Meaning Representation in Natural Language Categorization

Trevor Fountain (t.fountain@sms.ed.ac.uk) and
Mirella Lapata (mlap@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

A large number of formal models of categorization have been proposed in recent years. Many of these are tested on artificial categories or perceptual stimuli. In this paper we focus on categorization models for *natural language concepts* and specifically address the question of how these may be represented. Many psychological theories of semantic cognition assume that concepts are defined by features which are commonly elicited from humans. Norming studies yield detailed knowledge about meaning representations, however they are small-scale (features are obtained for a few hundred words), and admittedly of limited use for a general model of natural language categorization. As an alternative we investigate whether category meanings may be represented *quantitatively* in terms of simple co-occurrence statistics extracted from large text collections. Experimental comparisons of feature-based categorization models against models based on data-driven representations indicate that the latter represent a viable alternative to the feature norms typically used.

Introduction

Considerable psychological research has shown that people reason about novel objects they encounter by identifying the category to which these objects belong and extrapolating from their past experiences with other members of that category. This task of *categorization*, or grouping objects into meaningful categories, is a classic problem in the field of cognitive science, one with a history of study dating back to Aristotle. This is hardly surprising, as the ability to reason about categories is central to a multitude of other tasks, including perception, learning, and the use of language.

Numerous theories exist as to how humans categorize objects. These theories themselves tend to belong to one of three schools of thought. In the *classical* (or Aristotelian) view categories are defined by a list of "necessary and sufficient" features. For example, the defining features for the concept BACHELOR might be *male*, *single*, and *adult*. Unfortunately, this approach is unable to account for most ordinary usage of categories, as many real-world objects have a somewhat fuzzy definition and don't fit neatly into well-defined categories (Smith and Medin, 1981).

Prototype theory (Rosch, 1973) presents an alternative formulation of this idea, in which categories are defined by an idealized prototypical member possessing the features which are critical to the category. Objects are deemed to be members of the category if they exhibit enough of these features; for example, the characteristic features of FRUIT might include *contains seeds*, *grows above ground*, and *is edible*. Roughly speaking, prototype theory differs from the classical theory in that members of the category are not required to possess *all* of the features specified in the prototype.

Although prototype theory provides a superior and workable alternative to the classical theory it has been challenged by the *exemplar* approach (Medin and Schaffer, 1978). In this view, categories are defined not by a single representation but rather by a list of previously encountered members. Instead of maintaining a single prototype for FRUIT that lists the features typical of fruits, an exemplar model simply stores those instances of fruit to which it has been exposed (e.g., apples, oranges, pears). A new object is grouped into the category if it is sufficiently similar to one or more of the FRUIT instances stored in memory.

In the past much experimental work has tested the predictions of prototype- and exemplar-based theories in laboratory studies involving categorization and category learning. These experiments tend to use perceptual stimuli and artificial categories (e.g., strings of digit sequences such as 100000 or 0111111). Analogously, much modeling work has focused on the questions of how categories and stimuli can be represented (Griffiths et al., 2007a; Sanborn et al., 2006) and how best to formalize similarity. The latter plays an important role in both prototype and exemplar models as correct generalization to new objects depends on identifying previously encountered items correctly.

In this paper we focus on the less studied problem of categorization of *natural language concepts*. In contrast to the numerous studies using perceptual stimuli or artificial categories, there is surprisingly little work on how natural language categories are learned or used by adult speakers. A few notable exceptions are Heit and Barsalou (1996) who attempt to experimentally test an exemplar model within the context of natural language concepts, Storms et al. (2000) who evaluate the differences in performance between exemplar and prototype models on a number of natural categorization tasks, and Voorspoels et al. (2008) who model typicality ratings for natural language concepts. A common assumption underlying this work is that the meaning of the concepts involved in categorization can be represented by a set of features (also referred to as properties or attributes).

