Acquisition of common sense knowledge for basic level concepts

Eduard Barbu
Center for Mind/Brain Sciences
Rovereto
Trento, Italy
eduard.barbu@unitn.it

Abstract

Feature norms can be regarded as repositories of common sense knowledge for basic level concepts. We acquire from very large corpora feature-norm-like concept descriptions using a combination of a weakly supervised method and an unsupervised method. The success in identifying the specific properties listed in the feature norms as well as the success in acquiring the classes of properties present in the norms are reported.

Keywords

basic level categories, common-sense knowledge, feature norms

1 Introduction

The acquisition of common sense knowledge is the focus of a series of projects originated in AI like CYC [5] or Open Mind [10]. The aim of this paper is the acquisition of every day knowledge for a restricted category of concepts: basic level concepts denoting concrete objects.

One of the main criteria for concept organization in initial studies carried both in psychology and AI [2] was thought to be the taxonomic criteria. Early work in psychology [9] showed that not all levels of taxonomy are equal with respect to object categorization. There is a privileged level at which people consistently classify the objects in common speech called the basic level. For example, encountering an object (e.g. 19th century dinning table) in ordinary discussion we do not categorize it at its specific level (19th century dinning table) nor to its more general level (e.g. entity) but to its basic level (table). The basic level concept is the most inclusive level at which concepts share common features, it carves the world at its joints. Examples of basic level concepts are bird, dog, cat or car.

To acquire common-sense knowledge for basic level concepts we rely on an ongoing effort in cognitive psychology: the feature norms.

In a task called feature-generation subjects list what they believe the most important properties for a set of test concepts are. The experimenter processes the resulting conceptual descriptions and registers the final representation in the norm. Thus, a feature norm is a database containing a set of concepts and their most salient features (properties). The recorded properties are pieces of common sense knowledge. For example, in a norm one finds statements like:

- An **apple** (concept) *is a fruit* (property)[1].
- An **airplane** (concept) *is used for people transportation* (property).

In this paper we explore the possibility to acquire common-sense knowledge from very large corpora. The type of properties one finds in the norms guides the knowledge-extraction task. A double classification of the properties in the norms is used. At the morphological level the properties are grouped according to the part of speech of the words used to express them (noun properties, adjective properties, verb properties). At the semantic level we group the properties in semantic classes (taxonomic properties, part properties, etc.).

The properties in certain semantic classes are learnt using a pattern-based approach, while other classes of properties are learnt using a novel method based on co-occurrence associations.

The rest of the paper has the following organization. The second section discusses the structure of feature norms and presents the procedure for property learning. The third section reports and discusses the results. The fourth section puts our work in context briefly surveying the related work. The paper ends with the conclusions.

2 Feature Norm like Knowledge Acquisition

2.1 Property Classification

For our experiments we choose the feature norm obtained by McRae and colleagues [6]. The norm lists conceptual descriptions for 541 basic level concepts representing living and non-living things and was produced interviewing 725 participants.

We classify each property in the norm at two levels: morphological and semantic.

The morphological level contains the part of speech of the word representing the property. The semantic classification is inspired by a perceptually based taxonomy discussed later in this section. Table 1 shows a

[1] In this paper the concepts will be typed in **bold** and the properties in *italics*

part of the conceptual description for the focal concept **axe** (in this paper the focal concepts are the concept for which the subjects list properties in the feature generation task) and the double classification of the concept properties.

Property	Morphological Classification	Semantic Classification
Tool	Noun	Superordinate
Blade	Noun	Part
Chop	Noun	Action

Table 1: *The double classification of the properties of the concept axe*

The semantic classification is based on Wu and Barsalou (WB) taxonomy [12]. This taxonomy gives a perceptually oriented categorization of properties in the norms. WB taxonomy classifies the properties in 27 distinct classes. Some of these classes contain very few properties and therefore are of marginal interest. For example, the Affect Emotion class classifies only 11 properties. Therefore, we consider only the classes of properties with more than 100 members.

Unfortunately, we cannot directly use the WB taxonomy in the learning process because some of the distinctions it makes are too fine-grained. For example, the taxonomy distinguishes between external components of an object and its internal components. On this account the heart of an animal is an internal component whereas its legs are external components. Keeping these distinctions otherwise relevant from a psychological point of view will hinder the learning of feature norm concept descriptions . Therefore we remap the WB initial property classes on a new set of property classes more adequate for our task. Table 2 presents the new set of property classes together with the morphological classification of the properties in each class.

