2012.11.07

Constructing a hypothesis space from the Web for large-scale Bayesian word learning

Joshua T. Abbott (joshua.abbott@berkeley.edu)
Joseph L. Austerweil (joseph.austerweil@gmail.com)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Word Net

Xu and Tenenbaum (2007)

김형준

Introduction

What is the name(concept) of this object?

Introduction

Generalization

Word Learning

Bayesian Word Learning <- Bayes' rule

Main Problem?

$$P(y \in C|x) = ??$$

Contents

1. Introduction

2. Basic Concepts

3. Xu and Tenenbaum(2007) using Hierarchical Clustering

4. Abbott, Austerweil and Griffiths(2012) using Word Net

5. Discussion

Contents

2. Basic Concepts

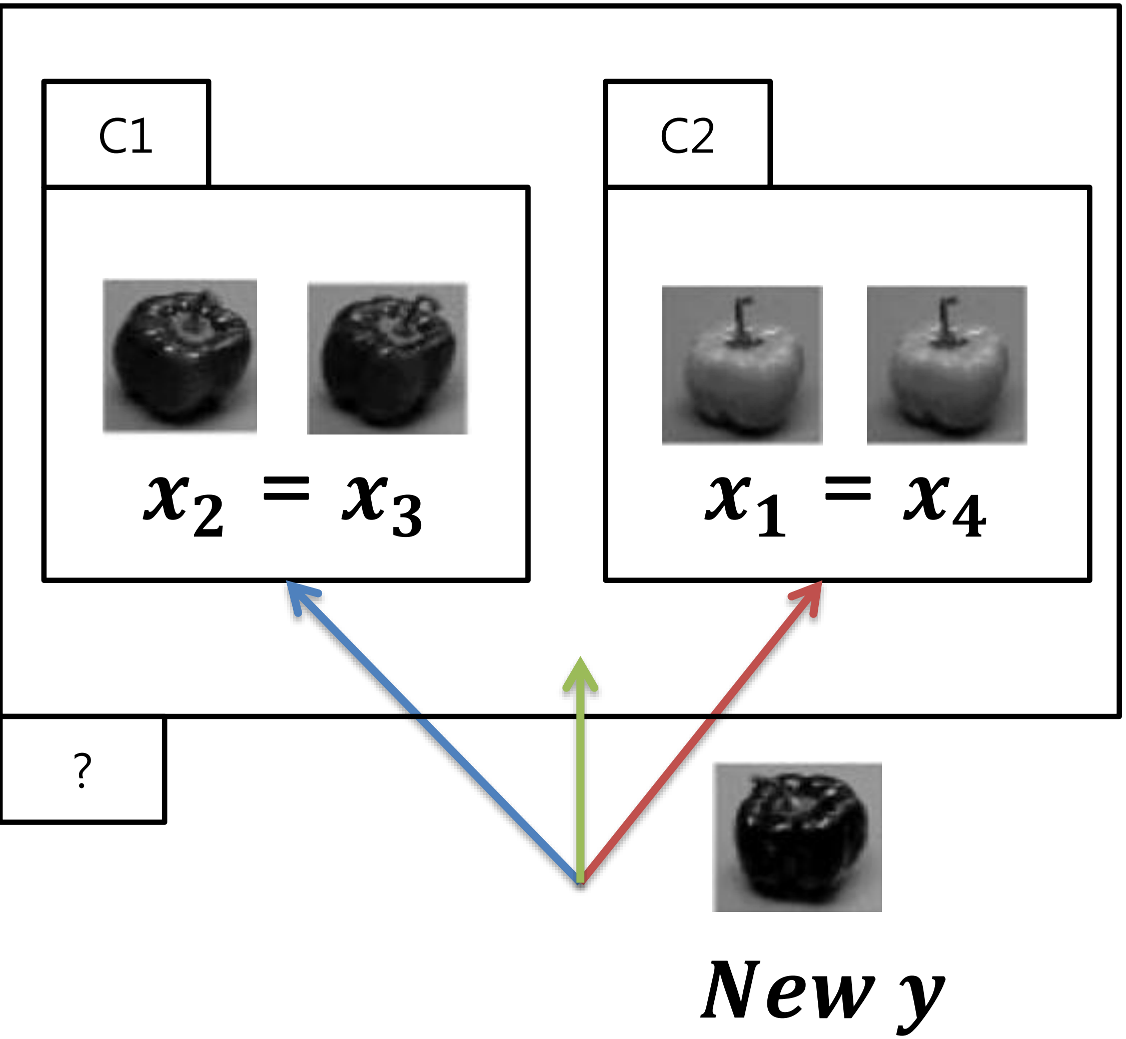
Basic Concepts

1) Bayesian Generalization Framework

- Bayesian Generalization Model

2) Hypothesis Space

Concepts



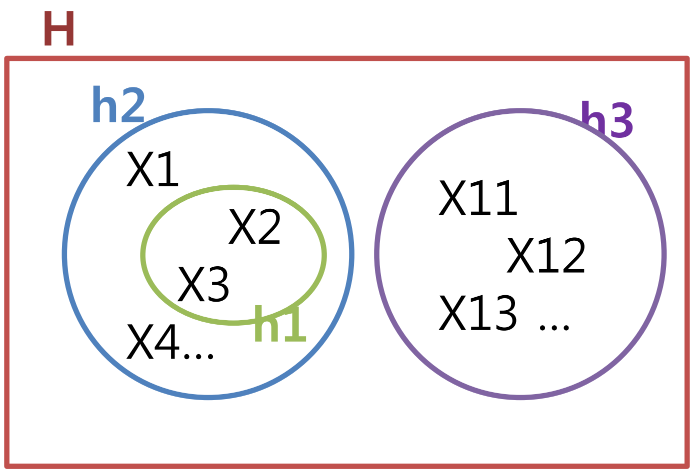
x = $\{x_1, \ldots, x_n\}$ of Concept C

p(y ∈ C|x) 를 계산하기 원함.
-> Using Hypothesis

Space H : A set of Hypothetical Concepts

The Bayesian Generalization Framework

x = $\{x_1, \ldots, x_n\}$ of Concept C



New Y

P(y ∈ C|x) 를 계산하기 원함.

-> Using Hypothesis

Space H : A set of Hypothetical Concepts

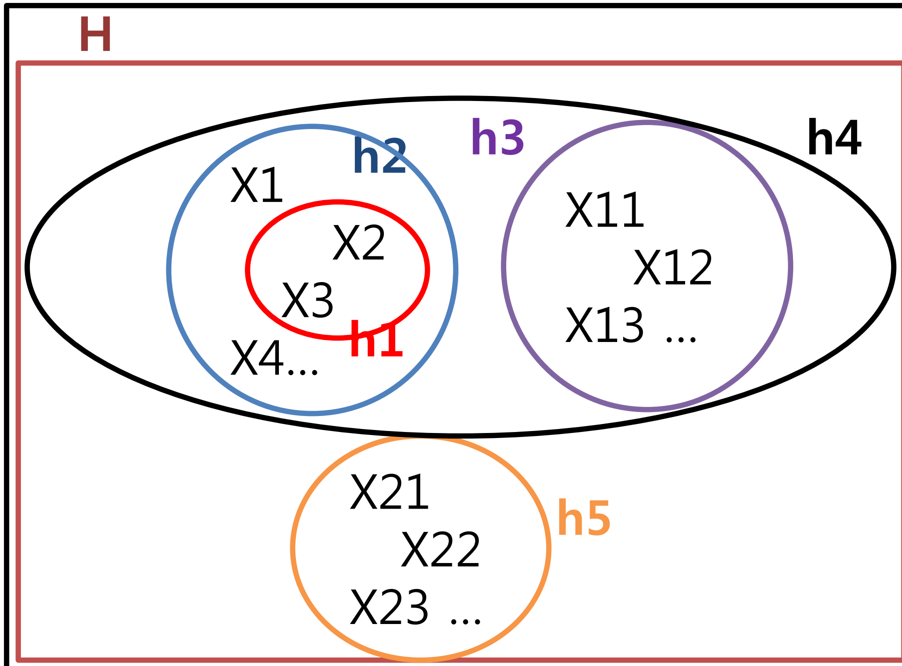
$$\mathbf{h_1} = \{x_2, x_3\} \text{ of } c_1$$

$$h_2 = \begin{cases} \{x_1, x_4\} \text{ of } c_2 \\ \{x_2, x_3\} \text{ of } c_1 \end{cases}$$

$$h_3 = \{x_{11}, x_{12}, x_{13}\} \text{ of } c_3$$

New Concept

P(y ∈ C|x) = ??