Indeed, featural representations have played a central role in psychological theories of semantic cognition and knowledge organization and many studies have been conducted to elicit detailed knowledge of features. In a typical procedure, participants are given a series of object names and for each object they are asked to name all the properties they can think of that are characteristic of the object. Although feature norms are often interpreted as a useful proxy of the structure of semantic representations, a number of difficulties arise

when working with such data (e.g., Sloman and Ripps 1998; Zeigenfusse and Lee 2009). For example, the number and types of attributes generated can vary substantially as a function of the amount of time devoted to each object. There are many degrees of freedom in the way that responses are coded and analyzed. It is not entirely clear how people generate features and whether all of these are important for representing concepts. Finally, multiple subjects are required to create a representation for each word, which limits elicitation studies to a small number of words and consequently the scope of any computational model based on these feature norms.

Even when the stimuli in question are of an abstract or linguistic nature, the features elicited are assumed to be representative of the underlying referents. As an alternative we propose to model the categorization of linguistic stimuli according to their distribution in corpora. Words whose referents exhibit differing features likely occur in correspondingly different contexts; our question is whether these differences in usage can provide a substitute for featural representations.

The idea that words with similar meaning tend to be distributed similarly across contexts is certainly not a novel one. *Semantic space* models, among which Latent Semantic Analysis (LSA, Landauer and Dumais 1997) is perhaps known best, operationalize this idea by capturing word meaning *quantitatively* in terms of simple co-occurrence statistics (between words and paragraphs or documents). More recently, *topic models* (Griffiths et al., 2007b) have arisen as a more structured representation of word meaning. In contrast to more standard semantic space models where word senses are conflated into a single representation, topic models assume that words observed in a corpus manifest some latent structure — word meaning is a probability distribution over a set of topics (corresponding to coarse-grained senses). Each topic is a probability distribution over words whose content is reflected in the words to which it assigns high probability.

In this work we investigate whether semantic representation models based on the statistical analysis of large text collections can provide a viable alternative to feature norms for natural language categorization. Specifically, we compare categorization models that represent concepts by features against LSA, Latent Dirichlet Allocation (LDA, Griffiths et al. 2007b; Blei et al. 2003), a well-known topic model, and a semantic space that takes syntactic information into account (Padó and Lapata, 2007). These semantic representations are used as input to two well-established categorization models, namely Nosofsky's (1988) generalized context model (GCM) and a prototype model derived from the GCM. We evaluate the performance of these models on three adult categorization tasks — category naming, typicality rating, and exemplar generation — which have been previously modeled using exclusively feature norms (Storms et al., 2000). Our results indicate that LSA-based meaning representations outperform more sophisticated alternatives across the board, whilst lagging behind feature norms only by a small margin.

Meaning Representation

In this section we briefly describe the feature norms used in our experiments. These were based on an existing general purpose database (McRae et al., 2005) which we augmented in several ways to suit our categorization tasks. We also describe three corpus-based models of meaning representation, highlight their differences, and motivate their selection.

Feature Norms

As mentioned earlier, many behavioral experiments have been conducted to elicit semantic feature norms across languages. One of the largest samples for English has been collected by McRae et al. (2005). Their norms consist of 541 basic-level concepts (e.g., DOG and CHAIR) with features collected in multiple studies over several years. For each concept several annotators were asked to produce a number of relevant features (e.g., *barks*, *has-four-legs*, and *used-for-sitting*). The production frequency of a feature given a particular concept can be viewed as a form of weighting indicating the feature's importance for that concept. A spatial representation of word meaning can be extracted from the norms by constructing a matrix in which each row represents a word and each column a feature for that word. Cells in the matrix correspond to the frequency with which a feature was produced in the context of a given word. An example of such a space is shown in Table 2 (a) (the numbers correspond to production frequencies, e.g., 12 participants thought *has-legs* is a feature of TABLE).

Unfortunately, McRae et al.'s (2005) norms do not include any explicit relational information. Because we are interested in using the norms in a model of categorization it was necessary for us to augment the concepts with category labels (e.g., 'dog' is an ANIMAL) and typicality ratings (e.g., 'dog' is a typical ANIMAL whereas 'Snoopy' isn't). We collected this information using Amazon Mechanical Turk[1], an online labor marketplace which has been used in a wide variety of elicitation studies and has been shown to be an inexpensive, fast, and (reasonably) reliable source of non-expert annotation for simple tasks (Snow et al., 2008).