Morphological Classification	Semantic Classification
Superordinate	Noun
Part	Noun
Stuff	Noun
Location	Noun
Action	Verb
Quality	Adjective

Table 2: *The semantic and morphological classification of properties in McRae feature norm*

The meaning of each semantic class of properties is the following:

- Superordinate. The superordinate properties are those properties that classify a concept from a taxonomic point of view. For example, the **dog** (focal concept) *is an animal* (taxonomic property).

- Part. The category part includes the properties denoting external and internal components of an object. For example *blade* (part property) is a part of an **axe** (focal concept).

- Stuff. The properties in this semantic class denote the stuff an object is made of. For example, **bottle** (focal concept) *is made of glass* (stuff property).

- Location. The properties in this semantic class denote typical places where instances of the focal concepts are found. For example, **airplanes** (focal concept) *are found in airports* (location property).

- Action. This class of properties represents the characteristic actions defining the behavior of an entity (the **cat** (focal concept) *meow* (action property)) or the function, instances of the focal concepts typically fulfill (the **heart** (focal concept) *pumps blood* (function property)).

- Quality. This class of properties denotes the qualities (color, taste, etc.) of the objects instances of the focal concepts. For example, the **apple** (focal concept) *is red* (quality property) or *is sweet* (quality property).

The most relevant properties produced by the subjects in the feature production experiments are in the categories presented above. Thus, asked to list the defining properties of the concepts representing concrete objects subjects will typically: classify the objects (Superordinate), list their parts and the stuff they are made from (Parts and Stuff), specify the location the objects are typically found in (Location), their intended functions, and their typical behavior (Action), or name their perceptual qualities (Quality).

3 Property Learning

To learn the property classes discussed in the preceding section we employ two different strategies. Superordinate, Part, Stuff and Location properties are learnt using a pattern-based approach. Quality and Action properties are learnt using a novel method that quantifies the strength of association between the nouns representing the focal concepts and the adjective and verbs co-occurring with them in a corpus. The learning decision is motivated by the following experiment. We took a set of concepts and their properties from McRae feature norm and extracted sentences from a corpus where a pair concept - property appears in the same sentence.

We noticed that, in general, the quality properties are expressed by the adjectives modifying the noun representing the focal concept. For example, for the concept property pair (**apple**, *red*) we find contexts like:

"She took the red apple" .

The action properties are expressed by verbs. The pair (**dog**, *bark*) is conveyed by contexts like:

"The ugly dog is barking".

where the verb expresses an action to which the dog (i.e. the noun representing the concept) is a participant.

The experiment suggests that to learn Quality and Action properties we should filter the adjectives and verbs co-occurring with the focal concepts.

For the rest of the property classes the extracted contexts suggest that the best learning strategy should be a pattern-based approach. Moreover with the exception of the Location relation, that, to our knowledge, has not been studied yet, for the relations Superordinate, Part and Stuff some patterns are already known. The properties we try to find lexico-syntactic patterns for are classified at the morphological level as nouns (see Table 2). The rest of the properties are classified as either adjectives (Qualities) or verbs (Action). To generate candidate patterns for Superordinate, Part, Stuff and Location relation we follow the procedure discussed in [1]. Basically the hypothesis we pursue is that the best lexico syntactic patterns are those highly associated with the instances representing the relation of interest. The idea is not new and was used in the past by other researchers.However, they used only frequency [8] or pointwise mutual information [7] to calculate the strength of association between patterns and instances. We improve previous work and employ two statistical association measures (Chi Squared and Log Likelihood) for the same task.

The precision of each candidate pattern is evaluated in the following way. A set of 50 concept-feature pairs is selected from a corpus using the devised pattern. For example, to evaluate the precision of the pattern: "Noun made of Noun " for the Stuff relation we extract concept feature pairs like **hammer** - *wood*, **bottle** - *glass*, **car** - *cheese*, etc. Then we label a pair as a hit if the semantic relation holds between the concept and the feature in the pair and a miss otherwise. The pattern precision is defined as the percent of hits. In the case of the three pairs in the example above we have two hits: **hammer** - *wood* and **bottle** - *glass* and one miss: **car** - *cheese*. Thus we have a pattern precision of 66 %.

The Quality and Action properties are learnt using an unsupervised approach. First the association strength between the nouns representing the focal concepts and the adjectives or verbs co-occurring with them in a corpus is computed. The co-occurring adjectives are those adjectives found one word at the left of the nouns representing the focal concepts. A co-occurring verb is a verb found one word at the right of the nouns representing the focal concepts or a verb separated from an auxiliary verb by the nouns representing the focal concepts.

The strongest 30 associated adjectives are selected as Quality properties and the strongest 30 associated verbs are selected as Action properties.

