

Extracting Entailing Words from Small Corpora for Ontology Building

Aurelie Herbelot

Computer Laboratory

University of Cambridge

J.J. Thompson Avenue, Cambridge, United Kingdom

aurelie.herbelot@cl.cam.ac.uk

Abstract

This paper explores the extraction of conceptual clusters from a small corpus, given user-defined seeds. We use the distributional similarity hypothesis (Harris, 1968) to gather similar terms, using semantic features as context. We attempt to preserve both precision and recall by using a bootstrapping algorithm with reliability calculations proposed by Pantel and Pennacchiotti (2006). Precision of up to 78% is achieved for our best query over a 16MB corpus. We find, however, that results are dependent on initial settings and propose a partial solution to automatically select appropriate seeds.

1 Introduction

The entailment task can be described as finding pairs of words so that one (the entailed word) can replace the other (the entailing word) in some contexts. Or in other words, it consists in finding words in a hyponym/hypernym relation. This is of course also the task of ontology extraction when applied to the is-a relationship. Hence, entailment tools such as those based on the distributional similarity hypothesis (Harris, 1968) can also be used in ontology extraction, as shown by Pantel and Lin (2002) in a clustering task.

The clustering method in ontology extraction was pioneered by Caraballo (1999). She showed that semantically-related words could be clustered together by mining coordinations and conjunctions.

Pantel and Lin (2002) followed in her path using distributional similarity to extract large clusters of related words from corpora.

The main advantage of the clustering method is that it allows users to find hyponymic relations that are not explicitly mentioned in the corpus. One major drawback is that the extracted clusters must then be appropriately named - a task that Pantel and Ravichandran (2004) showed has no simple solution. Furthermore, although mining a corpus for all its potential clusters may be a good way to extract large amounts of information, it is not a good way to answer specific user needs. For instance, if I wish to compile a list of all animals, cities or motion verbs in my corpus, I must mine the whole text, hope that my query will be answered by one of the retrieved clusters and identify the correct group. Finally, previous work has suggested that clustering is only reliable for large corpora: Pantel and Pennacchiotti (2006) claim that it is not adequate for corpora under 100 million words.

This paper proposes a user-driven approach to clustering where example seeds are given to the system, patterns extracted for those seeds and similar words subsequently returned, following the typical entailment scenario. The obvious difficulty, from an ontology extraction point of view, is to overcome data sparsity without compromising precision (the original handful of seeds might not produce many accurate patterns). We therefore investigate the use of bootstrapping on one hand in order to raise the number of extractions and of semantic features with reliability calculations on the other hand to help maintain precision to an acceptable level.

The next section reviews relevant previous work and includes a description of the piece of work which motivated the research presented here. We then describe our algorithm and experimental setup. Results for the whole corpus (16MB) and a 10% subset are presented and discussed. Problems relating to initial setting sensitivity are noted, and a partial solution proposed. We finally conclude with avenues for future work.

2 Previous Work and Motivation

The clustering method represents so far a marginal approach in a set of ontology extraction techniques dominated by the lexico-syntactic pattern-matching method (Hearst, 1992). Clustering was initially proposed by Carballo (1999) who used conjunction and coordination to cluster similar words. She obtained a precision of 33% on her hyponymy extraction task.

Pantel and Lin (2002), following from Carballo's work, proposed their 'clustering by committee' algorithm, using distributional similarity to cluster similar words. Their algorithm distinguishes between various senses of a word. Pantel and Ravichandran (2004) report that the algorithm has a precision of 68% over a 3GB corpus (the figure is calculated over the relation between clusters and their automatically generated names.)

On the entailment front, Geffet and Dagan (2005) also used distributional similarity over an 18 million word corpus and obtained up to 74% precision with a novel feature weighting function (RFF) and an Inclusion Testing algorithm which uses the k characteristic features common to two words in an entailment relation.

