Large-Scale Acquisition of Feature-Based Conceptual Representations from Textual Corpora

Barry Devereux (barry@csl.psychol.cam.ac.uk)[1], Nicholas Pilkington (ncvp2@cam.ac.uk)[2], Thierry Poibeau (thierry.poibeau@ens.fr)[3], Anna Korhonen (alk23@cam.ac.uk)[2]

[1] Centre for Speech, Language and the Brain, Department of Experimental Psychology, University of Cambridge

[2] Computer Laboratory, University of Cambridge

[3] Laboratoire LaTTiCe-CNRS, Paris

Abstract

Methods for estimating people's conceptual knowledge have the potential to be very useful to theoretical research on conceptual semantics. Traditionally, feature-based conceptual representations have been estimated using property norm data; however, computational techniques have the potential to build such representations automatically. The automatic acquisition of feature-based conceptual representations from corpora is a challenging task, given the unconstrained nature of what can constitute a semantic feature. Existing computational methods typically do not target the full range of concept-relation-feature triples occurring in human generated norms (e.g. *tiger have stripes*) but rather focus on concept-feature tuples (e.g. *tiger – stripes*) or triples involving specific relations only. We investigate the large-scale extraction of concept-relation-feature triples and the usefulness of encyclopedic, syntactic and semantic information in guiding the extraction process. Our method extracts candidate triples (e.g. *tiger have stripes*, *flute produce sound*) from parsed corpus data and ranks them on the basis of semantic information. Our investigation shows the usefulness of external knowledge in guiding feature extraction and highlights issues of methodology and evaluation which need to be addressed in developing models for this task.

Keywords: distributed conceptual representations; semantic features; corpus-based acquisition

Introduction

Concrete concepts like TIGER, APPLE and CHISEL constitute a fundamental part of people's coherent mental representations of the world around them. A key question in cognitive science is how these semantic representations are organised and accessed. Most theories of conceptual representation assume a distributed, feature-based model of conceptual knowledge (e.g. Cree, McNorgan, & McRae, 2006; Randall, Moss, Rodd, Greer, & Tyler, 2004; Tyler, Moss, Durrant-Peatfield, & Levy, 2000). According to such theories, conceptual knowledge is distributed across a network of interconnected feature units (such as *has_eyes*, *has_ears*, *has_stripes*) with concepts' meanings being represented as patterns of activation across these units. The relative prominence of this distributed, feature-based account of conceptual representation in the literature reflects the many perceived strengths of such a framework.

A key issue for all studies which aim to test distributed theories of concepts is the accurate estimation of the knowledge that people are likely to represent in such a system. Recent connectionist, behavioural and neuropsychological studies (e.g. Cree et al., 2006; Grondin, Lupker, & McRae, 2009; Randall et al., 2004; Tyler et al., 2000; Taylor, Salamoura, Randall, Moss, & Tyler, 2008) have relied on data derived from property norming studies. Currently, the largest set of norms available is that collected by Ken McRae and colleagues which contains features for 541 concrete concepts (McRae, Cree, Seidenberg, & McNorgan, 2005). Participants listed features for each concept word and McRae et al. normalised them by mapping different feature descriptions with the same meaning to the same feature label.

Feature-based representations of concepts based on property-norming studies have played an important role in testing theories of conceptual knowledge. However, property norms come with several important caveats (see e.g. Murphy, 2002, for a discussion). One issue is that participants tend to under-report features which are present in many of the concepts in a category (McRae et al., 2005; Murphy, 2002, p. 32); for TIGER for example, participants list salient features like *has_teeth* but not less salient features like *has_eyes*. Thus *has_eyes* is not listed for TIGER although presumably all McRae et al.'s participants knew that tigers have eyes. Another concern is the size of the currently available property norms. Although the largest collection of norms lists features for over 500 concepts, larger sets of norms would be useful given the number of confounding variables (word length, familiarity, etc) that need to be controlled for in studies of concepts and word meaning. Unfortunately, large scale property norming studies are costly and time consuming.

