A Cognitive Model of Semantic Network Learning

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson
Department of Computer Science
University of Toronto
{aida,afsaneh,suzanne}@cs.toronto.edu

Abstract

Child semantic development includes learning the meaning of words as well as the semantic relations among words. A presumed outcome of semantic development is the formation of a semantic network that reflects this knowledge. We present an algorithm for simultaneously learning word meanings and gradually growing a semantic network, which adheres to the cognitive plausibility requirements of incrementality and limited computations. We demonstrate that the semantic connections among words in addition to their context is necessary in forming a semantic network that resembles an adult's semantic knowledge.

1 Introduction

Child semantic development includes the acquisition of word-to-concept mappings (part of word learning), and the formation of semantic connections among words/concepts. There is considerable evidence that understanding the semantic properties of words improves child vocabulary acquisition. In particular, children are sensitive to commonalities of semantic categories, and this abstract knowledge facilitates subsequent word learning (Jones et al., 1991; Colunga and Smith, 2005). Furthermore, representation of semantic knowledge is significant as it impacts how word meanings are stored in, searched for, and retrieved from memory (Steyvers and Tenenbaum, 2005; Griffiths et al., 2007).

Semantic knowledge is often represented as a graph (a *semantic network*) in which nodes correspond to words/concepts[1], and edges specify the semantic relations (Collins and Loftus, 1975; Steyvers and Tenenbaum, 2005). Steyvers and Tenenbaum (2005) demonstrated that a semantic network that encodes adult-level knowledge of words exhibits a *small-world* and *scale-free* structure. That is, it is an overall sparse network with highly-connected local sub-networks, where these sub-networks are connected through high-degree hubs (nodes with many neighbours).

Much experimental research has investigated the underlying mechanisms of vocabulary learning and characteristics of semantic knowledge (Quine, 1960; Bloom, 1973; Carey and Bartlett, 1978; Gleitman, 1990; Samuelson and Smith, 1999; Jones et al., 1991; Jones and Smith, 2005). However, existing computational models focus on certain aspects of semantic acquisition: Some researchers develop computational models of word learning without considering the acquisition of semantic connections that hold among words, or how this semantic knowledge is structured (Siskind, 1996; Regier, 2005; Yu and Ballard, 2007; Frank et al., 2009; Fazly et al., 2010). Another line of work is to model formation of semantic categories but this work does not take into account how word meanings/concepts are acquired (Anderson and Matessa, 1992; Griffiths et al., 2007; Fountain and Lapata, 2011).

Our goal in this work is to provide a cognitively-plausible and unified account for both acquiring and representing semantic knowledge. The requirements for cognitive plausibility enforce some constraints on a model to ensure that it is comparable with the cognitive process it is formulating (Poibeau et al., 2013). As we model semantic acquisition, the first requirement is incrementality, which means that the model learns gradually as it processes the input. Also, there is a limit on the number of computations the model performs at each step.

In this paper, we present an algorithm for si-

[1]Here we assume that the nodes of a semantic network are word forms and its edges are determined by the semantic features of those words.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 244–254,
October 25-29, 2014, Doha, Qatar.

multaneously learning word meanings and growing a semantic network, which adheres to the cognitive plausibility requirements of incrementality and limited computations. We examine networks created by our model under various conditions, and explore what is required to obtain a structure that has appropriate semantic connections and has a small-world and scale-free structure.

2 Related Work

Models of Word Learning. Given a word learning scenario, there are potentially many possible mappings between words in a sentence and their meanings (real-world referents), from which only some mappings are correct (the *mapping problem*). One of the most dominant mechanisms proposed for vocabulary acquisition is *cross-situational learning*: people learn word meanings by recognizing and tracking statistical regularities among the contexts of a word's usage across various situations, enabling them to narrow in on the meaning of a word that holds across its usages (Siskind, 1996; Yu and Smith, 2007; Smith and Yu, 2008). A number of computational models attempt to solve the mapping problem by implementing this mechanism, and have successfully replicated different patterns observed in child word learning (Siskind, 1996; Yu and Ballard, 2007; Fazly et al., 2010). These models have provided insight about underlying mechanisms of word learning, but none of them consider the semantic relations that hold among words, or how the semantic knowledge is structured. Recently, we have investigated properties of the semantic structure of the resulting (final) acquired knowledge of such a learner (Nematzadeh et al., 2014). However, that work did not address how such structural knowledge might develop and evolve incrementally within the learning model.

