Online Word Games for Semantic Data Collection

David Vickrey Aaron Bronzan William Choi Aman Kumar
Jason Turner-Maier Arthur Wang Daphne Koller
Stanford University
Stanford, CA 94305-9010
{dvickrey,abronzan,aman,arthurex,koller}@cs.stanford.edu
{wchoi25,jasonptm}@stanford.edu

Abstract

Obtaining labeled data is a significant obstacle for many NLP tasks. Recently, online games have been proposed as a new way of obtaining labeled data; games attract users by being fun to play. In this paper, we consider the application of this idea to collecting semantic relations between words, such as hypernym/hyponym relationships. We built three online games, inspired by the real-life games of Scattergories™ and Taboo™. As of June 2008, players have entered nearly 800,000 data instances, in two categories. The first type of data consists of category/answer pairs ("Types of vehicle","car"), while the second is essentially free association data ("submarine","underwater"). We analyze both types of data in detail and discuss potential uses of the data. We show that we can extract from our data set a significant number of new hypernym/hyponym pairs not already found in WordNet.

1 Introduction

One of the main difficulties in natural language processing is the lack of labeled data. Typically, obtaining labeled data requires hiring human annotators. Recently, building online games has been suggested an alternative to hiring annotators. For example, von Ahn and Dabbish (2004) built the ESP Game[1], an online game in which players tag images with words that describe them. It is well known that there are large numbers of web users who will play online games. If a game is fun, there is a good chance that sufficiently many online users will play.

We have several objectives in this paper. The first is to discuss design decisions in building word games for collecting data, and the effects of these decisions. The second is to describe the word games that we implemented and the kinds of data they are designed to collect. As of June 2008, our games have been online for nearly a year, and have collected nearly 800,000 data instances. The third goal is to analyze the resulting data and demonstrate that the data collected from our games is potentially useful in linguistic applications. As an example application, we show that the data we have collected can be used to augment WordNet (Fellbaum, 1998) with a significant number of new hypernyms.

2 General Design Guidelines

Our primary goal is to produce a large amount of clean, useful data. Each of these three objectives ("large", "clean", and "useful") has important implications for the design of our games.

First, in order to collect large amounts of data, the game must be attractive to users. If the game is not fun, people will not play it. This requirement is perhaps the most significant factor to take into account when designing a game. For one thing, it tends to discourage extremely complicated labeling tasks, since these are more likely to be viewed as work. It would certainly be a challenge (although not necessarily impossible) to design a game that yields labeled parse data, for example.

In this paper, we assume that if people play a game in real life, there is a good chance they will play it online as well. To this end, we built online versions of two popular "real-world" games: Scattergories™ and Taboo™. Not only are these games fun, but there is also a preexisting demand for online versions of these games, driving search traffic to our site. We will go into more detail about these games in the next section.

An important characteristic of these games is that they involve more than one player. Interacting with another player increases the sense of fun. Another important feature these games share is that they are

[1] www.gwap.com/gwap/gamesPreview/espgame

timed. Timing has several advantages. First, timing helps make the games feel more "game-like", by adding a sense of urgency. Without timing, it risks feeling more like a labeling task than a game.

The next requirement is that the data be clean. First, the players must be capable of producing high-quality annotations. Second, the game should encourage users to enter relevant data. We award points as a motivating factor, but this can lead players to enter irrelevant data, or collude with other players, in order to get a higher score. In particular, collusion is more likely when players can freely communicate. An excellent technique for producing good data, used effectively in the ESP game, is to require the players to match on their inputs. Requiring players to match their partner's hidden answers discourages off-topic answers and makes it quite difficult to collude (requiring outside communication). We use this technique in all of our games.

Finally, the data must be useful. Ideally, it would be directly applicable to an NLP task. This requirement can come into conflict with the other goals. There are certainly many kinds of data that would be useful for NLP tasks (such as labeled parses), but designing a game to collect this data that people will play and that produces clean data is difficult.