In recent years, researchers have begun to develop methods which can automatically extract feature norm-like representations using corpus-based computational techniques (e.g. Almuhareb & Poesio, 2005; Barbu, 2008; Baroni, Murphy, Barbu, & Poesio, 2009). These approaches – and the approach we present in this paper – have their antecedents in early methods for extracting and organizing the semantic feature information implicit in dictionary definitions (e.g. Chodorow, Byrd, & Heidorn, 1985). The automatic approach is cost-effective and can gather large-scale frequency data from text corpora. As corpora contain words denoting concepts and their features in natural language, they provide ideal material for feature generation. However, current methods target concept-feature tuples only or are restricted to specific relations between concepts and their features. For example, Almuhareb and Poesio (2005) targeted *is-a* and *part-of* relations, whilst Barbu (2008) combined linguistic patterns with a co-occurrence based method to extract six types of features: *superordinate*, *part*, *stuff*, *location*, *quality* and *action*.

The Strudel model (Baroni et al., 2009) also uses linguis-

tic patterns, but more generally. Strudel uses "connector patterns" consisting of sequences of part-of-speech tags to look for candidate feature terms near a target concept. Properties are scored based on the number of distinct patterns connecting them to a concept, rather than on the overall number of corpus co-occurrences. When evaluated against the ESSLLI dataset that includes 44 concepts from the McRae norms (Baroni, Evert, & Lenci, 2008), Strudel yields the precision of 23.9% – which is the best state of the art result for unconstrained acquisition of concept feature tuples.

Due to the difficulty of the task, we believe that additional linguistic and world knowledge will be required to extract more accurate representations. Moreover, Strudel has the limitation that it produces concept-feature tuples – not concept-relation-feature triples similar to those in human generated norms (although the distribution of the connector patterns for a tuple does cue information about the broad class of semantic relation that holds between concept and feature).

In this paper, we investigate the challenges that need to be met in both methodology and evaluation when aiming to move towards unconstrained, large-scale extraction of concept-relation-feature triples in corpus data. The extraction of such realistic, human-like feature norms is extremely challenging and we do not predict a high level of accuracy in these first experiments. We investigate the usefulness of three types of external knowledge in guiding feature extraction: encyclopedic, syntactic and semantic knowledge. We first compile large automatically parsed corpora from Wikipedia which contains encyclopedic information. We then introduce a novel method which extracts concept-relation-feature triples from grammatical dependences produced by a parser. We use probabilistic information about semantic classes of features and concepts to guide the acquisition process. Our investigation shows that external knowledge can be useful in guiding the extraction of human-like norms.

Extraction Method

Corpora

We chose Wikipedia as our corpus as it is a freely available and comprehensive encyclopedia that includes basic information on many everyday topics. Almost all concepts in the norms have their own Wikipedia articles, and the articles often include facts similar to those elicited in norming studies (e.g. the article *Elephant* describes how elephants are large, are mammals, and live in Africa). By using Wikipedia, we investigate the usefulness of a smaller amount of more focused (encyclopedic) corpus data for the task.

The XML dump of Wikipedia was filtered to remove non-encyclopedic articles (e.g. talk pages), article sections that are unlikely to contain parsable text (e.g. bibliography sections), and inline references (e.g. book citations). The remaining content was preprocessed with Wikiprep (Gabrilovich & Markovitch, 2007), removing tables, unparsable elements (e.g. Wikipedia infoboxes) and the WikiMedia mark-up, yielding a plaintext version of each article. Two subcorpora were created from the resultant set of 1.84 million articles. The first of these (Wiki500) includes the Wikipedia articles that correspond to each of the McRae concepts. It contains *c*. 500 articles (1.1 million words). The second subcorpus consists of those articles which contain one of the McRae concept words in the title and the title is less than five words long.[1] This Wiki110K corpus includes 109,648 plaintext articles (36.5 million words).

