Semantic Models as a Combination of Free Association Norms and Corpus-based Correlations

Derrall Heath, David Norton, Eric Ringger, and Dan Ventura
Computer Science Department
Brigham Young University
Email: dheath@byu.edu, dnorton@byu.edu, ringger@cs.byu.edu, ventura@cs.byu.edu

Abstract—

We present computational models capable of understanding and conveying concepts based on word associations. We discover word associations automatically using corpus-based semantic models with Wikipedia as the corpus. The best model effectively combines corpus-based models with preexisting databases of free association norms gathered from human volunteers. We use this model to play human-directed and computer-directed word guessing games (games with a purpose similar to *Catch Phrase* or *Taboo*) and show that this model can measurably convey and understand some aspect of word meaning. The results highlight the fact that human-derived word associations and corpus-derived word associations can play complementary roles in semantic models.

I. Introduction

Language is a critical component of human intelligence, and the development of computer systems that can understand and communicate through language is an important problem in the field of artificial intelligence. Building computational models that provide meaning to words is a step in that direction. Most words are a representation of a concept, and it is the concept itself in which we are interested and thus, the terms 'concept' and 'word' will be used interchangeably throughout the paper.

The study of word meaning and conceptual knowledge is called *lexical semantics* (in linguistics), *semantic memory* (in cognitive psychology), or *cognitive semantics* (in cognitive linguistics). This question of what gives words meaning has been debated for years; however, it is commonly agreed that a word, at least in part, is given meaning by how the word is used in conjunction with other words (i.e., its context) [1], [2]. Many computational semantic models consist of building associations between words [3], [4]. These word associations essentially form a large graph that is typically referred to as a *semantic network*.

Word associations are commonly acquired in one of two ways: explicitly from people and automatically by inferring them from a corpus. ConceptNet [5] and WordNet [6] are examples of semantic networks that have been created explicitly by hand (or through crowd sourcing). In these networks, words are linked by specific types of relationships that are often intended for specific purposes. Although these networks have been applied to problems such as common sense reasoning [7], they are often either limited in their vocabulary, limited in their variety of word associations, or do not provide any notion of relationship strength.

Corpus-based semantic models (CSMs) are a class of computational models that attempt to learn semantic information from patterns of word co-occurrences in a corpus. These models are based on the idea that similar words will occur in similar contexts and words that are often associated together will often co-occur close together. CSMs have been successfully used on a variety of tasks such as information retrieval [8], multiple choice vocabulary tests [9], multiple choice synonym questions from the TOEFL test [10], and multiple choice analogy questions from the SAT test [11].

Free Association Norms (FANs) are a common means of gathering word associations from people and are considered to be one of the best methods for understanding how people, in general, associate words in their own minds [12]. Thus, the corpus-based models are often compared directly with FANs as a way to evaluate the quality of word associations discovered from corpora [13], [14]. FANs are rarely used themselves to help solve word similarity tasks and exist only as a baseline metric or to be analyzed directly by cognitive scientists.

Most CSMs have been applied to word similarity tasks such as the previously mentioned multiple choice synonymy test and clustering words into predefined groups. However, it would be beneficial for a semantic model to be able to determine a concept given a description of that concept without multiple choice options. For example, if the model were given a description of a word, it should be able to determine what that word is. Conversely, given a word, the model should be able to provide a description of that word that makes sense to humans.

We introduce a new task that involves using CSMs to play word guessing games (similar to *Catch Phrase* or *Taboo*) with people online. The idea is to evaluate how well a computer system that uses word associations can understand and convey concepts to humans. These word guessing games are designed for two purposes: for evaluation, and as a novel way of collecting new viable word associations. Thus, these guessing games are also contributions to the growing field of *Games with a Purpose* [15]. We play these word guessing games using Free Association Norms, using three common CSMs, and using a hybrid approach combining FANs with CSMs. We show that using a hybrid approach improves the ability of the system to play these guessing games and therefore can, at least partially, convey meaning to humans.

We will first describe our methodology in using FANs and our method for building the three CSMs used in this paper. We will then outline the initial experiments and results used to evaluate the models. Finally, we will detail the online word guessing game and discuss the results and applications.

II. Methodology

We want to create a system that can communicate and understand concepts. Given a concept, we want the system to provide a description that will enable a human to know what the given concept is. Conversely, given a description, we want the system to know to which concept the description is referring. For example, suppose the given concept is ‘space’, the system could provide a description in the form of other words associated with ‘space’ such as ‘planet’, ‘astronaut’, ‘star’, ‘rocket’, ‘dark’, and ‘mysterious’. Conversely, if the system is given a collection of words as a description such as ‘soldier’, ‘guns’, ‘bomb’, ‘death’, ‘fight’, ‘sorrow’, and ‘courage’, then the system should infer that the collection of words most likely represents the concept ‘war’.

