Minimally Supervised Learning for Unconstrained Conceptual Property Extraction

Colin Kelly (colin.kelly@cl.cam.ac.uk), Anna Korhonen (anna.korhonen@cl.cam.ac.uk)
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK

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

Abstract

We present a highly performant, minimally supervised system for the challenging task of unconstrained conceptual property extraction (e.g., *banana is fruit*, *spoon used for eating*). Our technique employs lightly supervised support vector machines to acquire promising features from our corpora (Wikipedia and UKWAC) and uses those features to anchor the search for plausible unconstrained relations in our corpus. We introduce a novel backing-off method to find the most likely relation for each concept/feature pair and produce a number of metrics which act as potential indicators of true relations, training our system using a stochastic search algorithm to find the optimal reweighting of these metrics. We also introduce a human semantic-similarity dataset; our output shows a strong correlation with human similarity judgements. Both our gold standard comparison and direct human evaluation results improve on those of previous approaches, with our human judgements evaluation showing a significant 20 percentage point performance increase.

Introduction

Recent theories in cognitive psychology attest a property-based, distributed and componential model of conceptual representation for concrete concepts (e.g., ***elephant***, ***screwdriver***) in the brain (Farah & McClelland, 1991; Tyler, Moss, Durrant-Peatfield, & Levy, 2000; Randall, Moss, Rodd, Greer, & Tyler, 2004). To explore the validity of these theories, researchers employ real-world knowledge taken from property norming studies where human volunteers are asked to list properties for concepts. McRae, Cree, Seidenberg, and McNorgan (2005) performed the largest such study to date, collecting properties for over 500 concrete nouns (we call these the 'McRae norms'). Some example properties from these norms can be found in Table 1.

However, as has been widely discussed (Murphy, 2002; McRae et al., 2005), these studies suffer from a number of weaknesses. For example, human participants often under-report certain properties, even when they are facts presumably known by the volunteers: though all participants are likely to have known that animals have hearts, *has* **heart** is not reported as a property for any animal concept. Similarly, *is* **animal** is listed as a property of all animals in the norms while **breathes** is only cited as a property for ***whale***. A related issue is inconsistency across similar concepts: *has* **legs** is listed as a property of ***leopard*** but is absent for ***tiger***.

Our task is to automatically extract such conceptual representations from large text corpora using NLP techniques. We hope to extract features for a given concept as well as those features' relationship with that concept; specifically, we aim to extract properties in the form of ***concept*** *relation* **feature** triples (e.g., ***knife*** *used for* **cutting**, ***pig*** *lives on* **farm**), where both the relation and the feature are unconstrained. Our task is particularly challenging because while we seek a very specific 'type' of information (namely, conceptual properties), there is an enormous amount of variation across the features and relations of properties which exhibit such characteristics.

Table 1: Top ten properties from McRae norms with production frequencies for ***knife*** and ***pig***.

knife		***pig***	
is sharp	29	an animal	21
used for cutting	25	lives on farms	20
is dangerous	14	is pink	20
has a handle	14	has a tail	17
has a blade	11	has a curly tail	15
a weapon	11	has a snout	12
a utensil	9	eaten as bacon	11
made of steel	8	oinks	9
is serrated	8	is fat	8
found in kitchens	8	is dirty	8

Previous approaches to our specific conceptual property extraction task (Baroni, Murphy, Barbu, & Poesio, 2009; Devereux, Pilkington, Poibeau, & Korhonen, 2009; Kelly, Devereux, & Korhonen, 2010, 2012) have been successful to varying degrees, however each has suffered from limitations. Baroni et al., for example, did not explicitly offer relations between their extracted concepts and features. The relations extracted by the Devereux et al., system were rather unsophisticated, with the relation corresponding to the verb found along the grammatical relation path linking concept to feature. The Kelly et al. (2010) system had reasonable performance but was founded on manually constructed rules and relied heavily on WordNet for its feature selection.

The system of Kelly et al. (2012) approached this task as one of relation classification. The relations generated were derived directly from its training set; it was therefore unable to posit new or unseen relationships between its extracted concepts and features. We believe their feature output, however, was promising and we extend and enhance their feature extraction method in the first component of our own system.