In this paper, we focus on a particular kind of linguistic data: semantic relationships between pairs of words and/or phrases. We do this for several reasons. First, this kind of data is relatively simple, leading to fun games which produce relatively clean data. Second, the real-world games we chose to emulate naturally produce this kind of data. Third, there are a number of recent works which focus on extracting these kinds of relationships, e.g. (Snow et al., 2006; Nakov & Hearst, 2008). Our work presents an interesting new way of extracting this type of data. Finally, at least one of these kinds of relationships, the hypernym, or "X is a Y" relation, has proven to be useful for a variety of NLP tasks.

3 Description of Our Games

We now describe our three games in detail.

3.1 Categorilla

Categorilla, inspired by Scattergories™, asks players to supply words or phrases which fit specific categories, such as "Things that fly" or "Types of fish". In addition, each game has a specific letter which all answers must begin with. Thus, if the current game has letter "b", reasonable answers would be "bird" and "barracuda", respectively. In each game, a randomly matched pair of players are given the same 10 categories; they receive points when they match with the other player for a particular category. Players are allowed to type as may answers for a given category as they wish (until a match is made for that category). After a match is made, the players get to see what word they matched on for that category. Each answer is supposed to fit into a specific category, so the data is automatically structured.

Our system contains 8 types of categories, many of which were designed to correspond to linguistic resources used in NLP applications. Table 1 describes the category types.

The purpose of the first three types of categories is to extract hypernym/hyponym pairs like those found in WordNet (e.g., "food" is a hypernym of "pizza"). In fact, the categories were automatically generated from WordNet, as follows. First, we assigned counts C_s to each synset s in WordNet using the SemCor[2] labeled data set of word senses. Let $desc(s)$ be the set of descendants of s in the hypernym hierarchy. Then for each pair of synsets s, d, where $d \in desc(s)$, we computed a conditional distribution $P(d|s) = \frac{C_d}{\sum_{d' \in desc(s)} C_{d'}}$, the probability that we choose node d from among the descendants of s. Finally, we computed the entropy of each node s in WordNet, $\sum_{d \in desc(s)} P(d|s) log P(d|s)$. Synsets with many different descendants occurring in SemCor will have higher entropies. Each node with a sufficiently high entropy was chosen as a category.

We then turned each synset into a category by taking the first word in that synset and plugging it into one of several set phrases. For nouns, we tried two variants ("Types of food" and "Foods"). Depending on the noun, either of these may be more natural (consider "Cities" vs. "Types of city"). "Types of food" tends to produce more adjectival answers than "Foods". We tried only one variation for verbs ("Methods of paying"). This phrasing is not perfect; in particular, it encourages non-verb answers like "credit card".

The second group of categories tries to capture selectional preferences of verbs – for example, "ba-

[2] Available at `www.cs.unt.edu/ rada/downloads.html`

Name	#	Description	Example	Good Answer
NHyp	269	Members of a class of nouns	"Vehicles"	"car"
NType	269	Members of a class of nouns	"Types of vehicle"	"car"
VHyp	70	Members of a class of verbs	"Methods of cutting"	"trimming"
VS	1380	Subjects of a verb	"Things that eat"	"cats"
VO	909	Direct objects of a verb	"Things that are abandoned"	"family"
VPP	77	Preposition arguments of a verb	"Things that are accused of"	"crime"
Adj	219	Things described by an adjective	"Things that are recycled"	"cans"
O	105	Other; mostly "Things found at/in ..."	"Things found in a school"	"teachers"

Table 1: Summary of category types. # indicates the number of categories of that type.

nana" makes sense as the object of "eat" but not as the subject. Our goal with these categories was to produce data useful for automatically labeling semantic roles (Gildea & Jurafsky, 2002), where selectional preferences play an important role. We tried three different types of categories, corresponding to subjects, objects, and prepositional objects. Examples are "Things that eat", "Things that are eaten", and "Things that are eaten with", to which good answers would be "animals", "food", and "forks". These categories were automatically generated using the labeled parses in Penn Treebank (Marcus et al., 1993) and the labeled semantic roles of PropBank (Kingsbury et al., 2002). To generate the object categories, for example, for each verb we then counted the number of times a core argument (ARG0-ARG5) appeared as the direct object of that verb (according to the gold-standard parses), and used all verbs with count at least 5. This guaranteed that all generated categories were grammatically correct and captured information about core arguments for that verb. Most of the prepositional object categories proved to be quite confusing (e.g., "Things that are acted as"), so we manually removed all but the most clear. Not surprisingly, the use of the Wall Street Journal had a noticeable effect on the types of categories extracted; they have a definite financial bias.