Our own investigation derives from a previous ontology extraction project (Herbelot and Copestake, 2006) on Wikipedia (<http://www.wikipedia.org/>). That project focused on uncovering taxonomic relationships in a corpus consisting of over 12,000 Wikipedia pages on animals. We extracted a semantic representation of the text in RMRS form (Copestake, 2004) and manually defined patterns characteristic of the taxonomic relationship, also in RMRS format. Matching those patterns to the text's semantics allowed us to return hyponymic re-

lationships which were filtered for taxonomic pairs. (The filtering simply consisted in checking whether the hyponym and hypernym were animal names, using a list compiled from Wikipedia article titles.) A careful evaluation was performed, both manually on a subset of the results and automatically on the whole extracted file using the NCBI¹ taxonomy. We reported a precision of 88.5% and a recall of 20%. This work highlighted the fact that the dictionaries of animal names that we had at our disposal (both the list extracted from Wikipedia and the NCBI itself) were far from comprehensive and therefore affected our recall.

In this paper, we attempt to remedy the shortcomings of our dictionaries and investigate a mining algorithm which returns conceptual clusters out of a small, consistent corpus. The fairly conventionalised aspect of Wikipedia articles (the structure and vocabulary become standardised with usage) tends to produce good, focused contexts for certain types of words or relationships, and this partly overcomes the data sparsity problem. We therefore propose the realistic task of finding clusters of terms that a reader of biological texts might be interested in. Here, specifically, we focus on animal names, geographical areas (i.e. potential animal habitats) and parts of the animal body. We reuse the Wikipedia corpus from our previous work - 16MB of plain text - and apply to it distributional similarity using semantic features, with a bootstrapping algorithm proposed by Pantel and Pennacchiotti (2006).

3 The Algorithm

The aim of the algorithm is to find words that are similar to the seeds provided by the user. In order to achieve this, we use the distributional similarity hypothesis (Harris, 1968) which states that words that appear in the same context are semantically related. Our 'context' consists here of the semantic triples in which a word appears, with the semantics of the text referring to its RMRS representation (Copestake, 2004). So for instance, in the sentence 'the cat chased a green mouse', the word 'mouse' has a context comprising two triples:

```
lemma:chase pos:v arg:ARG2 var:mouse pos:n  
which indicates that 'mouse' is object of 'chase' and
```

¹www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy

lemma:green pos:j arg:ARG1 var:mouse pos:n
 which indicates that the argument of ‘green’ is ‘mouse’.

Let’s assume that ‘mouse’ is one of the seeds provided to the system. We can transform its context into generic features by replacing the slots containing the seed with a hole: lemma:chase pos:v arg:ARG2 lemma:mouse pos:n becomes lemma:chase pos:v arg:ARG2 lemma:hole_ pos:n.

Then, every time a generic feature is encountered in the text, we can hypothesise that whichever word fills the hole position is semantically similar to our seed: if we encounter the triple lemma:chase pos:v arg:ARG2 var:bird pos:n, we hypothesise that ‘bird’ belongs to the same semantic class as ‘mouse’. (We assume that a match on any one feature - as opposed to the whole context - is sufficient to hypothesise the presence of an instance.)

We initially extract all the features that include one of the seeds presented to the system. We filter those features so that semantically weak relations, such as the one between a preposition and its argument, are discarded: triples containing a preposition or quantifier as the head or argument of the relation are deleted. At the moment, we are also leaving conjunction aside. Although we acknowledge, following Caraballo (1999), that conjunction is a good indicator of semantic similarity, it would mean expanding our features to two triples. For instance, ‘cat and mouse’ would be expressed by the pair lemma:and pos:x arg:L-INDEX var:cat pos:n and lemma:and pos:x arg:R-INDEX var:mouse pos:n. This is regarded as a necessary improvement but not implemented here. The features left after filtering are transformed into generic patterns by replacing the slot containing the seeds with a hole, as explained above.

We then calculate the reliability of those features using the calculation proposed by Pantel and Penacchiotti (2006): we calculate Pointwise Mutual Information between each feature and the instances it extracts:

$$pmi(f, i) = \log \left(\frac{P(f, i)}{P(f)P(i)} \right) \quad (1)$$

where $P(f)$ and $P(i)$ are the probabilities of occurrence of the feature and the instance respectively and

$P(f, i)$ is the probability that they appear together.

Pointwise Mutual Information is known for producing scores in favour of rare events. In order to counterbalance this effect, our figures are multiplied by the discount factor suggested in Pantel and Ravichandran (2004):

$$d = \frac{c_{if}}{c_{if} + 1} * \frac{\min(c_i, c_f)}{\min(c_i, c_f) + 1} \quad (2)$$

where c_{if} is the cooccurrence count of an instance and a feature, c_i the frequency count of instance i and c_f the frequency count of feature f .