Our system works by first employing a wealth of lexical, syntactic and semantic machine-learning attributes to train a support vector machine for feature-extraction. Unlike other

approaches, we make heavy use of unlabelled training data, rendering our system only very lightly supervised. Next, we return to our unlabelled corpus to find relations for the extracted features, using a novel, probabilistically motivated backing-off technique. In doing so, we are not constrained by relations found in the McRae norms: our method allows for the extraction of *any* relation.

Data

Recoded norms

We used the same set of recoded norms employed by Kelly et al. (2012) to train our system. This set, containing 510 concepts in total, is a coding of an anglicised version[1] of the McRae norms into a uniform ***concept*** *relation* **feature** format, where each **feature** and ***concept*** contain one word; the *relation* slot can contain one or more words.

Corpora

We used Wikipedia and the more general UKWAC corpus (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008), containing English-language webpages, as corpora. Together these offered a suitable balance of general and encyclopaedic text. We used the C&C-parser (Clark & Curran, 2007) to extract grammatical relations (GRs) and part of speech (POS) information from sentences, allowing us to construct a GR-POS graph for each. We trained our system on the corpora individually and in combination.

Chunking

We also used chunked versions of our two corpora. Chunking is a technique which identifies the constituent blocks of a sentence (verb phrase, noun phrase, prepositional phrase, etc.). To chunk our corpora, we used the Apache OpenNLP 1.5 suite (Baldridge, 2005), using the Tokenizer, POS Tagger and Chunker tools. The various components of the suite were trained using models supplied with the OpenNLP package.

Method

We trained our system with 466 of the 510 concepts in the anglicised McRae set to fix our training parameters and evaluated with the remaining 44 concepts, those in the ESSLLI expansion set (Baroni, Evert, & Lenci, 2008) (discussed later).

Feature derivation

In the first stage we focussed on extracting terms relevant to our concepts in order to generate a promising set of features, similar to those found in our norms.

Machine learning attributes Support vector machines (SVMs) are non-probabilistic binary linear classifiers which take a set of input data and predict, for each given input, which of two possible classes it corresponds to. This works by plotting training data points in a high-dimensional space and separating them with a hyperplane which has the largest distance (or margin) to the nearest training data points of each class. This plane is subsequently used to classify unseen data points. SVMs can also be extended to the multi-class case.

We trained an SVM by constructing paths through each sentence's GR-POS graph from the concept to prospective features and used the GR path labels, POS tags, relation verb instances and path-length as machine learning attributes. We augmented this (mostly syntactic) set of machine learning attributes to incorporate additional semantic and lexical information: bigrams and concept/feature clusters.[2] The intuition behind this was that similar types of concepts/features (as exhibited by cluster membership) might also exhibit similar types of relationships (e.g., 'tool' concepts and *used for* relations); the aim was to enable the SVM to detect the regularities that exist in the relationships between different semantic classes of concepts and features.

Every possible attribute across the training set corresponded to a distinct dimension of the vector space. The majority of the co-ordinates of the training data points took binary values depending on whether the dimension's corresponding attribute appeared in the path (except the clustering and path-length attributes which took integer values). Each training data point was labelled with its relation (or 'class').

Learning instances We applied the SVM Light software[3] (Joachims, 1999) to our learning attributes to extract an SVM score (the sum of absolute values of the decision function values, which can be interpreted as a measure of confidence of the SVM in its classification) for each concept-feature pair. We also calculated log-likelihood (LL) (Dunning, 1993) and pointwise mutual information (PMI) (Church & Hanks, 1990) statistics across the top 200 returned concept-feature pairs for each concept.

Previous work has ignored a large amount of potentially instructive training data by only examining sentences which link entities explicitly found in the training set. However, the use of 'negative' information could prove informative and therefore we trained on all GR-POS paths linking one of our concepts to *any* potential feature term[4] in each sentence. The size of our training set was 5.52 million instances for the Wikipedia corpus and 20.07 million instances for the UKWAC corpus.[5] As we were unaware of the nature of the relationship between these concept/feature terms, we labelled these unknown training paths as *unknownrel*.