Models of Categorization. Computational models of categorization focus on the problem of forming semantic clusters given a defined set of features for words (Anderson and Matessa, 1992; Griffiths et al., 2007; Sanborn et al., 2010). Anderson and Matessa (1992) note that a cognitively plausible categorization algorithm needs to be incremental and only keep track of one potential partitioning; they propose a Bayesian framework (the Rational Model of Categorization or RMC) that specifies the joint distribution on features and category labels, and allows an unbounded number of clusters. Sanborn et al. (2010) examine different categorization models based on RMC. In particular, they compare the performance of the approximation algorithm of Anderson and Matessa (1992) (local MAP) with two other approximation algorithms (Gibbs Sampling and Particle Filters) in various human categorization paradigms. Sanborn et al. (2010) find that in most of the simulations the local MAP algorithm performs as well as the two other algorithms in matching human behavior.

The Representation of Semantic Knowledge. There is limited work on computational models of semantic acquisition that examine the representation of the semantic knowledge. Steyvers and Tenenbaum (2005) propose an algorithm for building a network with small-world and scale-free structure. The algorithm starts with a small complete graph, incrementally adds new nodes to the graph, and for each new node uses a probabilistic mechanism for selecting a subset of current nodes to connect to. However, their approach does not address the problem of learning word meanings or the semantic connections among them. Fountain and Lapata (2011) propose an algorithm for learning categories that also creates a semantic network by comparing all the possible word pairs. However, they too do not address the word learning problem, and do not investigate the structure of the learned semantic network to see whether it has the properties observed in adult knowledge.

3 The Incremental Network Model

We propose here a model that unifies the incremental acquisition of word meanings and formation of a semantic network structure that reflects the similarities among those meanings. We use an existing model to learn the meanings of words (Section 3.1), and use those incrementally developing meanings as the input to the algorithm proposed here for gradually growing a semantic network (Section 3.2).

3.1 The Word Learner

We use the model of Fazly et al. (2010); this learning algorithm is incremental and involves limited calculations, thus satisfying basic cognitive plausibility requirements. A naturalistic language learning scenario consists of linguistic data in the context of non-linguistic data, such as the objects,

Utterance: {*let*, *find*, *a*, *picture*, *to*, *color* }
Scene: {LET, PRONOUN, HAS_POSSESSION, CAUSE, ARTIFACT, WHOLE, CHANGE, ... }

Table 1: A sample utterance-scene pair.

events, and social interactions that a child perceives. This kind of input is modeled here as a pair of an *utterance* (the words a child hears) and a *scene* (the semantic features representing the meaning of those words), as shown in Table 1 (and described in more detail in Section 5.1). The word learner is an instance of cross-situational learning applied to a sequence of such input pairs: for each pair of a word w and a semantic feature f, the model incrementally learns $P(f|w)$ from co-occurrences of w and f across all the utterance-scene pairs.

For each word, the probability distribution over all semantic features, $P(.|w)$, represents the word's *meaning*. The estimation of $P(.|w)$ is made possible by introducing a set of latent variables, *alignments*, that correspond to the possible mappings between words and features in a given utterance–scene pair. The learning problem is then to find the mappings that best explain the data, which is solved by using an incremental version of the expectation–maximization (EM) algorithm (Neal and Hinton, 1998). We skip the details of the derivations and only report the resulting formulas.

The model processes one utterance-scene pair at a time. For the input pair processed at time t, first the probability of each possible alignment (alignment probability) is calculated as:[2]

$$P(a_{ij}|u, f_i) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in u} P_{t-1}(f_i|w')} \quad (1)$$

where u is the utterance, and a_{ij} is the alignment variable specifying the word w_j that is mapped to the feature f_i. $P_{t-1}(f_i|w_j)$ is taken from the model's current learned meaning of word w_j. Initially, $P_0(f_i|w_j)$ is uniformly distributed. After calculating the alignment probabilities, the learned meanings are updated as:

$$P_t(f_i|w_j) = \frac{\sum_{u \in U_t} P(a_{ij}|u, f_i)}{\sum_{f' \in \mathcal{M}} \sum_{u \in U_t} P(a_{ij}|u, f')} \quad (2)$$

where U_t is the set of utterances processed so far, and $\mathcal{M}$ is the set of features that the model has observed. Note that for each w–f pair, the value of the summations in this formula can be incrementally updated after processing any utterance that contains w; the summation does not have to be calculated at every step.

[2]This corresponds to the expectation step of EM.

3.2 Growing a Semantic Network

In our extended model, as we learn words incrementally (as above), we also structure those words into a semantic network based on the (partially) learned meanings. At any given point in time, the network will include as its nodes all the word types the word learner has been exposed to. Weighted edges (capturing semantic distance) will connect those pairs of word types whose learned meanings at that point are sufficiently semantically similar (according to a threshold). Since the probabilistic meaning of a word is adjusted each time it is observed, a word may either lose or gain connections in the network after each input is processed. Thus, to incrementally develop the network, at each time step, our algorithm must both examine existing connections (to see which edges should be removed) and consider potential new connections (to see which edges should be added).