The third group of categories only has one type, which consists of adjective categories such as "Things that are large". While we did not have any specific task in mind for this category type, having a database of attributes/noun pairs seems potentially useful for various NLP tasks. To generate these categories, we simply took the most common adjectives in the SemCor data set. Again, the resulting set of adjectives reflect the corpus; for example, "Things that are green" was not generated as a category, while "Things that are corporate" was.

The final group of categories were hand-written. This group was added to make sure that a sufficient number of "fun" categories were included, since some of the category types, particularly the verb categories, are somewhat confusing and difficult. Most of the hand-written categories are of the form "Things found at/in X", where X is a location, such as "Japan" or "the ocean".

The starting letter requirement also has important consequences for data collection. It was designed to increase the variety of obtained data; without this restriction, players might produce a smaller set of "obvious" answers. As we will see in the results, this restriction did indeed lead to a great diversity of answers, but at a severe cost to data quality.

3.2 Categodzilla

Categodzilla is a slightly modified version of Categorilla, with the starting letter constraint relaxed. The combination of difficult categories and rare letters often leads to bad answers in Categorilla. To increase data quality, in Categodzilla for each category there are three boxes. In the first box you can type any word you want. Answers in the second box must start with a given "easy" letter such as "c". Answers in the third box must start with a given "hard" letter, such as "k". The boxes much be matched in order; guesses typed in the first box which match either of the other two boxes are automatically propagated.

3.3 Free Association

Free Association, inspired by Taboo™, simply asks players to type words related to a given "seed" word. Players are not allowed to type any of several words on a "taboo" list, specific to the current seed word.

As soon as a match is achieved, players move on to a new seed word.

The seed words came from two sources. The first was the most common words in SemCor. The second was the Google unigram data, which lists the most common words on the web. In both cases, we filtered out stop words (including all prepositions).

Unlike Categorilla, we found that nearly all collected Free Association data was of good quality, due to the considerably easier nature of the task. Of course, we do lose the structure present in Categorilla. As the name suggests, the collected data is essentially free word association pairs. We analyze the data in depth to see what kinds of relations we got.

4 Existing Word Games

Two notable word games already exist for collecting linguistic data. The first is the Open Mind Common Sense system[3] (Chklovski, 2003). The second is Verbosity[4] (von Ahn et al., 2006). Both these games are designed to extract common sense facts, and thus have a different focus than our games.

5 Bots

There may not always be enough players available online to match a human player with another human player. Therefore, one important part of designing an online game is building a bot which can function in the place of a player. The bots for all of our games are similar. Each has a simple random model which determines how long to wait between guesses. The bot's guesses are drawn from past guesses made by human players for that category/seed word (plus starting letter in the case of Categorilla). Just as with a human player, as soon as one of the bot's guesses matches one of the player's, a match is made.

If there are no past guesses, the bot instead makes "imaginary" guesses. For example, in Categorilla, we make the (obviously false) assumption that for every category and every starting letter there are exactly 20 possible answers, and that both the player's guesses and the bot's imaginary guesses are drawn from those 20 answers. Then, given the number of guesses made by the player and the number of imaginary guesses made by the bot, the probability of a match can be computed (assuming that all guesses are made independently). Once this probability passes a certain threshold, randomly generated for each category at the start of each game, the bot matches one of the player's guesses, chosen at random. The Free Association bot works similarly.

For Free Association, the bot rarely has to resort to generating these imaginary guesses. In Categorilla, due to the starting letter requirement, the bot has to make imaginary guesses much more often. Imaginary guessing can encourage poor behavior on the part of players, since they see that matches can occur for obviously bad answers. They may also realize that they are playing against a bot.

An additional complication for Categorilla and Categodzilla is that the bot has to decide which categories to make guesses for, and in what order. Our current guessing model takes into account past difficulty of the category and the current guessing of the human player to determine where to guess next.