Our system was therefore only very lightly supervised: only 6.8% of the UKWAC input and 8.7% of the Wikipedia input to the system was labelled with relations drawn from the McRae norms. Consequently, our SVM classified every

[1] See Taylor, Devereux, Acres, Randall, and Tyler (2011) for details.

[2] We generated 50 and 150 clusters for the concepts and features respectively using hierarchical clustering on WordNet.

[3] The multi-class implementation, SVM Multiclass (v. 2.20).

[4] Potential features were defined as all adjectives and singular/plural nouns in a sentence.

[5] Due to memory constraints associated with the very large number of training instances, we were only able to train our UKWAC models on one third of the UKWAC corpus; we selected every third learning pattern for training.

concept/feature pair into the *unknownrel* relation class. We therefore ignored the relation output from this stage of the system, instead using the top 200 returned concept/feature pairs ranked by their SVM scores as input to the next stage. In this way, we were interpreting a higher-rated SVM score as a proxy for the likelihood that a feature would have *some* kind of relationship with the concept at hand.

Relation extraction

The underlying hypothesis of our relation extraction stage was that if we found sequences of chunks in our corpus sentences which were anchored at each end by a known ***concept*** and **feature** (from the previous stage), and those chunks' labels matched the labels of our chunked property norms, then we could use the surface text of the chunk(s) between the anchors as the *relation* in our ***concept*** *relation* **feature** format.

Chunk pattern selection To decide which patterns of chunks were likely to be indicative of property norm relations, we turned to our training set. We passed the full text of the non-ESSLLI McRae norms through the chunker, and manually examined the output for chunk label patterns likely to indicate relations.

Using this output, we created a ruleset for selecting sentence fragments (chunk sequences) which were similar in structure to our property norms. We called a sequence of three labelled chunks a three-chunk, a sequence of four chunks a four-chunk, etc. We employed the first four most frequent label combinations (`NP VP NP`; `NP VP PP NP`; `NP VP ADJP`; and, `NP VP ADVP`) to form our ruleset; together these covered 95.6% of the three- and four-chunk label patterns generated from our training set. By using the `NP VP PP NP`-labelled four-chunks we were able to extract multi-word, prepositional verbs (e.g., *worn on*, *used for*) as potential relations: previous approaches to our task have not attempted this.

Chunk pre-selection We needed to select those chunks most relevant to our relation extraction task. To do this we passed through our chunked corpus, generating sets of 3 and 4 sequential chunks and pre-selecting those which were relevant to our concepts. Our criterion for relevancy at this stage was that the final term contained within the first chunk, when lemmatised, corresponded to a training concept.

Chunk to triple conversion Having pre-selected our chunks we generated triples from the chunk text. For three-chunks we did this by simply taking the final term in the first, second and third chunks and lemmatising each to give our ***concept***, *relation* and **feature** terms respectively. For four-chunks we followed the same process for the first and fourth chunks to yield our ***concept*** and **feature**. To extract the *relation* we took the final term of the second (`VP`) chunk and compounded it with the final term of the third (`PP`) chunk; the only exception to this was if the POS of the final term of the second chunk was `VBG`, in which case we lemmatised that term and compounded it with the third chunk's final term. For example:

- `[NP Mirrors_NNS] [VP are_VBP found_VBN] [PP in_IN] [NP the_DT bedroom_NN]` became ***mirror*** *found in* **bedroom**
- `[NP Most_JJS cats_NNS] [VP have_VBP] [NP furry_NN tails_NNS]` became ***cat*** *have* **tail**
- `[NP The_DT microwave_NN] [VP was_VBD running_VBG] [PP on_IN] [NP electricity_NN]` became ***microwave*** *run on* **electricity**

Relation selection

The third stage of our system worked by taking each ***concept***–**feature** pair from both the SVM and chunking output, and finding the best relation for that pair from the chunking output to generate a triple. It also assigned to that triple a number of metrics relating to its constituent parts, their relative frequency and association scores.