Recoding the McRae features

We recoded a British English version of the McRae norms to a uniform representation that is more appropriate for our computational work. Each concept-feature pair in the norms (e.g. TIGER *has_stripes*) was automatically recoded to a triple of the form *concept relation feature-head* where *concept* was the singular of the concept noun (e.g. 'tiger'), *relation* was the root form of a verb (e.g. 'have') and *feature-head* was always a singular noun or an adjective (e.g. 'stripe'). Feature-heads containing more complex information than could be captured with a single noun or adjective were split into two or more triples (for example, the norm feature *is a_musical instrument* for ACCORDION was recoded to the two triples *accordion be instrument* and *accordion be musical*). Where "beh" and "inbeh" appeared in features in the norms (indicating behaviour features of animate and inanimate concepts; e.g. DOG *beh_bark*) this was replaced with the verb "do". Prepositions and determiners were also removed when constructing the triples. Although this recoding involves a loss of information to some extent, it also enables us to clearly distinguish between the relation and feature-head parts in each feature norm. It is triples of this form that we aim to extract with our computational method.

Candidate feature extraction

Our method for extracting concept-relation-feature triples consists of two stages: we first extract large sets of candidate feature triples for each target concept from the corpus, and then re-rank and filter the triples with the aim of retaining only those triples which are most likely to be true semantic features.

For the first stage, the corpora are parsed using the Robust Accurate Statistical Parsing (RASP) system (Briscoe, Carroll, & Watson, 2006). For each sentence in the corpora, this yields the set of grammatical relations (GRs) for the most probable analysis returned by the parser. The GR sets for each sentence containing the target concept noun are then retrieved from the corpus. We construct an undirected acyclyic graph of the GRs that spans the sentence and which has the target concept word as its root node. The nodes are labelled by the words occurring in the sentence and an edge is present when a GR links those two words in the sentence. Edges can thus be labelled by the GR types. For example, the graph

[1] The subset was limited to articles with titles less than five words long in order to avoid articles on very specific topics which are unlikely to contain basic information about the target concept (e.g. *Coptic Orthodox Church of Alexandria* for CHURCH.)

constructed for the sentence *Tabby tigers can often have pale stripes* contains a path connecting *tiger*, *have* and *stripe*.

Our method considers the set of paths through the tree between the target concept root node and the other nodes which are either an adjective or a noun; these adjectives and nouns are the potential feature heads in the concept-relation-feature triples. If there is a verb in the path between the target concept and the feature head, we extract the candidate triple *concept verb feature-head*. The first stage of our method extracts all possible candidate triples from the set of paths. As this method is maximally greedy, the second stage evaluates the quality of these extracted candidates using semantic information, with the aim of filtering out the poor quality features.

Re-ranking based on semantic information

The more often a triple is extracted for a concept, the more likely it is that the triple corresponds to a feature related to the concept. However, production frequency alone is an inadequate measure of the quality of the feature term because concept terms and candidate feature terms can co-occur for all sorts of reasons. For example, one of the extracted triples for TIGER is *tiger have squadron* (because of the RAF squadron called the Tigers).

The probability of a feature being part of a concept's representation is dependent on the semantic category that the concept belongs to (*used for_cutting* should have low probability for animals, for example). We conducted an analysis of the norms to quantify this type of semantic information. Our aim was to identify higher-order structure in the distribution of semantic classes for features and concepts, with the goal of investigating whether this information is useful in feature extraction. More formally, we assume that there is a 2-dimensional probability distribution over concept and feature classes, $P(C,F)$, where C is a concept class (e.g. *Animal*) and F is a feature class (e.g. *Body-Part*). Knowing this distribution gives a way of evaluating how likely it is that a candidate feature f is true for a concept c, assuming that we know that $c \in C$ and $f \in F$. We can regard the McRae norms as being a sample drawn from this distribution, provided the concept and feature terms appearing in the norms can be assigned to suitable concept and feature classes. Clustering was used to identify such classes.