	Grla	Gdza	Free
Game Length	3min	3min	2min
Games Played	19656	2999	15660
Human-Human Games	428	45	401
Categories	3298	3298	9488
Guesses Collected	391804	78653	307963
Guesses/Categories	119	24	32
Unique Guesses	340433	56142	221874
Guesses: All/Unique	1.15	1.40	1.39
Guesses/Games	19.9	26.2	19.7
Guesses per minute	6.6	8.7	9.9

Table 2: Statistics for Categorilla, Categodzilla, and Free Association.

6 Users and Usage

Table 2 shows statistics of each of the games, as of late June 2008. While we have collected nearly 800,000 data instances, nearly all of the games were between a human and the bot. Over the course of a year, our site received between 40 and 100 visits a day; this was not enough to make it likely for human-human games to occur. The fact that we still collected this amount of data suggests that our bot is a satisfactory substitue for a human teammate. We have anecdotally found that most players do not realize they are playing against a bot. While most of the data comes from games between a human and a bot, our data set consists only of input by the human players.

[3] http://commons.media.mit.edu/en

[4] www.gwap.com/gwap/gamesPreview/verbosity

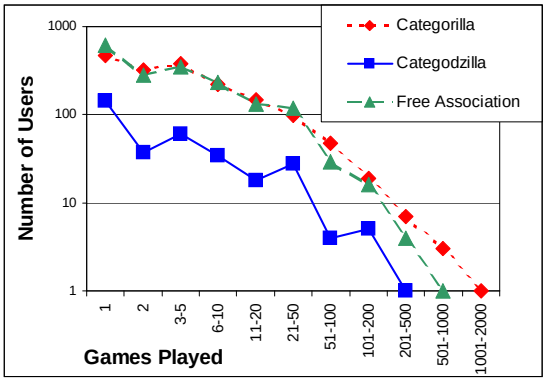

Figure 1: Users are grouped by number of games played. Note that this graph is on a double-log scale.

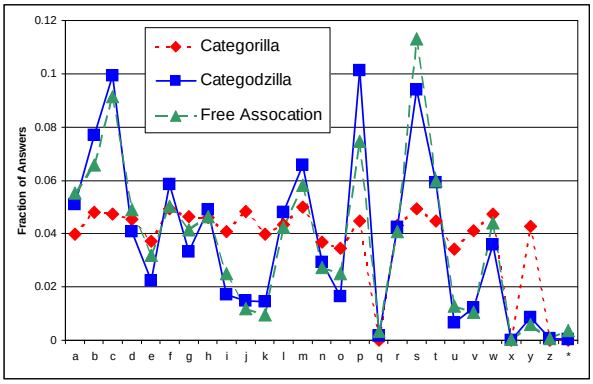

Figure 2: Fraction of answers with given initial letter. * denotes everything nonalphabetical.

Our main tool for attracting traffic to our site was Google. First, we obtained $1 a day in AdWords, which pays for between 7 to 10 clicks on our ad a day. Second, our site is in the top 10 results for many relevant searches, such as "free online scattergories".

Categorilla was the most popular of the games, with about 25% more games played than Free Association. Taking the longer length of Categorilla games into account (see Table 2), this corresponds to almost 90% more play time. This is despite the fact that Free Association is the first game listed on our home page. We hypothesize that this is because Scattergories^{TM} is a more popular game in real life, and so many people come to our site specifically looking for an online Scattergories^{TM} game. Categodzilla has been played signficantly less; it has been available for less time and is listed third on the site. Even for Categodzilla, the least played game, we have collected on average 24 guesses per category.

Several of our design decisions for the games were based on trying to increase the diversity of answers. Categorilla has the highest answer diversity. For a given category, each answer occurred on average only 1.15 times. In general, this average should increase with the amount of collected data. However, Categodzilla and Free Association have collected significantly fewer answers per category than Categorilla, but still have a higher average, around 1.4. The high answer diversity of Categorilla is a direct result of the initial letter constraint. For all three games, the majority of category/answer pairs occurred only once.

Figure 1 shows the distribution over users of the number of games played. Not surprisingly, it follows the standard Zipfian curve; there are a large number of users who have played only a few games, and a few users who have played a lot of games. The middle of the curve is quite thick; for both Categorilla and Free Association there are more than 100 players who have played between 21 and 50 games.