We then find the reliability of the feature as:

$$r_f = \frac{\sum_{i \in I} \left(\frac{pmi(i, f)}{\max_{pmi}} * r_i \right)}{|I|} \quad (3)$$

where r_f and r_i are the reliabilities of the feature and of an instance respectively, and I is the total number of instances extracted by f .

Initially, the seeds have reliability 1 and all the other words reliability 0. We then select features with n -best reliabilities. Those features are used to extract new instances, the reliability of which is calculated in the same fashion as for the initial patterns:

$$r_i = \frac{\sum_{f \in F} \left(\frac{pmi(i, f)}{\max_{pmi}} * r_f \right)}{|F|} \quad (4)$$

We select the instances with m -best reliabilities to replace the original seed list.

We then bootstrap. At each iteration, we increment n and m .

4 Experimental Setup

4.1 Corpus

The corpus is the one used in Herbelot and Copestake (2006); it consists of 12,200 Wikipedia articles on animals. The text was parsed using RASP3 (Briscoe and Carroll, 2002) and the RASP-to-RMRS converter (Copestake, 2004) was applied to the derivations to obtain the RMRS representation. In order to ascertain the limitations of the algorithm with regard to the size of the corpus, we experimented both on a subset of the corpus (1000 files) and on the entire text.

4.2 Initial Settings

All experiments are subject to the setting of three parameters:

- the original set of seeds given to the program
- the initial number of features retained for the extraction of new instances, after the feature reliability calculation phase (n_0)
- the initial number of instances retained to form the new set of seeds, after the instance reliability calculation phase (m_0).

At each iteration, we increment n by 1 and m by 5.

We attempted to extract various clusters out of the corpus. Our three queries were animal names, geographical areas/features and parts of the animal body. Small noun clusters were given to the system as seeds:

- (1) animal, mammal, fish, bird, insect, cat, snake
- (2) wetland, marshland, farmland, land, region, hill, forest, desert
- (3) whisker, hoof, scale, vertebra, fin, beak, abdomen, wing, shell

The subset corpus involved 5 iterations with, initially, $n_0=1$ and $m_0=5$. The whole corpus involved 10 iterations with, initially, $n_0=1$ and $m_0=10$. We recorded, at each iteration, the full list of instances extracted by the n -best features (before selection of the m -best words). This constituted our results.

4.3 Recall

As always in this type of project, the calculation of recall demands intensive manual work. Due to time restrictions, we mostly show here the number of instances extracted at each iteration of the algorithm. For comparison purposes, however, we also calculated true recall for the best iteration (the most precise) of the animal query. The figure was obtained over a 100 file subset of the whole corpus: we extracted all unique nouns out of the subset and manually annotated the animal names, yielding 527 entries. We then ran the instance extraction program over the 100 files, using the best feature list obtained by the system (the one with the best precision). Our recall figure was then taken as the number of extractions over the 100 files divided by the true number of animal names (527).

Table 1: Number of Extractions and Precision for the Three Queries, First Five Iterations

Animal Query		
Num Features	Num Extractions	Precision
1	97	86%
2	151	83%
3	174	70%
4	277	57%
5	289	54%
Geography Query		
Num Features	Num Extractions	Precision
1	1	100%
2	17	18%
3	27	26%
4	74	12%
5	74	12%
Parts of the Body Query		
Num Features	Num Extractions	Precision
1	1	100%
2	98	45%
3	313	25%
4	313	25%
5	313	25%

4.4 Precision

Precision is calculated, at each iteration, by one annotator. For larger files, we only show precision figures for the first 100, 200, 300 and 400 instances. The criterion for precision is whether the instance satisfies the entailment condition. So for instance, the word ‘creature’ is considered correct for the animal query because it entails ‘animal’ while the word ‘phylum’ is considered incorrect as it can entail any kind of organisms, including plants.

5 Results

5.1 Subset Corpus

Our results for the small 1000 file corpus experiments are shown in table 1.