$h_1 = \{x_2, x_3\}$ of c_1

$h_2 = \begin{cases} \{x_1, x_4\} \text{ of } c_2 \\ \{x_2, x_3\} \text{ of } c_1 \end{cases}$

$h_3 = \{x_{11}, x_{12}, x_{13}\}$ of c_3

$h_4 = \begin{cases} \{x_1, x_4\} \text{ of } c_2 \\ \{x_2, x_3\} \text{ of } c_1 \\ \{x_{11}, x_{12}, x_{13}\} \text{ of } c_3 \end{cases}$

$x_2 = x_3$: Sub-Ordinate Level (빨간 피망 = 빨간 피망) : h_1

$x_2 = x_1$: Basic-Ordinate Level (빨간 피망 = 파란 피망) : h_2

$x_2 = x_{11}$: Super-Ordinate Level (빨간 피망 = 보라 가지) : h_4

The Bayesian Generalization Framework

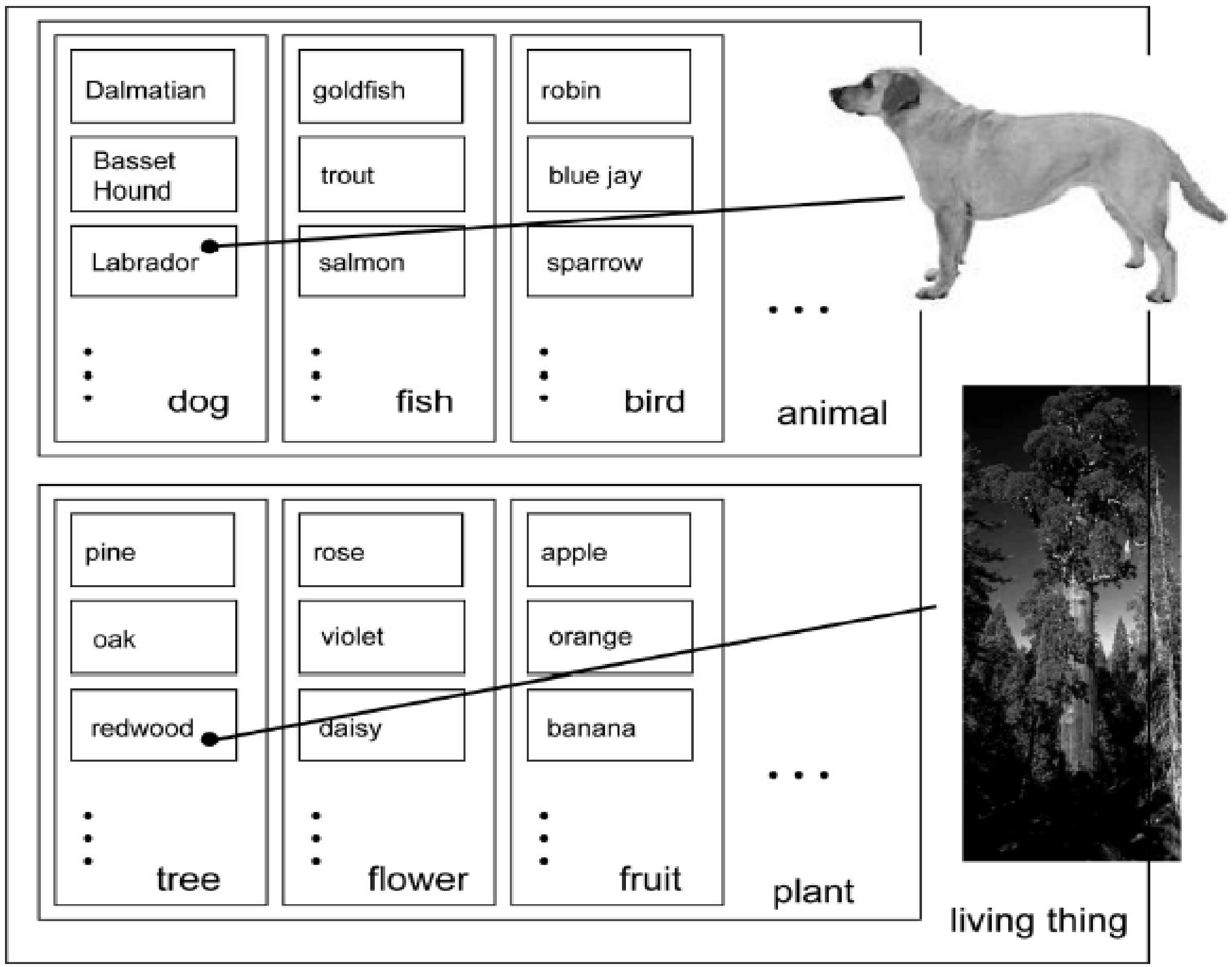


Figure 1. The extensions of words that label object-kind categories may overlap in a nested fashion, in accord with the tree-structured hierarchy of an object-kind taxonomy.

The Bayesian Generalization Framework

How to solve?

$$\mathbf{P(y \in C|x)}$$

Bayes rule

$$\underbrace{P(h|x)}_{\text{Posterior}} \propto \underbrace{p(x|h)}_{\text{Likelihood}}\underbrace{p(h)}_{\text{Prior}}$$

The Bayesian Generalization Framework

H

$h_1 = \{x_2, x_3\}$ of c_1

$h_2 = \begin{cases} \{x_1, x_4\} \text{ of } c_2 \\ \{x_2, x_3\} \text{ of } c_1 \end{cases}$

$h_3 = \{x_{11}, x_{12}, x_{13}\}$ of c_3

$h_4 = \begin{cases} \{x_1, x_4\} \text{ of } c_2 \\ \{x_2, x_3\} \text{ of } c_1 \\ \{x_{11}, x_{12}, x_{13}\} \text{ of } c_3 \end{cases}$

Bayes' Rule

$P(h|x) \propto p(x|h)p(h)$

→

$$P(y \in C|x) = \sum_{h \in H} P(y \in C|h)p(h|x)$$

where $P(y \in C) = 1$ if $y \in h$

and 0 otherwise

$$P(y \in C|x) = \sum_{h \ni y,x} p(h|x)$$

The Bayesian Generalization Framework

Prior

1) Uniform Distribution (Shepard, 1987)
2) A Stochastic Process over Structures(Kemp & Tenenbaum, 2009)
3) And so on.

The Bayesian Generalization Framework

Likelihood(Tenebaum and Griffiths(2001))

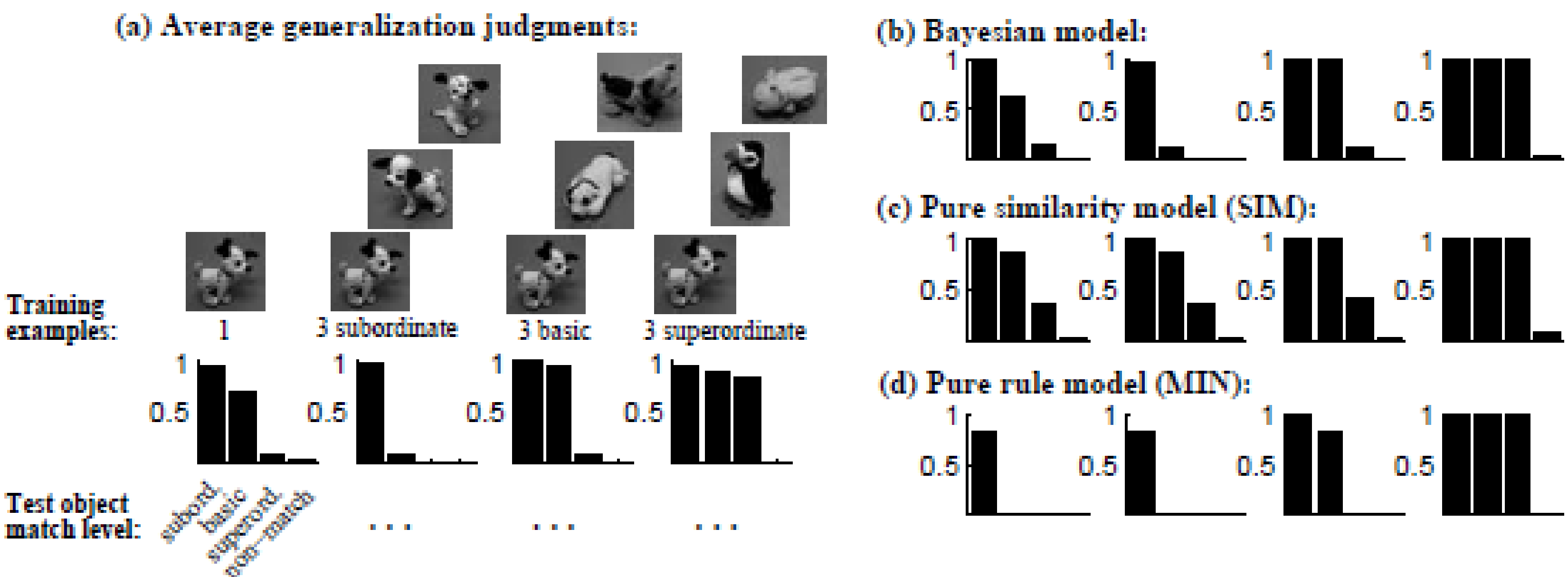


Figure 2: Data and model predictions for the word learning task.

각 Objects들은 각 가설로부터 랜덤적으로 Uniformly generated

The Bayesian Generalization Framework

Likelihood : Strong Sampling

$$P(x|h) = \begin{cases} \mathbf{1/|\boldsymbol{h}|^{\boldsymbol{n}}} & \boldsymbol{if\ x \subset h} \\ \mathbf{0} & \boldsymbol{otherwise} \end{cases}$$

Size Principle : If $|H| \uparrow$ then $L \downarrow$

If $|H| \downarrow$ then $L \uparrow$

Why Strong Sampling?