A simple approach to achieve this is to examine the current semantic similarity between a word w in the input and all the current words in the network, and include edges between only those word pairs that are sufficiently similar. However, comparing w to all the words in the network each time it is observed is computationally intensive (and not cognitively plausible).

We present an approach for incrementally growing a semantic network that limits the computations when processing each input word w; see Algorithm 1. After the meaning of w is updated, we first check all the words that w is currently (directly) connected to, to see if any of those edges need to be removed, or have their weight adjusted. Next, to look for new connections for w, the idea is to select only a small subset of words, $\mathcal{S}$, to which w will be compared. The challenge then is to select $\mathcal{S}$ in a way that will yield a network whose semantic structure reasonably approximates the network that would result from full knowledge of comparing w to all the words.

Previous work has suggested picking "important" words (e.g., high-degree words) independently of the target word w — assuming these might be words for which a learner might need to understand their relationship to w in the future (Steyvers and Tenenbaum, 2005). Our proposal is instead to consider for $\mathcal{S}$ those words that are

Algorithm 1 Growing a network after each input u.

```
for all w in u do
  update P(.|w) using Eqn. (2)
  update current connections of w
  select S(w), a subset of words in the network
  for all w' in S(w) do
    if w and w' are sufficiently similar then
      connect w and w' with an edge
    end if
  end for
end for
```

likely to be similar to w. That is, since the network only needs to connect similar words to w, if we can guess what (some of) those words are, then we will do best at approximating the situation of comparing w to all words.

The question now is how to find semantically similar words to w that are not already connected to w in the network. To do so, we incrementally track semantic similarity among words usages as their meanings are developing. Specifically we cluster word tokens (not types) according to their current word meanings. Since the probabilistic meanings of words are continually evolving, incremental clusters of word tokens can capture developing similarities among the various usages of a word type, and be a clue to which words (types) w might be similar to. In the next section, we describe the Bayesian clustering process we use to identify potentially similar words.

3.3 Semantic Clustering of Word Tokens

We use the Bayesian framework of Anderson and Matessa (1992) to form semantic clusters.[3] Recall that for each word w, the model learns its meanings as a probability distribution over all semantic features, $P(.|w)$. We represent this probability distribution as a vector F whose length is the number of possible semantic features. Each element of the vector holds the value $P(f|w)$ (which is continuous). Given a word w and its vector F, we need to calculate the probability that w belongs to each existing cluster, and also allow for the possibility of it forming a new cluster. Using Bayes rule we have:

$$P(k|F) = \frac{P(k)P(F|k)}{\sum_{k'} P(k')P(F|k')} \quad (3)$$

[3]The distribution specified by this model is equivalent to that of a Dirichlet Process Mixture Model (Neal, 2000).

where k is a given cluster. We thus need to calculate the prior probability, $P(k)$, and the likelihood of each cluster, $P(F|k)$.

Calculation of Prior. The prior probability that word $n+1$ is assigned to cluster k is calculated as:

$$P(k) = \begin{cases} \frac{n_k}{n+\alpha} & n_k > 0 \\ \frac{\alpha}{n+\alpha} & n_k = 0 \text{ (new cluster)} \end{cases} \quad (4)$$

where n_k is the number of words in cluster k, n is the number of words observed so far, and α is a parameter that determines how likely the creation of a new cluster is. The prior favors larger clusters, and also discourages the creation of new clusters in later stages of learning.

Calculation of Likelihood. To calculate the likelihood $P(F|k)$ in Eqn. (3), we assume that the features are independent:

$$P(F|k) = \prod_{f_i \in F} P(f_i = v|k) \quad (5)$$

where $P(f_i = v|k)$ is the probability that the value of the feature in dimension i is equal to v given the cluster k. To derive $P(f_i|k)$, following Anderson and Matessa (1992), we assume that each feature given a cluster follows a Gaussian distribution with an unknown variance σ^2 and mean μ. (In the absence of any prior information about a variable, it is often assumed to have a Gaussian distribution.) The mean and variance of this distribution are inferred using Bayesian analysis: We assume the variance has an inverse χ^2 prior, where σ_0^2 is the prior variance and a_0 is the confidence in the prior variance:

$$\sigma^2 \sim \text{Inv-}\chi^2(a_0, \sigma_0^2) \quad (6)$$

The mean given the variance has a Gaussian distribution with μ_0 as the prior mean and λ_0 as the confidence in the prior mean.

$$\mu|\sigma \sim \text{N}(\mu_0, \frac{\sigma^2}{\lambda_0}) \quad (7)$$

Given the above conjugate priors, $P(f_i|k)$ can be calculated analytically and is a Student's t distribution with the following parameters:

$$P(f_i|k) \sim t_{a_i}(\mu_i, \sigma_i^2(1 + \frac{1}{\lambda_i})) \quad (8)$$

$$\lambda_i = \lambda_0 + n_k \quad (9)$$

$$a_i = a_0 + n_k \quad (10)$$

$$\mu_i = \frac{\lambda_0 \mu_0 + n_k \bar{f}}{\lambda_0 + n_k} \quad (11)$$

$$\sigma_i^2 = \frac{a_0 \sigma_0^2 + (n_k - 1)s^2 + \frac{\lambda_0 n_k}{\lambda_0 + n_k}(\mu_0 + \bar{f})^2}{a_0 + n_k} \quad (12)$$

where $\bar{f}$ and s^2 are the sample mean and variance of the values of f_i in k.

