Semi-supervised learning for automatic conceptual property extraction

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

For a given concrete noun concept, humans are usually able to cite properties (e.g., *elephant is animal*, *car has wheels*) of that concept; cognitive psychologists have theorised that such properties are fundamental to understanding the abstract mental representation of concepts in the brain. Consequently, the ability to automatically extract such properties would be of enormous benefit to the field of experimental psychology. This paper investigates the use of semi-supervised learning and support vector machines to automatically extract concept-relation-feature triples from two large corpora (Wikipedia and UKWAC) for concrete noun concepts. Previous approaches have relied on manually-generated rules and hand-crafted resources such as WordNet; our method requires neither yet achieves better performance than these prior approaches, measured both by comparison with a property norm-derived gold standard as well as direct human evaluation. Our technique performs particularly well on extracting features relevant to a given concept, and suggests a number of promising areas for future focus.

1 Introduction

The representation of concrete concepts (e.g., **car**, **banana**, **spanner**) in the human brain has long been an important area of investigation for cognitive psychologists. Recent theories of this mental representation have proposed a componential, property-based and distributed model of conceptual knowledge (e.g., Farah and McClelland (1991), Randall et al. (2004), Tyler et al. (2000)).

In order to empirically test these cognitive theories, researchers have moved towards employing real-world knowledge in their experiments. This knowledge has usually been procured from human-derived lists of properties taken from property norming studies (Garrard et al., 2001; McRae et al., 2005). In such studies, human participants are asked to describe and note properties of a given concept (e.g., ***has*** *shell* for **turtle**). Synonymous responses are grouped together as a single property and those meeting a certain minimum response-frequency threshold are taken as valid properties. The most wide-ranging study to date was that conducted by McRae et al. (2005): some sample properties from this set are in Table 1.

As others have noted (Murphy, 2002; McRae et al., 2005), property norming studies are prone to a number of deficiencies. One such weakness is the incongruity of shared properties across even highly-related concepts: human respondents exhibit a lack of consistency when listing properties that are common to many similar concepts. For example, while ***has*** *legs* is listed as a property of **crocodile** in the McRae norms, it is absent as a property of **alligator**. A related issue is the non-comprehensive nature of the generated norms – although they may cover the most salient properties for a given concept, they are unlikely to comprise all of a concept's properties (e.g., ***has*** *heart* does not appear as a property of any of the 92 animal concepts).

Our research aims to use NLP techniques to create a system able to emulate the output of such studies, and overcome some of the aforementioned weaknesses. Our proposed system begins by searching dependency-parsed corpora for those sentences containing concept and feature terms which are also found in a McRae norm-derived training set of properties. For these sentences, the system generates grammatical relation/part-of-speech structural attributes and applies support vector machines (SVMs) to learn sets of attributes likely to indicate the instantiation of a property in a sentence. These

In: R. Levy & D. Reitter (Eds.), *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 11–20,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

turtle		**bowl**	
has a shell	25	is round	19
lays eggs	16	used for eating	12
swims	15	used for soup	11
is green	14	used for food	11
lives in water	14	used for liquids	10
is slow	13	used for eating cereal	10
an animal	11	made of plastic	8
walks	10	used for holding things	7
walks slowly	10	is curved	7
has 4 legs	9	found in kitchens	7

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

learned patterns of salient attributes are finally applied to a corpus to derive new properties for unseen concepts.

Our task is a challenging one: the properties we seek are extremely diverse in their form. They range from the simple (e.g., *banana is yellow*) to the complex (e.g., *bayonet found at the end of a gun*). Although the properties can broadly be divided into a number of categories (encyclopedic, taxonomic, functional, etc) there is not a great deal of regularity in the nature of the properties a given noun will likely possess: it is highly concept-dependent.

Furthermore, we hope to derive these properties from corpora, with the assumption that these properties will manifest themselves therein. Indeed, Andrews et al. (2005) discuss a theory of human knowledge which relies on a combination of both distributional (i.e., derived from spoken and written language) and experiential data (i.e., that derived from our interactions with the real world), claiming that the necessary contribution of each data-type for a comprehensive human semantic representation is non-trivial. Finally, there are difficulties associated with evaluating our system's output directly against a set of human-generated property norms: we discuss these in further detail later.