We assumed that each ***concept***–**feature** pair had one corresponding relation. We called the set of extracted triples generated by Stage 2, T (with triples $(c,r,f) \in T$) and the set of all extracted relations from Stage 2, R. For each concept, we also generated a final potential feature set, F_c, which, for a given concept, was the union of the top 200 features from Stage 1 (ranked by their SVM score) and the top 200 features from Stage 2 (ranked by their frequency in the extracted relations, but excluding features which appeared only once).

We defined Concept Feature Frequency (CFF) to be the number of times a concept, c, and feature, f, co-occurred across our extracted relations:

$$\mathrm{CFF}(c,f) = \sum_{r \in R} \mathrm{freq}(c,r,f) \tag{1}$$

We also calculated a Distinct Relation Score which measured the number of distinct relations linking c to f:

$$\mathrm{DRS}(c,f) = |D_{c,f}| \text{ for } D_{c,f} = \{r : (c,r,f) \in T\} \tag{2}$$

We next wanted to choose relations for our various ***concept***–**feature** pairs, $(c,f) \in C \times F_c$. We did this in three steps.

Step 1 For each concept, c, and feature, f, we iterated through all relations relating to that pair and calculated an Exact Match Score:

$$\mathrm{EMS}(c,f) = \max\{\mathrm{freq}(c,r,f) : r \in R\} \tag{3}$$

If $\mathrm{EMS}(c,f) > 0$ then we selected as our best relation, $\hat{r}$, the relation corresponding to that score. If there was more than one relation with the same score, then we chose the least common (i.e., that which had the lowest frequency across all our relations). If $\mathrm{EMS}(c,f) = 0$ then we left $\hat{r}$ undefined.

Step 2 Our first step only retrieved a relation if there was an exact match amongst our relation extraction output.

If there wasn't, we took a split approach; given a particular concept, c, and feature, f, we calculated separate probabilities across all our relations of c occurring with each relation, and of f occurring with each relation. We then calculated for each relation, r, a combined score for the combination of c, r and f by multiplying the constituent probabilities together. Our pairwise combination score was defined:

$$p(c,r) = \sum_{f \in F} \frac{\mathrm{freq}(c,r,f)}{\mathrm{freq}(c) \cdot \mathrm{freq}(r)} \tag{4a}$$

$$p(r,f) = \sum_{c \in C} \frac{\text{freq}(c,r,f)}{\text{freq}(r) \cdot \text{freq}(f)} \quad (4b)$$

$$\text{PCS}(c,f) = \begin{cases} p(c,\hat{r}) \cdot p(\hat{r},f) \text{ if } \hat{r} \text{ defined} \\ \max\{p(c,r) \cdot p(r,f) : r \in R\} \end{cases} \quad (4c)$$

If we had not already selected a best relation, $\hat{r}$, then we defined it as the relation, r, which corresponded to this pairwise combination score. Again, if there was more than one relation with the same score, then we chose the least common.

Step 3 Our final step assigned relations to concept/feature pairs which lacked an exact mutually linking relation. This occurred around 17% of the time and was usually due to both the concept and feature terms being relatively low frequency.[6]

To achieve this, we backed-off to semantic feature clusters: we defined $f_\star$ as the cluster for feature f, and $F_\star$ as the set of all feature clusters, and defined our Feature Cluster Score, $\text{FCS}(c, f_\star)$, analogously to our Pairwise Combination Score, merely substituting all instances of f for $f_\star$. Our best relation, $\hat{r}$, was defined as the relation corresponding to this FCS.

Reweighting

In our system's fourth and final stage we used the metrics derived above to assign an overall score for each triple using a weighting of parameters; we used our training set to derive the most optimal values for these parameters. We normalised our various metrics so that they all lay between 0 and 1.

Our relation selection stage had already fixed a relation, $\hat{r}$, for each concept and feature. Hence we calculated for each of our triples $t = (c, \hat{r}, f)$ an overall score:

$$\begin{aligned} \text{score}(t) = {} & \beta_{\text{PMI}} \cdot \text{PMI}(t) + \beta_{\text{LL}} \cdot \text{LL}(t) + \beta_{\text{SVM}} \cdot \text{SVM}(t) \\ & + \beta_{\text{CFF}} \cdot \text{CFF}(t) + \beta_{\text{DRS}} \cdot \text{DRS}(t) + \beta_{\text{EMS}} \cdot \text{EMS}(t) \\ & + \beta_{\text{PCS}} \cdot \text{PCS}(t) + \beta_{\text{FCS}} \cdot \text{FCS}(t) \end{aligned} \quad (5)$$

We wished to optimise our parameters for superior feature F-score performance against our training set. We employed a stochastic process to find best-possible values for our training parameters, using a random-restart hill-climbing algorithm, repeated 1000 times and selecting the output (and β values) offering the best F-score across these iterations.