Note that in the above equations, the mean and variance of the distribution are simply derived by combining the sample mean and variance with the prior mean and variance while considering the confidence in the prior mean (λ_0) and variance (a_0). This means that the number of computations to calculate $P(F|K)$ is limited as w is only compared to the "prototype" of each cluster, which is represented by μ_i and σ_i of different features.

Adding a word w to a cluster. We add w to the cluster k with highest posterior probability, $P(k|F)$, as calculated in Eqn. (3).[4] The parameters of the selected cluster (k, μ_i, λ_i, σ_i, and a_i for each feature f_i) are then updated incrementally.

Using the Clusters to Select the Words in $\mathcal{S}(w)$. We can now form $S(w)$ in Algorithm 1 by selecting a given number of words n_s whose tokens are probabilistically chosen from the clusters according to how likely each cluster k is given w: the number of word tokens picked from each k is proportional to $P(k|F)$ and is equal to $P(k|F) \times n_s$.

4 Evaluation

We evaluate a semantic network in two regards: The semantic connectivity of the network – to what extent the semantically-related words are connected in the network; and the structure of the network – whether it exhibits a *small-world* and *scale-free* structure or not.

4.1 Evaluating Semantic Connectivity

The distance between the words in the network indicates their semantic similarity: the more similar a word pair, the smaller their distance. For word pairs that are connected via a path in the network, this distance is the weighted shortest path length between the two words. If there is no path between a word pair, their distance is considered to be ∞ (which is represented with a large number). We refer to this distance as the "learned" semantic similarity.

[4]This approach is referred to as local MAP (Sanborn et al., 2010); because of the incremental nature of the algorithm, it maximizes the current posterior distribution as opposed to the "global" posterior.

To evaluate the semantic connectivity of the learned network, we compare these learned similarity scores to "gold-standard" similarity scores that are calculated using the WordNet similarity measure of Wu and Palmer (1994) (also known as the WUP measure). We choose this measure since it captures the same type of similarity as in our model: words are considered similar if they belong to the same semantic category. Moreover, this measure does not incorporate information about other types of similarities, for example, words are not considered similar if they occur in similar contexts. Thus, the scores calculated with this measure are comparable with those of our learned network.

Given the gold-standard similarity scores for each word pair, we evaluate the semantic connectivity of the network based on two performance measures: coefficient of correlation and the median rank of the first five gold-standard associates. Correlation is a standard way to compare two lists of similarity scores (Budanitsky and Hirst, 2006). We create two lists, one containing the gold-standard similarity scores for all word pairs, and the other containing their corresponding learned similarity scores. We calculate the Spearman's rank correlation coefficient, ρ, between these two lists of similarity scores. Note that the learned similarity scores reflect the semantic distance among words whereas the WordNet scores reflect semantic closeness. Thus, a negative correlation is best in our evaluation, where the value of -1 corresponds to the maximum correlation.

Following Griffiths et al. (2007), we also calculate the median learned rank of the first five gold-standard associates for all words: For each word w, we first create a "gold-standard" associates list: we sort all other words based on their gold-standard similarity to w, and pick the five most similar words (associates) to w. Similarly, we create a "learned associate list" for w by sorting all words based on their learned semantic similarity to w. For all words, we find the ranks of their first five gold-standard associates in their learned associate list. For each associate, we calculate the median of these ranks for all words. We only report the results for the first three gold-standard associates since the pattern of results is similar for the fourth and fifth associates; we refer to the median rank of first three gold-standard associates as

1^{st}, 2^{nd}, and 3^{rd}.

4.2 Evaluating the Structure of the Network

A network exhibits a small-world structure when it is characterized by short path length between most nodes and highly-connected neighborhoods (Watts and Strogatz, 1998). We first explain how these properties are measured for a graph with N nodes and E edges. Then we discuss how these properties are used in assessing the small-world structure of a graph.[5].

Short path lengths. Most of the nodes of a small-world network are reachable from other nodes via relatively short paths. For a connected network (*i.e.*, all the node pairs are reachable from each other), this can be measured as the average distance between all node pairs (Watts and Strogatz, 1998). Since our networks are not connected, we instead measure this property using the median of the distances (d_{median}) between all node pairs (Robins et al., 2005), which is well-defined even when some node pairs have a distance of ∞.