Given their provenance, the properties found in property norms are free-form. To simplify our task we apply a more rigid representation to the properties we already have and to those we aim to seek. We delineate each property into a **concept** ***relation*** *feature* triple (see Section 2.2) and our task becomes one of finding valid ***relation*** *feature* pairs given a particular **concept**. This recoding renders our task more well-defined and makes evaluation of our method more comparable to previous and related work.

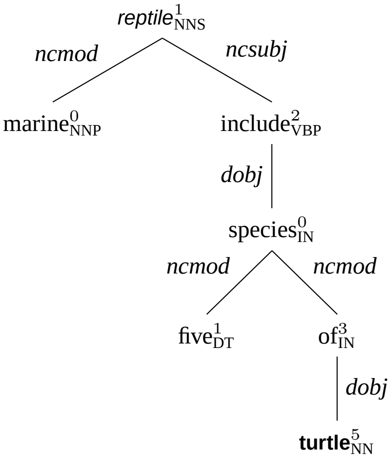

Figure 1: C&C-derived GR-POS graph for the sentence *Marine reptiles include five species of turtle*.

Having framed our task in this way, there is an obvious parallel with relation extraction: both necessitate the selection/classification of relationships between individual entities (in our case, between **concept** and *feature*). Hearst (1992) was the first to propose a pattern-based approach to this task using lexico-syntactic patterns to automatically extract hyponyms and this technique has frequently been used for ontology learning. For example, Pantel and Pennacchiotti (2008) linked instantiations of a set of semantic relations into existing semantic ontologies and Davidov et al. (2007) employed seed concepts from a given semantic class to discover relations shared by concepts in that class.

Our task is more complex than classic relation extraction for two main reasons: 1) the relations which we aim to extract are not limited to a small set of just a few well-defined relations (e.g., ***is-a*** and ***part-of***) nor to the relations of a specific semantic class (e.g., ***capital-is*** for countries). Indeed the relations can be as many and diverse as the concepts themselves (e.g., each concept could possess a unique and distinguishing relation and feature). 2) We are attempting to simultaneously extract two pieces of information: features of the concept and those features' defining relationship with the concept, but only those relations and features which would be classified as 'common-sense', something which is easy for humans to recognise but difficult (if not impossible) to describe rigorously or formally.

There has recently been work on the automatic extraction of binary relations that scale to a web corpus, for example the ReVerb (Etzioni et al., 2011) and WOE (Wu and Weld, 2010) systems. These systems are designed to extract legitimate relations from a given sentence. In contrast, our aim is to capture more general relationships which are 'common-sense'; just because an extracted relation is correct in a given context does not automatically make it true in general. Previous reasoned approaches to our task have taken their lead from Hearst and her successors, employing manually-created rulesets to extract such properties from corpora (e.g., Baroni et al. (2009), Devereux et al. (2010), and our comparison system (Kelly et al., 2010)). Baroni et al. extract relational information in the form of 'type-sketches', which give an approximate, implicit description of the relationship whereas we are aiming to extract explicit relations between the target concept and its corresponding features. Devereux et al. and Kelly et al. have attempted this, but both employ WordNet (Fellbaum, 1998) to extract semantic relatedness information.

We use semi-supervised learning as it offers a flexible technique of harnessing small amounts of labelled data to derive information from unlabelled datasets/corpora and allows us to guide the extraction towards our desired 'common-sense' output. We chose SVMs as they have been used for a variety of tasks in NLP (e.g., Joachims et al. (1998), Giménez and Marquez (2004)). We will demonstrate that our system's performance exceeds that of Kelly et al. (2010) and Etzioni et al. (2011). It is, as far as we are aware, the first work to employ semi-supervised learning for this task.