We use the term ‘description’ here to mean ‘a collection of other words’. In this paper we are not concerned with how words are structured together. We acknowledge that structuring of sentences, relationship types, and word order contribute to the meaning of concepts. Indeed much of the recent work on CSMs deal with trying to automatically infer additional semantic information such as word order, sentence structure, and word relationship types [16], [17], [18]. However, we are interested in the degree to which simple word associations can accomplish the task at hand and leave these more advanced CSMs to future work.

We present a computational semantic model that combines human free association norms with common corpus-based approaches. The idea is to use the FANs to capture general knowledge and then fill in the gaps using a CSM. Here we describe each individual model, initial testing results, and how the individual models will be combined.

1) Lemmatization and Stop Words: We use the standard practice of removing stop words (words like ‘the’ and ‘of’) and lemmatizing (representing different forms of the same word with the word’s morphological lemma) as we build word associations. WordNet maintains a database of word forms and hence, we use WordNet to perform the lemmatization [6]. It should be noted, however, that lemmatization with WordNet has its limits. For one, we cannot lemmatize a word across different parts of speech (noun, verb, adjective, etc). For example, ‘redeem’ and ‘redeeming’ will remain separate words because ‘redeeming’ could be the gerund form of the verb ‘redeem’ or it could be an adjective (i.e., ‘a redeeming quality’). Since the part of speech is not provided for individual words encountered, we must account for all parts of speech, hence words like ‘relax’, ‘relaxing’ and ‘relaxation’ remain separate words.

A. Free Association Norms

We use two preexisting databases of human word associations: The Edinburgh Associative Thesaurus [19] and University of Florida’s Word Association Norms [12]. These databases were built by asking hundreds of human volunteers to provide the first word that comes to mind when given a cue word. This technique, called *free association*, is able to capture many different types of word associations including word co-ordination (pepper, salt), collocation (trash, can), super-ordination (insect, butterfly), synonymy (starving, hungry), and antonymy (good, bad).

We build a semantic model from this data as follows: The association strength between two words is simply a count of the number of volunteers that said the second word given the first word. We also consider word associations to be undirected. In other words, if word A is associated with word B, then word B is associated with word A. Hence, when we encounter data in which word A is a cue for word B and word B is also a cue for word A, we combine them into a single association pair by adding their respective association strengths. Between these two databases, there are a total of 19,310 (lemmatized) unique words and 288,069 unique associations. From now on, we will refer to this model as *FANs*.

B. Corpus-based Semantic Models

One of the most popular CSMs is *Latent Semantic Analysis* (LSA) [20], [2]. LSA is based on the idea that similar words will appear in similar documents (or contexts). LSA builds either a term × document or a term × term matrix from a corpus and then performs Singular Value Decomposition (SVD), which reduces the given large sparse matrix to a low-rank approximation of that matrix along with a set of vectors, each representing a word (as well as a set of vectors for each document). These vectors also represent points in semantic space, and the closer a word’s vector is to another in this space, the closer they are in meaning (and the stronger the association between words). Starting with a term × term matrix is considered advantageous because the size of the matrix is invariant to the size of the corpus. It is also argued by some that it is more congruent to human cognition than the term × document matrix used in some implementations of LSA [13], [21]. We implement the same version of LSA that is used in [10], which uses a term × term matrix and a co-occurrence window of ± 2.

Another popular method is the *Hyperspace Analog to Language* (HAL) model [22]. This model is based on the same idea as LSA, except the notion of word order is partially captured in the co-occurrence matrix (with a co-occurrence window of ± 10), and HAL then uses the co-occurrence counts directly as vectors representing each word in semantic space. We use the same implementation as specified in the original paper [22].

The third CSM we use is constructed from the direct co-occurrence counts (DCC) obtained from the corpus. We build a term × term co-occurrence matrix M using a co-occurrence window of ±30. To account for the fact that common words will have generally higher co-occurrence counts, we scale these counts by weighting each element of the matrix by the inverse of the total frequency of both words at each element. This is done by considering each element $M_{i,j}$, then adding the total number of occurrences of each word (i and j), subtracting out the value at $M_{i,j}$ (to avoid counting it twice), then dividing $M_{i,j}$ by this computed number, as follows:

$$M_{i,j} \leftarrow \frac{M_{i,j}}{(\sum\limits_{i} M_{i,j} + \sum\limits_{j} M_{i,j} - M_{i,j})} \tag{1}$$

The result could be a very small number and hence, we then also normalize the values between 0 and 1. Once the co-occurrence matrix is built from the corpus, we use the weighted/normalized co-occurrence values themselves as association strengths between words.

