achieve a mean pairwise tree correlation of 0.511. This indicates that there is a fair amount of agreement with respect to the taxonomic organization of the words in question. The HRG comes close achieving a mean tree correlation of 0.412, followed by Agglo, and Brown. In general, we observe that the HRG manages to produce hierarchies that resemble those generated by humans to a larger extent than competing algorithms. The results in Table 3 also hint at the fact that the taxonomy induction task is relatively hard as participants do not achieve perfect agreement despite the fact that they are asked to taxonomize only 12 words.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tree Correlation</th>
<th>Min</th>
<th>Max</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRG</td>
<td>0.412</td>
<td>-0.039</td>
<td>0.799</td>
<td>0.166</td>
</tr>
<tr>
<td>Brown</td>
<td>0.181</td>
<td>0.006</td>
<td>0.510</td>
<td>0.121</td>
</tr>
<tr>
<td>Agglo</td>
<td>0.274</td>
<td>-0.056</td>
<td>0.603</td>
<td>0.121</td>
</tr>
<tr>
<td>Agreement</td>
<td>0.511</td>
<td>-0.109</td>
<td>1.000</td>
<td>0.267</td>
</tr>
</tbody>
</table>

**Table 3:** Model performance on a subset of the target words used in Experiments 1–2, applied to a subset of the semantic network used in Experiment 2. Instead of a WordNet-derived hierarchy, models were evaluated against hierarchies manually produced by participants in an online study. Tree correlation values are means; we also report the minimum (Min), maximum (Max), and standard deviation (Std) of the mean.

4.4 Experiment 4: Taxonomy Induction from a Small-world Network

**Method** In Experiment 2 we hypothesized that a small-world input graph would be more advantageous for the HRG. In order to explore this further, we imposed something of a small-world structure on the BNC, see Table 2.

To evaluate the HRG, along with the baselines from Experiment 2, against the resulting hierarchies we constructed a semantic network over the subset of concepts using similarities derived from the BNC; this network was a subgraph of that used in Experiment 2. We compute *inter-annotator agreement* as the mean pairwise tree-height correlation between the hierarchies our participants produced. We also report for each model the mean tree-height correlation between the hierarchy it produced and those created by human annotators.

The modified weight \( \hat{W}_{A,B} \) between a pair of terms \( A, B \) was computed according to Equation (2), where \( s \) indicates the proportion of edge weight drawn from the clustering, \( W_{A,B} \) is the edge weight in the original (BNC) semantic network, and \( C_{A,B} \) is a binary value indicating that \( A \) and \( B \) belong to the same cluster (i.e., \( C_{A,B} = 1 \) if \( A \) and \( B \) share a cluster; \( C_{A,B} = 0 \) otherwise).

\[
\hat{W}_{A,B} = (1 - s)W_{A,B} + sC_{A,B}
\] (2)

The value of the \( s \) parameter was tuned empirically on held-out development data and set to \( s = 0.4 \) for both CW and Brown algorithms. Each re-weighted network was then used as input to an HRG, and the resulting taxonomies were evaluated in the same manner as in Experiments 1 and 2.

**Results** Table 4 shows results for cluster F-score and tree-height correlation evaluation for taxonomies inferred by the HRG using semantic network derived from the BNC and re-weighted using CW and Brown.

<table>
<thead>
<tr>
<th>Method</th>
<th>F-score</th>
<th>Tree Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRG</td>
<td>0.276</td>
<td>0.104</td>
</tr>
<tr>
<td>HRG + CW</td>
<td>0.291</td>
<td>0.161</td>
</tr>
<tr>
<td>HRG + Brown</td>
<td>0.255</td>
<td>0.173</td>
</tr>
</tbody>
</table>

**Table 4:** Cluster F-score and tree-height correlation evaluation for taxonomies inferred by the HRG using semantic network derived from the BNC and re-weighted using CW and Brown.

The BNC semantic network, using a combination of the baseline clustering methods evaluated in Experiment 2. Specifically, we first obtain a (flat) clustering using either CW or Brown, which we then use to re-weight the BNC graph given as input to the HRG.

Note that, as the clustering algorithms used are unsupervised this procedure does not introduce any outside supervision into the overall taxonomy induction task.

As can be seen, re-weighting improves tree-height correlation substantially: HRG with CW and Brown is significantly better than HRG on its own (\( p < 0.05 \)). In the case of CW, cluster F-score also yields a slight improvement. Interestingly, the tree-height correlations obtained with CW and Brown are comparable to those attained by the HRG.
Figure 4: An excerpt from a hierarchy induced by the HRG, using the BNC semantic network with Brown re-weighting. The HRG does not provide category labels for internal nodes of the hierarchy, but subtrees within this excerpt correspond roughly to (0) TEXTILES, (1) CLOTHING, (2) GENDERED CLOTHING, (3) MEN’S CLOTHING, and (4) WOMEN’S CLOTHING.

when using the human-produced feature norms (differences in correlations are not statistically significant). An excerpt of a HRG-induced taxonomy is shown in Figure 4.

5 Conclusions

In this paper we have presented a novel method for automatically inducing lexical taxonomies based on Hierarchical Random Graphs. The approach is conceptually simple, taking a graph representation as input and fitting a taxonomy via combination of a maximum likelihood approach with a Monte Carlo Sampling algorithm. Importantly, the approach does not operate on corpora directly, instead it relies on an abstract, interim representation (a semantic network) which we argue is advantageous, as it allows to easily encode additional information in the input. Furthermore, the model presented here is largely parameter-free, as both the input graph and the inferred taxonomy are derived empirically in an unsupervised manner (minimal tuning is required when graph re-weighting is employed, the parameter $s$).

Our experiments have shown that both the input semantic network and the representation of its nodes influence the quality of the induced taxonomy. Representing the terms of the taxonomy as vectors in a human-produced feature space yields more coherent semantic classes compared to a corpus-based vector representation (see the F-score in Tables 1 and 4). This is not surprising, as feature norms provide more detailed and accurate knowledge about semantic representations than often noisy and approximate corpus-based distributions.\(^4\) It may be possible to obtain better performance when considering more elaborate representations. We have only experimented with a simple semantic space, however variants that utilize syntactic information (e.g., Padó and Lapata (2007)) may be more appropriate for the taxonomy induction task. Our experiments have also shown that the topology of the input semantic network is critical for the success of the HRG. In particular edge re-weighting plays an important role and generally improves performance. We have adopted a simple method based on flat clustering; it may be interesting to compare how this fares with more involved weighting schemes such as those described in Navigli et al. (2011). Finally, we have shown that naive participants are able to perform the taxonomy induction task relatively reliably and that the HRG approximates human performance on a small-scale experiment. We have evaluated model output using F-score and tree-height correlation which we argue are complementary and allow to assess hierarchical clustering more rigorously.

Avenues for future work are many and varied. Besides exploring the performance of our algorithm on more specialized domains (e.g., mathematics or geography) we would also like to create an incremental version that augments an existing taxonomy with missing information. Additionally, the taxonomies inferred with the HRG do not currently admit term ambiguity which we could remedy by modifying our technique for constructing a consensus hierarchy to reflect the sampled frequency of observed subtrees.

\(^4\)Note that as multiple participants are required to create a representation for each word, norming studies typically involve a small number of items, consequently limiting the scope of any computational model based on normed data.
References