2 Method

We will use SVMs to learn lexico-syntactic patterns in our corpora corresponding to known properties in order to find new ones. Training an SVM requires a labelled training set. To generate this set we harness our already-known concepts/features (and their relationships) from the McRae norms to find instantiations of said relationships within our corpora. We use parsed sentence information from our corpora to create a set of attributes describing each relationship, our learning patterns. In doing so, we are assuming that across sentences in our corpora containing a concept/feature pair found in the McRae norms, there will be a set of consistent lexico-syntactic patterns which indicate the same relationship as that linking the pair in the norms.

Thus we iterate over our chosen corpora, parsing each concept-containing sentence to yield grammatical relation (GR) and part-of-speech (POS) information from which we can create a GR-POS graph relating the two. Then for each triple, we find any/all paths through the graph which link the **concept** to its *feature* and use the corresponding ***relation*** to label this path. We collect descriptive information about the path in the form of attributes describing it (e.g., path nodes, labels, length) to create a training pattern specific to that **concept** ***relation*** *feature* triple and sentence. It is these lists of attributes (and their ***relation*** labels) which we employ as the labelled training set and as input for our SVM.

2.1 Corpora

We employ two corpora for our experiments: Wikipedia and the UKWAC corpus (Ferraresi et al., 2008). These are both publicly available and web-based: the former a source of encyclopedic information and the latter a source of general text. Our Wikipedia corpus is based on a Sep 2009 version of English-language Wikipedia and contains around 1.84 million articles (>1bn words). Our UKWAC corpus is an English-language corpus (>2bn words) obtained by crawling the `.uk` internet domain.

2.2 Training data

Our experiments use a British-English version of the McRae norms (see Taylor et al. (2011) for details). We needed to recode the free-form McRae properties into relation-classes and features which would be usable for our learning algorithm. As we will be matching the features from these properties with individual words in the training corpus it was essential that the features we generated contained only one lemmatised word. In contrast, the relations were merely labels for the relationship described (they did not need to occur in the sentences we were training from) and therefore needed only to be single-string relations. This allowed prepositional verbs as distinct relations, something which has not been attempted in previous work yet can be semantically significant (e.g., the relations ***used-in***, ***used-for*** and ***used-by*** have dissimilar meanings).

We applied the following sequential multi-step process to our set of free-form properties to distill them to triples of the form **concept *relation*** *feature*, where ***relation*** can be a multi-word string and *feature* is a single word:

1. Translation of implicit properties to their correct relations (e.g., *pig an animal* → *pig is an animal*).
2. Removal of indefinite and definite articles.
3. Behavioural properties become "does" properties (e.g., *turtle beh eats* → *turtle does eats*).
4. Negative properties given their own relation classes (e.g., *turkey does cannot fly* → *turkey doesnt fly*).
5. All numbers are translated to named cardinals (e.g., *spider has 8 legs* → *spider has eight legs*).
6. Some of the norms already contained synonymous terms: these were split into separate triples for each synonym (e.g., *pepper tastes hot/spicy* → *pepper tastes hot* and *pepper tastes spicy*).
7. Prepositional verbs were translated to one-word, hyphenated strings (e.g., *made of* → *made-of*).
8. Properties with present participles as the penultimate word were split into one including the verb as the feature and one including it in the relation (e.g., *envelope used for sending letters* → *envelope used-for-sending letters* and *envelope used-for sending*).
9. Any remaining multi-word properties were split with the first term after the concept acting as the relation (e.g., *bull has ring in its nose* → *bull has ring*, *bull has in*, *bull has its* and *bull has nose*).
10. All remaining stop-words were removed; properties ending in stop-words (e.g., *bull has in* and *bull has its*) were removed completely.

This yielded 7,518 property-triples with 254 distinct relations and an average of 14.7 triples per concept.

2.3 Parsing

We parsed both corpora using the C&C parser (Clark and Curran, 2007) as we employ both GR and POS information in our learning method. To accelerate this stage, we process only sentences containing a form (e.g., singular/plural) of one of our training/testing concepts. We lemmatise each word using the WordNet NLTK lemmatiser (Bird, 2006). Parsing our corpora yields around 10Gb and 12Gb of data for UKWAC and Wikipedia respectively.

