Acquiring Human-like Feature-Based Conceptual Representations from Corpora

Colin Kelly
Computer Laboratory
University of Cambridge
Cambridge, CB3 0FD, UK
colin.kelly
@cl.cam.ac.uk

Barry Devereux
Centre for Speech,
Language, and the Brain
University of Cambridge
Cambridge, CB2 3EB, UK
barry@csl.psychol.cam.ac.uk

Anna Korhonen
Computer Laboratory
University of Cambridge
Cambridge, CB3 0FD, UK
anna.korhonen
@cl.cam.ac.uk

Abstract

The automatic acquisition of feature-based conceptual representations from text corpora can be challenging, given the unconstrained nature of human-generated features. We examine large-scale extraction of concept-relation-feature triples and the utility of syntactic, semantic, and encyclopedic information in guiding this complex task. Methods traditionally employed do not investigate the full range of triples occurring in human-generated norms (e.g. *flute produce sound*), rather targeting concept-feature pairs (e.g. *flute – sound*) or triples involving specific relations (e.g. *is-a*, *part-of*). We introduce a novel method that extracts candidate triples (e.g. *deer have antlers*, *flute produce sound*) from parsed data and re-ranks them using semantic information. We apply this technique to Wikipedia and the British National Corpus and assess its accuracy in a variety of ways. Our work demonstrates the utility of external knowledge in guiding feature extraction, and suggests a number of avenues for future work.

1 Introduction

In the cognitive sciences, theories about how concrete concepts such as ELEPHANT are represented in the mind have often adopted a distributed, feature-based model of conceptual knowledge (e.g. Randall et al. (2004), Tyler et al. (2000)). According to such accounts, conceptual representations consist of patterns of activation over sets of interconnected semantic feature nodes (e.g. *has_eyes*, *has_ears*, *is_large*). To test these theories empirically, cognitive psychologists require an accurate estimate of the kinds of knowledge that people are likely to represent in such a system. To date, the most important sources of such knowledge are property-norming studies, where a large number of participants write down lists of features for concepts. For example, McRae et al. (2005) collected a set of norms listing features for 541 concrete concepts. In that study, the features listed by different participants were normalised by mapping different feature descriptions with identical meanings to the same feature label.[1] Table 1 gives the ten most frequent normed features for two concepts in the norms.

elephant		banana	
Relation	*Feature*	*Relation*	*Feature*
is	large	is	yellow
has	a trunk	is	a fruit
is	an animal	is	edible
is	grey	is	soft
lives	in Africa	grows on	trees
has	ears	eaten by	peeling
has	tusks	-	grows
has	legs	eaten by	monkeys
has	four legs	is	long
has	large ears	tastes	good

Table 1: Sample triples from McRae Norms

However, property norm data have certain weaknesses (these have been widely discussed; e.g. Murphy (2002), McRae et al. (2005)). One issue is that participants tend to under-report features that are present in many of the concepts in a given category (McRae et al., 2005; Murphy, 2002). For example, for the concept ELEPHANT, participants list salient features like *has_trunk*, but not less salient features such as *breathes_air*, even though presumably all McRae et al.'s participants knew that elephants breathe air. Although the largest collection

[1] For example, for CAR, "used for transportation" and "people use it for transportation" were mapped to the same *used_for_transportation* feature.

Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, pages 61–69,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

of norms lists features for over 500 concepts, the relatively small size of property norm sets still gives cause for concern. Larger sets of norms would be useful to psycholinguists; however, large-scale property norming studies are time-consuming and costly.