We obtained category labels as follows. We presented each participant with twenty unrelated, randomly selected concepts from McRae et al.'s (2005) data set and asked them to label each with the category to which it best belonged. Responses were in the form of free text, i.e., participants were asked to key in a label rather than select one from a list. Each concept was labeled by ten participants; concepts were then grouped according to the resulting categories. Because annotations collected from Mechanical Turk can be noisy we then discarded those categories containing fewer than five unique concepts, leaving 41 categories for 541 exemplars. These category labels are listed in Table 1. To fully integrate them into the norms it was necessary to collect semantic features for them. To do this, we replicated the norming study of McRae et al. (2005), again using Mechanical Turk. Participants were presented with a single concept (drawn from the set of category labels collected in our previous study) and asked to generate ten relevant features. Instructions and examples were taken from McRae et al. (2005). For each category label we collected features from 30 participants, resulting in a large number of features per item. These features were then mapped into the features already present in the norms; as in McRae et al. (2005) this mapping was performed manually.[2]

[1] http://www.mturk.com

[2] The extended database can be downloaded from http://homepages.inf.ed.ac.uk/s0897549/data/.

INSTRUMENT	keyboard	FURNITURE	chair	HOUSING	apartment	DEVICE	stereo
REPTILE	rattlesnake	CONTAINER	bin	VEHICLE	bike	TRANSPORTATION	van
CLOTHING	jeans	STRUCTURE	building	VEGETABLE	carrot	FOOD	bread
HARDWARE	drill	APPLIANCE	stove	BIRD	seagull	GARMENT	coat
HOUSE	cottage	PLANT	vine	TOOLS	hammer	FISH	trout
EQUIPMENT	football	UTENSIL	ladle	THING	doll	ENCLOSURE	fence
TOY	surfboard	KITCHEN	dish	RODENT	rat	INSECT	grasshopper
BUG	beetle	HOME	house	FRUIT	grapefruit	SPORTS	helmet
MAMMAL	horse	OBJECT	door	ACCESSORIES	necklace	COOKWARE	pan
STORAGE	cabinet	BUILDING	apartment	ANIMAL	cat	WEAPON	bazooka

Table 1: Category labels with most typical exemplars produced by participants in category naming and typicality rating study.

This augmented dataset could be used as-is to evaluate a model of categorization on either a category naming or an exemplar generation task (we describe these tasks in detail in the following section). We further wished to use typicality rating as an additional means for evaluation (Voorspoels et al., 2008). We therefore elicited typicality ratings again via Mechanical Turk. Participants were presented with a single category (e.g., FRUIT) along with twenty randomly selected exemplars belonging to the category (e.g., 'cherry', 'apple', and 'tomato') and asked to rate the typicality of each exemplar among members of the category. Typicality ratings for each exemplar-category pair were collected from 20 participants and an overall rating for each exemplar was computed by taking their mean. The highest rated exemplar for each category is shown in Table 1.

We assessed the quality of the data obtained from Mechanical Turk by calculating their *reliability*, namely the likelihood of a similarly-composed group of participants presented with the same task under the same circumstances producing identical results. We split the collected typicality ratings randomly into two halves and computed the correlation between them; this correlation was averaged across three random splits. These correlations were adjusted by applying the Spearman-Brown prediction formula (Storms et al., 2000; Voorspoels et al., 2008). The reliability of the ratings averaged over 41 concepts was 0.64 with a standard deviation of 0.03. The minimum reliability was 0.52 (INSTRUMENT); the maximum was 0.75 (FURNITURE). Reliability on the category naming task was computed similarly, with an average of 0.72, a maximum of 0.91 (INSTRUMENT), and a minimum of 0.13 (STRUCTURE). These reliability figures may seem low compared with Storms et al. (2000) who perform a similar study. However, note that they conduct a smaller scale experiment; they only focus on eight common natural language concepts (whereas we include 41), and 12 exemplars for each concept (our exemplars are 541).

