Qualia Structures and their Impact on the Noun Categorization Task

Sophia Katrenko Pieter Adriaans

University of Amsterdam, the Netherlands

ESSLLI workshop, August 6 2008

Outline

Previous Work

1. Lexical representation/categorization in cognitive science
   - a lexical concept is represented by a set of features (Rapp & Caramazza, 1991; Gonnerman et. al., 1997)
   - lexical concepts are atomic representations and "conceptual relations . . . can be captured by the sets of inferential relations drawn from elementary and complex concepts" (Almeida, 1999), the thesis of conceptual atomism (Fodor, 1990)
2. Categorization in computational lingustics
   - word-space models (Sahlgren, 2006; Lenci, Baroni, and others)

Data

- 44 concrete nouns to be categorized in

Data

- 44 concrete nouns to be categorized in
  - 2 categories (natural kind and artifact)

Data

- 44 concrete nouns to be categorized in
  - 2 categories (natural kind and artifact)
  - 3 categories (vegetable, animal and artifact)

Data

- 44 concrete nouns to be categorized in
  - 2 categories (natural kind and artifact)
  - 3 categories (vegetable, animal and artifact)
  - 6 categories (green, fruitTree, bird, groundAnimal, vehicle and tool) the entity derived from the origin.

Generative Lexicon Theory

Pustejovsky (1998) proposed a linguistically motivated approach to modelling categories. Semantic descriptions use 4 levels of linguistic representations such as

- argument structure ("specification of number and a type of logic arguments")
- event structure ("definition of the event type of an expression")
- qualia structure ("a structural differentiation of the predicative force for a lexical item")
- lexical inheritance structure ("identification of how a lexical structure is related to other structures in the type lattice")

Generative Lexicon Theory (cont'd)

$$\begin{bmatrix} \alpha \\ \text{ARGSTR} : \begin{bmatrix} \text{ARG1} : x \\ \ldots \end{bmatrix} \\ \text{EVSTR} : \begin{bmatrix} \text{EV1} : e_1 \\ \ldots \end{bmatrix} \\ \text{QUALIA} : \begin{bmatrix} \text{CONST} : \textbf{what } x \textbf{ is made of} \\ \text{FORMAL} : \textbf{what } x \textbf{ is} \\ \text{TELIC} : \textbf{function of } x \\ \text{AGENTIVE} : \textbf{how } x \textbf{ came into being} \end{bmatrix} \end{bmatrix}$$

Generative Lexicon Theory (cont'd)

What features/properties are important to classify the concrete nouns correctly? According to the Generative Lexicon Theory, lexical expressions are represented by the following roles:

- formal (how to distinguish a given object from the other, *is-a* information)

$$Formal(\lambda x[\alpha(x)]) = \lambda x[Q(x)] \leftrightarrow \alpha \subseteq Q$$

- constitutive (*part-whole* information, parts of the object)

$$Const([\alpha(x)]) = \lambda y[Q(y)] \leftrightarrow \forall x[\alpha(x) \rightarrow \exists y[Q(y) \wedge made\_of(x,y)]]$$

Generative Lexicon Theory (cont'd)

- telic (a purpose of an object, what it is used for)

$$Telic(\lambda x[\alpha(x)]) = \lambda y \lambda e \exists x[\phi(e,y,x)] \leftrightarrow$$
$$\lambda y \forall x \forall e \forall y[\psi_\alpha(e,y,x) > \exists e'[\phi(e',y,x) \wedge e < e']]$$

- agentive (origin, how it came into being)

$$Agentive(\lambda x[\alpha(x)]) = \lambda e[\psi(e)] \leftrightarrow$$
$$\forall x, e[\alpha(e,x) \rightarrow \exists e' \exists y[\psi(e') \wedge e' \prec e \wedge make(e',y,x)]]$$

Generative Lexicon Theory (cont'd)

a. $$\left[\begin{array}{lll} \textbf{snowball} & & \\ \text{ARGSTR} & = & \left[\begin{array}{lll} \text{ARG1} & = & x \\ \text{D-ARG1} & = & y \end{array}\right] \\ \text{QUALIA} & = & \left[\begin{array}{lll} \text{FORMAL} & = & ball(x) \\ \text{CONST} & = & snow(y) \end{array}\right] \end{array}\right] \Longrightarrow$$

b. $$\lambda x[ball(x) \wedge const(x) = \exists y[snow(y)]]$$

Approach (cont'd)