In NLP, researchers have developed methods for extracting and classifying generic relationships from data, e.g. Pantel and Pennacchiotti (2008), Davidov and Rappoport (2008a, 2008b). In recent years, researchers have also begun to develop methods which can automatically extract feature norm-like representations from corpora, e.g. Almuhareb and Poesio (2005), Barbu (2008), Baroni et al. (2009). The automatic approach is capable of gathering large-scale distributional data, and furthermore it is cost-effective. Corpora contain natural-language instances of words denoting concepts and their features, and therefore serve as ideal material for feature generation tasks. However, current methods are restricted to specific relations between concepts and their features, or target concept-feature pairs only. For example, Almuhareb and Poesio (2005) proposed a method based on manually developed lexico-syntactic patterns that extracts information about attributes and values of concepts. They used these syntactic patterns and two grammatical relations to create descriptions of nouns consisting of vector entries and evaluated their approach based on how well their vector descriptions clustered concepts. This method performed well, but targeted *is-a* and *part-of* relations only. Barbu (2008) combined manually defined linguistic patterns with a co-occurrence based method to extract features involving six classes of relations. He then split learning for the property classes into two distinct paradigms. One used a pattern-based approach (four classes) with a seeded pattern-learning algorithm. The other measured strength of association between the concept and referring adjectives and verbs (two classes). His pattern-based approach worked well for properties in the *superordinate* class, had reasonable recall for *stuff* and *location* classes, but zero recall for *part* class. His approach for the other two classes used various association measures which he summed to establish an overall score for potential properties.

The recent Strudel model (Baroni et al., 2009) relies on more general linguistic patterns, "connector patterns", consisting of sequences of part-of-speech (POS) tags to look for candidate feature terms near a target concept. The method assumes that "the variety of patterns connecting a concept and a potential property is a good indicator of the presence of a true semantic link". Thus, properties are scored based on the count of distinct patterns connecting them to a concept. When evaluated against the ESSLLI dataset (Baroni et al. (2008); see section 3.1), Strudel yields a precision of 23.9% – this figure is the best state-of-the-art result for unconstrained acquisition of concept-feature pairs.

It seems unlikely that further development of the shallow connector patterns will significantly improve accuracy, as these already broadly cover most POS sequences that are concept-feature connectors. Because of the difficult nature of the task, we believe that extraction of more accurate representations necessitates additional linguistic and world knowledge. Furthermore, the utility of Strudel is limited because it only produces concept-feature pairs, and not concept-relation-feature triples similar to those in human generated norms (although the distribution of the connector patterns for a extracted pair does offer clues about the broad class of semantic relation that holds between concept and feature).

In this paper, we explore issues of both methodology and evaluation that arise when attempting unconstrained, large-scale extraction of concept-relation-feature triples in corpus data. Extracting such human-like features is difficult, and we do not anticipate a high level of accuracy in these early experiments. We examine the utility of three types of external knowledge in guiding feature extraction: syntactic, semantic and encyclopedic. We build three automatically parsed corpora, two from Wikipedia and one from the British National Corpus. We introduce a method that (i) extracts concept-relation-feature triples from grammatical dependency paths produced by a parser and (ii) uses probabilistic information about semantic classes of features and concepts to re-rank the candidate triples before filtering them. We then assess the accuracy of our model using several different methods, and demonstrate that external knowledge can help guide the extraction of human-like features. Finally, we highlight issues in both methodology and evaluation that are important for further progress in this area of research.

2 Extraction Method

2.1 Corpora

We used Wikipedia to investigate the usefulness of world knowledge for our task. Almost all concepts in the McRae norms have their own Wikipedia articles, and the articles often include facts similar to those elicited in norming studies.[2] Extraneous data were removed from the articles (e.g. infoboxes, bibliographies) to create a plaintext version of each article. The 1.84 million articles were then compiled into two subcorpora. The first of these (Wiki500) consists of the Wikipedia articles corresponding to each of the McRae concepts. It contains *c*. 500 articles (1.1 million words). The second subcorpus is comprised of those articles where the title is fewer than five words long and contains one of the McRae concept words.[3] This corpus, called Wiki110K, holds 109,648 plaintext articles (36.5 million words).

We also employ the 100-million word British National Corpus (BNC) (Leech et al., 1994) which contains written (90%) and spoken (10%) English. It was designed to represent a broad cross-section of modern British English. This corpus provides an interesting contrast with Wikipedia, since we assume that any features contained in such a wide-ranging corpus would be presented in an incidental fashion rather than explicitly. The BNC may contain useful features which are encoded in everyday speech and text but not in Wikipedia, perhaps due to their ambiguity for encyclopedic purposes, or due to their non-scientific but rather common-sense nature. For example, *eaten by monkeys* is listed as a feature of BANANA in the McRae norms, but the word *monkey* does not appear in the Wikipedia *banana* article.