The precision for the animal query is acceptable, although it decreases quickly with the introduction of noisy features. The precision for the other two queries is extremely low. We find, however, that set-

Table 2: Number of Extractions and Precision for the First Five Iterations, $n_0=2$, geography query

Number of Features	Number of Extractions	Precision
1	1	100%
2	35	77%
3	59	56%
4	105	33%
5	106	33%

ting $n_0=2$ and $m_0=10$ for the landscape query produces far better results (see table 2). This is due to the second feature extracted at iteration 1 producing better instances than the first and pushing correct seeds at the top of the reliability list.

5.2 Whole Corpus

Our experiments on the whole corpus (12,200 files) highlight that on this task, the algorithm is also very sensitive to our choice of seeds. For the animal query, using the same seeds as for the subset leads to a number of extractions no better than in the initial experiments and to worse precisions (see table 3). The body and geography queries produce results similar to the ones obtained in the subset experiment (with $n_0=1$ for the parts of the body, $n_0=2$ for geographical features) but using the initial settings $n_0=1$ and $m_0=10$ for both queries. This seems to indicate that the only way the algorithm is affected by a minimal corpus size is by being more sensitive to the setting of n_0 and m_0 . The precision itself is not linked to size, in this experiment at least.

We investigated the role of the seeds in this result and attempted to improve our figures by lengthening the animal list to 16 initial seeds. This did not change the results. Shortening the list again, taking away the more general nouns and lower frequency words (*animal, fish, insect, bird, cat*) led however to an acceptable precision again (see table 4). The final seed list consists of *spider, snake, cetacean, cattle, ant, crustacean, reptile, beetle, monkey, dinosaur, hamster, tortoise*. This brings up the question of defining a ‘good seed’, which is discussed in the next section.

The first 25 instances for the best results are shown in table 5.

Table 3: Number of Extractions and Precision for the First 10 Iterations, Animal Query, Whole Corpus, Initial Set of Seeds

Number of Features	Number of Extractions	Precision
1	16	63%
2	69	39%
3	110	54%
4	140	48%
5	189	44%
6	201	44%
7	202	44%
8	221	43%
9	251	44%
10	269	45%

Table 5: First 25 Instances out of Best Results for All Three Experiments

Animal query	Geography query	Parts of the body query
muricoid	marshland	whisker
flickerus	region	rao
acoel	hillock	inability
hamster	embankment	electroreceptor
ungulate	ditch	cue
cattle	terrain	hearing
blesmol	upland	existence
monkey	savannah	smell
reptile	key	sight
crocodile	slope	effort
dinosaur	cliff	skeleton
predator	dune	wood
duck	wetland	belly
megalosaur	scrub	dinosaur
palimation	field	order
scyphomedusa	environment	chemoreceptor
sophistication	vegetation	ending
coral	meter	apparatus
balitorina	patch	receptor
haeckelium	savanna	taste
kurilensis	island	barbel
lepiotacea	location	bristle
paramelemorph	woodland	membrane
porcelanid	country	tissue
sea-scorpion	plain	tortoise

Table 4: Number of Extractions and Precision for the First 10 Iterations, Animal Query, Whole Corpus, New Set of Seeds

Num Features	Num Extractions	Precision 100	Precision 200	Precision 300	Precision 400
1	233	41%	-	-	-
2	467	62%	46%	50%	58%
3	653	69%	60%	64%	65%
4	797	75%	60%	57%	59%
5	859	74%	74%	60%	57%
6	970	76%	73%	63%	59%
7	1622	78%	72%	64%	59%
8	1717	75%	70%	64%	57%
9	1719	74%	72%	64%	60%
10	1804	75%	72%	67%	62%

5.3 Recall Considerations

We calculated recall over a 100 file subset, using the features obtained after iteration 7. We obtained 185 extractions, yielding a recall of 35%. This figure is similar to the recall obtained in our previous RMRS-based project, when attempting automatic pattern extraction (see Herbelot and Copestake, 2006 Section 5 reporting a 37% recall). We have not yet fully investigated the cause for the low recall in this task but it can partially be attributed to the incomplete rule coverage mentioned in our previous work. This is being remedied with a new version of the RASP-RMRS converter. The problem is also inherent to the algorithm: using only a handful of best features for the extraction of new instances is bound to limit coverage.