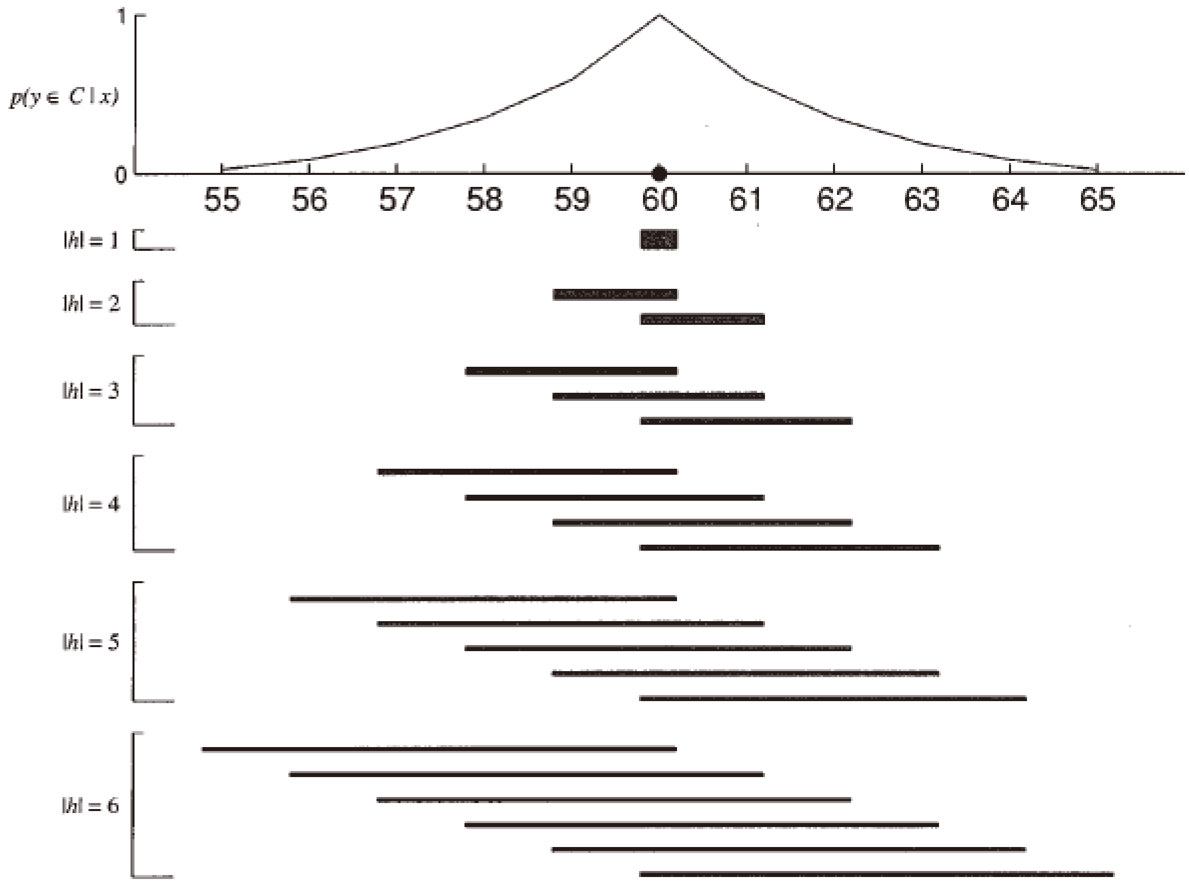


Figure 1. An illustration of the Bayesian approach to generalization from $x = 60$ in a one-dimensional psychological space (inspired by Shepard 1989, August). For the sake of simplicity, only intervals with integer-valued endpoints are shown. All hypotheses of a given size are grouped together in one bracket. The thickness (height) of the bar illustrating each hypothesis h represents $p(h|x)$, the learner's degree of belief that h is the true consequential region given the observation of x. The curve at the top of the figure illustrates the gradient of generalization obtained by integrating over just these consequential regions. The profile of generalization is always concave regardless of what values $p(h|x)$ takes on, as long as all hypotheses of the same size (in one bracket) take on the same probability.

Why Strong Sampling?

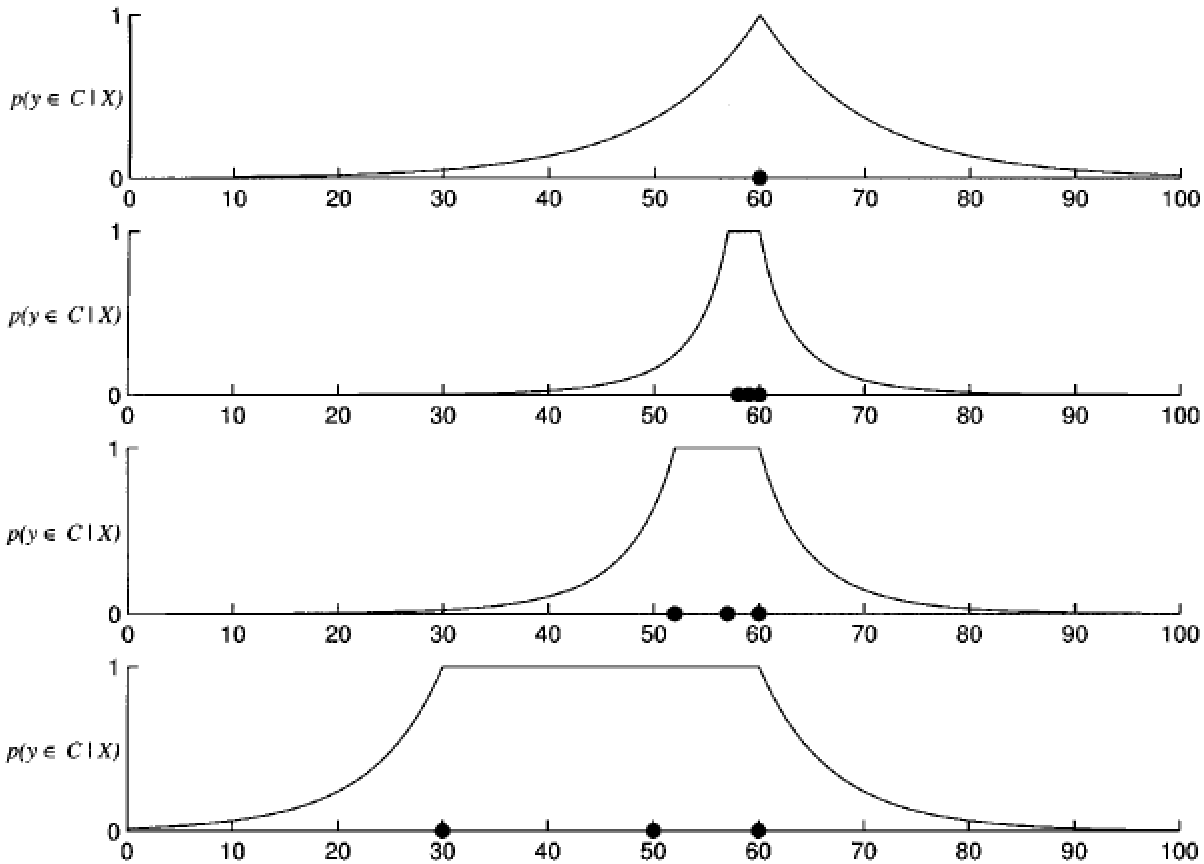


Figure 2. The effect of example variability on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization increases as a function of the range of examples.

Why Strong Sampling?