Figure 2 shows the distribution of initial letters of collected answers for each game. Categorilla is nearly flat over all letters besides 'q', 'x', and 'z' which are never chosen as the inital letter constraint. This means players make a similar number of guesses even for difficult initial letters. In contrast, the distribution of initial letters for Free Association data reflects the relatively frequency of initial letters in English. Even though Categodzilla does have letter constraints in the 2nd and 3rd columns, its statistics over initial letter are very similar to Free Association.

7 Categorilla and Categodzilla Data

In our analyses, we take ALL guesses made at any time, whether or not they actually produced a match. This greatly increases the amount of usable data, but also increases the amount of noise in the data.

The biggest question about the data collected from Categorilla and Categodzilla is the quality of the data. Many categories can be difficult or somewhat confusing, and the initial letter constraint further increases the difficulty.

To evaluate the quality of the data, we asked three volunteer labelers to label 1000 total category/answer pairs. Each labeler labeled every pair with one of three labels, 'y', 'n', or 'k'. 'y' means that the answer fit the category. 'n' means that it

Annotator	y	k	n
#1	72	13	115
#2	77	27	96
#3	88	42	70
Majority	76	29	95

Table 3: Comparison of annotators

Data Set	y	k	n
Control	30	14	156
Categorilla	76	29	95
Categodzilla	144	23	33

Table 4: Overall answer accuracy

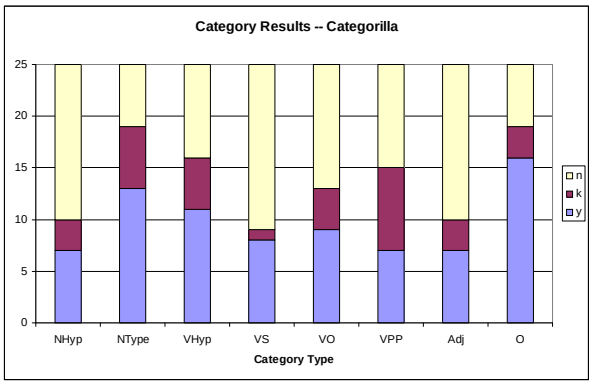

Figure 3: Categorilla accuracy by category type

does not fit. 'k' means that it "kind of" fits. This was mostly left up to the labelers; the only suggestion was that one use of 'k' could be if the category was "Things that eat" and the answer was "sandwich." Here, the answer is clearly related to the category, but doesn't actually fit.

The inter-annotator agreement was reasonable, with a Fleiss' kappa score of .49. The main difference between annotators was how permissive they were; the percentage of answers labeled 'n' ranged from 58% for the first annotator to 35% for the third. The labeled pairs were divided into 5 subgroups of 200 pairs each (described below); Table 3 shows the number of each label for the Categorilla-Random subset. We aggregated the different annotations by taking a majority vote; if all three answers were different, the item was labeled 'k'. Table 3 also shows the statistics of the majority vote on the same subset.

Overall Data Quality. We compared results for three random subsets of answers, Control-Random, Categorilla-Random, and Categodzilla-Random. Categorilla-Random was built by selecting 200 random category/answer pairs from the Categorilla data. Note that category/answer pairs that occurred more than once were more likely to be selected. Categodzilla-Random was built similarly. Control-Random was built by randomly selecting two sets of 200 category/answer pairs each (including data from both Categorilla and Categodzilla), and then combining the categories from the first set with the answers from the second to generate a set of random category/answer pairs.

Table 4 shows results for these three subsets. The chance for a control answer to be labeled 'y' was 15%. Categorilla produces data that is significantly better than control, with 38% of answers labeled 'y'. Categodzilla, which is more relaxed about initial letter restrictions, is significantly better than Categorilla, with 72% of answers labeled 'y'. This relaxation has an enormous impact on the quality of the data. Note however that these statistics are not adjusted for accuracy of individual players; it may be that only more accurate players play Categodzilla.