How can we acquire qualia information? Some of the methods proposed in the past:

- Hearst, 1992 (hyperonymy)
- Girju, 2007 (part-whole relations)
- Cimiano and Wenderoth, 2007
  - predefined patterns for all 4 roles
  - ranking results according to some measures
- Yamada et al., 2007
  - fully supervised
  - focuses on acquisition of telic information

Approach (cont'd)

We make use of the patterns defined by Cimiano and Wenderoth, 2007

role	pattern
formal	x_NN is_VBZ (a_DT|the_DT) kind_NN of_IN
	x_NN is_VBZ
	x_NN and_CC other_JJ
	x_NN or_CC other_JJ
telic	purpose_NN of_IN (a_DT)* x_NN is_VBZ
	purpose_NN of_IN p_NNP is_VBZ
	(a_DT|the_DT)* x_NN is_VBZ used_VVN to_TO
	p_NNP are_VBP used_VVN to_TO

Table: Patterns: some examples

Approach (cont'd)

role	pattern
constitutive	(a_DT|the_DT)* x_NN is_VBZ made_VVN (up_RP)*of_IN
	(a_DT|the_DT)* x_NN comprises_VVZ
	(a_DT|the_DT)* x_NN consists_VVZ of_IN
	p_NNP are_VBP made_VVN (up_RP)*of_IN
	p_NNP comprise_VVP
agentive	to_TO * a_DT new_JJ x_NN
	to_TO * a_DT complete_JJ x_NN
	to_TO * new_JJ p_NNP
	to_TO * complete_JJ p_NNP
	a_DT new_JJ x_NN has_VHZ been_VBN
	a_DT complete_JJ x_NN has_VHZ been_VBN

Table: Patterns: some examples

Approach (cont'd)

- Categorization procedure consists of the following steps
  - extraction of the passages containing candidates for the role fillers using patterns (Google, 50 snippets per pattern)

Approach (cont'd)

- Categorization procedure consists of the following steps
  - extraction of the passages containing candidates for the role fillers using patterns (Google, 50 snippets per pattern)
  - PoS tagging of all passages

Approach (cont'd)

- Categorization procedure consists of the following steps
  - extraction of the passages containing candidates for the role fillers using patterns (Google, 50 snippets per pattern)
  - PoS tagging of all passages
  - actual extraction of the candidates for the role fillers using patterns

Approach (cont'd)

- Categorization procedure consists of the following steps
  - extraction of the passages containing candidates for the role fillers using patterns (Google, 50 snippets per pattern)
  - PoS tagging of all passages
  - actual extraction of the candidates for the role fillers using patterns
  - building a word-space model where rows correspond to the words provided by the organizers of the challenge and columns are the qualia elements for a selected role (clustering using CLUTO toolkit)

Results

clustering	entropy	purity
2-way	0.59	0.80
3-way	**0.00**	**1.00**
6-way	0.13	0.89
2-way$_{>1}$	0.70	0.77
3-way$_{>1}$	0.14	0.96
6-way$_{>1}$	0.23	0.82

Table: Performance using *formal* role only

What are the most representative elements in the clusters?

The similarity between elements in a cluster is measured as follows:

$$z_l = \frac{s_j^l - \mu_l^l}{\delta_l^l} \qquad (1)$$

s_j^l stands for the average similarity between the object j and the rest objects in the same cluster, μ_l^l is the average of s_j^l values over all objects in the lth cluster, and δ_l^l is the standard deviation of the similarities.

What are the most representative elements in the clusters?

- the core of the cluster respresenting tools is formed by *chisel* followed by *knife* and *scissors* as they have the largest internal z-score (the same cluster wrongly contains *rocket* but according to the internal z-score, it is an outlier (with the lowest z-score in the cluster))
- *bowl*, *cup*, *bottle* and *kettle* all have the lowest internal z-scores in the cluster of vehicles. The core of the cluster is formed by a *truck* and *motorcycle*

Descriptive and discriminative features: 3-way clustering

Cl	Features
VEG	fruit (41.3%), vegetables (28.3%), crop (14.6%), food (3.4%), plant (2.5%)
ANI	animal (43.3%), bird (23.0%), story (6.6%), pet (3.5%), waterfowl (2.4%)
ART	tool (31.0%), vehicle (15.3%), weapon (5.4%), instrument (4.4%), container (3.9%)
VEG	fruit (21.0%), vegetables (14.3%), animal (11.6%), crop (7.4%), tool (2.5%)
ANI	animal (22.1%), bird (11.7%), tool (10.1%), fruit (7.4%), vegetables (5.1%)
ART	tool (15.8%), animal (14.8%), bird (7.9%), vehicle (7.8%), fruit (6.8%)