Data-driven Approaches

In addition to feature norms, we obtained semantic representations for categories and exemplars from natural language corpora. We compared three computational models: Latent Semantic Analysis (LSA; Landauer and Dumais 1997), Latent Dirichlet Allocation (LDA; Griffiths et al. 2007b; Blei et al. 2003), and Dependency Vectors (DV; Padó and Lapata 2007). LSA has historically been a popular method of extracting meaning from corpora, and has been successful at explaining a wide range of behavioral data — examples include lexical priming, deep dyslexia, text comprehension, synonym selection, and human similarity judgments (see Landauer and Dumais 1997 and the references therein). LSA provides a simple procedure for constructing spatial representations of word meanings. The same is true for dependency vectors where co-occurrence statistics are computed between words attested in specific syntactic relations (e.g., *object-of*, *subject-of*). The assumption here is that syntactic information provides a linguistically informed context, and therefore a closer reflection of lexical meaning. LDA, in contrast, imposes a probabilistic model onto those distributional statistics, under the assumption that hidden topic variables drive the process that generates words. Both spatial and topic models represent the meanings of words in terms of an n-dimensional series of values, but whereas semantic spaces treat those values as defining a vector with spatial properties, topic models treat them as a probability distribution.

Latent Semantic Analysis To create a meaning representation for words LSA constructs a word-document co-occurrence matrix from a large collection of documents. Each row in the matrix represents a word, each column a document, and each entry the frequency with which the word appeared within that document. Because this matrix tends to be quite large it is often transformed via a singular value decomposition (Berry et al., 1995) into three component matrices: a matrix of word vectors, a matrix of document vectors, and a diagonal matrix containing singular values. Remultiplying these matrices together using only the initial portions of each (corresponding to the use of a lower dimensional spatial representation) produces a tractable approximation to the original matrix. This dimensionality reduction can be thought of as a means of inferring latent structure in distributional data whilst simultaneously making sparse matrices more informative. The resulting lower-dimensional vectors can then be used to represent the meaning of their corresponding words; example representations in LSA space are shown in Table 2 (b) (vector components represent tf-idf scores).

Dependency Vectors Analogously to LSA, the dependency vectors model constructs a co-occurrence matrix in which each row represents a single word; unlike LSA, the columns of the matrix correspond to other words in whose syntactic context the target word appears. These dimensions may be either the context word alone (e.g., *walks*) or the context word paired with the dependency relation in which it occurs (e.g., *subj-of-walks*). Many variants of syntactically aware semantic space models have been proposed in the literature. We adopt the framework of Padó and Lapata (2007) where a semantic space is constructed over dependency paths, namely

sequences of dependency edges extracted from the dependency parse of a sentence. Three parameters specify the semantic space: (a) the *content selection function* determines which paths contribute towards the representation (e.g., paths of length 1), (b) the *path value function* assigns weights to paths (e.g., it can be used to discount longer paths, or give more weight to paths containing subjects and objects as opposed to determiners or modifiers.), and (c) the *basis mapping function* creates the dimensions of the semantic space by mapping paths that end in the same word to the same dimension. A simple dependency space in shown in Table 2 (c) (vector components represent co-occurrence frequencies).

Latent Dirichlet Allocation Unlike LSA and DV, LDA is a probabilistic model of text generation. Each document is modeled as a distribution over K topics, which are themselves characterized as distribution over words. The individual words in a document are generated by repeatedly sampling a topic according to the topic distribution and then sampling a single word from the chosen topic. Under this framework the problem of meaning representation is expressed as one of statistical inference: give some data — words in a corpus, for instance — infer the latent structure from which it was generated. Word meaning in LDA is represented as a probability distribution over a set of latent topics. In other words, the meaning of a word is a vector whose dimensions correspond to topics and values to the probability of the word given these topics; the likelihood of seeing a word summed over all possible topics is always one. Example representations of words in LDA space appear in Table 2 (d) (vector components are topic-word distributions).

Implementation All three models of word meaning were trained on the British National Corpus. For the LSA model we used the implementation provided in the Infomap toolkit[3], with words represented as vectors in a 100-dimensional space; for the DV model we used the implementation[4] of Padó and Lapata (2007) with dependency paths up to length 3 and a length-based path value function that assigns each path a value inversely proportional to its length, thus giving more weight to shorter paths corresponding to more direct relationships. We obtained dependency information from the output of MINIPAR, a broad coverage dependency parser (Lin, 2001). Infrequent dependencies attested less than 500,000 times in the BNC were discarded. The LDA model used the implementation[5] of Phan et al. (2008) with 100 topics. Inference in this model is based on a Gibbs sampler which we ran for 2,000 iterations. Additionally, LDA has two hyperparameters α and β which were set to 0.5 and 0.1, respectively.