To quantify the attraction strength between the concept and the potential properties of type adjective or verb we use the log-likelihood measure.

4 Results and discussion

4.1 Experimental setup

The corpus used for learning feature-norm-like concept descriptions is ukWaC [3]. UkWaC is a very large corpus of British English, containing more than 2 billion words, constructed by crawling the web. For evaluating the success of our method we have chosen a test set of 44 concepts from McRae feature norm. In the next two subsections we report and discuss the results obtained for Superordinate, Stuff, Location and Part properties and Quality and Action properties respectively. All our experiments were performed using the CWB and UCS toolkits (`http://www.collocations.de/software.html`).

4.2 Results for Superordinate, Stuff, Location and Part properties

For the concepts in the test set we extract properties using the manually selected patterns reported in table 3.

Relation	**Pattern**
Superordinate	Noun [JJ]-such [IN]-as Noun
	Noun [CC]-and [JJ]-other Noun
	Noun [CC]-or [JJ]-other Noun
Stuff	Noun [VVN]-make [IN]-of Noun
Location	Noun [IN]-from [DT]-the Noun
Part	Noun [VVP]-comprise Noun
	Noun [VVP]-consists [IN]-of Noun

Table 3: *The selected patterns*

The results of property extraction phase are reported in Table 4. The columns of the table represent in order: the name of the class of semantic properties to be extracted, the recall of our procedure and the pattern precision. The recall tells how many properties in the test set are found using the patterns in Table 3. The pattern precision states how precise the selected pattern is in finding the properties in a certain semantic class and it is computed as shown at the end of the section 2.2. In case more than one pattern have been selected, the pattern precision is the average precision for all selected patterns.

Property Class	**Recall**	**Pattern Precission**
Superordinate	87%	85%
Stuff	21%	70%
Location	33%	40%
Part	0%	51%

Table 4: *The results for each property class*

As one can see from Table 4, the recall for the superordinate relation is very good and the precision of the patterns is not bad either (average precision 85%). However, many of the extracted superordinate properties are roles and not types. For example, **banana**, one of the concepts in the test set, has the superordinate property: *is a fruit* (type). Using the patterns for superordinate relation we find that **banana** *is a fruit* (a type) but also *is an ingredient* and *is a product* (roles). The lexico-syntactic patterns for the superordinate relation blur the type-role distinction. Other extracted pairs for the superordinates relation include (the left side of the pair contains a concept from the test set, while the right side lists its extracted superordinates): **cat**- (*pet*, *animal*), **potato**-(*vegetable*, *food*),

chicken-(*bird*, *product*). In general, as we see from the pattern precision, the extracted taxonomic knowledge is accurate.

The pattern used to represent the Stuff relation has a bad recall (21 %) and an estimated precision of 70 %. To be fair, the pattern expresses better than the estimated precision the substance an object is made of. The problem is that in many cases constructions of type "Noun made of Noun" are used in a metaphoric way as in: "car made of cheese". In the actual context the car was not made of cheese but the construction is used to show that the respective car was not resistant to impact. Other examples of extracted relations are: **bottle**-(*glass*, *aluminum*), **ship** -(*oak*, *metal*), **cup**-(*stone*, *paper*). The extracted information should be carefully assessed because many times the properties extracted are highly contextual and do not qualify as common-sense knowledge.

The pattern for Location relation has bad precision and bad recall. The properties of type Location listed in the norm represent typical places where objects can be found. For example, in the norm it is stated that **bananas** *are found in tropical climates* (the tropical climate being the typical place where banana-trees grow). However what one can hope from a pattern-based approach is to find patterns representing with good precision the concept of Location in general. We found a more precise Location pattern than the selected one: "N is found in N". Unfortunately, this pattern has 0% recall for our test set. The extracted properties are in general imprecise: **duck**- (*exploit*), **hammer**-(*north*).

The patterns for Part relation have 0% recall for the concepts in the test set and their precision for the general domain is not very good either. As others have shown [4] a pattern based approach is not enough to learn the part relation and one needs to use a supervised approach to achieve a relevant degree of success.

4.3 Results for Quality and Action properties

We computed the association strength between the concepts in the test set and the co-occurring verbs and adjectives using the log-likelihood measure. Some of the extracted properties for the concepts in the test set are shown in Table 5.

The results for Quality and Action properties are presented in Table 6. The columns of the table represent in order: the name of the class of semantic properties, the Recall and the Property Precision. The Recall represents the percent of properties in the test set our procedure found. The Property Precision computes the precision with which our procedure finds properties in a semantic class. The property precision is the percent of quality and action properties found among the strongest 30 adjectives and verbs associated with the focal concepts.Because the number of potential properties is reasonable for hand checking, the validation for this procedure was performed manually.