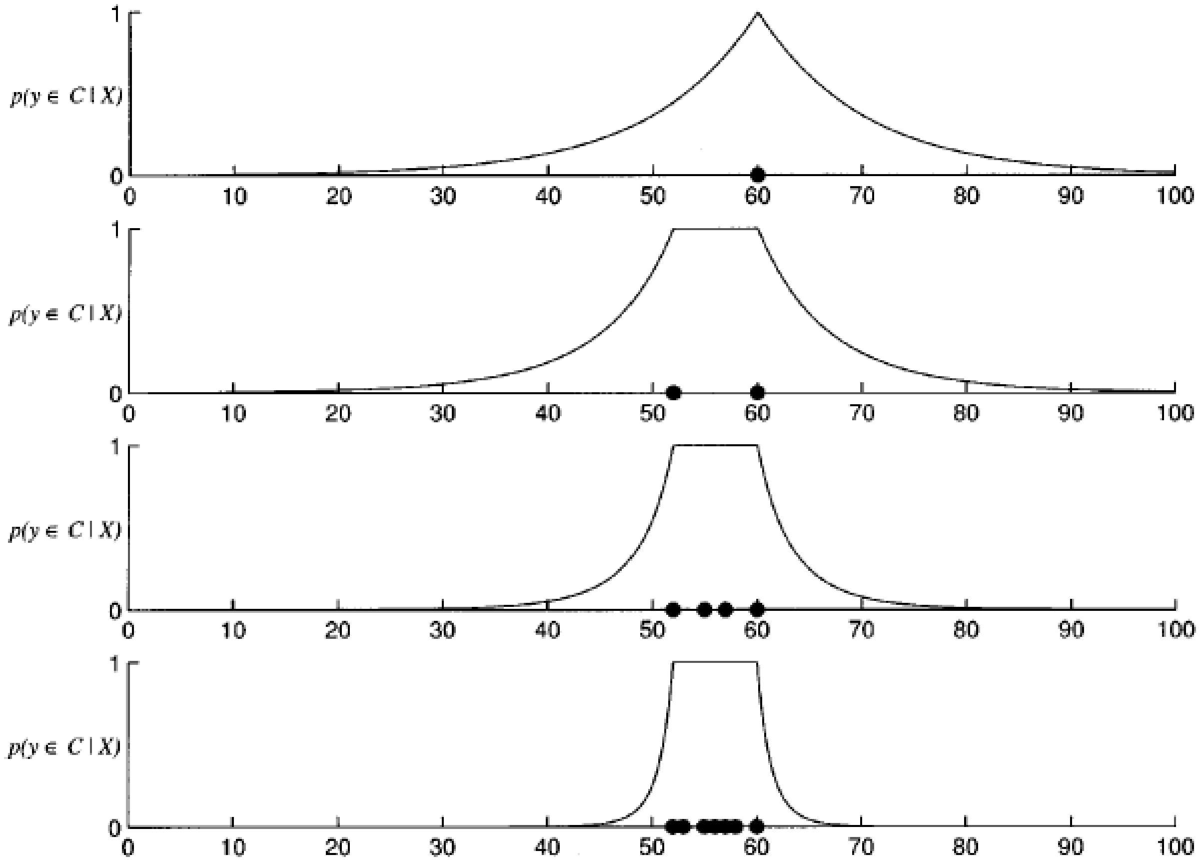
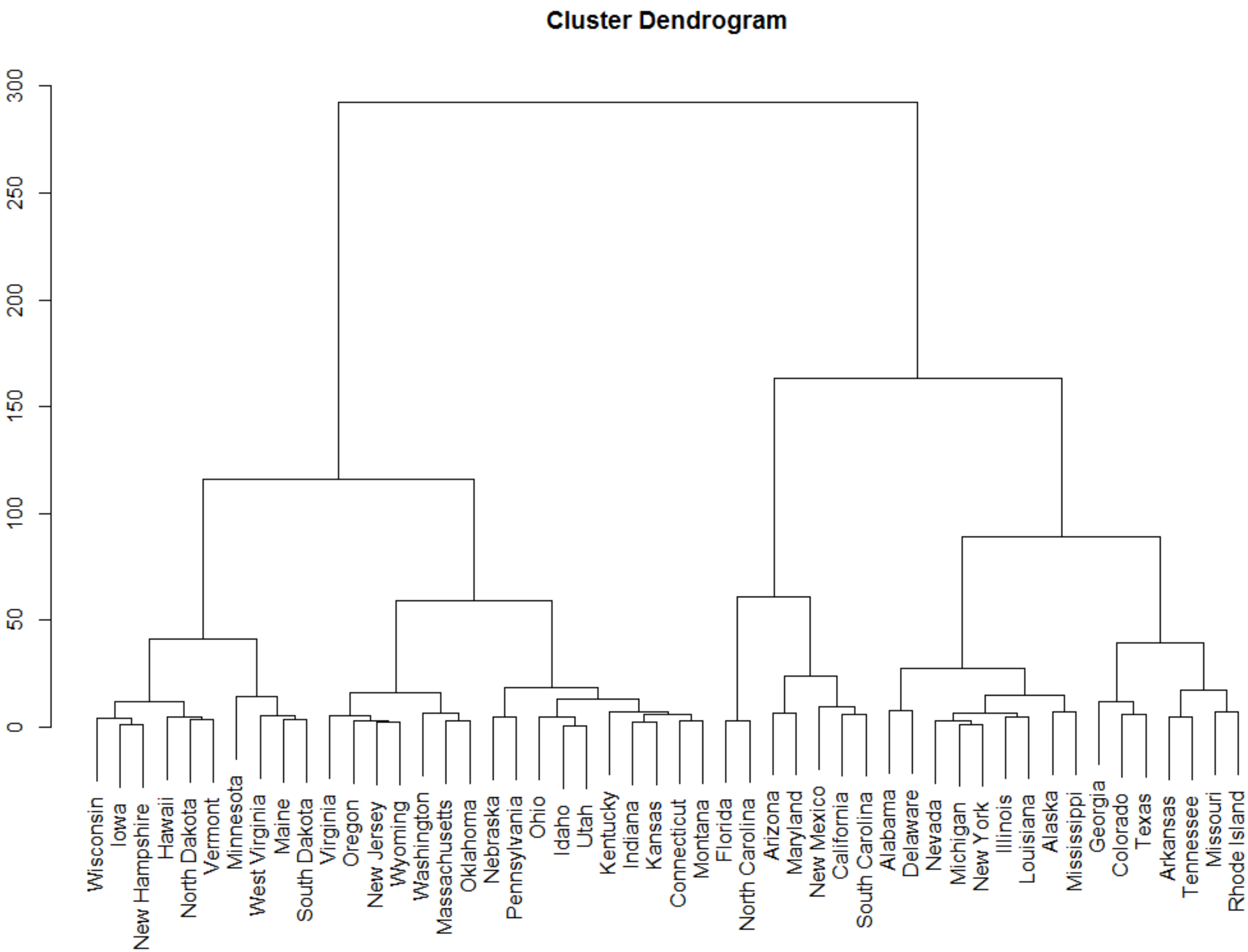


Figure 3. The effect of the number of examples on Bayesian generalization (under the assumptions of strong sampling and an Erlang prior, $\mu = 10$). Filled circles indicate examples. The first curve is the gradient of generalization with a single example, for the purpose of comparison. The remaining graphs show that the range of generalization decreases as a function of the number of examples.

Contents

3. Hierarchical Clustering
→ Bayesian Inference

EX) Hierarchical Clustering




```
        Murder Assault
Alabama   13.2     236
Alaska    10.0     263
```

```
           [,1]
 [1,]  0.600000
 [2,]  1.004988
 [3,]  1.414214
 [4,]  2.088061
 [5,]  2.332381
 [6,]  2.758623
 [7,]  2.879236
 [8,]  2.973214
 [9,]  3.001666
[10,]  3.124100
[11,]  3.310589
[12,]  3.448188
[13,]  4.031129
[14,]  4.472136
[15,]  4.609772
[16,]  4.700000
[17,]  4.833218
[18,]  5.000000
[19,]  5.281098
[20,]  5.348832
[21,]  5.660389
[22,]  6.000000
[23,]  6.177378
[24,]  6.539113
[25,]  6.800000
[26,]  6.847627
[27,]  6.881860
[28,]  7.049113
[29,]  7.294518
[30,]  7.569016
[31,]   9.314505
[32,]  11.800424
[33,]  12.070211
[34,]  13.092364
[35,]  14.043148
[36,]  15.005332
[37,]  16.243152
[38,]  17.089178
[39,]  18.248288
[40,]  24.109956
[41,]  27.188968
[42,]  39.560081
[43,]  41.109610
[44,]  59.052942
[45,]  61.131007
[46,]  89.244384
[47,] 116.155069
[48,] 163.282455
[49,] 292.254752
```

Word Learning as Bayesian Inference

Xu and Tenenbaum(2007)
Hierarchical Clustering 이용

- **15 images per concept**
- **Pair of 2 objects**
  403회 유사 판단

(1 ~ 9) Scale: 45분 소요..

Nodes <- Potential Words(Hypotheses)[Alphabet]
Leaves <- the Domain of Possible Objects[Number]

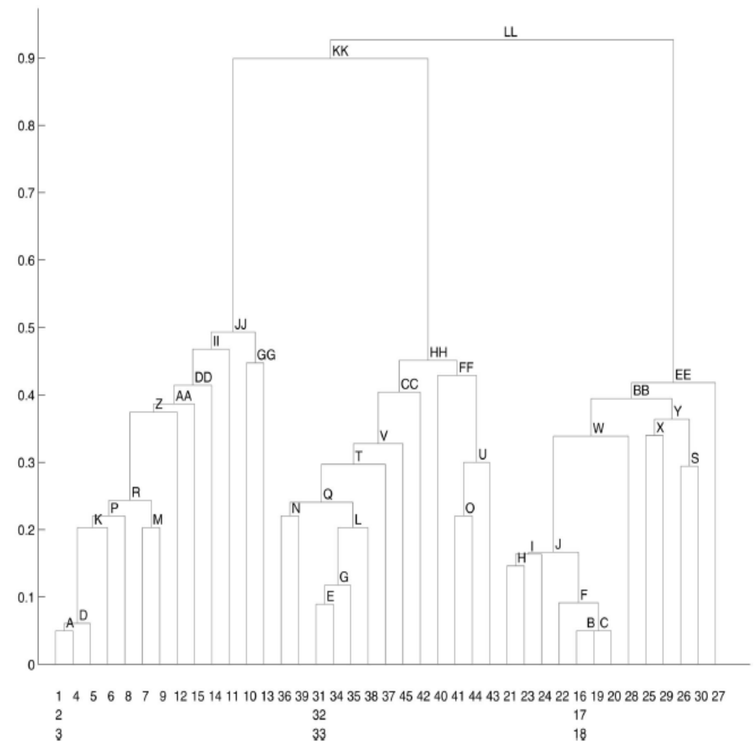
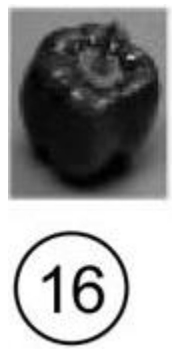


Figure 7. Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for Bayesian word learning. Letter codes refer to specific clusters (hypotheses for word meaning): vegetable (EE), vehicle (HH), animal (JJ), pepper (J), truck (T), dog (R), green pepper (F), yellow truck (G), and Dalmatian (D). The clusters labeled by other letter codes are given in the text as needed. Numbers indicate the objects located at each leaf node of the hierarchy, keyed to the object numbers shown in Figures 3 and 4. The height of a cluster, as given by the vertical axis on the left, represents the average within-cluster dissimiliarity of objects within that cluster.