(a) Feature Norms

	has_4_legs	*used_for_eating*	*is_a_pet*	...
TABLE	12	9	0	...
DOG	14	0	15	...

(b) LSA

	Document 1	Document 2	Document 3	...
TABLE	0.02	0.98	-0.12	...
DOG	0.73	-0.02	0.01	...

(c) DV

	subj-of-walk	*subj-of-eat*	*obj-of-clean*	...
TABLE	0	3	28	...
DOG	36	48	19	...

(d) LDA

	Topic 1	Topic 2	Topic 3	...
TABLE	0.02	0.73	0.04	...
DOG	0.32	0.01	0.02	...

Table 2: Semantic representations for 'table' and 'dog' using feature norms, Latent Semantic Analysis (LSA), Dependency Vectors (DV), and Latent Dirichlet Allocation (LDA).

Categorization

Models

The semantic representations described above served as the input to two categorization models, representative of the exemplar-based and prototype-based approaches. In the generalized context model (GCM, Nosofsky 1988; Medin and Schaffer 1978) categories are represented by a list of stored exemplars and inclusion of an unknown item in a category is determined by the net similarity between the item and each of the category's exemplars. Specifically, the similarity $\eta_{w,j}$ of a novel item w to the category c is calculated by summing its similarity to all stored items i belonging to c:

$$\eta_{w,c} = \sum_{i \in c} \eta_{w,i} \tag{1}$$

To calculate the inter-item similarity $\eta_{w,i}$ we compute the cosine of the angle between the vectors representing w and i:

$$\eta_{w,i} = cos(\theta) = \frac{v_w \cdot v_i}{||v_w||\ ||v_i||} \tag{2}$$

Following Vanpaemel et al. (2005), we can modify Equation (1) into a prototype model by replacing the list of stored exemplars with a single 'prototypical' exemplar c_j:

$$\eta_{w,c} = \eta_{w,c_j} \tag{3}$$

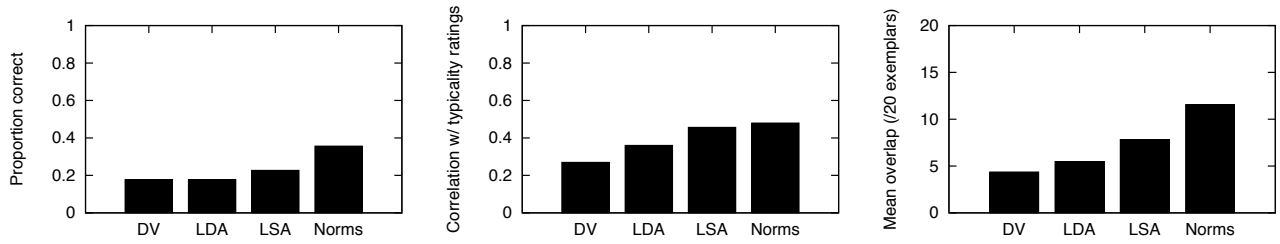
For the category prototype c_j we use the representation of the category label, e.g., the prototype for the category FRUIT is the semantic representation of the word 'fruit'. The similarity between an item and a category thus reduces to the cosine distance between the item and prototype representations.

Tasks

We evaluated the performance of our models on three categorization tasks introduced in Storms et al. (2000): category naming, typicality rating, and exemplar generation.