This process offered a reasonable approximation of the best possible F-scores our system could produce and their corresponding β values; following this process, our best F-scores were 0.2739, 0.2803 and 0.2996 for our Wikipedia, UKWAC and combined corpora respectively.

Evaluation

We evaluated our system using gold standard, human semantic-similarity and direct human evaluations.

Gold standard evaluation

We began by comparing our top twenty output using the ESSLLI gold standard set. This 'expansion' set comprises the top ten lemmatised properties for each of 44 concepts from the recoded McRae norms, together with a feature expansion set generated for each ***concept*** *relation* **feature** triple. One of the reasons for using this set is that McRae et al. normalised their features by channelling synonymous properties into a single representation. The ESSLLI set undoes some of these normalizations, expanding the feature terms to a set of synonyms. In this way, **loud**, **noise** and **noisy** (for example) can all be counted as matches against the property *is* **loud**. The relations were not expanded.

Table 2: Our best precision, recall and F-scores against the synonym-expanded ESSLLI norms across our corpora, found using the training β parameters.

Relation	Corpus	Prec.	Recall	F
With	Wikipedia	0.1131	0.2265	0.1509
	UKWAC	0.1000	0.2005	0.1335
	Combined	0.1214	0.2431	0.1620
	Kelly et al.	0.1238	0.2493	0.1654
With (aug.)	Wikipedia	0.1214	0.2431	0.1620
	UKWAC	0.1048	0.2101	0.1398
	Combined	0.1298	0.2598	0.1731
Without	Wikipedia	0.2798	0.5603	0.3732
	UKWAC	0.2560	0.5132	0.3416
	Combined	0.2798	0.5606	0.3733
	Kelly et al.	0.2417	0.4847	0.3225

Our results can be found in Table 2. We also assessed our system using the full text of the relations found in the original McRae norms as additional 'relation synonyms'; these augmented results can be found under the 'With (aug.)' relation heading. We have exceeded the performance of Kelly et al. (2012) (best F-score of 0.1654) with a best overall F-score of 0.1731 for the combined corpus.

We also note that performing these evaluations on the top ten properties returned further improved the situation (perhaps unsurprising since the ESSLLI set contains only ten properties per concept); for example, evaluating our top ten triples against the relation synonyms set returned a precision of 0.2215 for the combined corpus. Furthermore, the precision on the combined corpus for the top ten evaluation of features-only was 0.4409, surpassing Baroni et al. (2009) who offer a best score of 0.239 on the same evaluation.

Human-generated semantic similarity

Comparison with the ESSLLI gold standard is still an incomplete evaluation: not all conceptual properties for a given concept are contained therein, and lexical variation can mark valid relations as wrong. Furthermore, one of the primary advantages of our computational approach is its ability to extract a large number of properties for a given concept. Hence, we introduced an alternative approach to calculate how semantically meaningful our output was by evaluating the triples' capacity to predict human-rated similarity between words.

We asked five native English speakers to rate the similarity of 90 concept pairs, where concepts in the pairs were all drawn from the ESSLLI set. The raters were given instructions explaining the task and then presented with each concept pair, one by one, a scale of 1 to 7 and asked to rate how

[6] Only a small proportion of our triples derived their relations in this way; at this point, in our training sets we had assigned relations to over 94% of triples from our Wikipedia corpus, and 97% from the UKWAC corpus.

Table 3: Pearson correlation (r) results with confidence intervals between our V_{Human} vector and our similarity vectors V (with dimensionality D and derived from the top n properties) from our system.