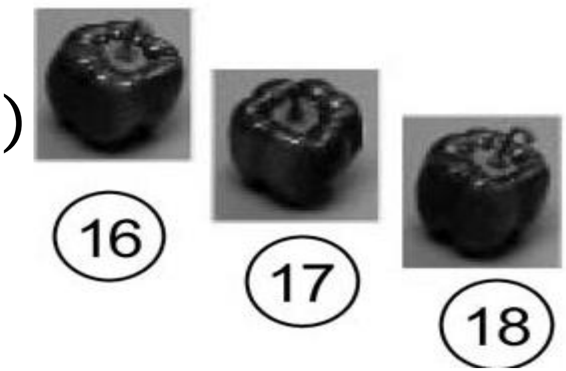
The Bayesian Generalization Framework

Con 1)



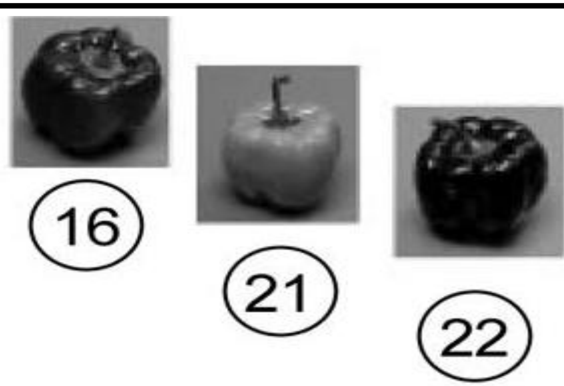
$$h*_1 > h*_2 > h*_4$$

Con 2)



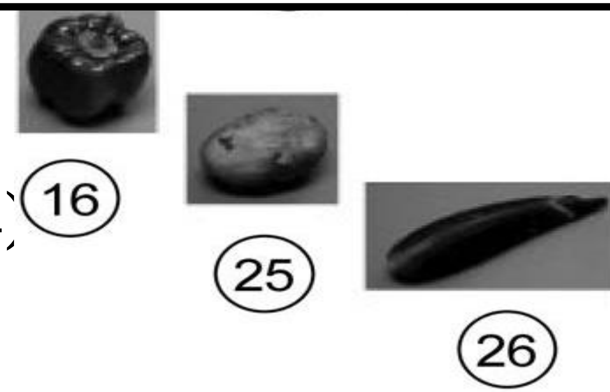
$$h*_1 \gg h*_2 \gg h*_4$$

Con 3)



$$h*_1 = h*_2 > h*_4$$

Con 4)



$$h*_1 = h*_2 = h*_4$$

Word Learning as Bayesian Inference

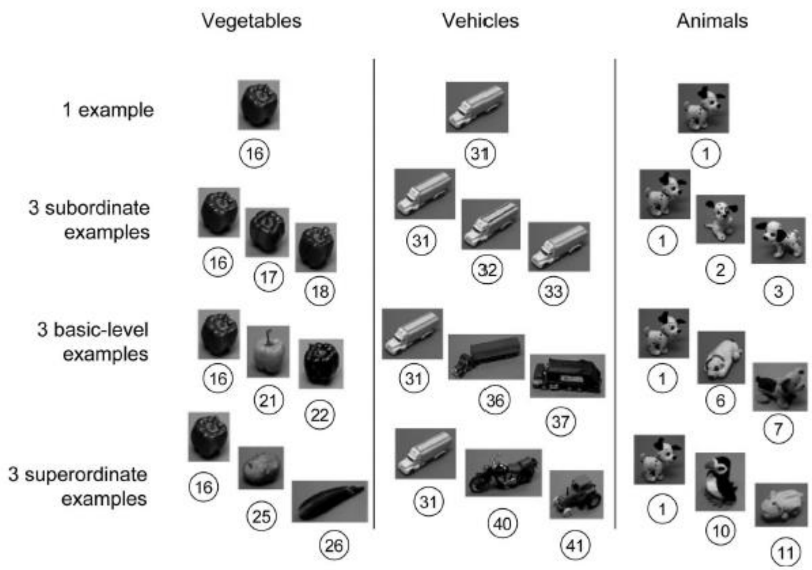


Figure 3. Twelve training sets of labeled objects used in Experiment 1, drawn from all three domains (animals, vegetables, and vehicles) and all four test conditions (one example, three subordinate examples, three basic-level examples, and three superordinate examples). The circled number underneath each object is used to index that object's location in the hierarchical clustering shown in Figure 7.

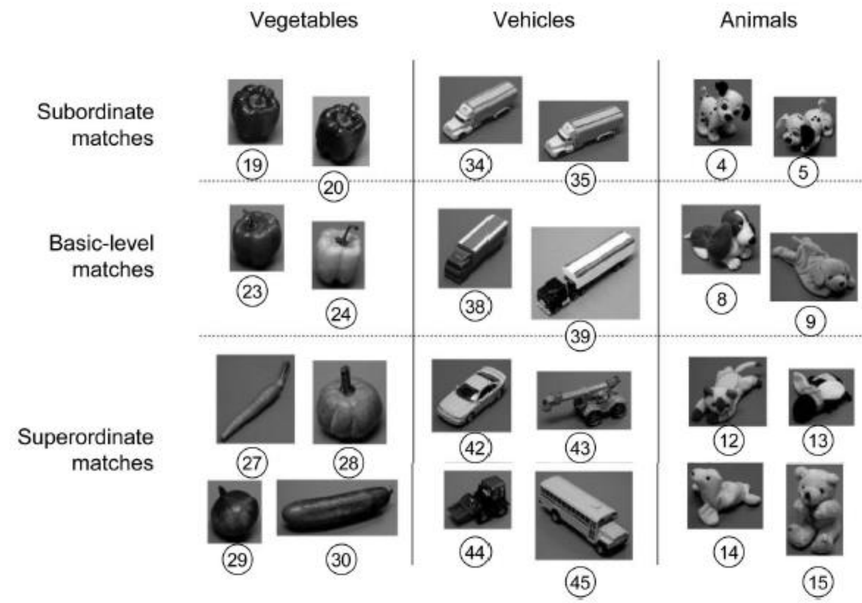


Figure 4. The test set of 24 objects used to probe generalization of word meanings in Experiment 1. For each training set in Figure 3, this test set contains two subordinate matches, two basic-level matches, and four superordinate matches. The circled number underneath each object is used to index that object's location in the hierarchical clustering shown in Figure 7.

Word Learning as Bayesian Inference

Design and Procedure

Training : 12 Trials (Domains*3, Levels*4)

Each Trials ->

Each Test(sub*2 + Basic*2 + Super*4)

Word Learning as Bayesian Inference

Nodes <- Potential Words(Hypotheses)[Alphabet]
Leaves <- the Domain of Possible Objects[Number]

Height : Minimal Distance form the node to a leaf

Prior P(h) $\propto$ height(parent(h)) – height(h)

차이가 클 수록 사전 믿음이 높음!

Likelihood p(x|h) $\propto$ $\left[\dfrac{1}{height(h)+\varepsilon}\right]^{n}$

.05

WORD LEARNING AS BAYESIAN INFERENCE 261

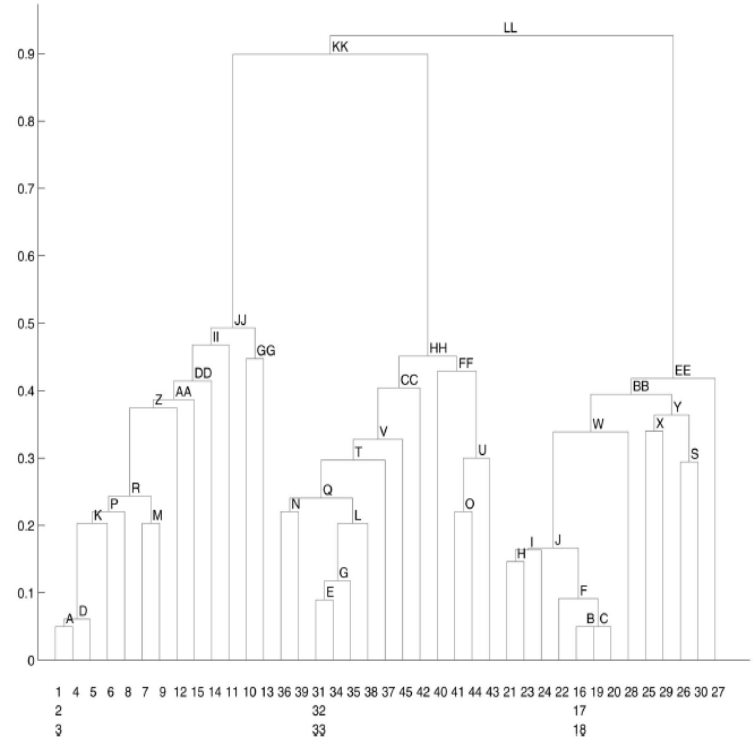
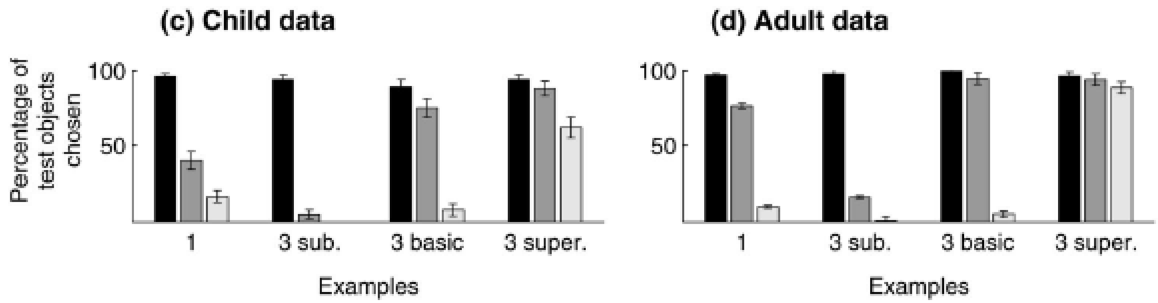
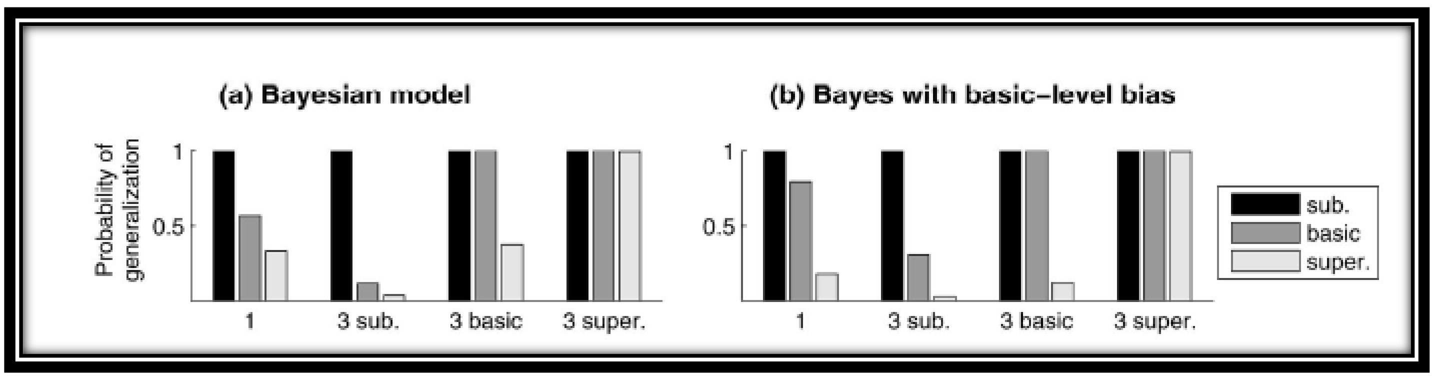
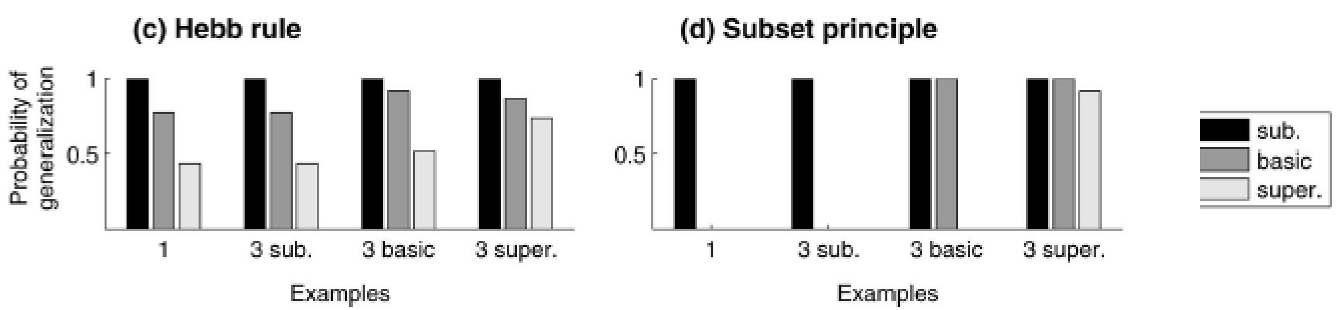