The C&C dependency parse output contains, for a given sentence, a set of GRs forming an acyclic graph whose nodes correspond to words from the sentence, with each node also labelled with the POS of that word. Thus the GR-POS graph interrelates all lexical, POS and GR information for the entire sentence. It is therefore possible to construct a GR-POS graph rooted at our target term (the concept in question), with POS-labelled words as nodes, and edges labelled with GRs linking the nodes to one another. An example graph can be seen in Figure 1.

2.4 Support vector machines

We use SVMs (Cortes and Vapnik, 1995) for our experiments as they have been widely used in NLP and their properties are well-understood, showing good performance on classification tasks (Meyer et al., 2003). In their canonical form, 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.

There are more than two possible relation-labels to learn for our input patterns, so ours is a multi-class classification task. For our experiments we use the SVM Light Multiclass (v. 2.20) software (Joachims, 1999) which applies the fixed-point SVM algorithm described by Crammer and Singer (2002) to solve multi-class problem instances. Joachims' software has been widely used to implement SVMs (Vinokourov et al., 2003; Godbole et al., 2002).

2.5 Attribute selection

Previous techniques for our task have made use of lexical, syntactic and semantic information. We are deliberately avoiding the use of manually-created semantic resources, so we rely only on lexical and syntactic attributes for our learning stage (i.e., the GR-POS paths described earlier).

A table of all the categories of attributes we extract for each GR-POS path are in Table 2.4, together with attributes from the path linking **turtle** and *reptile* in our example sentence (see Figure 1).

We ran our experiments with two vector-types which we call our 'verb-augmented' and our 'non-augmented' vector-types. The sets are identical except the verb-augmented vector-type will also contain an additional attribute category containing an attribute for every instance of a relation verb (i.e., a verb which is found in our training set of relations, e.g., *become*, *cause*, *taste*, *use*, *have* and so on) in the lexical path. We do this to ascertain whether this additional verb-information might be more informative to our system when learning relations (which tend to be composed of verbs).

Attribute category	Example attribute(s)
GR path-length	LEN
lemmatised anchor node	LEM=turtle
POS of anchor node	POS=NN

GR path labels
from anchor
(indexed)
	GR1=dobjR
	GR2=ncmodR
	GR3=dobjR
	GR4=ncsubjN

GR path labels
from target
(indexed)
	GR1=ncsubjR
	GR2=dobjN
	GR3=ncmodN
	GR4=dobjN

POS of path nodes
from anchor
(indexed)
	POS1=IN
	POS2=NNS
	POS3=VBP
	POS4=NNS

POS of path nodes
from target
(indexed)
	POS1=NNS
	POS2=VBP
	POS3=NNS
	POS4=IN

lemmatised path nodes
(bag of words)
	LEM=include
	LEM=species
	LEM=of

POS of all path nodes
(set)
	POS=IN
	POS=NNS
	POS=VBP

Relation verbs	N/A

GR path labels
(set)
	GR=dobjR
	GR=ncmodN
	GR=ncsubjN

lemmatised target node	LEM=reptile
POS of target node	POS=NNS

Table 2: An example vector for an instance of the relation-label ***is***. The attributes are distinguished from one another by their attribute category. Relation verbs only appear in the verb-augmented vector-type and no such verbs appear in our example sentence, so this category of attribute is empty. All attributes in the table will receive the value 1.0 except the LEN attribute which will have the value 0.2 (the reciprocal of the path length, 5).

We considered allocating a 'no-rel' relation label to those sets of attributes corresponding to paths through the GR-POS graph which did *not* link the concept to a feature found in our training data; however our initial experiments indicated the SVM model would assign every pattern we tested to the 'no-rel' relation. Therefore we used only positive instances in our training pattern data.