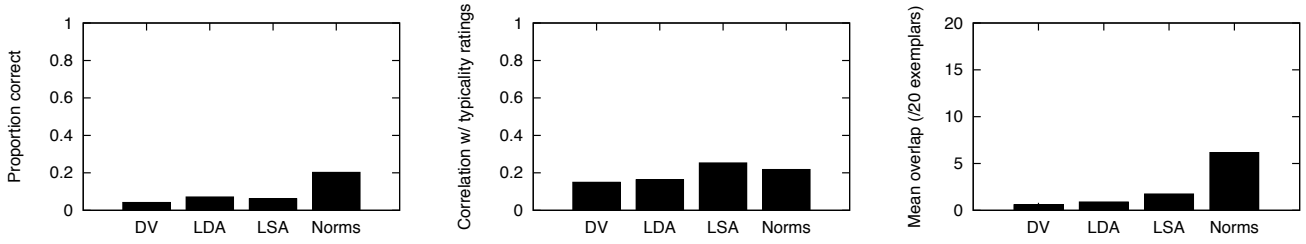
In *category naming* the model is presented with a previously unencountered word and must predict the most appropriate category to which it belongs, e.g., the exemplar 'apple' would be most correctly identified as a member of the category FRUIT, or (with lesser likelihood) FOOD or TREE. In the exemplar model (see (1)), we measure the similarity $\eta_{w,c}$

[3] http://infomap.stanford.edu/
[4] http://www.nlpado.de/~sebastian/dv.html
[5] http://gibbslda.sourceforge.net/



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 1: Performance of exemplar model using feature norms and data-driven meaning representations.



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 2: Performance of prototype model using feature norms and data-driven meaning representations.

of the novel word against all previously encountered exemplars and select the category with the highest *net* similarity between its exemplars and the word in question; for the prototype model (see (3)) this is the category with the highest similarity between the word and the category's label. Performance on the category naming task was determined in a leave-one-out fashion: a single exemplar was removed from the training examples and then categorized. This was repeated for each exemplar in the training set. The latter consisted of 41 subject-produced category labels each with an average of 30 exemplars.

In a *typicality rating task* the model is presented with both an exemplar and label of the category to which it belongs, and must predict the degree to which it is common amongst members of that category. For the category FOOD, for example, 'pizza' or 'bread' would be considered highly typical exemplars, while 'lutefisk' or 'black pudding' would likely be considered much more atypical. The predicted typicality rating for a word and a category is simply the similarity between the two. In the exemplar model this is the sum similarity between the word and each of the category's exemplars; in the prototype model this is the similarity between the category's label and the word. Performance on the typicality rating task was evaluated by computing the correlation between the models' predicted typicality ratings and the average value predicted by the participants of our rating study. The dataset included typicality ratings for 1,228 exemplar-category pairs.

In an *exemplar generation* task the model is given a category label and must generate exemplars typical of the category, e.g., for FOOD we might generate 'pizza', 'bread', 'chicken', etc. Given a category the model selects from the exemplars known to belong those that are most typical; typicality is again approximated by word-category similarities as determined by the model-specific $\eta_{w,c}$. We evaluate performance on the exemplar generation task by computing the average overlap (across categories) between the exemplars generated by the model and those ranked as most typical of the category by our participants.

Results

Figure 1 summarizes our results with the exemplar model and four meaning representations: McRae et al.'s (2005) feature norms (Norms), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Dependency Vectors (DV). Results are shown for category naming (Figure 1(a)) typicality rating (Figure 1(b)) and exemplar generation (Figure 1(c)). We examined performance differences between models using a χ^2 test (category naming and exemplar generation) and Fisher's r-to-z transformation (to compare correlation coefficients for the typicality rating task).

On category naming the exemplar model performs significantly better with the feature norms than when using any of the three corpus-derived representations ($p < 0.01$); however, LSA performs significantly better ($p < 0.05$) than DV or LDA. On typicality rating there is no significant difference between the feature norms and LSA. The norms are significantly better ($p < 0.01$) than either DV or LDA, while LSA surpasses both of the other two corpus-derived representations ($p < 0.01$). Additionally, LDA performs significantly better than DV ($p < 0.05$). On the exemplar generation task the feature norms are significantly better ($p < 0.01$) than any of the corpus-based representations; similarly, LSA performs significantly better than LDA or DV ($p < 0.01$), while LDA again outperforms the dependency space ($p < 0.05$).