2.2 Candidate feature extraction

Using a modified, British English version of the published norms, we recoded them to a uniform *concept-relation-feature* representation suitable for our experiments – it is triples of this form that we aim to extract. Our method for extracting concept-relation-feature triples consists of two main stages. In the first stage, we extract large sets of candidate concept-relation-feature triples for each target concept from parsed corpus data. In the second stage, we re-rank and filter these triples with the intention of retaining only those triples which are likely to be true semantic features.

In the first stage, the corpora are parsed using the Robust Accurate Statistical Parsing (RASP) system (Briscoe et al., 2006). For each sentence in the corpora, this yields the most probable analysis returned by the parser in the form of a set of grammatical relations (GRs). The GR sets for each sentence containing the target concept noun are then retrieved from the corpus. These GRs form an undirected acyclic graph, whose nodes are labelled with words in the sentence and their POS, and whose edges are labelled with the GR types linking the nodes together. Using this graph we generate all possible paths which are rooted at our target concept node using a breadth-first search.

We then examine whether any of these paths match prototypical feature-relation GR structures according to our manually-generated rules. The rules were created by first extracting features from the McRae norms for a small subset of the concepts and extracting those sentences from the Wiki500 corpus which contained both concept and feature terms. For each sentence, we then examined each path through the graph (containing the GRs and POS tags) linking the concept, the feature, and all intermediate terms, and (providing no other rule already generated the concept-relation-feature triple) manually generated a rule based on each path.

For example, the sentence *There are also aprons that will cover the sleeves* should yield the triple *apron cover sleeve*. We examine the tree structure of the sentence rooted at the concept (*apron*):

```
apron+s:17_NN2
cmod-that cover:34_VV0
     L--- dobj sleeve+s:44_NN2
          L--- det the:40_AT
     L--- aux will:29_VM
cmod-that cover:34_VV0
xcomp be+:8_VBR
     L--- ncmod also:12_RR
     L--- ncsubj There:2_EX
```

Here, the relation is relatively simple – we merely

[2] e.g. The article *Elephant* describes how elephants are large, are mammals, and live in Africa.

[3] This was done in order to avoid articles on very specific topics which are unlikely to contain basic information about the target concept.

create a rule which requires that the relation is a verb (i.e. has a `V` POS tag), the feature has an `NN` tag and that there is a `dobj` GR linking the feature to the concept. Our rules are effectively a constraint on (a) which paths should be followed through the tree, and (b) which items in that path should be noted in our concept-relation-feature triple. By creating several such rules and applying them to a large number of sentences, we extract potential features and relations for our concepts.

We avoided specifying too many POS tags and GRs in rules since this could have resulted in too few matching paths. In the above example, we could have required also a `cmod-that` relation linking the feature and concept – but this would have excluded sentences like *the apron covered the sleeves*. Conversely, we avoided making our rules too permissive. For example, eliminating the `dobj` requirement would have yielded the triple *apron be steel* from the sentence *the apron hooks were steel.*

The application of this method to a number of concepts in the Wiki500 corpus yielded 15 rules which we employed in our experiments. We extract triples using both singular and plural occurrences of both the concept term and the feature term. We show the first three of our rules in Table 2. The first stage of our method uses the 15 rules to extract a very large number of candidate triples from corpus data.

Rule:	relation of concept has a VVN tag, feature has a NN tag and they are linked by an xcomp GR
S:	*This is an **anchor** which relies solely on being a heavy weight.*
T:	anchor be weight
Rule:	relation of concept is a verb, feature is an adjective and they are linked by an xcomp GR
S:	*Sliced **apples** turn brown with exposure to air due to the conversion of natural phenolic substances into melanin upon exposure to oxygen.*
T:	apple turn brown
Rule:	feature of concept has a VV0 tag, relation is a verb and they are linked by an aux GR
S:	*Grassy bottoms may be good holding, but only if the **anchor** can penetrate the foliage.*
T:	anchor can penetrate

Table 2: Three sample rules for a given **concept**, with example sentence (S) and corresponding triple (T).