Effect of Category Type on Data Quality. Within each type of category (see Table 1), certain categories appear much more often than others due to the way categories are selected (at least two "easy" categories are guaranteed every game). To adjust for this, we built a subset of 200 category/answer pairs by selecting 25 different categories randomly from each type of category. We then selected an answer at random from among the answers submitted for that category. In addition, we built a control set using the same 200 categories but instead using answers selected at random from the entire Categorilla data set. Results for Categorilla data are shown in Figure 3; we omit the corresponding graph for control for lack of space. For most categories, the Categorilla data is significantly better than the control. The hand-written category type, O, has the best data quality, which is not surprising because these categories allow the most possible answers, and thus are easiest of think of answers for. These categories also have the highest number of 'y' labels for the control. Next best are the hypernym categories, NType. NType is much higher than the other noun hypernym category NHyp because the "Type of" phrasing is generally more natural and allows for adjectival answers. The VPP category type, which tries to extract prepositional objects, contains

Data Set	Letters	Size	y	k	n
Control	Easy	127	.14	.08	.78
Control	Hard	72	.15	.06	.79
Categorilla	Easy	106	.45	.14	.41
Categorilla	Hard	94	.30	.15	.55

Table 5: Accuracy of easy letters vs. hard letters. Size is the number of answers for that row.

the most number of 'k' annotations; this is because players often put answers that are subjects or objects of the verb, such as "pizza" for "Things that are eaten with". The adjective category type, Adj, has the lowest increase over the control; this is likely due to the nature of the extracted adjectives.

Effect of Initial Letter on Data Quality. In general, we would expect common initial letters to yield better data since there are more possible answers to choose from. We did not have enough labeled data to do letter by letter statistics. Instead, we broke the letters into two groups, based on the empirical difficulty of obtaining matches when given that initial letter. The easy letters were 'abcfhlmnprst', while the hard letters were 'degijkouvwy'. Table 5 shows the results on Categorilla-Random and Control-Random on these two subsets. First, note that the results on Control-Random are the same for hard letters and easy letters. This means that words starting with common letters are not more likely to fit in a category. For both hard letters and easy letters, the accuracy is considerably better on the Categorilla data. However, the increase in the number of 'y' labels for easy letters is twice that for hard letters. The quality of data for hard letters is considerably worse than that for easy letters.

8 Free Association Data

In contrast to Categorilla and even Categodzilla, we found that the Free Association data was quite clean. However, it is also not structured; we simply get pairs of related words. Thus, the essential question for this game is what kind of data we get.

To analyze the types of relationships between words, the authors labeled 500 randomly extracted unique pairs with a rich set of word-word relations, described in Table 6. This set of relations was designed to capture the observed relationships encountered in the Free Association data. Unlike our Categorilla labeled set, pairs that occurred more than once were NOT more likely to be selected than pairs that occurred once (i.e., the category/answer pairs were aggregated prior to sampling). Sampling in this way led to more diversity in the pairs extracted.

To label each pair, the authors found a sequence of relationships which connected the two words. In many cases, this was a single link. For example, "dragon" and "wing" are connected by a single link, "wing" IS PART OF "dragon". In others, multiple links were required. For the seed word "dispute" and answer "arbitrator", we can connect using two links: "dispute" IS OBJECT OF "resolve", "arbitrator" IS SUBJECT OF "resolve". There were two other possible ways to label a pair. First, they might be totally unrelated (i.e., a bad answer). Second, they might be related, but not connectable using our set of basic relations. For example, "echo" is clearly related to "valley", but in a complicated way.

The quality of the data is considerably higher than Categorilla and Categodzilla; under 10% of words are unrelated. Slightly over 20% of the pairs are labeled Misc, i.e., the words are related but in a complicated way. 3% of the pairs can be linked with a chain of two simple relations. The remaining 67% of all pairs were linked with a single simple relation.

The category Desc deserves some discussion. This category included both simple adjective descriptions, such as "creek" and "noisy", and also qualifiers, such as "epidemic" and "typhoid", where one word specifies what kind of thing the other is. The distinction between Desc and Phrase was simply based on to what extent the combination of the two words was a set phrase (such as "east" and "Germany").