Note that the DCC model only captures first order relationships between words, while HAL and LSA capture higher order relationships. That is to say, DCC will only associate words that co-occur often (e.g., ‘dog’, ‘kennel’), while HAL and LSA can associate words that co-occur with similar words even if those words never co-occur directly (e.g., ‘dog’, ‘puppy’). For each of the models (LSA, HAL, and DCC), we use the Wikipedia corpus as it is large, easily accessible, and covers a wide range of human knowledge [23]. Initially, for comparison we limit the vocabulary to the same 19,310 (lemmatized) words that exist in the FANs database.

C. TOEFL Synonymy Test

We take a detour to conduct an initial test to compare the performance of each model on the standard TOEFL multiple choice synonym test [2]. We consider three versions of the results. The first is simply the number of questions correctly answered out of the 80 total questions (AllQ). The second is limited to only the questions in which the word being considered and the correct answer exist in the vocabulary (VocabQ). The third is limited to only the questions in which the model has an association between the word being considered and the correct answer (AssocQ). Note that if a model cannot answer the question, it is counted as wrong; we do not adjust the score for random guessing. The results for FANs, HAL, DCC, and LSA can be seen in the first four rows of Table I.

The first thing to note is that for all questions (AllQ), the four models perform poorly (the human standard for actual TOEFL test takers is 0.645). The obvious reason for the poor results is that the vocabulary is limited to the 19,310 lemmatized words. When throwing out questions that are not in the vocabulary (VocabQ), the scores improve considerably. The vocabulary for the CSMs can be easily expanded and since LSA and DCC perform the best, we expand their vocabulary to 43,578 lemmatized words, called LSA2 and DCC2. When we do this, the TOEFL scores for AllQ are increased to 0.888 and 0.738 respectively. LSA2’s score is comparable to previous TOEFL results for this implementation of LSA, which achieved a score of 0.925 (the differences likely due to different corpus, and not accounting for random guessing) [10].

	AllQ	**VocabQ**	**AssocQ**
FANs	0.300(24/80)	0.511(24/47)	**0.923(24/26)**
HAL	0.388(31/80)	0.660(31/47)	0.660(31/47)
DCC	0.438(35/80)	0.745(35/47)	0.745(35/47)
LSA	0.513(41/80)	0.872(41/47)	0.872(41/47)
DCC2	0.738(59/80)	0.747(59/79)	0.747(59/79)
LSA2	0.888(71/80)	0.899(71/79)	0.899(71/79)
FAN-LSA2	**0.900(72/80)**	**0.911(72/79)**	0.911(72/79)

TABLE I

THE TOEFL SYNONYM TEST SCORES FOR THE DIFFERENT MODELS. ALLQ IS THE RAW SCORE FOR ALL 80 QUESTIONS. VOCABQ IS THE SCORE BASED ON THE LIMITED VOCABULARY AND ASSOCQ IS THE SCORE BASED ON EXISTING ASSOCIATIONS IN EACH MODEL. LSA2 USES AN EXTENDED VOCABULARY AND PERFORMS THE BEST WHEN COMBINED WITH FANS (FAN-LSA2). FANS PERFORMS THE BEST WHEN THERE EXISTS AN ASSOCIATION.

FANs perform the worst except when considering only questions in which an association score between the words exists (AssocQ). The AssocQ scores for the three CSM models are the same as the VocabQ scores because a similarity score can be computed between any pair of words. FANs are limited in their number of associations because obtaining them is a tedious process of receiving input from people. However, when an association does exist, FANs achieve the best score of 0.923.

We can use the FANs to augment the LSA2 model and build a hybrid model (FAN-LSA2) that improves the results (see last row in Table I). FAN-LSA2 simply defers to the FANs model first and if no association exists between the words in question, then the LSA2 model is used.

D. Combining Models

From the TOEFL test we can see that FANs and CSMs have different strengths and weaknesses. FANs are limited in vocabulary, are limited in the number of associations between words, and are difficult to acquire. However, the associations that do exist are meaningful and they capture the most relevant associations. CSMs, on the other hand, can automatically discover associations with a large vocabulary, but it is difficult to tell how meaningful the associations are. Other studies have shown that FANs and CSMs each provide different types of word associations [13]. A combination of these methods into a single model has the potential to take advantage of the strengths of each method, as indicated by the improved performance of FAN-LSA2 in the TOEFL test. The hypothesis is that the combined model will better communicate meaning to a person than either model individually because it presents a wider range of associations.