Clustering Our cluster analysis used Lin's (1998) similarity metric, which uses the WordNet ontology as the basis for calculating similarity. Such a measure is appropriate for our purposes as we are interested in generating suitable superordinate classes for which we can calculate the distributional statistics. The concepts and feature-head terms appearing in the recoded norms were each clustered independently into 50 clusters using hierarchical clustering. Table 1 presents three concept clusters and three feature clusters with five representative members of each cluster (we have given intuitive labels to the clusters for explanatory purposes). In general, semantically similar concepts and features clustered together.

We calculated the conditional probability $P(F|C)$ of a feature cluster given a concept cluster using the data in the McRae norms. Table 2 gives the conditional probability for each of the three feature clusters given each of the three concept clusters that were presented in Table 1. For example, $P(\textit{Body Parts}|\textit{Reptiles})$ is higher than $P(\textit{Body Parts}|\textit{Vehicles})$: given a concept in the *Reptiles* cluster the probability of a *Body_Part* feature is relatively high whereas given a concept in the *Vehicle* cluster the probability of a *Body_Part* feature is low. The cluster analysis therefore supports our hypothesis that the likelihood of a particular feature for a particular concept is not independent of the semantic categories that the concept and feature belong to.

Clusters	Example Members
Concept clusters	
Reptiles	alligator, crocodile, iguana, rattlesnake
Fruit/Veg	cucumber, honeydew, mushroom, plum
Vehicles	ambulance, helicopter, car, rocket, jet
Feature clusters	
Body Parts	ear, foot, fuzz, nose, tongue
Plant Parts	bark, berry, blade, grape, prune
Activities	cluck, drip, emergency, flow, funeral

Table 1: Example members of concept and feature clusters

	Reptiles	Fruit/Veg	Vehicles
Body Parts	0.164	0.031	0.023
Plant Parts	0.009	0.130	0.014
Activities	0.100	0.060	0.140

Table 2: $P(F|C)$ for $C \in \{\text{Reptiles, Fruit/Veg, Vehicles}\}$ and $F \in \{\text{Body Parts, Plant Parts, Activities}\}$

Reranking We used this distributional semantic information to improve the quality of the *concept relation feature* candidate triples, by using the conditional probabilities of the appropriate feature cluster given the concept cluster as a weighting factor. To get the probabilities for a triple, we first find the clusters that the concept and the feature-head words belong to. When the feature-head word of the extracted triple appears in the norms, its cluster membership is looked up directly; when it is not in the norms we assign the feature-head to the feature cluster with which it has the highest average similarity. Given the concept and feature clusters determined for the concept and feature in the triple, we reweight the triple's frequency by multiplying it by the conditional probability. This helps downgrade incorrect triples that occur frequently in the data and boost the evidence for correct triples.

Baseline model For the purposes of evaluation, we also implemented a co-occurrence-based model based on the "SVD" (Singular Value Decomposition) model described by Baroni et al. (2009). A word-by-word co-occurrence matrix was constructed for both our corpora, storing how often each target word co-occurred in the same sentence as each context word. Context words were defined to be the 5,000 most frequent content words in the corpora. Target words were the concept names in the recoded norms, supplemented with the 10,000

most frequent content words in the corpora (with the exception of the 10 most frequent words). The dimensionality of the co-occurrence matrix was reduced to 150 columns by singular value decomposition. Cosine similarity between pairs of target words was calculated and, for each concept word, we chose the 200 most similar target words to be the feature-head terms extracted by the model.

Experimental Evaluation

Methods of Evaluation

We considered several methods for evaluating the quality of the extracted feature triples. One method is to calculate precision and recall for the extracted triples with respect to the McRae norms "gold standard". However, direct comparison with the recoded norms is problematic since an extracted feature which is semantically equivalent to a triple in the norms may have a different lexical form. For example, *avocado have stone* appears in the recoded norms whilst *avocado contain pit* is extracted by our method; direct comparison of these two triples results in *avocado contain pit* being incorrectly counted as an error. To deal with the fact that semantically identical features can be lexically different, we followed the approach taken in the ESSLLI 2008 Workshop on semantic models (Baroni et al., 2008). The gold standard for the ESSLLI task was the top 10 features for 44 of the McRae concepts: for each feature an expansion set was given, listing words that were synonyms of the feature term that appeared in the norms. For example, the feature *lives on water* was expanded to the set {*aquatic*, *lake*, *ocean*, *river*, *sea*, *water*}.