Relation	V	n	D	r	Conf. Int.
With	Wikipedia	20	654	0.598	[0.446, 0.716]
	UKWAC		712	0.629	[0.486, 0.740]
	Combined		692	0.671	[0.539, 0.771]
	Wikipedia	300	3585	0.693	[0.568, 0.787]
	UKWAC		3442	0.683	[0.555, 0.780]
	Combined		3380	0.723	[0.606, 0.809]
Without	Wikipedia	20	478	0.720	[0.603, 0.807]
	UKWAC		456	0.754	[0.649, 0.832]
	Combined		475	0.742	[0.632, 0.822]
	Wikipedia	300	7324	0.782	[0.685, 0.851]
	UKWAC		8698	0.806	[0.719, 0.868]
	Combined		8727	0.807	[0.721, 0.869]
With	McRae		410	0.785	[0.691, 0.854]
Without			355	0.787	[0.693, 0.855]
	LSA		300	0.708	[0.586, 0.798]

similar the two concepts were.

To compare our system with these ratings we constructed a vector space of dimension D, where D was the number of distinct properties across our triples. For each of our 44 concepts, we generated a concept-score vector with non-zero entries by inserting the triple scores, score(t), into their correct entries in the concept-score vector. We then constructed a 44×44 symmetric pairwise similarity matrix across our concepts by calculating the cosine similarity between their concept-score vectors. From this we extracted a similarity vector, V, for our 90 pairwise comparisons.

We calculated twelve such matrices (using the top twenty and top 300 extracted triples, across three corpora and excluding and including the relation term). We also generated two such matrices using both the feature-heads and the full text of the McRae property norms, using the norm production frequencies as entries in each concept's vector, as well as comparing our ratings with LSA-predicted (Landauer, Foltz, & Laham, 1998) similarities.[7] Our results are in Table 3.[8]

Our systems' performance, evaluating with and without relation and when using the top twenty triples, was comparable to LSA (correlation 0.708) with average correlations across our corpora of 0.754 and 0.671 respectively. Including the top 300 extracted triples brought our correlations up to 0.807 and 0.754 respectively, an extremely strong result given that the average Pearson coefficient of correlation across the five judges (considering all pairwise combinations) was 0.820.

Human evaluation

In our final evaluation, we asked two native English speaking human judges to assess the accuracy of our triples. Following the methodology of Devereux et al. (2009), we asked them to classify output triples for 15 concepts into four categories: 'correct' (c), 'plausible' (p), 'related' (r) and 'wrong' (w). Our judges were unaware of the aims of the evaluation. We concatenated their ratings using the methodology of Devereux et al.[9] however our instructions reflected the fact that, unlike previous systems, our output contained prepositional relations and we therefore did not wish our volunteers to allow for absent prepositions. This evaluation offers an important insight into the viability of our method as a property extraction system. Our results are in Table 4, and Table 5 shows a sample of our output and the corresponding judgements.

Table 4: Inter-annotator agreement and judgements for our extraction system applied to our three corpora.

Corpus		Judge A	Judge B	Avg	% c / p	Kappa (Agree)
Wikipedia	c / p	202	204	203	67.7	0.6343
	r / w	98	96	97		(252)
UKWAC	c / p	193	204	198.5	66.2	0.7398
	r / w	107	96	101.5		(265)
Combined	c / p	212	216	214	71.3	0.7229
	r / w	88	84	86		(266)

It is clear that our best results were again in the combined corpus, where an impressive 71.3% of our returned triples were marked as either plausible or correct with a Kappa (Fleiss, 1971) score of 0.7229 indicating substantial agreement between annotators. This constitutes a major improvement over Kelly et al. (2012) who evaluated on the same set of concepts and whose corresponding score was just 51.1%.

Discussion

As the first system to offer viable unconstrained property norm-like extraction, this paper brings research into conceptual property extraction to the next level. Our system employs both full parsing and chunking to extract features and relations respectively and introduces a novel multi-step backing-off method for relation selection. Our gold standard performance exceeded that of previous approaches, and our human evaluation indicated that we have outperformed the system of Kelly et al. (2012) by a significant margin. We also introduced a semantic similarity evaluation for this task, showing a strong Pearson correlation of 0.754 with human ratings when employing just 20 extracted properties per concept, with the correlation rising to 0.807 when using 300 properties. In this latter case, the predicted similarities were almost as correlated with human judgements as the human judgements are with each other.