1) Combining Method: This method merges two separate databases of word associations into a single database before querying it for associations. This method assumes that FANs contain more valuable word associations than the CSMs because FANs are typically used as the gold standard in the literature. However, CSMs do contain some valuable associations not present in the FANs. The idea is to add the top n

associations for each word from one of the CSMs to the FANs but to weight the association strength low. This is beneficial for two reasons. First, if there are any associations that overlap, adding them again will strengthen the association in the combined database. Second, new associations not present in the FANs will be added to the combined database and provide a greater variety of word associations. We keep the association strength low because we want the CSM data to reinforce, but not dominate, the FANs.

We first copy all word associations from the FANs to the combined database. Next, let W be the set of all unique words in the vocabulary, let $A_{i,n} \subseteq W$ be the set of the top n words associated with word $i \in W$ from the CSM, let $score_{i,j}$ be the association strength between words i and j from the CSM, let max_i be the maximum association score present in the FANs for word i, and let θ be a scaling factor. Now for each $i \in W$ and for each $j \in A_{i,n}$, the new association score between words i and j is computed as follows:

$$new\_score_{i,j} \leftarrow (max_i \cdot \theta) \cdot score_{i,j} \quad (2)$$

This equation scales $score_{i,j}$ (which is already normalized) to lie between 0 and a certain percentage (θ) of max_i. The n associated words from the CSM are then added to the combined database with the updated scores ($new\_score_{i,j}$). If the word pair is already in the database, then the updated score is added to the score already present. For the results presented in this paper we use $n = 20$ and $\theta = 0.2$, which were determined based on preliminary experiments.

III. Word Guessing Game

The models are evaluated by playing a word guessing game called *Wordlery* (similar to *Catch Phrase* or *Taboo*), which is accessed through an online interface (http://darci.cs.byu.edu/Wordlery/). There are two modes: one in which the user must guess the word (user_guess) and one in which the system (or model) must guess the word (system_guess). In the user_guess mode, the system presents the user with a set of eight words. The user then has seven chances to guess the concept that the words represent. We record whether or not the user is able to guess the word and how many guesses it took. Figure 1 shows the user interface for user_guess mode. The eight words presented by the system (on the right in the figure) are the top n word associations for the hidden word, where $n = 8$. This mode is similar to one of the evaluation metrics used in another study, in which human volunteers had a single chance to guess the word that generated a list of associated words [22].

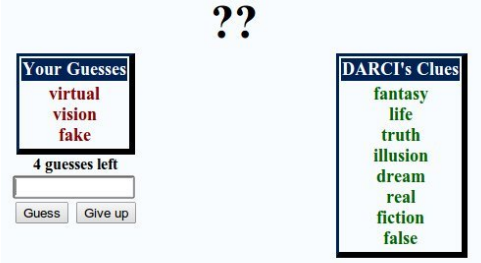

Fig. 1. User interface for the user_guess mode. On the left the user attempts to guess the concept the system is trying to communicate through the word associations on the right.

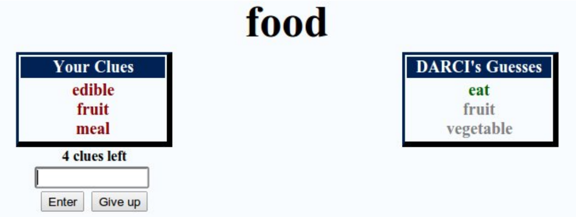

Fig. 2. User interface for the system_guess mode. The system attempts to guess the concept ('food') from user provided word association clues on the left.

In the system_guess mode, the user is presented with a word/concept and can then provide up to seven other words one at a time as clues to the system. Figure 2 shows the user interface for this mode. After each word provided by the user (on the left), the system gives its current guess (on the right along with its previous guesses) until either it guesses the correct word or the number of allotted clues has been reached. System guesses are generated in a few simple steps. First, the system retrieves the words associated with each user-provided clue. Second, for each associated word retrieved, the system counts the number of user-provided words with which it is associated. The system also sums the association strengths between each associated word and each of the user-provided clues for a total association strength. In the third step, the system ranks the associated words first by their frequency, then by their total association strength. Finally, the top word (that hasn't already been guessed) is returned. See Figure 3 for an example of this process.