2.3 Re-ranking based on semantic information

The second stage of our method evaluates the quality of the extracted candidates using semantic information, with the aim of filtering out the poor quality features generated in the first stage. We would expect the number of times a triple is extracted for a given concept to be proportional to the likelihood that the triple represents a true feature of that concept. However, production frequency alone is not a sufficient indicator of quality, because concept terms can produce unexpected candidate feature terms.[4]

One may attempt to address this issue by introducing semantic categories. In other words, the probability of a feature being part of a concept's representation is dependent on the semantic category to which the concept belongs (for example, *used_for-cutting* would be expected to have low probability for animal concepts). We analysed the norms to quantify this type of semantic information with the aim of identifying higher-order structure in the distribution of semantic classes for features and concepts. The overarching goal was to determine whether this information can indeed improve the accuracy of feature extraction.

In formal terms, 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. *Apparel*) and F is a feature class (e.g. *Materials*). Knowing this distribution provides us with a means of assessing how likely it is that a candidate feature f is true for a concept c, assuming that we know that $c \in C$ and $f \in F$. The McRae norms may be considered to be a sample drawn from this distribution, if the concept and feature terms appearing in the norms can be assigned to suitable concept and feature classes. These classes were identified by way of clustering. The reranking step employed the McRae norms so we could establish an upper bound for the semantic analysis, although we could also use other knowledge resources, e.g. the Open Mind Common Sense database (Singh et al., 2002).

2.3.1 Clustering

We utilised Lin's similarity measure (1998) for our similarity metric, employing WordNet (Fell-

[4] For example, one of the extracted triples for TIGER is *tiger have squadron* because of the RAF squadron called the Tigers.

	k-means	
banjo	biscuit	blackbird
bat	cup	ox
beehive	kettle	peacock
birch	sailboat	prawn
bookcase	shoe	prune
	NMF	
ashtray	bouquet	eel
bayonet	cabinet	grapefruit
cape	card	guppy
cat	cellar	moose
catfish	chandelier	otter
	Hierarchical	
Fruit/Veg	*Apparel*	*Instruments*
apple	apron	accordion
avocado	armour	bagpipes
banana	belt	banjo
beehive	blouse	cello
blueberry	boot	clarinet

Table 3: First five elements alphabetically from three sample clusters for the three clustering methods.

	Hierarchical Clustering	
Plant Parts	*Materials*	*Activities*
berry	cotton	annoying
bush	fibre	listening
core	nylon	music
plant	silk	showing
seed	spandex	looking

Table 4: Example members of feature clusters for hierarchical clustering.

	Fruit/Veg	Apparel	Instruments
Plant Parts	0.144	0.037	0.008
Materials	0.006	0.148	0.008
Activities	0.009	0.074	0.161

Table 5: $P(F|C)$ for $C \in \{$Fruit/Veg, Apparel, Instruments$\}$ and $F \in \{$Plant Parts, Materials, Activities$\}$

baum, 1998) as the basis for calculating similarity. This metric is suitable for our task as we would like to generate appropriate superordinate classes for which we can calculate distributional statistics. We could merely cluster on the most frequent sense of concept and feature words in WordNet, but the most frequent sense in WordNet may not correspond to the intended sense in our feature norm data.[5] So we consider also other senses of words in WordNet by employing a manually-annotated list to choose the correct sense in WordNet. This is only possible for concept clustering since we don't possess a manual WordNet sense annotation for the 7000 McRae features; for the feature clustering, we simply use the most frequent sense in WordNet.

The concepts and feature-head terms appearing in the recoded norms were each clustered independently into 50 clusters using three methods: hierarchical clustering, *k*-means clustering and non-negative matrix factorization (NMF). We show the first five alphabetical elements from three of the clusters produced by our clustering methods in Table 3. The hierarchical clustering seems to be producing the most intuitive clusters.