Highly-connected neighborhoods. The neighborhood of a node n in a graph consists of n and all of the nodes that are connected to it. A neighborhood is maximally connected if it forms a complete graph —*i.e.*, there is an edge between all node pairs. Thus, the maximum number of edges in the neighborhood of n is $k_n(k_n-1)/2$, where k_n is the number of neighbors. A standard metric for measuring the connectedness of neighbors of a node n is called the *local clustering coefficient* (C) (Watts and Strogatz, 1998), which calculates the ratio of edges in the neighborhood of n (E_n) to the maximum number of edges possible for that neighborhood:

$$C = \frac{E_n}{k_n(k_n-1)/2} \tag{13}$$

The *local clustering coefficient* C ranges between 0 and 1. To estimate the connectedness of all neighborhoods in a network, we take the average of C over all nodes, *i.e.*, C_{avg}.

Small-world structure. A graph exhibits a small-world structure if d_{median} is relatively small and C_{avg} is relatively high. To assess this for a graph g, these values are typically compared to those of a random graph with the same number of nodes and edges as g (Watts and Strogatz, 1998; Humphries and Gurney, 2008). The random graph is generated by randomly rearranging the edges of the network under consideration (Erdos and Rényi, 1960). Because any pair of nodes is equally likely to be connected as any other, the median of distances between nodes is expected to be low for a random graph. In a small-world network, this value d_{median} is expected to be as small as that of a random graph: even though the random graph has edges more uniformly distributed, the small-world network has many locally-connected components which are connected via *hubs*. On the other hand, C_{avg} is expected to be much higher in a small-world network compared to its corresponding random graph, because the edges of a random graph typically do not fall into clusters forming highly connected neighborhoods.

Given these two properties, the "small-worldness" of a graph g is measured as follows (Humphries and Gurney, 2008):

$$\sigma_g = \frac{\frac{C_{avg}(g)}{C_{avg}(random)}}{\frac{d_{median}(g)}{d_{median}(random)}} \tag{14}$$

where $random$ is the random graph corresponding to g. In a small-world network, it is expected that $C_{avg}(g) \gg C_{avg}(random)$ and $d_{median}(g) \geq d_{median}(random)$, and thus $\sigma_g > 1$.

Note that Steyvers and Tenenbaum (2005) made the empirical observation that small-world networks of semantic knowledge had a single connected component that contained the majority of nodes in the network. Thus, in addition to σ_g, we also measure the relative size of a network's largest connected component having size N_{lcc}:

$$\text{size}_{\text{lcc}} = \frac{N_{lcc}}{N} \tag{15}$$

Scale-free structure. A scale-free network has a relatively small number of *high-degree* nodes that have a large number of connections to other nodes, while most of its nodes have a small degree, as they are only connected to a few nodes. Thus, if a network has a scale-free structure, its degree distribution (*i.e.*, the probability distribution of degrees over the whole network) will follow a power-law distribution (which is said to be "scale-free"). We evaluate this property of a network by plotting its degree distribution in the logarithmic scale, which (if a power-law distribution) should appear as a straight line. None of our networks ex-

[5]We take the description of these measures from Nematzadeh et al. (2014)

hibit a scale-free structure; thus, we do not report the results of this evaluation, and leave it to future work for further investigation.

5 Experimental Set-up

5.1 Input Representation

Recall that the input to the model consists of a sequence of utterance–scene pairs intended to reflect the linguistic data a child is exposed to, along with the associated meaning a child might grasp. As in much previous work (Yu and Ballard, 2007; Fazly et al., 2010), we take child-directed utterances from the CHILDES database (MacWhinney, 2000) in order to have naturalistic data. In particular, we use the Manchester corpus (Theakston et al., 2001), which consists of transcripts of conversations with 12 British children between the ages of $1;8$ and $3;0$. We represent each utterance as a bag of lemmatized words (see Utterance in Table 1).

For the scene representation, we have no large corpus to draw on that encodes the semantic portion of language acquisition data.[6] We thus automatically generate the semantics associated with an utterance, using a scheme first introduced in Fazly et al. (2010). The idea is to first create an input generation lexicon that provides a mapping between all the words in the input data and their associated *meanings*. A scene is then represented as a set that contains the meanings of all the words in the utterance. We use the input generation lexicon of Nematzadeh et al. (2012) because the word meanings reflect information about their semantic categories, which is crucial to forming the semantic clusters as in Section 3.3.