Figure 7. Hierarchical clustering of similarity judgments yields a taxonomic hypothesis space for Bayesian word learning. Letter codes refer to specific clusters (hypotheses for word meaning): vegetable (EE), vehicle (HH), animal (JJ), pepper (J), truck (T), dog (R), green pepper (F), yellow truck (G), and Dalmatian (D). The clusters labeled by other letter codes are given in the text as needed. Numbers indicate the objects located at each leaf node of the hierarchy, keyed to the object numbers shown in Figures 3 and 4. The height of a cluster, as given by the vertical axis on the left, represents the average within-cluster dissimiliarity of objects within that cluster.

Word Learning as Bayesian Inference





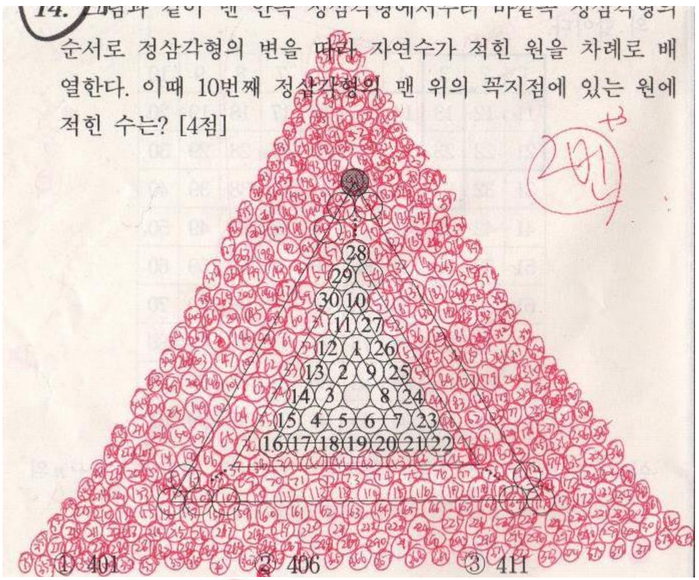


Contents

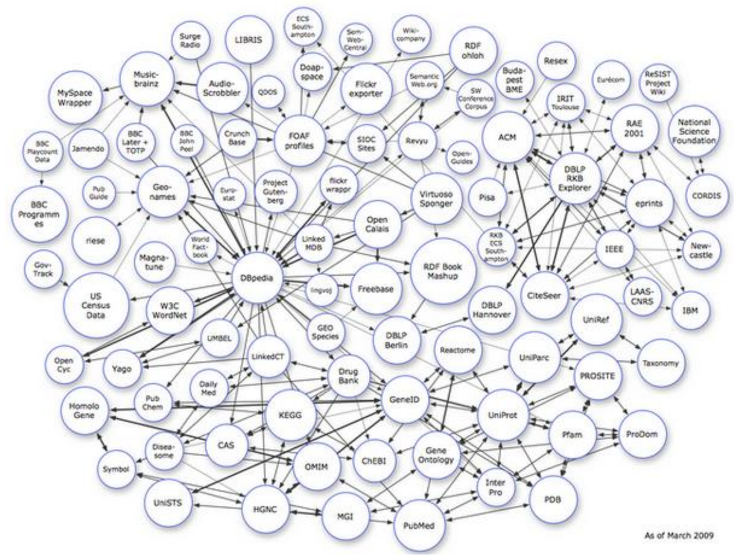
4. Word-Net

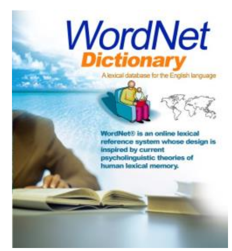
→ Bayesian Inference

Word Learning as Bayesian Inference



VS





Large-Scale Word Learning

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Large-Scale Word Learning

1) 82,115 noun nodes를 추출하여 Tree를 구성

Nodes == Hypothesis.

2) Binary Matrix, H 구성.

Rows == Objects (64,958)

Columns == Hypotheses (82,115 = 17,157 + 64,958)

Inner nodes + $\boxed{\text{Leaf nodes}}$

H_{ij} (i = leaf node, j = Hypothesis node) =

$$\begin{cases} 1 & i \subset j \\ 0 & \text{otherwise} \end{cases}$$

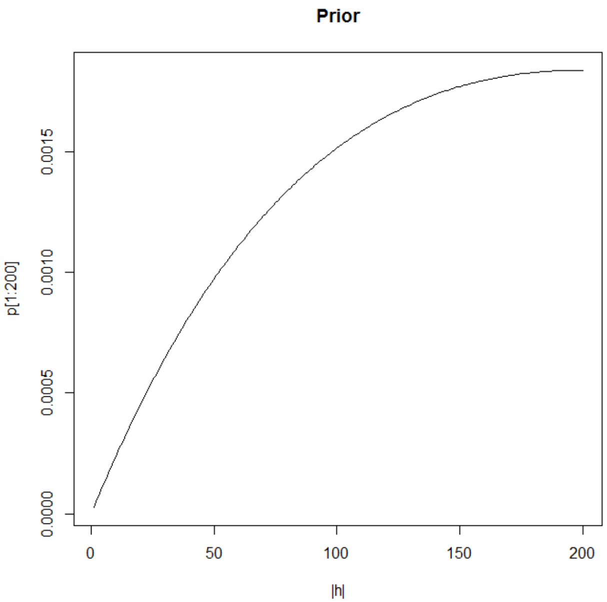
Subordinate Objects 구별 가능

Large-Scale Word Learning

1) Bayesian Model

Prior : $P(h) \propto (|h|/\sigma^2)\exp\{-|h|/\sigma\}$

σ = 200 by hand



Likelihood : P(x|h) $= \begin{cases} 1/|h|^n & \text{if } x \subset h \\ 0 & \text{otherwise} \end{cases}$

$$\text{Bscore}(y) = P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x})$$

Erlang Distribution
= Distribution of sum of exponential variates

Prior

Shepard (1987)

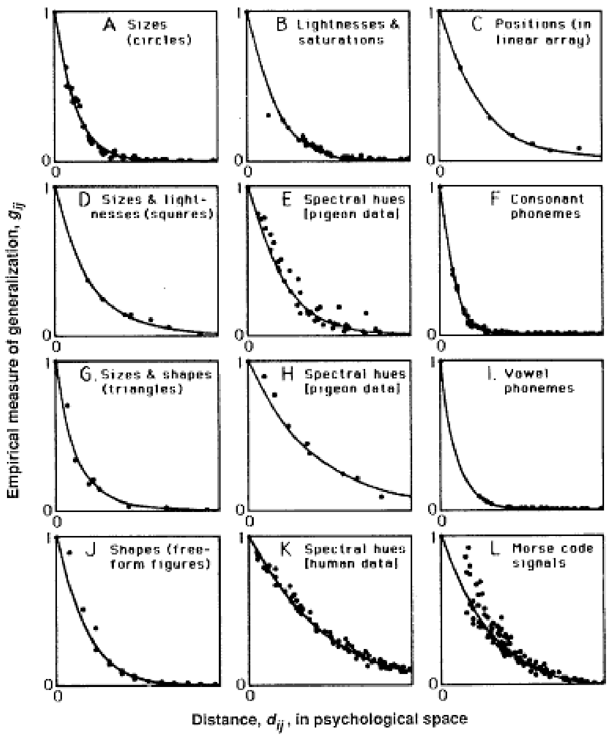
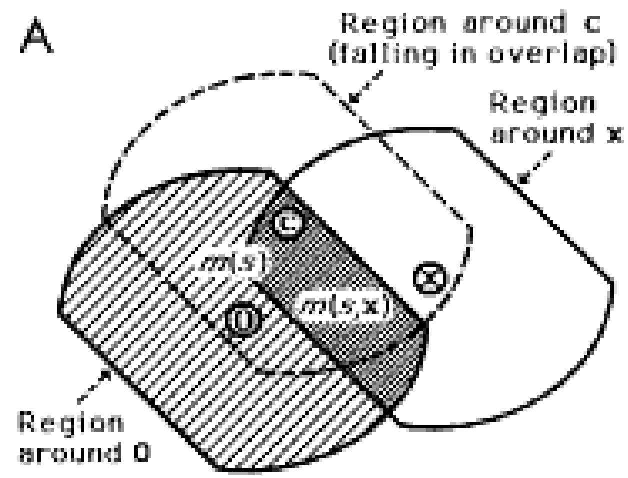