We calculated the conditional probability $P(F|C)$ of a feature cluster given a concept cluster using the data in the McRae norms. Table 5 gives the conditional probability for each of the three feature clusters given each of the three concept clusters that were presented in Tables 3 and 4 for hierarchical clustering. For example, $P(\mathit{Materials}|\mathit{Apparel})$ is higher than $P(\mathit{Materials}|\mathit{Fruit/Veg})$: given a concept in the *Apparel* cluster the probability of a *Materials* feature is relatively high whereas given a concept in the *Fruit/Veg* cluster the probability of a *Materials* feature is low. The cluster analysis therefore supports our hypothesis that the likelihood of a particular feature for a particular concept is dependent on the semantic categories that both belong to.

2.3.2 Reranking

We investigated whether this distributional semantic information could be used to improve the quality of the candidate triples, by using the conditional probabilities of the appropriate feature cluster given the concept cluster as a weighting factor. To obtain the probabilities for a triple, we first find the clusters that the concept and feature-head words belong to. If the feature-head word of the extracted triple appears in the norms, its cluster membership is drawn directly from there; if not, we assign the feature-head to the feature cluster with which it has the highest average similarity.[6] Having determined the concept and fea-

[5] e.g. the first and second most frequent definitions of *kite* refer to a slang meaning for the word *cheque* – only the third most frequent meaning refers to *kite* as a toy, which most people would understand to be its predominant sense.

[6] We use average-linkage for hiearchical and *k*-means clustering, and mean cosine similarity for NMF.

ture clusters for the triple, we reweight its raw corpus occurrence frequency by multiplying it by the conditional probability. In this way, incorrect triples that occur frequently in the data are downgraded and more plausible triples have their ranking boosted.

2.3.3 Baseline model

We also implemented as a baseline a co-occurrence-based model, based on the "SVD" model described by Baroni and colleagues (Baroni and Lenci, 2008; Baroni et al., 2009) – it is a simple, word-association method, not tailored to extracting features. A context-word-by-target-word frequency co-occurrence matrix was constructed for both corpora, with a sentence-sized window. Context words and target words were defined to be the 5,000 and 10,000 most frequent content words in the corpus respectively. The target words were supplemented with the concept words from the recoded norms. The co-occurrence matrix was reduced to 150 dimensions by singular value decomposition, and cosine similarity between pairs of target words was calculated. The 200 most similar target words to each concept acted as the feature-head terms extracted by this model.

3 Experimental Evaluation

3.1 Methods of Evaluation

We considered a number of methods for evaluating the quality of the extracted feature triples. One possibility would be 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 there may be extracted features which are semantically equivalent to a triple in the norms but possessing a different lexical form.[7]

Since 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 concept-feature pair an expansion set was generated containing synonyms of the feature terms appearing in the norms. For example, the feature *lives on water* was expanded to the set $\{$*aquatic*, *lake*, *ocean*, *river*, *sea*, *water*$\}$.

We would expect to find in corpus data correct features that do not appear in our "gold standard" (e.g. *breathes_air* is listed for WHALE but for no other animal). We therefore aim to attain high recall when evaluating against the ESSLLI set (since ideally all features in the norms should be extracted) but we are somewhat less concerned about achieving 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 that did not appear in the norms.

Extraction set	Corpus	Prec.	Recall
SVD Baseline	Wiki500	0.0235	0.4712
	Wiki110K	0.0140	0.2798
	BNC	0.0131	0.2621
Method - unfiltered	Wiki500	0.0242	0.6515
	Wiki110K	0.0039	0.8944
	BNC	0.0042	0.8813
Method - top 20 (unweighted)	Wiki500	0.1159	0.2326
	Wiki110K	0.0761	0.1523
	BNC	0.0841	0.1692
Method - top 20 (hierarchical clustering)	Wiki500	0.1693	0.3394
	Wiki110K	0.1733	0.3553
	BNC	0.1943	0.3896
Method - top 20 (*k*-means clustering)	Wiki500	0.1159	0.2323
	Wiki110K	0.1000	0.2008
	BNC	0.1216	0.2442
Method - top 20 (NMF clustering)	Wiki500	0.1375	0.2755
	Wiki110K	0.1409	0.2826
	BNC	0.1500	0.3010

Table 6: Results when matching on features only.

3.2 Evaluation

Previous large-scale models of feature extraction have been evaluated on pairs rather than triples e.g. Baroni et al. (2009). Table 6 presents the results of our method 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). Results for six sets of extractions are presented. The first set is the set of features extracted by the SVD baseline.