Schulte im Walde et al. (2008) address very similar issues to those discussed in this section. They built a free association data set containing about 200,000 German word pairs using a combination of online and offline volunteers (but not a game). They then analyze the resulting associations by comparing the resulting pairs to a large-scale lexical resource, GermaNet (the German counterpart of WordNet). Our data analysis was by hand, making it comparatively small scale but more detailed. It would be interesting to compare the data sets to see whether the use of a game affects the resulting data.

9 Filtering Bad Data

In this section, we consider a simple heuristic for filtering bad data: only retaining answers that were

Name	#	Description	Example
Misc	103	Words related, but in a complicated way	“echo”, “valley”
Desc	76	One of the words describes the other	“cards”, “business”
None	47	Words are not related	“congress”,“store”
Syn	46	The words are synonyms	“downturn”, “dip”
Obj	33	One word is the object of the other	“exhale”,“emission”
Hyp	30	One word is an example of the other	“cabinet”,“furniture”
≈Syn	29	The words are “approximate” synonyms	“maverick”,“outcast”
Cousin	21	The words share a common hypernym (is-a) relation	“meter”,“foot”
Has	18	One word “has” the other	“supermarket”,“carrots”
2-Chain	15	Words are linked by a chain of two simple relations	“arbitrator”,“dispute”
Phrase	13	Words make a phrase; similar to Desc	“East”, “Germany”
Part	11	One is a part of the other	“dragon”,“wings”
At	10	One is found at the other	“harbor”, “lake”
Subj	8	One is the subject of the other	“actor”, “pretend”
Form	7	One is a form of the other	“revere”,“reverence”
Def	7	One defines the other	“blind”,”unable to see”
Opp	7	The two are opposites	“positive”,“negative”
Sound	6	The two words sound similar	“boutique”,“antique”
Sub	5	One is a subword of the other	“outlet”, “out”
Unit	2	One is a unit of the other	“reel”,“film”
Made	2	One is made of the other	“knee”,“bone”

Table 6: Relation types for 500 hand-labeled examples. # indicates the number of pairs with that label.

guessed some minimum number of times. Note that in this section all answers were stemmed in order to combine counts across plurals and verb tenses.

For the Categorilla data, filtering out category/answer pairs that only occurred once from Categorilla-Random left a total of 64 answers (from an original 200), of which 36 were labeled ’y’ and 8 were labeled ’k’. The fraction of ’y’ labels in the reduced set is 56%, up from 38% in the original set. This gain in quality comes at the cost of losing slightly over two-thirds of the data.

For Categodzilla-Random, a similar filter left 88 (out of 200), with 79 labeled ’y’ and 7 labeled ’k’. For the hand-labeled Free Association data, applying this filter yielded a total of 123 pairs (out of an original 500), with only 2 having no relation[5]. In these two games, this filter eliminates nearly all bad data while keeping a reasonable fraction of the data.

Clearly, this filter is less effective for Categorilla than the other two games. One of the main reasons for this is that the letter constraints cause people to try to fit words starting with that letter into all categories that they even vaguely relate to, rather than thinking of words that really fit that category. Examples include {“Art supplies”,“jacket”}, {“Things found in Chicago”,“king”} and {“Things that are African”,“yak”}. Of course, we can further increase the quality of the data by making the filter more restrictive, at the cost of losing more data. For example, removing answers occuring fewer than 5 times from Categorilla-Random leaves only 8 answers (out of 200), 7 labeled ’y’ and 1 labeled ’n’.

There are other ways we could filter the data. For example, suppose we are given an outside database of pairs of words which are known to be semantically related. We could apply the following heuristic: if an answer to a particular category is similar to many other answers for that category, then that answer is likely to be a good one. Preliminary experiments using distributional similarity of words as the similarity metric suggest that this heuristic captures complimentary information to the guess frequency heuristic. We leave as future work a full integration of the two heuristics into a single improved filter.

[5]The higher fraction of lost pairs for Free Association is primarily due to the method of sampling pairs for evaluation, as discussed in Section 8.

Classified Type	#	Example
Real hypernyms	96	"equipment","racquet"
Compound hypernyms	32	"arrangement","flower"
Adjectives	25	"building","old"
Sort-of hypernyms	14	"vegetable","salad"
Not hypernyms	33	"profession","money"

Table 7: Breakdown of potential hypernym pairs

10 Using the Data

Categorilla and Categodzilla produce structured data which is already in a usable or nearly usable form. For example, the NHyp and NType categories produce lists of hypernyms, which could be used to augment WordNet. We looked at this particular application in some detail.