The manual comparison between the corpus extracted properties and the norm properties confirm the hypothesis regarding the relation between the association strength of features of type adjective and verbs and their degree of relevance as properties of concepts.

Concept	Quality	Action
Duck	*wild*, *tufted* *lame*, *ruddy*	*waddle*, *fly* *swim*, *quack*
Eagle	*golden*, *bald* *white-tailed*, *spotted*	*soar*, *fly* *perch*, *swoop*
Turtle	*marine*, *green* *giant*, *engendered*	*dive*, *nest* *hatch*, *crawl*

Table 5: *Some quality and action properties for the concepts in the test set*

Property Class	Recall	Property Precission
Quality	60%	60%
Action	70%	83%

Table 6: *The results for Quality and Action property classes*

For each concept in the test set roughly 18 adjectives and 25 verbs in the extracted set of potential properties represent qualities and action respectively (see Property Precision column in Table 6). This can be explained by the fact that all concepts in the test set denote concrete objects. Many of the adjectives modifying nouns denoting concrete objects express the objects qualities, whereas the verbs usually denote actions different actors perform or to which various objects are subject.

Many of the properties found using this method encode pieces of common sense knowledge not present in the norms. For example, the semantic representation of the concept **turtle** has the following Quality properties listed in the norm: *green*, *hard*, *small.* The strongest adjectives associated in the UkWaC corpus with the noun turtle ordered by the loglikelihood score are: *marine*, *green*, *giant.* The property *marine* carries a greater distinctiveness than any of similar feature listed in the norms.

Likewise, the actions typically associated with the concept **turtle** in the McRae feature norm are: *lays eggs*, *swims*, *walks slowly.* The strongest verbs associated in the UkWaC corpus with the noun turtle are: *dive*, *nest*, *hatch.* The *dive* action is more specific and therefore more distinct than the *swim* action registered in the feature norm. The *hatch* property is characteristic to reptiles and birds and thus a good candidate for the representation of the concept turtle.

5 Related Work

The need of acquiring common-sense knowledge to enable computers understand and reason with natural language was recognized long time ago. The first large-scale effort for acquisition of common sense knowledge is the project CYC. Human users codify by hand millions of rules representing every-day knowledge (in CYC one finds concepts like cat and mammal and assertions like the cat is a mammal).

A more up to date effort to acquire knowledge about daily life is the project OpenMind. It attempts at building a huge database of common sense knowledge exploiting the wisdom of crowds. Thousands of non-expert contributors introduce knowledge inside a set of predefined scenarios like: Story telling, Typical arguments of verbs or the Listing of objects appearing usually together.

An interesting method to gather the common-sense knowledge is von Ahns work, who draws on the data collected with the help of online games [11].

The work reported here uses an alternative basis for common-sense property acquisition, it builds on the effort in cognitive psychology to extract kinds of properties people are likely to know about the concepts. Of course, as the experience of CYC shows, there is much more to common sense knowledge than the acquisition of concept properties. However we think that our work, having a sound empirical basis, is a step in the right direction.

6 Conclusions

The presented method for acquiring common-sense knowledge based on feature-norm concept description has been successful at learning semantic property classes Superordinate, Quality and Action. For learning the superordinates of the focal concepts one needs to use a high precision pattern. For Quality and Action properties one needs to apply the method based on co-occurrence association presented in section 2.2.

To learn all other property classes other methods (probably a supervised approach) must be devised.

Acknowledgments

The author would like to thank Verginica Barbu Mititelu and three anonymous reviewers for their suggestions and corrections.

References

[1] E. Barbu. Combining methods to learn feature-norm-like concept descriptions. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 9–17, August 2008.

[2] A. M. Collins and M. R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:241–248, 1969.

[3] A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *WAC4 Workshop Conference Proceedings*, pages 84–89, 2008.

[4] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, March 2006.

[5] D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[6] K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, Nov. 2005.

[7] P. Pantel and M. Pennacchiotti. Espresso: A bootstrapping algorithm for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, 2006.

[8] D. Ravichandran and E. Hovy. Learning learning surface text patterns for a question answering system. In *Proceedings of ACL*, 2002.

[9] E. Rosch and C. Mervis. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.

[10] P. Singh. The public acquisition of commonsense knowledge. In *AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, 2002.

[11] L. von Ahn. Games with a purpose. *IEEE Computer Magazine*, 39(6):96–98, June 2006.

[12] L. Wu and L. Barsalou. Perceptual simulation in conceptual combination. *Acta Psychologica*, page In press, 2009.