We cycle through all training concepts/features, finding sentences containing both. For each such sentence, our system generates the attributes from the GR-POS path linking the concept to the feature (the linking-path) to create a pattern for that pair, in the form of a relation-labelled vector containing real-valued attributes. The system assigns 1.0 to all attributes occurring in a given path and the LEN value receives the reciprocal of the path-length.[1] Each linking-path is collected into a ***relation***-labelled, sparse vector in this manner. In the larger UKWAC corpus this corresponds to over 29 million unique attributes across all found linking-paths (this figure corresponds to the dimensionality of our vectors). We then pass all vectors to the learning module[2] of SVM Light to generate a learned model across all training concepts.

2.6 Extracting candidate patterns

Having trained our model, we must now find potential features and relations for our test concepts in our corpora. We again only examine sentences which contain at least one of our test concepts. Furthermore, to avoid a combinatorial explosion of possible paths rooted at those concepts we only permit as candidates those paths whose anchor node is a singular or plural noun and whose target node is either a singular/plural noun or adjective. This filtering corresponds to choosing patterns containing one of the three most frequent anchor node POS tags (NN, NNS and NNP) and target node POS tags (NN, JJ and NNS) found during our training stage. These candidate patterns constitute 92.6% and 87.7% of all the vectors, respectively, from our training set of patterns (on the UKWAC corpus). This pattern pre-selection allows us to immediately ignore paths which, despite being rooted at a test concept, are unlikely to contain property norm-like information.

2.7 Generating and ranking triples

We next classified our test concepts' candidate patterns using the learned model. SVM Light assigns each pattern a relation-class from the training set and outputs the values of the decision functions from the learned model when applied to that particular pattern. The sign of these values indicates the binary decision function choice, and their magnitude acts as a measure of confidence. We wanted those vectors which the model was most confident in across all decision functions, so we took the sum of the absolute values of the decision values to generate a pattern score for each vector/relation-label.

[1] All other possible attributes are assigned the value 0.0.

[2] Using a regularisation parameter (C) value of 1.0 and default parameters otherwise.

	Vector-type	Corpus	β_{LL}	β_{PMI}	β_{SVM}	Prec.	Recall	F
Ignoring relation.	Non-augmented	Wikipedia	0.3	0.00	1.00	0.2214	0.3197	0.2564
		UKWAC	0.10	0.05	0.60	0.2279	0.3330	0.2664
		UKWAC-Wikipedia	0.35	0.00	0.75	0.2422	0.3533	0.2829
	Verb-augmented	Wikipedia	0.20	0.00	0.65	0.2217	0.3202	0.2568
		UKWAC	0.30	0.00	0.95	0.2326	0.3400	0.2720
		UKWAC-Wikipedia	**0.40**	**0.05**	**1.00**	**0.2444**	**0.3577**	**0.2859**
With relation.	Non-augmented	Wikipedia	0.05	0.00	1.00	0.1199	0.1732	0.1394
		UKWAC	0.05	0.00	1.00	0.1126	0.1633	0.1312
		UKWAC-Wikipedia	0.05	0.00	0.65	0.1241	0.1808	0.1449
	Verb-augmented	Wikipedia	0.05	0.00	1.00	0.1215	0.1747	0.1410
		UKWAC	0.05	0.00	1.00	0.1190	0.1724	0.1387
		UKWAC-Wikipedia	**0.05**	**0.00**	**0.70**	**0.1281**	**0.1860**	**0.1494**

Table 3: Parameter estimation both with and without relation, using our augmented and non-augmented vector-types and across our two corpora and the combined corpora set.

From these patterns we derived an output set of triples where the concept and feature of a triple corresponded to the anchor and target nodes of its pattern and the relation corresponded to the pattern's relation-label. Identical triples from differing patterns had their pattern scores summed to give a final 'SVM score' for that triple.

2.8 Calculating triple scores

A brief qualitative evaluation of our system's output indicates that although the higher-ranked (by SVM score) features and relations were, for the most part, quite sensible, there were some obvious output errors (e.g., non-dictionary strings or verbs appearing as features). Therefore we restricted our features to those which appear as nouns or adjectives in WordNet and excluded features containing an NLTK (Bird, 2006) corpus stop-word. Despite these exclusions, some general (and therefore less informative) relation/feature combinations (e.g., ***is** good*, ***is** new*) were still ranking highly. To mitigate this, we extract both log-likelihood (LL) and pointwise mutual information (PMI) scores for each concept/feature pair to assess the relative saliency of each extracted feature, with a view to downweighting common but less interesting features. To speed up this and later stages, we calculate both statistics for the top 1,000 triples extracted for each concept only.