We expect to find correct features in corpus data which are not in the "gold standard" (e.g. *breathes_air* is listed for WHALE but for no other animal). We therefore aim for high recall in the evaluation against the ESSLLI set (since all features in the norms should ideally be extracted) but not necessarily high precision (since extracted features that are not in the norms may still be correct; e.g. *breathes_air* for TIGER). To evaluate the ability of our model to generate such novel features, we also conducted a manual evaluation of the highest ranked extracted features which did not appear in the norms. Finally, we introduce a novel evaluation method which makes no direct use of McRae norms. This is based on analysis of the extracted feature-based semantic reprentations in terms of conceptual structure properties. Conceptual structure statistics such as feature distinctiveness, sharedness and correlation strength have an important role to play in testing distributed theories of conceptual knowledge (e.g. see Randall et al., 2004; Taylor et al., 2008). Therefore, we were interested in the accuracy of the conceptual structure statistics that can be calculated from the extracted features. If the conceptual structure statistics calculated for the extracted features resemble those obtained from human-generated norms, it provides evidence that the extracted features capture important aspects of the semantics of concrete concepts.

Extraction set	Corpus	Prec.	Recall
SVD Baseline	Wiki500	0.0235	0.4712
	Wiki110K	0.0140	0.2798
Method - unfiltered	Wiki500	0.0239	0.5081
	Wiki110K	0.0068	0.8083
Method - top 25% unweighted	Wiki500	0.0470	0.2735
	Wiki110K	0.0179	0.6260
Method - top 25% weighted	Wiki500	0.0814	0.4167
	Wiki110K	0.0230	0.6851

Table 3: Results for the baseline model and the extraction method, when matching on features but not relations.

Precision and Recall

The *recall score* for a concept is defined as the number of extracted features for the concept that appear in the recoded norms divided by the total number of features for that concept in the norms. High recall indicates that a high proportion of the McRae features are being extracted. The *precision score* for a concept is defined as the number of extracted features for that concept that appear in the norms divided by the total number of features extracted for the concept.[2] As discussed above, we aim to maximize recall.

Table 3 presents the results when we evaluate using the feature-head term alone (i.e. in calculating precision and recall we disregard the relation verb and require only a match between the feature-head terms in the extracted triples and the recoded norms). Evaluating tuples (rather than triples) is how large-scale models of feature extraction have typically been evaluated in the past (e.g. Baroni et al., 2009).

Results for four sets of extractions are presented. The first set is the set of features extracted by the SVD baseline. The second set of extracted triples are the full set of triples extracted by our method, prior to the reweighting stage. "Top 25% unweighted" gives the results when all but the top 25% most frequently extracted triples for each concept are filtered out. Note that the filtering criteria here is raw extraction frequency, without reweighting by conditional probabilities. "Top 25% weighted" are the corresponding results when the features are weighted by the conditional probability factors prior to filtering; that is, using the top 25% reranked features. The effectiveness of using the semantic class-based analysis data in our method can thus be assessed by comparing the filtered results with and without feature weighting.

For the baseline implementation, the results are better using the smaller Wiki500 corpus than the larger Wiki110K corpus. This is not surprising, since the smaller corpus contains only the articles corresponding to the concepts in the norms. This smaller corpus thus minimizes sources of noise such as word polysemy that are more apparent in the larger corpus (e.g. "tiger" almost always refers to the animal in the Wiki500 corpus, but can have other meanings in larger or general cor-

[2]Since we define precision over the whole set of extracted features, our precision score is not comparable to Baroni et al. (2009), where the top 10 extracted features are used.

pora (the RAF squadron called the Tigers, etc)).