In this lexicon, the "true" meaning for each word w is a vector over a set of possible semantic features for each part of speech; in the vector, each feature is associated with a *score* for that word (see Figure 1). Depending on the word's part of speech, the features are extracted from various

[6] Yu and Ballard (2007) created a corpus by hand-coding the objects and cues that were present in the environment, but that corpus is very small. Frank et al. (2013) provide a larger manually annotated corpus (5000 utterances), but it is still very small for longitudinal simulations of word learning. (Our corpus contains more than 100,000 utterances.) Moreover, the corpus of Frank et al. (2013) is limited because a considerable number of words are not semantically coded. (Only a subset of concrete objects in the environment are coded.)

apple: { FOOD:1, SOLID:.72, $\cdots$, PLANT-PART:.22, PHYSICAL-ENTITY:.17, WHOLE:.06, $\cdots$ }

Figure 1: Sample true meaning features & their scores for *apple* from Nematzadeh et al. (2012).

lexical resources such as WordNet[7], VerbNet[8], and Harm (2002). The score for each feature is calculated using a measure similar to tf-idf that reflects the association of the feature with the word and with its semantic category: term frequency indicates the strength of association of the feature with the word, and inverse document frequency (where the documents are the categories) indicates how informative a feature is for that category. The semantic categories of nouns (which we focus on in our networks) are given by WordNet lex-names[9], a set of 25 general categories of entities. (We use only nouns in our semantic networks because the semantic similarity of words with different parts of speech cannot be compared, since their semantic features are drawn from different resources.)

The input generation lexicon is used to generate a scene representation for an utterance as follows: For each word w in the utterance, we probabilistically sample features, in proportion to their score, from the full set of features in its true meaning. The probabilistic sampling allows us to simulate the noise and uncertainty in the input a child perceives by omitting some meaning features from the scene. The scene representation is the union of all the features sampled for all the words in the utterance (see Scene in Table 1).

5.2 Methods

We experiment with our network-growth method that draws on the incremental clustering, and create "upper-bound" and baseline networks for comparison. Note that all the networks are created using our Algorithm 1 (page 4) to grow networks incrementally, drawing on the learned meanings of words and updating their connections on the basis of this evolving knowledge. The only difference in creating the networks resides in how the comparison set $\mathcal{S}(w)$ is chosen for each target word w that is being added to the growing network at each time step. We provide more details in the paragraphs below.

[7] http://wordnet.princeton.edu

[8] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[9] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html

Upper-bound. Recall that one of our main goals is to substantially reduce the number of similarity comparisons needed to grow a semantic network, in contrast to the straightforward method of comparing each w to all current words. At the same time, we need to understand the impact of the increased efficiency on the quality of the resulting networks. We thus need to compare the target properties of our networks that are learned using a small comparison set $\mathcal{S}$, to those of an "upper-bound" network that takes into account all the pair-wise comparisons among words. We create this upper-bound network by setting $\mathcal{S}(w)$ to contain all words currently in the network.

Baselines. On the other hand, we need to evaluate the (potential) benefit of our cluster-driven selection process over a more simplistic approach to selecting $\mathcal{S}(w)$. To do so, we consider three baselines, each using a different criteria for choosing the comparison set $\mathcal{S}(w)$: The Random baseline chooses the members of this set randomly from the set of all observed words. The Context baseline can be seen as an "informed" baseline that attempts to incorporate some semantic knowledge: Here, we select words that are in the recent context prior to w in the input, assuming that such words are likely to be semantically related to w. We also include a third baseline, Random+Context, that picks half of the members of $\mathcal{S}$ randomly and half of them from the prior context.

Cluster-based Methods. We report results for three cluster-based networks that differ in their choice of $\mathcal{S}(w)$ as follows: The Clusters-only network chooses words in $\mathcal{S}(w)$ from the set of clusters, proportional to the probability of each cluster k given word w (as explained in Section 3.3). In order to incorporate different types of semantic information in selecting $\mathcal{S}$, we also create a Clusters+Context network that picks half of the members of $\mathcal{S}$ from clusters (as above), and half from the prior context. For completeness, we include a Clusters+Random network that similarly chooses half of words in $\mathcal{S}$ from clusters and half randomly from all observed words.

We have experimented with several other methods, but they all performed substantially worse than the baselines, and hence we do not report them here. E.g., we tried picking words in $\mathcal{S}$ from the best cluster. We also tried a few methods inspired by (Steyvers and Tenenbaum, 2005): E.g., we examined a method where if a member of $\mathcal{S}(w)$ was sufficiently similar to w, we added the direct neighbors of that word to $\mathcal{S}$. We also tried to grow networks by choosing the members of $\mathcal{S}$ according to the degree or frequency of nodes in the network.