[7] 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 marked as an error.

The second set of extracted triples consists of the full set of triples extracted by our method, prior to the reweighting stage. “Top 20 unweighted” gives the results when all but the top 20 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 20 (*clustering type*)” are the corresponding results when the features are weighted by the conditional probability factors (derived from our three clustering methods) prior to filtering; that is, using the top 20 reranked features. The effectiveness of using the semantic class-based analysis data in our method can be assessed by comparing the filtered results with and without feature weighting.

For the baseline implementation, the results are better when we use the smaller Wiki500 corpus compared to the larger Wiki110K corpus. This is not surprising, since the smaller corpus contains only those articles which correspond to the concepts found in the norms. This smaller corpus thus minimises noise due to phenomena such as word polysemy which are more apparent in the larger corpus.

The results for the baseline model and the unfiltered method are quite similar for the Wiki500 corpus, whilst the results for the unfiltered method using the Wiki110K corpus give the maximum recall achieved by our method; 89.4% of the features are extracted, although this figure is closely followed by that of the BNC at 88.1%. As the unfiltered method is deliberately greedy, a large number of features are being extracted and therefore precision is low.

Extraction set	Corpus	Prec.	Recall
Method - top 20 (hierarchical clustering)	Wiki500	0.1011	0.2028
	Wiki110K	0.1102	0.2210
	BNC	0.0955	0.1917

Table 7: Results for our best method when matching on features and relations.

For the results of the filtered method, where all but the top 20 of features were discarded, we see the benefit of reranking, with the reranked frequencies for all three clustering types yielding much higher precision and recall scores than the unweighted method. Our best performance is achieved using the BNC and hierarchical clustering, where we obtain 19.4% precision and 38.9% recall. Thus both general and encyclopedic corpus data prove useful for the task. An interesting question is whether these two data types offer different, complementary feature types for the task. We discuss this point further in section 3.3.

Using exactly the same gold standard, Baroni et al. (2009) obtained precision of 23.9%. However, this result is not directly comparable with ours, since we define precision over the whole set of extracted features while Baroni et al. considered the top 10 extracted features only.

The innovation of our method is that it uses information about the GR-graph of the sentence to also extract the relation 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. We therefore 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). The results for our best method are shown in Table 7. Unsurprisingly, because this task is more difficult, precision and recall are reduced. However, since we enforce no constraints on what the relation may be and since we do not have expanded synonym sets for our relations (as we do for our features) it is actually impressive to have both the exact relation verb and feature matching with the recoded norms almost one in every five times. To our knowledge, our work is the first to try to compare extracted features to the full relation and feature norm parts of the triple.

3.3 Qualitative analysis

Since a key aim of our work is to learn novel features in corpus data, we also performed a qualitative evaluation of the extracted features and relations. This analysis revealed that many of the errors were not true errors but potentially valid triples missing from the gold standard. Table 8 shows the top 10 features for two concepts extracted by our best method from the Wiki500 corpus and the BNC corpus. We label those features that are correct according to the norms as Correct (C), those which do not appear in our norms but we believe to be plausible as Plausible (P), and those that do not appear in the norms and are also implausible as Incorrect (I). We can see that our method has detected several plausible features not appearing in the norms (and thus our gold standard), e.g. *swan have chick* and *screwdriver be*

		swan			
Wiki500			BNC		
be	bird	C	have	number	I
be	black	P	have	water	C
have	chick	P	have	lake	C
have	plumage	C	be	bird	C
have	feather	C	be	white	C
restrict	water	C	have	neck	C
be	mute	P	be	wild	P
eat	grass	P	have	duck	I
turn	elisa	I	have	song	I
have	neck	C	have	pair	I
		screwdriver			
Wiki500			BNC		
use	handle	C	have	tool	C
have	blade	P	have	end	P
use	tool	C	have	blade	P
remedy	problem	P	have	hand	I
have	size	P	be	sharp	P
have	head	C	have	bit	P
rotate	end	P	have	arm	I
have	plastic	P	be	large	P
achieve	goal	I	be	sonic	P
have	hand	I	have	range	P

Table 8: Top 10 returned features and relations for *swan* and *screwdriver*.

sharp. Indeed, it could be argued that some 'incorrect' features (e.g. *screwdriver achieve goal*) could be considered to be at least broadly accurate. We recognise that the ideal evaluation for our method would involve having human participants assess the extracted features for a diverse cross-section of our concepts, but this is beyond the scope of this paper.