The results for the baseline model and the unfiltered experimental method are quite similar for the Wiki500 corpus. As our extraction method is deliberately greedy, extracting many candidate features per sentence, it is not surprising that its performance is comparable to a purely co-occurrence-based method. The innovation of our method is that it uses information about the GR-graph of the sentence to also extract the verb which appears in the path linking the concept and feature terms in the sentence, which is not possible in a purely co-occurrence-based model.

The results for the unfiltered model using the Wiki110K corpus give the maximum recall achieved by our method; 81% of the features are extracted. Precision is low (because of the large number of features being extracted) although, as discussed above, we are less interested in precision, particularly for the unfiltered model. For the results of the filtered feature sets, where all but the top 25% of features were discarded, we see the benefit of reranking, with the reranked frequencies yielding higher precision and recall scores than the method using the unweighted extracted frequencies.

We also evaluated the extracted triples using the full relation + feature-head pair (i.e. both the feature and the relation verb have to be correct). Previous researchers have typically only compared extracted features to the feature-head term; to our knowledge our work is the first to try and compare extracted features to the full relation + feature norm. Unsurprisingly, this reduces recall and precision compared to the case where only the feature-head terms need match. For example, for the Wiki110K corpus recall falls from 69% to 35% for the filtered re-ranked model. However, given that we impose no constraints on what the relation verb can be and that we do not have expanded synonym sets for verbs it is actually impressive that the verb agrees with what is in the recoded norms about 50% of the time.

Manual Evaluation Analysis

Inspection of the extracted triples reveals that some of them are correct although they do not appear in the gold standard norms. One motivation for developing NLP technology for feature extraction is the need to enrich existing models of conceptual representation with novel features. To evaluate the method's ability to learn this type of novel data, 10 concepts were selected at random from among the McRae concepts and the top 20 extracted triples not present in the norms were selected. Two judges evaluated whether these were genuine errors or valid data missing from the norms. The judges rated each "erroneous" triple as correct, plausible, wrong, or wrong but related. The judges worked first independently and then discussed the results to reach consensus. Across the 10 concepts, 23% and 26% of the relation+feature pairs were considered correct and plausible respectively, indicating roughly half of the errors were not true errors but potentially valid triples missing from the norms. This demonstrates the potential of NLP methods in enriching existing models of conceptual representation.

Measure	Correl	p
Number of features	0.203	< 0.001
Number of distinctive features	0.168	< 0.001
Number of shared features	0.113	0.983
Mean distinctiveness	0.167	< 0.001
Proportion of shared features	0.155	< 0.001
Mean correlational strength	-0.118	0.014

Table 4: Evaluation in terms of CSA variables

Evaluation in terms of conceptual structure

Of particular interest to distributed, feature-based theories of conceptual knowledge is how relationships which exist between the features of concepts influence conceptual processing. Statistics capturing such relationships have proven useful in testing theories of distributed semantic representation, including the conceptual structure account (Randall et al., 2004; Tyler et al., 2000). Researchers have calculated several variables from norm data which capture various aspects of the structural organization of the semantic space (e.g. McRae et al., 2005; Randall et al., 2004). Here, we propose a novel method for evaluating feature extraction methods which is based on testing whether conceptual structure statistics calculated from the extracted features exhibit similar qualities to those calculated on the McRae norms.

Various kinds of conceptual structure variables can be calculated. The simplest is the *number of features* in the concept (i.e. the number of features with non-zero production frequency). Features can also be distinguished by whether they are shared or distinctive. Highly shared features occur in many concepts (e.g. *has_legs*); highly distinctive features occur in few concepts (e.g. *has_an udder*). The reciprocal of the number of concepts that a feature occurs in is a measure of the feature's *distinctiveness* (so a feature occurring in two concepts has distinctiveness of 0.5). In particular, a feature is defined to be *distinguishing* if it occurs in one or two concepts and *shared* if it occurs in more than two concepts. For each concept, we can then define the mean distinctiveness of its features, the number of shared and distinguishing features it has, and the proportion of shared features. We can also define a measure of the strength of interconnection between a pair of features. For example, *has_eyes* and *has_ears* co-occur together in concepts more often than do the features *is_gray* and *has_teeth*. The correlation strength for a pair of features is calculated as the Pearson correlation of their production frequencies across concepts. We can then calculate the *mean correlational strength* of a concept's constituent features (using only the shared features; see Cree et al., 2006; Taylor et al., 2008).We therefore define a total of six conceptual structure variables, summarized in Table 4.