Fig. 1. Twelve gradients of generalization. Measures of generalization between stimuli are plotted against distances between corresponding points in the psychological space that renders the relation most nearly monotonic. Sources of the generalization data (g) and the distances (d) are as follows. (**A**) g, McGuire (*33*); d, Shepard (*7, 18*). (**B**) g, Shepard (*7, 17*); d, Shepard (*7, 18*). (**C**) g, Shepard (*17*); d, Shepard (*8*). (**D**) g, Attneave (*25*); d, Shepard (*8*). (**E**) g, Guttman and Kalish (*4*); d, Shepard (*11*). (**F**) g, Miller and Nicely (*34*); d, Shepard (*35*). (**G**) g, Attneave (*25*); d, Shepard (*8*). (**H**) g, Blough (*36*); d, Shepard (*11*). (**I**) g, Peterson and Barney (*37*); d, Shepard (*35*). (**J**) g and d, Shepard and Cermak (*38*). (**K**) g, Ekman (*39*); d, Shepard (*18*). (**L**) g, Rothkopf (*40*); d, Cunningham and Shepard (*41*). The generalization data in the bottom row are of a somewhat different type. [See (*39*) and the section "Limitations and Proposed Extensions."]

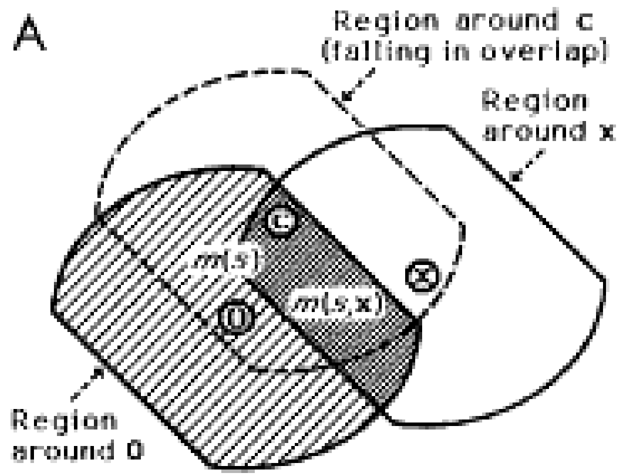


S : Particular size s

Why Erlang?

Prior

Shepard (1987)



$$g(\mathbf{x}) = \int_0^\infty p(s)\,\frac{m(s,\mathbf{x})}{m(s)}\,ds$$

Monte Carlo Stimulation

Prior

Shepard (1987)

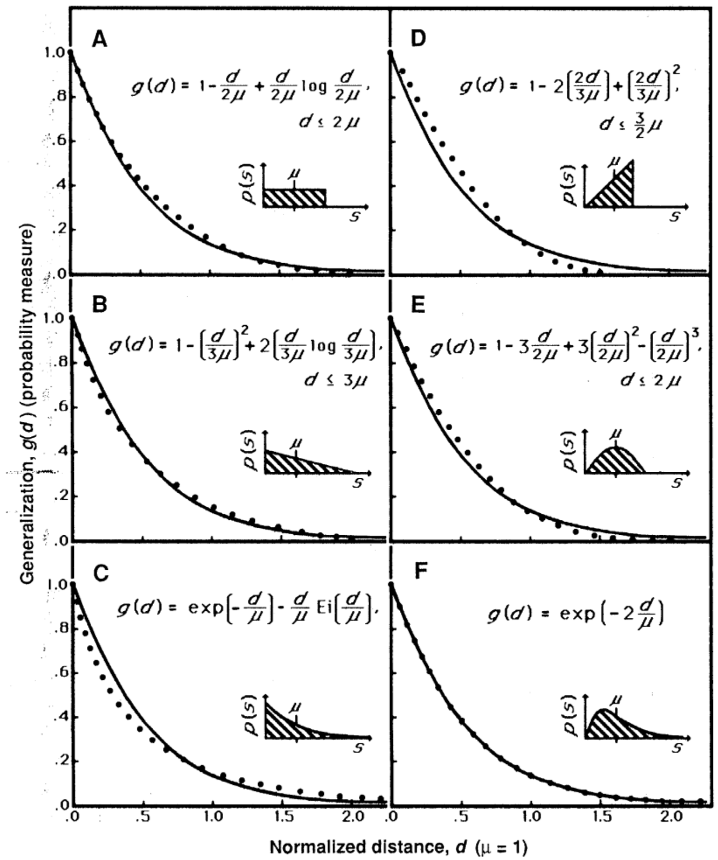


Fig. 3. Six generalization functions, $g(d)$, relating probability of generalization to normalized distance in psychological space, derived by substituting into Eq. 6 the functions $p(s)$ shown in the shaded insets, and integrating (dotted curve); and the corresponding simple exponential decay function (smooth curve). In (C), the function Ei is defined as follows

Minkowski Power Metric Formula

$$d_{ij} = \left(\sum_{k=1}^{K} |x_{ik} - x_{jk}|^r \right)^{1/r}$$

K- dimensional space

i – i번째 자극

j – j번째 자극

공간에서 거리개념이 다를 수 있음.

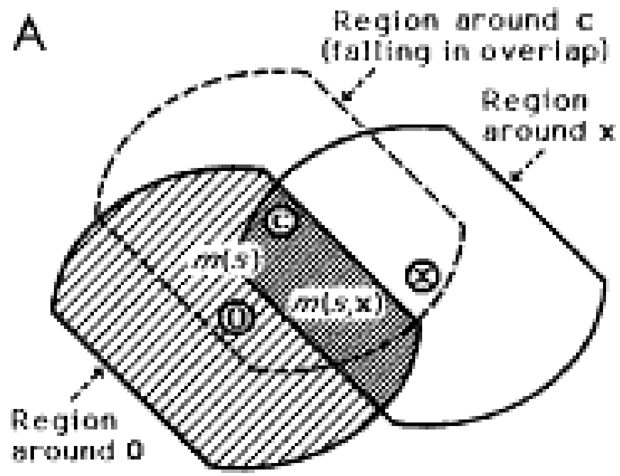
r- 1 Rhombic

r- 2 Circular

수학적 증명

Prior

Shepard
(1987)



$$p(s) = C \cdot m(s) \cdot q(s)$$

$$\int_0^\infty p(s)\, ds = 1 \qquad (3)$$

$$\int_0^\infty s \cdot p(s)\, ds = \mu < \infty \qquad (4)$$

q(s) = Before encountering the first stimulus, It's PDF

-> revise p(s) ~ Erlang PDF $p(s) = \left(\frac{2}{\mu}\right)^2 s \cdot \exp\left(-\frac{2}{\mu} s\right) \qquad (9)$

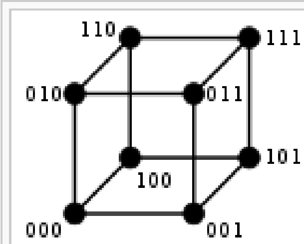
Large-Scale Word Learning

2) Prototype Model

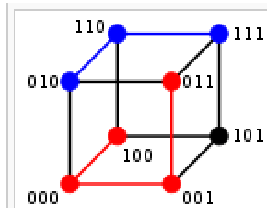
1] Define the Prototype of a set of objects, X_{proto}
to have those features owned by a majority of the objects in the set

$$\text{Pscore}(y) = \exp\{-\lambda_p \text{ dist}(y, x_{\text{proto}})\}$$

Dist = Hamming Distance between the two vectors
Λ = free Parameter , here 0.15 by hand using half-interval search
Pscore was normalized over all objects y



3-bit binary cube for finding Hamming distance



Two example distances: 100->011 has distance 3 (red path); 010->111 has distance 2 (blue path)

Half – Interval Search

Binary Search or Half-Interval Search

Ex) Telephone Book

Search for Smith

Rogers .... Thomas
Samson .... Thomas

Large-Scale Word Learning

3) Exemplar Model

1] Define the Exemplar model using similar scoring metric

Compute a distance for each item X_j

$$\text{Escore}(y) = \sum_{x_j \in \mathbf{x}} \exp\{-\lambda_e \, \text{dist}(y, x_j)\}, \qquad (8)$$

Dist = Hamming Distance between the two vectors

Λ = free Parameter , here 0.20 by hand

using half-interval search

Escore was normalized over all objects y

Behavioral Experiments

Exp 1. Validating Out Approach – Xu and Tenenbaum(2007)

Within-Subjects Design

$0.05 for each training set, 34 participants

1) Training

4 Conditions : 각각 3 분류(Vegetables, Vehicles, Animals)

4 Levels : 총 12회 무선.

Behavioral Experiments

Test Sets were the same

8 Objects : two subordinate examples(e.g two other Dalmatians)
two basic-level examples (e.g a Cocker Spaniel and a Corgi)
four superordinate examples (e.g a cat, a bear, a sea lion, and a horse).
+ 16 non-matching objects

각 Trial 마다 다른 새로운 단어 (e.g “dak”)의 one or more examples 을 보여줌.-> test set에 있는 물체로 부터 dak라는 것들을 선택.