At each round of the Wordlery game, the system randomly selects a word from the vocabulary of available words. The system then randomly selects a mode, either user_guess or system_guess. Finally, the system randomly selects one of the semantic models being evaluated, and the word guessing game is played by the user. To evaluate the effectiveness of each method we use the *win/loss record*, or the proportion of games

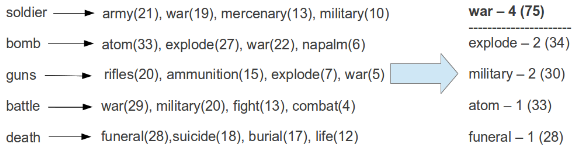
soldier
army(21), war(19), mercenary(13), military(10)
bomb
atom(33), explode(27), war(22), napalm(6)
guns
rifles(20), ammunition(15), explode(7), war(5)
battle
war(29), military(20), fight(13), combat(4)
death
funeral(28),suicide(18), burial(17), life(12)
war – 4 (75)
explode – 2 (34)
military – 2 (30)
atom – 1 (33)
funeral – 1 (28)

Fig. 3. A simplified example of how the model guesses a concept given a set of words. The words on the left are the user-provided clues, while the words to the right of the arrows are the lists of associated words for each clue, with their association strength. The system then sorts the associated words by their frequency (the number of different "clue-relation" lists in which a word appears), then by their total association strength. The top word is returned as the guess.

1 Guess/Clue	**HAL**	**DCC**	**DCC2**	**LSA**	**LSA2**	**FAN**	**FAN-DCC**	**FAN-DCC2**	**FAN-LSA**	**FAN-LSA2**
Overall	0.063	0.186	0.188	0.178	0.171	0.329	0.347	**0.353**	0.344	0.344
User_guess	0.081	0.146	0.145	0.173	0.165	0.295	**0.314**	0.313	0.311	0.311
System_guess	0.045	0.227	0.231	0.182	0.178	0.364	0.380	**0.392**	0.377	0.378

7 Guesses/Clues	**HAL**	**DCC**	**DCC2**	**LSA**	**LSA2**	**FAN**	**FAN-DCC**	**FAN-DCC2**	**FAN-LSA**	**FAN-LSA2**
Overall	0.153	0.374	0.365	0.381	0.368	0.563	0.578	0.577	0.582	**0.589**
User_guess	0.196	0.315	0.309	0.349	0.346	0.499	0.508	0.509	0.515	**0.527**
System_guess	0.111	0.433	0.421	0.413	0.391	0.627	0.649	0.645	0.649	**0.650**

TABLE II
THE WIN/LOSS RECORD FOR SEVERAL SEMANTIC MODELS FOR EACH MODE OF THE WORDLERY GAME (HIGHER IS BETTER). THE COMBINED MODELS PERFORM THE BEST FOR EACH MODE OF THE GAME. THIS TABLE CORRESPONDS TO THE RESULTS IN FIGURE 4.

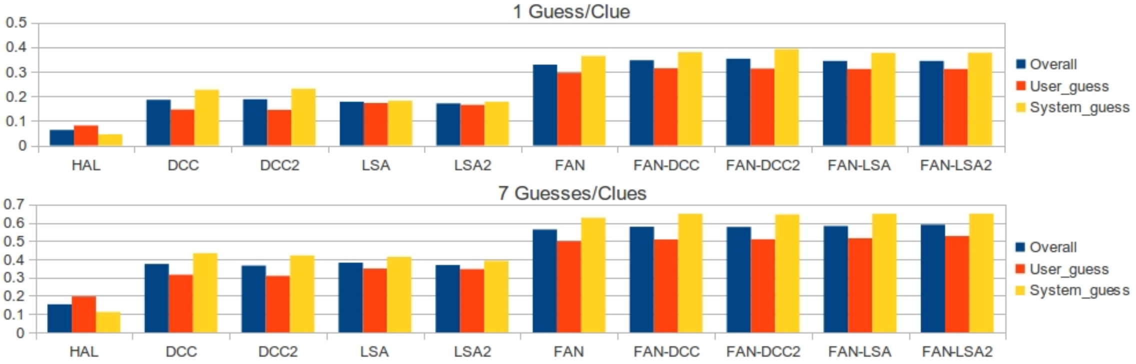

Fig. 4. The win/loss record for several semantic models for each mode of the Wordlery game (higher is better). The combined models perform the best for each mode of the game. This chart corresponds to the data in Table II.

in which the correct word was guessed to the total number of games played. We consider the win/loss record allowing all seven guesses/clues as well as the win/loss record allowing for only the first guess/clue.

In addition to evaluation, another purpose for playing these word guessing games is to provide a new method for gathering word associations from people. In the system_guess mode, the user provides clue words that are associated with the given word and each word entered is saved in a separate database as being associated with the given word. The idea is to gather word associations that more accurately reflect how a person thinks about a concept as opposed to simply the first other word that comes to mind as is done with FANs. We have opted to not save any word associations for the user_guess mode as user guesses tend to be more inconsistent, especially when the user has no definite idea of what to guess.