5.3 Experimental Parameters

We use $20,000$ utterance–scene pairs as our training data. Recall that we use clustering to help guide our semantic network growth algorithm. Given the clustering algorithm in Section 3.3, we are interested to find the set of clusters that best explain the data. (Other clustering algorithms can be used instead of this algorithm.) We perform a search on the parameter space, and select the parameter values that result in the best clustering, based on the number of clusters and their average F-score. The value of the clustering parameters are as follows: $\alpha = 49$, $\lambda_0 = 1.0$, $a_0 = 2.0$, $\mu_0 = 0.0$, and $\sigma_0 = 0.05$. Two nouns with feature vectors F_1 and F_2 are connected in the network if $cosine(F1, F2)$ is greater than or equal to 0.6. (This threshold was selected following empirical examination of the similarity values we observe among the "true" meaning in our input generation lexicon.) The weight on the edge that connects these nouns specifies their semantic distance, which is calculated as $1 - cosine(F1, F2)$.

Because we aim for a network creation method that is cognitively plausible in performing a limited number of word-to-word comparisons, we need to ensure that all the different methods of selecting the comparison set $\mathcal{S}(w)$ yield roughly similar numbers of such comparisons. Keeping the size of $\mathcal{S}$ constant does not guarantee this, because each method can yield differing numbers of connections of the target word w to other words. We thus parameterize the size of $\mathcal{S}$ for each method to keep the number of computations similar, based on experiments on the development data. In development work we also found that having an increasing size of $\mathcal{S}$ over time improved the results, as more words were compared as the knowledge of learned meanings improved. To achieve this, we use a percentage of the words in the network as the size of $\mathcal{S}$. In practice, the setting of this parameter yields a number of comparisons across all methods that is about 8% of the maximum possible word-to-word comparisons that would be performed in the naive (computationally intensive) approach.

Note that all the Cluster-based, Random and Random+Context methods include a random selection mechanism; thus, we run each of these methods 50 times and report the average ρ, median ranks and size_{lcc} (see Section 4). For the networks (out of 50 runs) that exhibit a small-world structure (small-worldness greater than one), we report the average small-worldness. We also report the percentage of runs whose resulting network exhibit a small-world structure.

6 Experimental Results and Discussion

Table 2 presents our results, including the evaluation measures explained above, for the Upper-bound, Baseline, and Cluster-based networks created by the various methods described in Section 5.2.[10]

Recall that the Upper-bound network is formed from examining a word's similarity to all other (observed) words when it is added to the network. We can see that this network is highly connected (0.85) and has a small-world structure (5.5). There is a statistically significant correlation of the network's similarity measures with the gold standard ones (-0.38). For this Upper-bound structure, the median ranks of the first three associates are between 31 and 42. These latter two measures on the Upper-bound network give an indication of the difficulty of learning a semantic network whose knowledge matches gold-standard similarities.

Considering the baseline networks, we note that the Random network is actually somewhat better (in connectivity and median ranks) than the Context network that we thought would provide a more informed baseline. Interestingly, the correlation value for both networks is no worse than for the Upper-bound. The combination of Random+Context yields a slightly lower correlation, and no better ranks or connectivity than Random. Note that none of the baseline networks exhibit a small world structure ($\sigma_g \ll 1$ for all three, except for one out of 50 runs for the Random method).

Recall that the Random network is not a network resulting from randomly connecting word pairs, but one that incrementally compares each target word with a set of randomly chosen words when considering possible new connections. We suspect that this approach performs reasonably well because it enables the model to find a broad range of similar words to the target; this might be effective especially because the learned meanings of words are changing over time.

Turning to the Cluster-based methods, we see that indeed some diversity in the comparison set for a target word might be necessary to good performance. We find that the measures on the Clusters-only network are roughly the same as on the Random one, but when we combine the two in Clusters+Random we see an improvement in the ranks achieved. It is possible that the selection from clusters does not have sufficient diversity to find some of the valid new connections for a word.

We note that the best results overall occur with the Clusters+Context network, which combines two approaches to selecting words that have good potential to be similar to the target word. The correlation coefficient for this network is at a respectable 0.36, and the median ranks are the second best of all the network-growth methods. Importantly, this network shows the desired small-world structure in most of the runs (77%), with the highest connectivity and a small-world measure well over 1.

The fact that the Clusters+Context network is better overall than the networks of the Clusters-only and Context methods indicates that both clusters and context are important in making "informed guesses" about which words are likely to be similar to a target word. Given the small number of similarity comparisons used in our experiments (only around 8% of all possible word-to-word comparisons), these observations suggest that both the linguistic context and the evolving relations among word usages (captured by the incremental clustering of learned meanings) contain information crucial to the process of growing a semantic network in a cognitively plausible way.

7 Conclusions

We propose a unified model of word learning and semantic network formation, which creates a network of words in which connections reflect structured knowledge of semantic similarity between words. The model adheres to the cognitive plausibility requirements of incrementality and use of limited computations. That is, when incrementally adding or updating a word's connections in the network, the model only looks at a subset of words rather than comparing the target word to all the nodes in the network. We demonstrate that

[10] All the reported co-efficients of correlation (ρ) are statistically significant at $p < 0.01$.