Results: telic role

seed	extractions
helicopter	to rescue
rocket	to propel
chisel	to cut, to chop, to clean
hammer	to hit
kettle	to boil, to prepare
bowl	to serve
pencil	to draw, to create
spoon	to serve
bottle	to store, to pack

Table: Some extractions for the *telic* role

Results: constitutive role

seed	extractions
helicopter	a section, a body
rocket	a section, a part, a body
motorcycle	a frame, a part, a structure
truck	a frame, a segment, a program, a compartment
telephone	a tranceiver, a handset, a station
kettle	a pool, a cylinder
bowl	a corpus, a piece
pen	an ink, a component
spoon	a surface, a part
chisel	a blade
hammer	a handle, a head
bottle	a container, a component, a wall, a segment, a piece

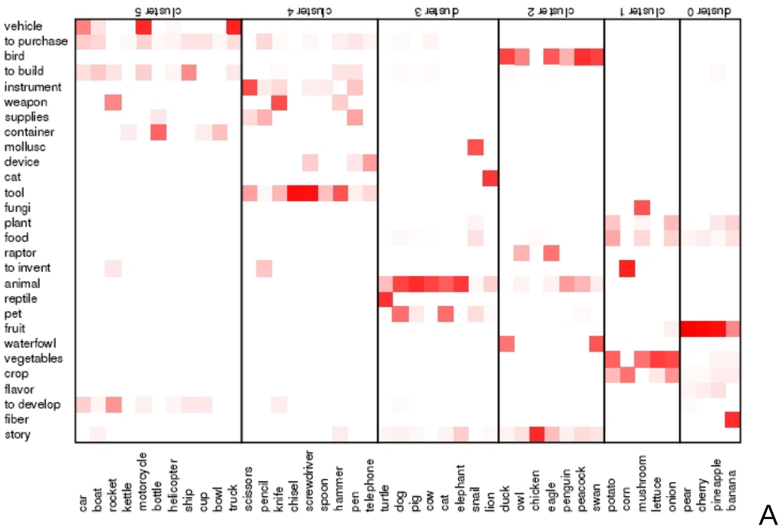
Table: Some extractions for the *constitutive* role

Results per role

role	clustering	entropy	purity	comments
formal	6-way	0.13	0.89	all 44 words
agentive	6-way	0.54	0.61	43 words
constitutive	6-way	0.51	0.61	28 words

Table: Performance using one role only

Results: formal and agentive roles combined



The best performance

The best results are obtained by combining formal role with the agentive one

clustering	entropy	purity
2-way	0.59	0.80
3-way	0.00	1.00
6-way	0.09	0.91

Table: Performance using *formal* and *agentive* roles

Interestingly, the worst performance on 2-way clustering is achieved by combining formal and constitutive roles (entropy of 0.92, purity of 0.66)

Error analysis

1. Errors due to the extraction procedure
   - incorrect PoS tagging/sentence boundary detection
   - patterns do not always provide correct extractions/features ("chicken and other stories")
2. Ambiguous words ("in fact, scottish gardens are starting to see many more butterflies including peacocks")
3. Features that do not suffice to discriminate among all categories

Error analysis (cont'd)

1. 6-way clustering always fails to discriminate between tools and vehicles well. Containers (a bowl, a kettle, a cup, a bottle) are always placed in the cluster of vehicles (instead of tools). This is the only type of errors for the 6-way clustering.
2. In 2-way clustering, vegetables are usually not considered natural objects

Conclusions

1. formal role is already sufficient for identification of vegetables, animals and artifacts (perfect clustering)
2. a combination of formal and agentive roles provides the best performance on 6-way clustering (in line with Pustejovsky, 2001)
3. no combination of roles accounts well for natural objects and artifacts

Possible directions?

- filtering out the false extractions (by hand, ranking) and re-clustering
- changing a set of patterns (ideally, learning patterns)

Possible directions?

- filtering out the false extractions (by hand, ranking) and re-clustering
- changing a set of patterns (ideally, learning patterns)

Thanks!