PMI was proposed by Church and Hanks (1990) to estimate word association. We will use it to measure the strength of association between a concept and its feature. We hope that emphasising concept-feature pairs with high mutual information will render our triples more relevant/informative.

We also employ the LL measure across our set of concept-feature pairs. Proposed by Dunning (1993), LL is a measure of the distribution of linguistic phenomena in texts and has been used to contrast the relative corpus frequencies of words. Our aim is to highlight features which are particularly distinctive for a given concept, and hence likely to be features of that concept alone.

We calculate an overall score for a triple, t, by a weighted combination of the triple's SVM, PMI and LL scores using the following formula:

$$\text{score}(t) = \beta_{\text{PMI}} \cdot \text{PMI}(t) + \beta_{\text{LL}} \cdot \text{LL}(t) + \beta_{\text{SVM}} \cdot \text{SVM}(t)$$

where the PMI, SVM and LL scores are normalised so they are in the range [0, 1]. The relative β weights thus give an estimate of the three measures' importance relative to one another and allows us to gauge which combination of these scores is optimal.

2.9 Datasets

We also wanted to ascertain the extent to which the output from both our corpora could be combined to improve results, balancing the encyclopedic but somewhat specific nature of Wikipedia with the generality and breadth of the UKWAC corpus. We combined the output by summing individual SVM scores of each triple from both corpora to yield a combined SVM score. PMI and LL scores were then calculated as usual from this combined set of triples.

3 Experimental Evaluation

3.1 Evaluation methodology

We employ ten-fold cross-validation to ascertain optimal SVM, LL and PMI β parameters for our final system. We exclude 44 concepts from our set of

	Relation	Prec.	Recall	F
Kelly et al.	Without	0.1943	0.3896	0.2592
	With	0.1102	0.2210	0.1471
ReVerb	Without	0.1142	0.2258	0.1514
	With	0.0431	0.0864	0.0576
Our method	Without	0.2417	0.4847	0.3225
	With	0.1238	0.2493	0.1654

Table 4: Our best scores on the ESSLLI set compared to Kelly et al. (2010) and the ReVerb system (Etzioni et al., 2011). Our results are from the verb-augmented vector-type, using the combined UKWAC-Wikipedia corpus and using the β parameters highlighted in Table 3.

510 to use in our final system testing and split the remaining 466 concepts randomly and evenly into 10 folds. We apply the training steps above to nine of the folds, generating predictions for the single held-out fold. We repeat this for all ten folds, yielding relations and features with SVM, LL and PMI scores for our full set of 466 training concepts on the UKWAC, Wikipedia and combined corpora.

We varied the β values from our scoring equation in the range [0,1] (interval 0.05) and compared the top twenty triples for each concept directly against the held-out training set. The best F-scores and their corresponding β values (evaluating on full triples and concept-feature pairs alone) are in Table 3. We can see that our best results employ the verb-augmented vector-type and the combined corpus, with a best F-score of 0.2859 when ignoring the relation term and 0.1494 when including it in the evaluation. The main difference between these two results is the relative contribution of the reweighting factors: the SVM score is the most important overall, but the LL and PMI scores come into play when evaluating without the relation. This could be explained by the fact that the PMI and LL scores do not use any relation terms in their calculations.

3.2 Quantitative evaluation

The unseen subset of the McRae norms is a set of human-generated common-sense properties with which our extracted properties can be compared. However, an issue with the McRae norms is that semantically identical properties can be represented by lexically different triples. This problem was acknowledged by Baroni et al. (2008) who created a synonym-expanded set of properties for 44 concepts (selected evenly across six semantic classes; the 44 concepts we excluded for testing) to partially solve it. This expansion set comprises the concepts' top ten properties from the McRae norms with semi-automatically generated synonyms for each of the ten distinct features. For example, the triple **turtle** ***has*** *shell* was expanded to also include **turtle** ***has*** *shield* and **turtle** ***has*** *carapace*.