Potential criticisms of our system include the fact that our chunk to triple conversion process won't necessarily always yield a true reflection of the sentence's original meaning. It is, for example, possible for the final chunk to contain adjectives which modify the final noun. These could have importance from a conceptual representation perspective (e.g., features such as **long neck** for ***giraffe*** *has* **long neck**). Also, the modifying portion of a chunk may be semantically significant, altering the final term's meaning (e.g., a ***tea bag*** is quite different from a ***bag***). It should be possible to have more gen-

[7] 300 factors, using the TASA corpus at `lsa.colorado.edu`.

[8] The correlation confidence intervals, calculated using Fisher transformations (Fisher, 1915), are given at the 95% level of confidence, and two-tailed $p < 0.05$.

[9] i.e. both 'correct' and 'plausible' triples were counted as correct, while 'related' or 'wrong' triples were considered incorrect.

Table 5: Judges' assessments of the top twenty extracted relation/feature pairs (combined corpus) for two concepts.

knife	Judge A	Judge B	*pig*	Judge A	Judge B
sharpened by **hand**	c	c	*eat* **piglet**	c	p
based on **design**	c	c	*get* **fat**	c	c
made of **steel**	c	c	*produce* **pork**	r	c
be **small**	c	p	*breed* **farm**	r	r
pick on **fork**	r	r	*put into* **sausage**	c	c
be **make**	p	r	*be* **large**	p	p
crafted from **metal**	c	c	*have* **baby**	c	c
scaled for **use**	p	p	*be* **different**	p	p
make **cut**	c	c	*stunned through* **use**	r	w
be **sharp**	c	c	*be* **bacon**	c	r
be **weapon**	c	c	*be* **welfare**	r	r
have **edge**	c	c	*discover* **sheep**	c	c
have **handle**	c	c	*killed for* **meat**	c	c
be **serrated**	c	c	*used for* **food**	c	c
made of **stainless**	w	r	*label* **cattle**	w	w
is for **cutting**	c	c	*be* **animal**	c	c
have **blade**	c	c	*shackled by* **ham**	r	r
be **useful**	p	c	*chew* **tail**	c	c
be **tool**	c	c	*have* **disease**	c	c
be **dangerous**	c	c	*found in* **guinea**	c	c

eral chunk to triple extraction (e.g., by using a larger corpus to mitigate the sparsity associated with multi-word terms).

Finally, a major issue is our lack of comprehensive training/testing data; our norms are incomplete insofar as there were a large number of 'correct' properties absent from our gold standard. In future work we hope to implement large-scale evaluation of our system's output (e.g., using Amazon Turk) which would allow us to rapidly obtain large amounts of human-generated feedback. We could then use active-learning to introduce a feedback loop of human-annotation to better pinpoint inaccurate features or relations. Feedback which strongly indicated that certain properties were uninteresting could prove invaluable in getting even closer to a conceptual structure-like representation of concepts.

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 comments and feedback.

References

Baldridge, J. (2005). *The Apache OpenNLP project.*

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.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clark, S., & Curran, J. (2007, 1). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, *33*(4), 493–552.

Devereux, B., Pilkington, N., Poibeau, T., & Korhonen, A. (2009). Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language & Computation*, 1–34.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Farah, M., & McClelland, J. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339–357.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating UKWAC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (wac-4) – Can we beat Google?* (pp. 47–54).

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Joachims, T. (1999). Making large scale SVM learning practical.

Kelly, C., Devereux, B., & Korhonen, A. (2010). Acquiring human-like feature-based conceptual representations from corpora. In *First Workshop on Computational Neurolinguistics* (p. 61). Association for Computational Linguistics.

Kelly, C., Devereux, B., & Korhonen, A. (2012). Semi-supervised learning for automatic conceptual property extraction. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 11–20). Association for Computational Linguistics.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

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

Murphy, G. (2002). *The Big Book of Concepts*. The MIT Press.

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

Taylor, K., Devereux, B., Acres, K., Randall, B., & Tyler, L. (2011). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, *122*(3), 363-74.

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