6 Discussion

As the algorithm is sensitive to the original seeds, we would like our program to be able to select good seeds without relying on the user’s choice. Intuitively, a good seed is one that has a medium frequency in the corpus and is situated midway in its conceptual hierarchy:

- a rare word makes the program suffer from data sparsity while a very frequent one is too general and returns patterns unrelated to the subject matter.

- a general concept appears in a wider variety of features than a less generic one and is therefore less ‘focused’. For instance, ‘animal’ appears with the nouns ‘rights’ (animal rights) and with the verbs

‘swim’, ‘walk’, ‘climb’, etc while ‘fish’ only select for ‘swim’ and is therefore more likely to give a good quality pmi with the swimming-related features.

This intuition is supported by some linguistic observations in prototype semantics. A prototype is a linguistic object which is considered a particularly good exemplar of the conceptual category it belongs to; i.e., it can be described using elements which have a strong semantic tie to its category or hypernym. For instance, a dog is a good prototype for the category ‘animal’ because it exhibits a physical appearance and behaviour that we commonly associate with animals. By contrast, some people would argue that humans or corals are not animals on the basis of behavioural patterns and shape. We are not interested here in the philosophical implication of linking humans to the animal concept but rather in the potential of good prototypes to also be good seeds. It is indeed expected that a convincing prototype will frequently appear in contexts highly characteristic of the concept under consideration.

Rosch et al (1976) introduce the notion of ‘basic level category’ to refer to the level in a conceptual hierarchy which best gathers the characteristic elements of a concept, that is, the categorical level of the best prototype. It is found that in the animal taxonomy, for example, the genus level is the basic level category (as opposed to higher or lower levels such as kingdom or species). The notion usually refers to levels halfway through the hierarchy,

those most commonly used when naming an object belonging to the concept under consideration (compare *Look at that vertebrate!*/*Look at that dog!*/*Look at that Yorkshire terrier!*). These findings tally with our intuition that good seeds can be found midway through the WordNet hierarchy.

As a preliminary investigation of this idea, we selected five seeds from the top three levels of the animal hierarchy in WordNet (the data is sparse at those levels, hence the use of the first three levels instead of one only) and five seeds from the fourth level, under the synset ‘vertebrate’. The occurrence counts of the ‘good’ seeds in our second experiment on the animal query (see Section 5) have an average of 758. We estimated that seeds with a frequency between 420 and 1100 were acceptable (roughly 758 +/- 50% – this could be expressed as a percentage of the number of words in the corpus) and used this range to constrain our second set of words.

Following this heuristic, our two sets of seeds are:

- (1) bird, animal, reptile, mammal, insect
- (2) parrot, passerine, bat, salamander, rodent

We observe that the number of extractions for the first test after 10 iterations is of 8 entries only with a 50% precision. The number of extractions for the second test is of 50 entries with a 64% precision.

Another test on our worst query, the body parts, had similar effects: we selected the ‘external body parts’ level as our more general level and the level ‘appendage’ as the lower one (those are parent and child in WordNet - the body hierarchy is a lot flatter than the animal one). This time, we did not constrain the seeds’ frequency count, as the WordNet data was sparse. After 10 iterations, we obtained 4190 terms with the higher level - the result of a noisy generic pattern involving the identity copula - and 260 terms with the lower level. Precisions were calculated as 13% and 23% respectively on the first 100 terms, showing that, even though results are extremely low, the more specific seeds perform better than the general ones. This seems to support our hypothesis, although further experiments are obviously needed.

Regarding the overall sensitivity of the program to initial settings (see the choice of n_0 in the previous section), our aim for further work is to fully test the algorithm over a range of settings and seeds. If it is not sufficiently robust to a variety of initial

conditions, it may need revisiting.

We also note, looking at the feature list for the animal query, that the patterns extracted are very much corpus-dependent. We give the list of the best seven features at the 7th iteration of the animal query in table 6.

We would not normally expect features including ‘modern’ or ‘appear’ to be indicative of the presence of an animal, but they work well on this particular set of texts where species are likely to ‘appear’ or to be the ‘modern’ form of an ancient animal.