When considering the top 20 features extracted using our best method applied to the Wiki500 corpus versus the BNC corpus, the overlap of features is relatively low at 22.73%. When one also takes the extracted relations into account, this figure descends to 6.45%. It is clear that relatively distinct groups of features are being extracted from the encyclopedic and general corpus data. Future work could investigate combining these for improved performance e.g. using the intersection of the best features from the BNC and Wiki110k corpora to improve precision and the union to improve recall.

4 Discussion

This paper examined large-scale, unconstrained acquisition of human-like feature norms from corpus data. Our work was not limited to only a subset of concepts, relation types or concept-feature pairs. Rather, we investigated concepts, features and relations in conjunction, and extracted property norm-like concept-relation-feature triples.

Our investigation shows that external knowledge is highly useful in guiding this challenging task. Encyclopedic information proved useful for feature extraction: although our Wikipedia corpora are considerably smaller than the BNC, they performed almost equally well. We also demonstrated the benefits of employing syntactic information in feature extraction: our base extraction method operating on parsed data outperforms the co-occurrence-based baseline and permits us to extract relation verbs. This underscores the usefulness of parsing for semantically meaningful feature extraction. This is consistent with recent work in the field of computational lexical semantics, although GR data has not previously been successfully applied to feature extraction.

We showed that semantic information about co-occurring concept and feature clusters can be used to enhance feature acquisition. We employed the McRae norms for our analysis, however we could also employ other knowledge resources and cluster relation verbs using recent methods, e.g. Sun and Korhonen (2009), Vlachos et al. (2009).

Our paper has also investigated methods of evaluation, which is a critical but difficult issue for feature extraction. Most recent approaches have been evaluated against the ESSLLI sub-set of the McRae norms which expands the set of features in the norms with their synonyms. Yet even expansion sets like the ESSLLI norms do not facilitate adequate evaluation because they are not complete in the sense that there are true features which are not included in the norms. Our qualitative analysis shows that many of the errors against the recoded norms are in fact correct or plausible features. Future work can aim for larger-scale qualitative evaluation using multiple judges as well as investigating other task-based evaluations. For example, we have demonstrated that our automatically-acquired feature representations can make predictions about fMRI activity associated with concept stimuli that are as powerful as those produced by a manually-selected set of features (Devereux et al., 2010).

Acknowledgments

This research was supported by EPSRC grant EP/F030061/1 and the Royal Society University Research Fellowship, UK. We are grateful to McRae and colleagues for making their norms publicly available, and to the anonymous reviewers for their input.

References

Abdulrahman Almuhareb and Massimo Poesio. 2005. Concept learning and categorization from the web. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pages 103–108.

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

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

Marco Baroni, Stefan Evert, and Alessandro Lenci, editors. 2008. *ESSLLI 2008 Workshop on Distributional Lexical Semantics*.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2009. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, pages 1–33.

Edward J. Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Interactive Demo Session of COLING/ACL-06*, pages 77–80.

D. Davidov and A. Rappoport. 2008a. Classification of semantic relationships between nominals using pattern clusters. *ACL.08*.

D. Davidov and A. Rappoport. 2008b. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. *ACL.08*.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL-HLT Workshop on Computational Neurolinguistics*.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

G. Leech, R. Garside, and M. Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML'98*, pages 296–304.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

Gregory Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.

Patrick Pantel and Marco Pennacchiotti. 2008. Automatically harvesting and ontologizing semantic relations. In Paul Buitelaar and Philipp Cimiano, editors, *Ontology learning and population*. IOS press.

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

P. Singh, T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237.

Lin Sun and Anna Korhonen. 2009. Improving Verb Clustering with Automatically Acquired Selectional Preferences. *Empirical Methods on Natural Language Processing*.

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

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82, Athens, Greece.