Behavioral Experiments

Exp 1. Validating Out Approach

Result

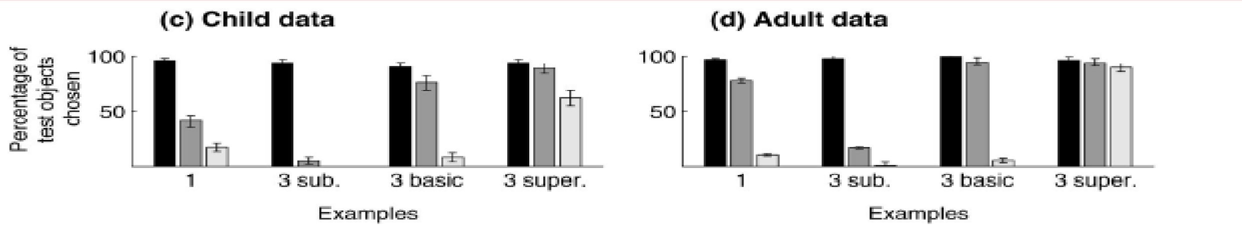


Figure 8. Predictions of the Bayesian model, both with and without a basic-level bias, compared with the data from adults in Experiment 1 and those from children in Experiment 3. Sub. = subordinate; super. = superordinate.

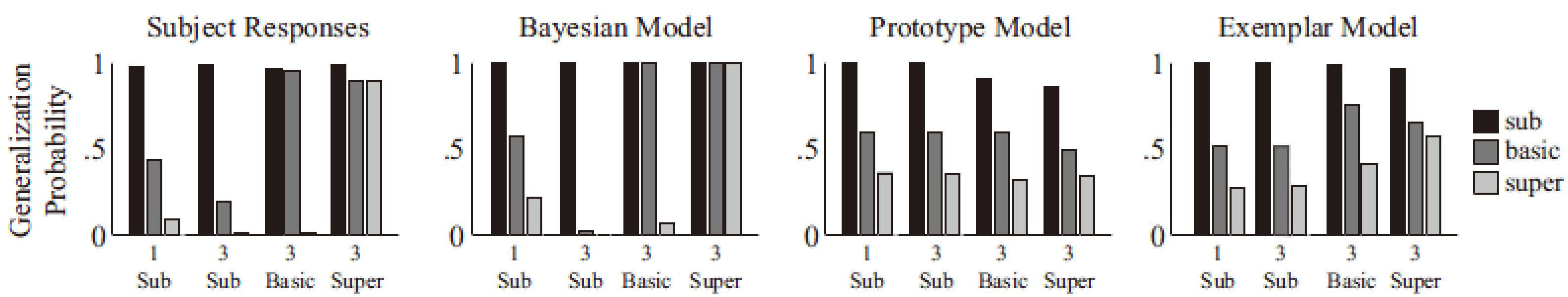


Figure 1: Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the three domains in Experiment 1. The generalizations for non-matching items are omitted for brevity (neither the participants chose nor the Bayesian model predicted non-matching objects, while the prototype and exemplar models predicted non-matches less than 4% of the time for each condition).

$\boxed{r^2 = .98}$ $\boxed{r^2 = .66}$ $\boxed{r^2 = .84}$

Behavioral Experiments

Exp2) Novel Domains

	Clothing		Containers		Seats	
Object level	1	2	1	2	1	2
Subordinate						
Basic						
Superordinate						

Table 1: Training images for Experiment 2.

36 Participants, $0.05 for each trial

Behavioral Experiments

Exp2) Novel Domains

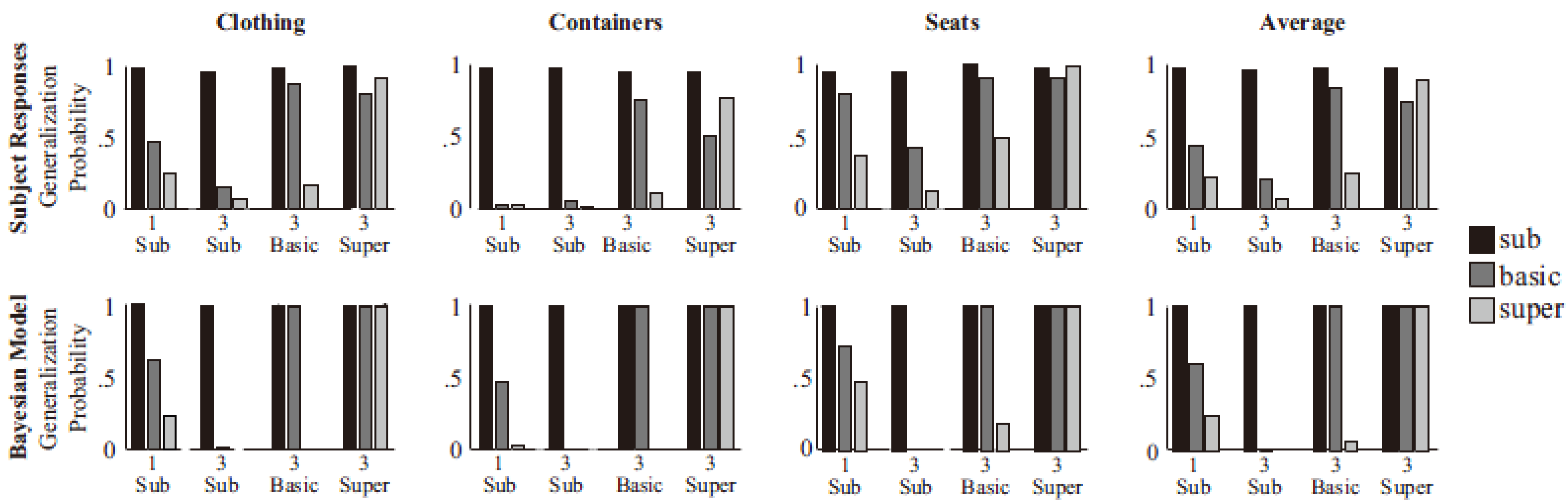


Figure 2: Participant generalization judgments and the predictions of the Bayesian model for Experiment 2. From left to right, the columns present the results for the three taxonomies (clothing, containers, and seats) and average results.

$r^2 = .97$

$r^2 = .88$

$r^2 = .91$

$r^2 = .95$

$Pr^2 = .80$

$Er^2 = .90$

Contents

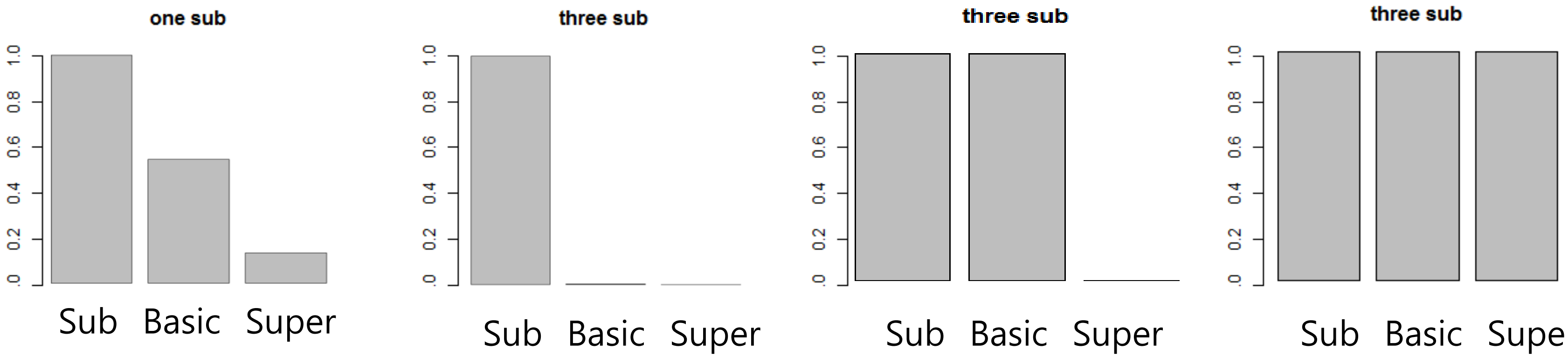
5. Discussion

Discussion

**** 의의**
1) Hypothesis Space 손 -> 기계
2) Bayesian Model의 재검증
3) New Domain 적용

**** 미래 연구**
1) One Subordinate-level and One Basic-level Object
- Heterogeneous Training Sets
2) Training과 Test 자극의 시각적 유사성에 따른 상호작용 효과
[BWLM로는 불가]
3) 상식 추론에도 적용가능 from ConcepNet of OpenCyc

Simulation



1) Dump truck : |h| = 1

2) Truck : |h| = 24

3) Motor Vehicle : |h| = 78

References

J.T. Abbott, J.L. Austerweil, and T.L. Griffiths. "Constructing a hypothesis space from the web for large-scale Bayesian word learning". *Proceeddings of the 34th Annual Conference of the Cognitive Science Society*, 2012

R.N. Shepard "Toward a Universal Law of Generalization for Psychological Science", *Science,* Vol. 237, No.4820. 1987. pp.1317-1323

J.B. Tenenbaum "Rules and Similarity in Concept Leaning". *Advances in Neural Information Processing Systems 12*. 2000. pp59-65

J.B. Tenenbaum and T. L. Griffiths. "Generalization, Similarity, and Bayesian inference". *Behavioral and Brain Sciences 24*. 2001. pp 629-640

F. Xu and J.B. Tenenbaum. "Word Learning as Bayesian Inference". *Psychological Review.* Vol. 114. No.2. 2007. pp 245-272