The disappointing precision on the body query can be explained by the fact that many terms used to describe parts of organs are borrowed from general vocabulary: *area*, *tip*, *side*, etc. On one hand, these terms are not likely to have a strong pmi with focused body-related features. On the other hand, when extracted, they then return too general features at pattern-extraction stage. It is also likely that both the geography and body queries suffer from data sparsity (many features return less than 10 instances). This demonstrates that, although clustering can be used on a small-size corpus for queries related to the main topic - here, animals - it may not be so successful when it comes to concepts outside of the text’s focus.

7 Conclusion and Further Work

We have presented here a clustering technique to extract and gather words under user-chosen concepts from a small size corpus (16MB). Using semantic features as characteristic context of each concept and a bootstrapping algorithm previously proposed by Pantel and Pennacchiotti (2006), we achieve 78% precision on the top 100 terms of our best query (72% on the top 200). We note that the performance of the algorithm is very sensitive to initial settings, in particular to the initial number of characteristic features retained and to the choice of seeds. Accordingly, we would like to investigate ways to improve the robustness of the program to user-defined choices. We propose a partial solution to the issue of the seeds by selecting mid-frequency terms out of WordNet levels situated half-way through the conceptual hierarchy.

The less successful queries suffer not only from original settings but also from data sparsity. We aim

Table 6: Seven Best Features For Animal Query

lemma::appear	lempos::v	arg::ARG1	var::hole	pos::n
lemma::like	lempos::v	arg::ARG2	var::hole	pos::n
lemma::species	lempos::n	arg::ARG1	var::hole	pos::n
lemma::go	lempos::v	arg::ARG1	var::hole	pos::n
lemma::modern	lempos::j	arg::ARG1	var::hole	pos::n
lemma::allow	lempos::v	arg::ARG2	var::hole	pos::n
lemma::call	lempos::v	arg::ARG2	var::hole	pos::n

to investigate how to keep only the best seeds at the top of the returned instance list at each iteration. We would like to propose a separate clustering step on the extracted instances, relying on shared extracting features (i.e. it is possible that the ‘good’ instances share the same features while the ‘bad’ seeds are extracted by less similar feature lists).

We have also shown that ontology extraction by clustering is possible on a small topic-focused corpus. We would hope to extend these results to other texts.

8 Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC: EP/P502365/1 project). I would also like to thank my supervisor, Dr Ann Copestake, for her support and comments throughout this project, as well as the anonymous reviewers for their helpful feedback. Finally, special thanks to the people who provided typing support for me and were confronted with the esotericism of the command line.

References

Briscoe, Ted and Carroll, John. 2002. *Robust Accurate Statistical Annotation of General Text*. In Proceedings of the third international conference on language resources and evaluation (LREC 2002), Las Palmas, Canary Islands.

Caraballo, Sharon. 1999. *Automatic Construction of a Hyponym-Labeled Noun Hierarchy from Text*. In Proceedings of ACL99, Baltimore, MD, pp 120-126.

Copestake, Ann. 2004. *Robust Minimal Recursion Semantics*. <http://www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf>.

Geffet, Maayan and Dagan, Ido. 2005. *The Distributional Inclusion Hypotheses and Lexical Entailment*. In Proceedings of ACL2005, Ann Arbor, MI.

Harris, Zelig. 1968. *Mathematical Structures of Language*. Wiley, 1968.

Hearst, Marti. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proceedings of COLING-92, Nantes, France, pp 539-545.

Herbelot, Aurelie and Copestake, Ann. 2006. *Acquiring Ontological Relationships from Wikipedia Using RMRS*. In Proceedings of Workshop on Web Content Mining with Human Language Technologies, ISWC06, Athens, GA.

Pantel, Patrick and Lin, Dekang. 2002. *Discovering Word Senses from Text*. In Proceedings of SIGKDD02, Edmonton, Alberta, Canada.

Pantel, Patrick and Pennacchiotti, Marco. 2006. *Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations*. In Proceedings of COLING/ACL06, Sydney, Australia, pp 113-120.

Pantel, Patrick and Ravichandran, Deepak. 2004. *Automatically Labelling Semantic Classes*. In Proceedings of HLT/NAACL04, Boston, MA, pp 321-328.

Rosch Eleanor, Mervis, C., Gray, W., Johnson, D. and Boyes-Bram P. 1976. *Basic Objects in Natural Categories*. *Cognitive Psychology*, 8:382-439.