	Judge			Judge	
turtle	A	B	**bowl**	A	B
is *green*	c	c	***is*** *large*	p	p
is *small*	c	c	***used for*** *food*	c	c
is *species*	c	c	***used for*** *mixing*	c	c
is *marine*	c	c	***used for storing*** *food*	c	c
used for *sea*	r	r	***used for storing*** *soup*	r	r
is *animal*	c	c	***is*** *ceramic*	c	c
is *many*	p	c	***is*** *small*	p	p
has *shell*	c	c	***used for storing*** *cereal*	r	r
is *large*	c	p	***used for storing*** *spoon*	r	r
is *reptile*	c	c	***used for storing*** *sugar*	p	c

Table 5: Our judges' assessments of the correctness of the top ten relation/feature pairs for two concepts extracted from our best system.

We use the two best systems (i.e., including and excluding the relation; highlighted in Table 3) to generate two sets of top twenty output triples for our 44 concepts. We then calculate precision, recall and F-scores for each against our synonym-expanded set.[3] Using this expanded set allows us to compare our work with that of Kelly et al. (2010). We also compare with the top twenty output of the Reverb system Etzioni et al. (2011) using their publicly available relations derived from the ClueWeb09 corpus, employing their normalized triples ranked by frequency. All sets of results are in Table 4. We note that even though Kelly et al. optimised their algorithm on the ESSLLI set to yield a theoretical best-possible score—we are evaluating 'blind'—our performance still shows an advance on theirs: the improvement on both sets when comparing the population of F-scores across all 44 concepts is statistically significant at the 0.5% level.[4]

3.3 Human evaluation

The above does not quite offer the full picture: unlike the features, the relations are not synonym-expanded. Furthermore, it is possible that there

[3] We note that we are incorporating an upper bound for precision of 0.500 by comparing with only the top ten properties.

[4] Paired t-tests. 'With relation': $t = 3.524$, d.f.$= 43$, $p = 0.0010$. 'Without relation': $t = 3.503$, d.f.$= 43$, $p = 0.0011$.

Relation		A	B	κ	Agreements
With	c / p	146	161	0.7421	261 (87%)
	r / w	153	138		
Without	c / p	226	235	0.5792	255 (85%)
	r / w	74	65		

Table 6: Inter-annotator agreement for our best system, both including and excluding the relation.

are correct properties being generated which simply don't appear in the ESSLLI evaluation set.

In order to address these concerns, we also performed a human evaluation on 15 of our concepts.[5] We asked two native English-speaking judges to decide whether a given triple was *correct*,[6] *plausible*,[7] *wrong but related*,[8] or *wrong*.[9] We executed the human evaluation on our two best systems (as described above). As there were shared triples and concept-feature pairs across the two output sets, each triple and pair was evaluated only once. The judges were aware of the purposes of the study but were blind to the source sets. Some example judgements are in Table 5.

The agreement results across all 15 concepts together with their κ coefficients (Cohen, 1960) are in Table 6. In our evaluation we conflated the *correct*/*plausible* and *wrong but related*/*wrong* categories (see also Kelly et al. (2010) and Devereux et al. (2010)). We did this because of the subjective nature of the judgements, and because we are seeking properties which are indeed correct or at least plausible. These results indicate that our system is extracting correct or plausible triples 51.1% of the time (rising to 76.8% when considering features only). They also demonstrate a marked discrepancy between the results for our two evaluations, reflecting the necessity of human evaluation when assessing our particular task.

4 Discussion

In this paper we have shown that semi-supervised learning techniques can automatically learn lexico-syntactic patterns indicative of property norm-like relations and features. Using these patterns, our system can extract relevant and accurate properties from any parsed corpus and allows for multi-word relation labels, allowing greater semantic precision.