There exist other online games for the purpose of collecting semantic information from people. For example, a game called *Wikispeedia* is an online game for inferring semantic distances between concepts [24]. Wikispeedia is played by randomly selecting two unrelated Wikipedia articles and having the user reach one of the articles from the other by clicking through hyperlinks in the articles encountered. The path the user takes is analyzed to derive a semantic distance between the concepts represented by the starting and ending articles.

IV. RESULTS

Initially over 2500 games were played by a variety of anonymous individuals through the online interface. This resulted in 7868 word associations gathered from the game. This data provides a snapshot of how people play the online game, and we can use this data to objectively evaluate the models by having each model play against the collected data (reenacting the human input). We first evaluate several variations of the models using this collected data. We then select the best models from this initial phase and have them play against humans.

A. *Wordlery with Collected Data*

We randomly selected 1500 unique words from the collected database and had each model play both modes of the word guessing game using the collected data to simulate the human input. Table II and Figure 4 shows the results for FANs, DCC, DCC2 (extended vocabulary), HAL, LSA, and LSA2 (extended vocabulary) as well as the results for the combination of FANs with DCC, DCC2, LSA, and LSA2 (named FAN-DCC, FAN-DCC2, FAN-LSA, and FAN-LSA2).

When considering only the individual models, FANs perform the best by a considerable margin, as expected, since this version of the game consists of simulating real human responses. However, all the combined models perform better than the FANs. This shows that combining FANs and CSMs successfully takes advantage of their respective strengths.

1 Guess/Clue	**DCC**	**LSA**	**FANs**	**FAN-DCC2**	**FAN-LSA2**
Overall	0.232	0.325	0.354	**0.440**	0.418
User_guess	0.421	0.472	**0.577**	0.576	0.511
System_guess	0.089	0.113	0.161	0.296	**0.308**

7 Guesses/Clues	**DCC**	**LSA**	**FANs**	**FAN-DCC2**	**FAN-LSA2**
Overall	0.678	0.656	0.756	**0.837**	0.811
User_guess	0.684	0.607	0.722	**0.812**	0.728
System_guess	0.673	0.726	0.786	0.864	**0.910**

TABLE III

THE WIN/LOSS RECORD FOR EACH OF THE FIVE SEMANTIC MODELS FOR EACH MODE OF THE WORDLERY GAME (HIGHER IS BETTER). THE FAN-DCC2 MODEL PERFORMS THE BEST OVERALL, WHILE FAN-LSA2 IS A CLOSE SECOND. HOWEVER, FANS PERFORMS BETTER ON THE SINGLE GUESS USER_GUESS MODE. UNDERLINED SCORES DENOTE STATISTICAL SIGNIFICANCE COMPARED TO THE FANS MODEL USING THE z PROPORTIONALITY TEST. THIS TABLE CORRESPONDS TO THE RESULTS IN FIGURE 5.

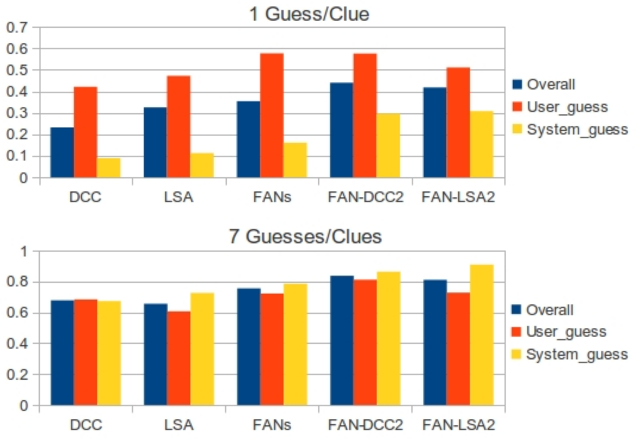

Fig. 5. The win/loss record for each of the five semantic models for each mode of the Wordlery game (higher is better). The FAN-DCC2 model performs the best overall, while FAN-LSA2 is a close second. However, FANs performs better on the single guess User_guess mode. This chart corresponds to the data in Table III.

Surprisingly, increasing the vocabulary size (for LSA and DCC) has very little influence on the performance of the CSMs. This result is likely due to the fact that people tend to provide more common words as guesses/clues. These common words are likely to be included in the 19,310 words from the smaller vocabulary, and extending the vocabulary makes little difference.