The results show a significant correlation between the norms and the extracted triples for five of the six conceptual structure variables. This is important as it indicates that the semantic representations generated from the extracted features are capturing some aspects of the conceptual structure that is present in the norms. However, the correlations are

quite weak, and we do not see expected differences between living and non-living domains that are observed in the McRae norms. What we wish to highlight here is the potential usefulness of conceptual structure statistics as a means for evaluating models: improvements to the extraction method should yield better quality conceptual structure statistics.

Discussion

The feature acquisition method that we have presented above aims to extract semantically unconstrained concept-relation-feature triples from corpus data. High accuracy extraction of such general representations from corpora is unrealistic given the state of the art. The main goal of our experiment was to investigate issues in both methodology and evaluation which need to be addressed when aiming towards higher accuracy feature extraction in the future. In particular, we examined the usefulness of three types of knowledge for guiding feature extraction: encyclopedic, syntactic, and lexical-semantic. We have also compared different approaches to evaluation: direct evaluation against existing norms, qualitative analysis, and evaluation against conceptual structure variables.

Our extraction method performs better than the co-occurrence-based baseline, demonstrating the benefits of using syntactic information for feature extraction. Using GRs also allows us to extract a relation verb for each concept-feature pair, which is not possible using a purely co-occurrence-based approach like the SVD baseline. Performance was improved further by using semantic constraints calculated from the concept and feature clusters: the re-weighting of features based on distributional data increased the rank of higher-quality features.

Our paper highlights the difficulties inherent in evaluating the quality of extracted features. Evaluation that tests against existing property norms is problematic, since participants in property norming studies list features in unsystematic ways. Furthermore, as property norms are created by normalizing participants' responses to a set of feature labels, direct lexical comparison with property norms is not necessarily meaningful. Although the ESSLLI sub-set of the norms which expands the set of features in the norms with their synonyms goes some way towards addressing the latter issue, the former issue remains: norms are not complete in the sense that there are true features which are not included in the norms.

We therefore considered other forms of evaluation. Our qualitative analysis shows that about 50% of the errors against the recoded norms are in fact correct or plausible features. Our novel evaluation in terms of the conceptual structure variables acts as a valuable task-based evaluation that avoids direct comparison with the norms, and instead compares higher-level structural properties of concepts. Future work can aim for larger-scale qualitative evaluation using multiple judges as well as investigate other task-based evaluations.

Acknowledgments

This research was supported by EPSRC grant EP/F030061/1. We thank McRae et al. for making their norms available.

References

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the web. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 103–108).

Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 9–16).

Baroni, M., Evert, S., & Lenci, A. (Eds.). (2008). *ESSLLI 2008 Workshop on Distributional Lexical Semantics.*

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2009). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 1–33.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the Interactive Demo Session of COLING/ACL-06* (pp. 77–80).

Chodorow, M. S., Byrd, R. J., & Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics* (pp. 299–304).

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(4), 643–58.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI'07* (pp. 1606–1611).

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of ICML'98* (p. 296-304).

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: The MIT Press.

Randall, B., Moss, H. E., Rodd, J. M., Greer, M., & Tyler, L. K. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*(2), 393–406.

Taylor, K. I., Salamoura, A., Randall, B., Moss, H., & Tyler, L. K. (2008). Clarifying the nature of the distinctiveness by domain interaction in conceptual structure: comment on cree, McNorgan, and McRae (2006). *Journal of Experimental Psychology: Learning, Memory & Cognition*, *34*(3), 719–725.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231.