Comparing all Pairs

Method	Semantic Connectivity				Small World	
	ρ	1^{st}	2^{nd}	3^{rd}	size_{lcc}	σ_g (%)
Upper-bound	-0.38	31	41	42	0.85	5.5
Baselines						
Random	-0.38	56	76.9	68.9	0.6	5.2 (2)
Context	$\mathbf{-0.39}$	97	115	89	0.5	0
Random+Context	-0.36	63.3	87.2	79.1	0.6	0 (0)
Cluster-based Methods						
Clusters-only	-0.32	58.6	72.0	71.6	**0.7**	5.5 (43)
Clusters+Context	-0.36	53.9	67.6	64.8	**0.7**	**7.2** (**77**)
Clusters+Random	-0.35	**48.1**	**61.2**	**58.1**	**0.7**	6.9 (48)

Table 2: Connectivity and small-worldness measures for the Upper-bound, Baseline, and Cluster-based network-growth methods; best performances across the Baseline and Cluster-based methods are shown in bold. ρ: co-efficient of correlation between similarities of word pairs in network and in gold-standard; 1^{st}, 2^{nd}, 3^{rd}: median ranks of corresponding gold-standard associates given network similarities; size_{lcc}: proportion of network in the largest connected component; σ_g: overall "small-worldness", should be greater than 1; %: the percentage of runs whose resulting networks exhibit a small-world structure. Note there are 1074 nouns in each network.

using the evolving knowledge of semantic connections among words as well as their context of usage enables the model to create a network that shows the properties of adult semantic knowledge. This suggests that the information in the semantic relations among words and their context can efficiently guide semantic network growth.

Acknowledgments

We would like to thank Varada Kolhatkar for valuable discussion and feedback. We are also grateful for the financial support from NSERC of Canada, and University of Toronto.

References

John R. Anderson and Michael Matessa. 1992. Explorations of an incremental, bayesian algorithm for categorization. *Machine Learning*, 9(4):275–308.

Lois Bloom. 1973. *One word at a time: The use of single word utterances before syntax*, volume 154. Mouton The Hague.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word.

Allan M. Collins and Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Eliana Colunga and Linda B. Smith. 2005. From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 112(2):347–382.

Paul Erdos and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063.

Trevor Fountain and Mirella Lapata. 2011. Incremental models of natural language category acquisition. In *Proceedings of the 32st Annual Conference of the Cognitive Science Society*.

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers referential intentions to model early cross-situational word learning. *Psychological Science*.

Michael C. Frank, Joshua B. Tenenbaum, and Anne Fernald. 2013. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, 9(1):1–24.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

Michael W. Harm. 2002. Building large scale distributed semantic feature sets with WordNet. Technical Report PDP.CNS.02.1, Carnegie Mellon University.

Mark D. Humphries and Kevin Gurney. 2008. Network small-world-ness: a quantitative method for determining canonical network equivalence. *PLoS One*, 3(4):e0002051.

Susan S. Jones and Linda B. Smith. 2005. Object name learning and object perception: a deficit in late talkers. *J. of Child Language*, 32:223–240.

Susan S. Jones, Linda B. Smith, and Barbara Landau. 1991. Object properties and knowledge in early lexical learning. *Child Development*, 62(3):499–516.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 2: The Database. Erlbaum, 3rd edition.

Radford M. Neal and Geoffrey E. Hinton. 1998. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Radford M. Neal. 2000. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. 2012. Interaction of word learning and semantic category formation in late talking. In *Proc. of CogSci'12*.

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. 2014. Structural differences in the semantic networks of simulated word learners.

Thierry Poibeau, Aline Villavicencio, Anna Korhonen, and Afra Alishahi, 2013. *Computational Modeling as a Methodology for Studying Human Language Learning*. Springer.

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press.

Terry Regier. 2005. The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29:819–865.

Garry Robins, Philippa Pattison, and Jodie Woolcock. 2005. Small and other worlds: Global network structures from local processes1. *American Journal of Sociology*, 110(4):894–936.

Larissa K. Samuelson and Linda B. Smith. 1999. Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition*, 73(1):1 – 33.

Adam N. Sanborn, Thomas L. Griffiths, and Daniel J. Navarro. 2010. Rational approximations to rational models: alternative algorithms for category learning.

Jeffery Mark Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91.

Linda B. Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78.

Anna L. Theakston, Elena V. Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *J. of Child Language*, 28:127–152.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Chen Yu and Dana H. Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(1315):2149 – 2165. Selected papers from the 3rd International Conference on Development and Learning (ICDL 2004), Time series prediction competition: the CATS benchmark.

Chen Yu and Linda B. Smith. 2007. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420.