Another surprising result to note is that DCC performs slightly better than LSA for the system_guess mode and the FAN-DCC combination performs slightly better than the FAN-LSA combination for both modes of the 1st guess version of the game. This suggests the possibility that the 1st order word correlations that DCC captures (e.g., ‘dog’,‘kennel’) are better than (or at least comparable to) the higher order word correlations that LSA captures (e.g., ‘dog’,‘puppy’) for this type of semantic task. The differences are not significant, and LSA has the edge for the 7 guesses version of the game, but the suggestion is there. Most semantic tasks in the literature deal with semantic similarity (like the TOEFL test), in which models like LSA usually perform well because they explicitly try to capture word synonymy/similarity. The word guessing game requires the ability to come up with an answer (free response) as opposed to multiple choice, which additional types of word associations (beyond just synonymy) can help facilitate.

B. Wordlery with People

We selected FANs, DCC, LSA, FAN-DCC2, and FAN-LSA2 to play Wordlery with people online since they represented the top models from the collected data experiments. Approximately 900 games were played by a variety of anonymous individuals through the online interface. On average, each of the five semantic models participated in 180 of the total rounds played. Table III and Figure 5 shows the results for the overall win/loss record, the user_guess win/loss record and the system_guess win/loss record.

The first thing to note is that, overall, the FAN-DCC2 model achieves the best win/loss score (for both 1 guess and 7 guesses). The FAN-LSA2 model performs second best, with FANs not far behind. The CSMs (LSA and DCC) by themselves are significantly inferior to the FANs and combined models. The FANs do well presumably because those associations come directly from humans and hence can convey concepts back to human players. The combined methods take advantage of those human associations and then supplement them with the corpus inferred associations, which results in better performance. However, when allowing for only one guess on the the User_guess mode, FANs perform the best. We note again that the DCC-based models slightly outperform (or are close to) the LSA-based models on the user_guess mode, which promotes the usefulness of 1st-order correlations for certain semantic tasks such as this game.

When allowing for all 7 guesses/clues, the User_guess scores are generally lower than System_guess which suggests that, of the two modes, user_guess is harder. This makes sense since this mode only provides a static set of words as clues from which the user has to make a finite number of guesses. The interactive nature of the System_guess mode is likely one of the reasons for the better performance. However, the reason could also be that humans are good at providing relevant words as clues. Hence, we are collecting these word/clue pairs provided by the users for future studies. Perhaps the user_guess mode could be enhanced to allow the system to adapt its clues based on the user’s guesses as one might do in a human-human game. Note that when only 1 guess/clue is allowed, the scores for the System_guess mode are very low. This is expected since the model is allowed only a single guess based on one clue, for which the FAN-LSA2 and FAN-DCC2 models performs the best by a significant margin.

V. APPLICATIONS

The results of these games show that word associations can convey some aspect of the meaning of words. Such

	DCC	LSA	FANs	FAN-DCC2	FAN-LSA2	Human
Score	0.09	0.15	0.18	0.18	0.18	**0.41**

TABLE IV
THE RESULTS FOR DETERMINING THE CORRECT WORD GIVEN ONLY ITS DEFINITION (HIGHER IS BETTER). THE LOW SCORES CONFIRM THE DIFFICULTY OF THE TASK.

associations allow the system to communicate concepts to humans and allow humans to communicate concepts to the system, which is an important step in building a computational model that is capable of understanding language, processing input, making decisions, and interacting with people.

For example, in information retrieval, a user may not know what word to use in a query. The user could formulate a query using other words that describe (i.e., are associated with) the concept. In the query expansion process, the system could then automatically infer the concept and provide search results accordingly with possible improvements to both recall and precision.

To provide another example, and as an additional experiment, we randomly selected 100 word definitions from a dictionary. The system tokenized each definition into individual words (ignoring stop words), which essentially became "clue" words for the definition's corresponding word. Using the same method as in the System_guess mode of the Worldery game, each semantic model had to guess the word (or a synonym) that corresponded to each definition (allowing only one guess). For a baseline comparison, a couple of human volunteers performed the same task for all 100 definitions. Table IV shows the results for the same models used in the Wordlery game and the average score for the human volunteers.

Keep in mind that the purpose of this experiment is not for rigorous testing but is a proof of concept to demonstrate a difficult task that requires semantic modeling. Even the human volunteers were not able to correctly determine the word for nearly 60% of the definitions. Although the semantic models performed relatively poorly, they do show potential. The ability of a computer system to recall a concept given a description shows some level of language understanding. On a larger scale this is analogous to topic modeling. For example, instead of the system answering the question, "to what is this definition referring?", it could answer the question, "what is this document about?".

Future applications could go beyond word-to-word associations and build associations between words and other objects (such as images), which can potentially expand the ability of the system to communicate and understand meaning in a variety of ways. For example, we plan to build a system that is capable of communicating ideas through visual art. For a concept such as 'freedom' the system will use the word and image associations to automatically compose an image that conveys the meaning of 'freedom' to the viewer.