As already mentioned, the work of Baroni et al. (2009) is relevant to our own. Their approach achieves a precision score of 0.239 on the top ten returned features evaluated against the ESSLLI set: our best system offers precision of 0.370 on the same evaluation. Moreover, Baroni et al. do not explicitly derive relation terms. We better the performance of a comparable system (Kelly et al., 2010), even when evaluating against an unseen set of concepts, and our system does not use manually-generated rules or semantic information. Furthermore, human evaluation shows over half of our extracted properties are correct/plausible.

For future work, we have already mentioned that we are ignoring a large amount of potentially instructive training data, specifically those GR-POS paths in our corpus which don't terminate on one of our training features, as well as those paths through sentences containing one of our concepts but none of our training features. It might therefore be worthwhile investigating the use of this "negative" information. Another potential avenue for exploration would be the expansion of the learning vector-types. Although we already use a significant number of learning attributes (an average of 37.9 per training pattern), we could include more: there may be additional information not directly on the GR-POS path linking a concept and feature (e.g., nodes adjacent to said path) which might be indicative of their relationship. We would also consider using active-learning, introducing a feedback loop and human-annotation to better distinguish between relations which our algorithm tends to classify incorrectly. For example, we could supplement input pattern data with disambiguating POS-GR graphs, drawing a distinction between valid and non-valid relations.

Finally, our system could also be evaluated in the context of a psycholinguistic experiment. For example, we could use our system output to predict concept similarity by using our extracted triples to create vector representations of each concept, calculating the distance between those vectors and comparing these similarity ratings with human judgements.

[5] The 44 evaluation concepts had been separated into superordinate categories for unrelated psycholinguistic research and we selected our 15 proportionally and at random from these superordinate categories.

[6] A correct, valid, feature.

[7] A triple which is plausible but only in a specific set of circumstances or a feature which was correct but very general.

[8] The triple is incorrect but there existed some sort of relationship between the concept and relation and/or feature.

[9] When the triple is simply wrong.

Acknowledgements

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

References

M. Andrews, G. Vigliocco, and D. Vinson. 2005. Integrating attributional and distributional information in a probabilistic model of meaning representation. In Timo Honkela et al., editor, *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 15–25, Espoo, Finland: Helsinki University of Technology.

M. Baroni, S. Evert, and A. Lenci, editors. 2008. *ESSLLI 2008 Workshop on Distributional Lexical Semantics*.

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

S. Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

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

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

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

K. Crammer and Y. Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292.

D. Davidov, A. Rappoport, and M. Koppel. 2007. Fully unsupervised discovery of concept-specific relationships by web mining. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 232.

B. Devereux, N. Pilkington, T. Poibeau, and A. Korhonen. 2010. Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language & Computation*, pages 1–34.

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

O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M.T. Center. 2011. Open information extraction: The second generation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

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

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 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*, pages 47–54.

P. Garrard, M.A.L. Ralph, J.R. Hodges, and K. Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2):125–174.

J. Giménez and L. Marquez. 2004. Svmtool: A general pos tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*. Citeseer.

S. Godbole, S. Sarawagi, and S. Chakrabarti. 2002. Scaling multi-class support vector machines using interclass confusion. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 513–518. ACM.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

T. Joachims, C. Nedellec, and C. Rouveirol. 1998. Text categorization with support vector machines: learning with many relevant. In *Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany*, pages 137–142. Springer.

T. Joachims. 1999. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19.

C. Kelly, B. Devereux, and A. Korhonen. 2010. Acquiring human-like feature-based conceptual representations from corpora. In *First Workshop on Computational Neurolinguistics*, page 61. Citeseer.

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

D. Meyer, F. Leisch, and K. Hornik. 2003. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186.

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

P. Pantel and M. Pennacchiotti. 2008. Automatically harvesting and ontologizing semantic relations. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 171–195. IOS Press.

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

K.I. Taylor, B.J. Devereux, K. Acres, B. Randall, and L.K. Tyler. 2011. Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*.

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.

A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. 2003. Inferring a semantic representation of text via cross-language correlation analysis. *Advances in neural information processing systems*, 15:1473–1480.

F. Wu and D.S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.