Our results with the prototype model are shown in Figure 2 and broadly follow a similar pattern. On category naming the feature norms outperform any of the corpus-based representations ($p < 0.01$), LSA is significantly better than LDA which

in turn is better than DV ($p < 0.05$). On typicality rating there is no significant difference between the feature norms and LSA; the difference between LSA and either of the other two representations is significant ($p < 0.01$). On the exemplar generation task feature norms significantly outperform all other representations ($p < 0.01$); LSA is significantly better ($p < 0.01$) than LDA or DV.

Discussion

In this work we have quantitatively evaluated feature norms and alternative corpus-based meaning representations on three natural language categorization tasks. Perhaps unsurprisingly our results indicate that feature norms are more accurate representations when compared to corpus-based models. As feature norms rely on explicit human judgment, they are able to capture the dimensions of meaning that are psychologically salient. Corpus-based models on the other hand learn in an unsupervised fashion and require no human involvement or external sources of knowledge.

Overall we find LSA to be a reasonable approximation of feature norms, superior to both LDA and the syntactically more aware dependency vectors. This result is consistent across models (exemplar vs. prototype) and tasks. Importantly, the LSA model is language-independent and capable of extracting representations for an arbitrary number of words. By contrast, feature norms tend to cover a few hundred words and involve several subjects over months or years. Albeit in most cases better than our models, feature norms themselves yield relatively low performance on all three tasks we attempted using either an exemplar or prototype model (see Figures 1 and 2). We believe the reasons for this are twofold. Firstly, McRae et al.'s 2005 norms were not created with categorization in mind, we may obtain better predictions with some form of feature weighting (see Storms et al. 2000). Secondly, the tasks seem hard even for humans as corroborated by our reliability ratings.

The differences in performance between LSA, LDA, and DV can be explained by differences between the notion of similarity implicit in each. Closely related words in LDA appear in the same *topics*, which are often corpus-specific and difficult to interpret; words belonging to different categories may be deemed similar yet be semantically unrelated. By contrast, the poor performance of the DV model is somewhat disappointing. Our experiments used a large number of dependency relations; it is possible that a more focused semantic space with a few target relations may be more appropriate.

Finally, our simulation studies reveal that an exemplar model is a better predictor of categorization performance than a prototype one. This result is in agreement with previous studies (Voorspoels et al., 2008; Storms et al., 2000) showing that exemplar models perform consistently better across a broad range of natural language concepts from different semantic domains. This finding is also in line with studies involving artificial stimuli (e.g., Nosofsky 1992).

Directions for future work are two-fold. Firstly, we wish to explore alternative meaning representations more suited to the categorization task. A potential candidate is the feature-topic model (Steyvers, 2009; Andrews et al., 2009), in which documents are represented by a mixture of learned topics in addition to predefined topics derived from feature norms. Secondly, we expect that developing specialized models for natural language categorization that are tailored to data-driven meaning representations would improve performance.

References

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Berry, M., Dumais, S., and O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying rational models of categorization via the heriarchical dirichlet process. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007b). Topics in semantic representation. *Psychological Review*, 114:2007.

Heit, E. and Barsalou, L. (1996). The instantiation principle in natural language categories. *Memory*, (4):413–451.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Lin, D. (2001). LaTaT: Language and text analysis tools. In *Proceedings of the 1st Human Language Technology Conference*, pages 222–227, San Francisco, CA.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavioral Research Methods Instruments & Computers*, 37(4):547–559.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:700–708.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In Healy, A. F., Josslyn, S. M., and Shiffrin, R. M., editors, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, volume 1, pages 149–167. Hillsdale, NJ: Erlbaum.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of The 17th International World Wide Web Conference (WWW 2008)*, pages 91–100.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, pages 328–350.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2006). A more rational model of categorization. *In Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Sloman, S. A. and Ripps, L. J. (1998). Similarity as an explanatory construct. *Cognition*, (65):87–101.

Smith, E. and Medin, D. (1981). *Categories and Concepts*. Harvard University Press.

Snow, R., OĆonnor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP 2008*.

Steyvers, M. (2009). Combining feature norms and text data with topic models. *Acta Psychologica*. (in press).

Storms, G., Boeck, P. D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, 42:51–73.

Vanpaemel, W., Storms, G., and Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637.

Zeigenfusse, M. D. and Lee, M. D. (2009). Finding the features that represent stimuli. *Acta Psychological*. (in press).