First, in order to remove noisy data, we used only Categodzilla data and removed answers which occurred only once. We took all category/answer pairs where the category was of type either NHyp or NType, and where the answer was a noun. This resulted in 1604 potential hypernym/hyponym pairs. Of these, 733 (or 46%) were already in WordNet. The remaining 871 were not found in WordNet. We then hand-labeled a random subset of 200 of the 871 to determine how many of them were real hypernym/hyponym pairs. The results are shown in Table 7. Counting compound hyponyms, nearly two-thirds of the pairs are real hypernym/hyponym pairs. These new pairs could directly augment WordNet. For example, for the word "crime", WordNet has as hyponyms "burglary" and "fraud". However, it doesn't have "arson", "homicide", or "murder", which are among the 871 new pairs. WordNet lists "wedding" as being an "event", but not "birthday".

The verb subject, object, and prepositional object categories were designed to collect data about the selectional preferences of verbs. These categories turned out to be problematic for several reasons. First, statistics about selectional preferences of verbs are not too difficult to extract from the web (although in some cases they might be somewhat noisy). Thus, the motivation for extracting this data using a game is not as apparent. Second, providing arguments of verbs out of the context of a sentence may be too difficult. For example, for the category "Things that are accumulated", there a couple of obvious answers, such as "wealth" or "money", but beyond these it becomes more difficult. In the context of an actual document, quite a lot of things can accumulate, but outside of that context it is difficult to think of them.

One solution to this problem would be to provide context. For example, the category "Things that accumulate in your body" is both easier to think of answers for and probably collects more useful data. However, automatically creating categories with the right level of specificity is not a trivial task; our initial experiments suggested that it is easy to generate too much context, creating an uninteresting category.

The Free Association game produces a lot of very clean data, but does not classify the relationships between the words. While a web of relationships might be useful by itself, classifying the pairs by relation type would clearly be valuable. Snow et al. (2006) and Nakov and Hearst (2008), among others, look at using a large amount of unlabeled data to classify relations between words. One issue with extracting new relations from text, for example meronyms (part-of relationships), is that they tend to occur fairly rarely. Thus, it is very easy to get a large number of spurious pairs. Using our data as a set of candidate pairs for relation extraction could greatly reduce the resulting noise. We believe that application of existing techniques to the data from the Free Association game could lead to a clean, classified set of word-word relations, but leave this as future work.

11 Discussion and Future Work

One way to extend Categorilla and Categodzilla would be to add additional types of categories. For example, a meronym category type (e.g. "Parts of a car") would work well. Further developing the verb categories (e.g., "Things that accumulate in your body") is another challenging but interesting direction; these categories would produce phrase-word relationships rather than word-word relationships.

Probably the most interesting direction for future work is trying to increase the complexity of the data collected from a game. There are two significant difficulties: keeping the game fun, and making sure the collected data is not too noisy. One interesting question for future research is whether different game architectures might be better suited to certain kinds of data. For example, a "telephone" style game, where players relay a phrase or sentence through some noisy channel, might be an interesting way to obtain paraphrase data.

References

Chklovski, T. (2003). Using analogy to acquire commonsense knowledge from human contributors. *Thesis*.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database*. MIT Press.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*.

Kingsbury, P., Palmer, M., & Marcus, M. (2002). Adding semantic annotation to the penn treebank. *Proceedings of the Human Language Technology Conference (HLT'02)*.

Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*.

Nakov, P., & Hearst, M. (2008). Solving relational similarity problems using the web as a corpus. *Proceedings of ACL*.

Schulte im Walde, S., Melinger, A., Roth, M., & Weber, A. (2008). An empirical characterisation of response types in german association norms. *To appear, Research on Language and Computation*.

Snow, R., Jurafsky, D., & Ng, A. (2006). Semantic taxonomy induction from heterogenous evidence. *Proceedings of COLING/ACL*.

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *ACM CHI*.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. *Proceedings of the SIGCHI conference on Human Factors in computing systems*.