VI. Conclusions

We have experimented with two methods for obtaining word associations: through human free association norms (FANs) and by inferring them from a corpus (CSMs). We have also introduced a new semantic task to evaluate word associations by playing word guessing games. We compare the word associations from these methods and conclude that the FANs generally provide better quality associations than CSMs alone. Obviously, corpus-based approaches are heavily dependent on the corpus used. In future work, this influence could be assessed by evaluating word associations using the word guessing game from a variety of corpora. Our findings seem to be consistent with other studies that also show the superiority of FANs [13], [14]. It would seem that a universal corpus would be needed to discover word associations that are of the same quality as free association norms. But does a universal corpus exist? Or is it possible to create a model that can extract quality word associations from a standard corpus such as Wikipedia?

We have outlined a way to combine FANs with corpus-based semantic models. We use the word guessing game to show that combining the two methods of forming word associations is superior to each of the methods in isolation. This tells us that the CSMs have value and can complement the FANs. Perhaps for domain-specific tasks, preexisting databases of free association norms could provide a core of common human knowledge, while a domain-specific corpus and a CSM could be used to enhance the associations.

The word guessing games also provide a new way of gathering word associations from people. Once enough data has been collected, we will reevaluate the associations generated from the game comparing them with free association norms to see if they provide better quality associations for communicating meaning. In future work, we intend to incorporate more advanced corpus-based semantic models that take into account additional semantic information such as word order, sentence structure, and relationship types. We believe this will improve the results on certain semantic tasks such as the definition-based word recall experiment.

References

[1] K. Erk, "What is word meaning, really?: (and how can distributional models help us describe it?)," in *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 17–26.

[2] T. Landauer and S. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition induction and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[3] R. Sun, *The Cambridge Handbook of Computational Psychology*, 1st ed. New York, NY, USA: Cambridge University Press, 2008.

[4] S. De Deyne and G. Storms, "Word associations: Norms for 1,424 Dutch words in a continuous task," *Behavior Research Methods*, vol. 40, no. 1, pp. 198–205, feb 2008.

[5] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, pp. 211–226, 2004.

[6] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[7] H. Liu and P. Singh, "Commonsense reasoning in and over natural language," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, M. Negoita, R. Howlett, and L. Jain, Eds. Springer Berlin / Heidelberg, 2004, vol. 3215, pp. 293–306.

[8] G. Salton, *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

[9] G. Denhière and B. Lemaire, "A computational model of children's semantic memory," in *Proceedings of the 26th Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 2004, pp. 297–302.

[10] R. Rapp, "Word sense discovery based on sense descriptor dissimilarity," in *Proceedings of the Ninth Machine Translation Summit*, 2003, pp. 315–322.

[11] P. D. Turney, "Similarity of semantic relations," *Computational Linguistics*, vol. 32, no. 3, pp. 379–416, Sep. 2006.

[12] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The University of South Florida word association, rhyme, and word fragment norms," http://www.usf.edu/FreeAssociation/, 1998.

[13] T. Wandmacher, E. Ovchinnikova, and T. Alexandrov, "Does latent semantic analysis reflect human associations?" in *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 2008, pp. 63–70.

[14] Y. Peirsman and D. Geeraerts, "Predicting strong associations on the basis of corpus data," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 648–656.

[15] L. von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, June 2006.

[16] M. N. Jones and D. J. K. Mewhort, "Representing word meaning and order information in a composite holographic lexicon," *Psychological Review*, vol. 114, pp. 1–37, 2007.

[17] M. Baroni, B. Murphy, E. Barbu, and M. Poesio, "Strudel: A corpus-based semantic model based on properties and types," *Cognitive Science*, vol. 34, pp. 222–254, 2010.

[18] M. Baroni and A. Lenci, "Distributional memory: A general framework for corpus-based semantics," *Computational Linguistics*, vol. 36, no. 4, pp. 673–721, Dec. 2010.

[19] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper, "An associative thesaurus of English and its computer analysis," in *The Computer and Literary Studies*, A. J. Aitkin, R. W. Bailey, and N. Hamilton-Smith, Eds. Edinburgh, UK: University Press, 1973.

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[21] C. Burgess, "From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model," *Behavior Research Methods, Instruments, & Computers*, vol. 30, pp. 188–198, 1998.

[22] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods, Instruments, & Computers*, vol. 28, pp. 203–208, 1996.

[23] L. Denoyer and P. Gallinari, "The Wikipedia XML corpus," in *INEX Workshop Pre-Proceedings*, 2006, pp. 367–372.

[24] R. West, J. Pineau, and D. Precup, "Wikispeedia: an online game for inferring semantic distances between concepts," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, San Francisco, CA, USA, 2009, pp. 1598–1603.