Online Entropy-based Model of Lexical Category Acquisition

Grzegorz Chrupała  Afra Alishahi

Spoken Language Systems
and
Department of Computational Linguistics
Saarland University

CoNLL 2010
1. Lexical category acquisition in humans

2. Online information-theoretic model

3. Task-based evaluation
Outline

1. Lexical category acquisition in humans
2. Online information-theoretic model
3. Task-based evaluation

Chrupala and Alishahi (UdS)
Human category acquisition

- Humans incrementally learn lexical categories from exposure to language
  - Children form robust lexical categories early on
    [Gelman and Taylor, 1984, Kemp et al., 2005]

- Distributional properties of words provide cues about its category
  - Children are sensitive to co-occurrence statistics
    [Aslin et al., 1998]
  - Child-directed speech provides contextual evidence for learning categories
    [Redington et al., 1998, Mintz, 2002]
Unsupervised category induction

- Many unsupervised models use distributional information to learn categories
  - [Brown et al., 1992, Clark, 2003, Goldwater and Griffiths, 2007]

- But most are not cognitively plausible
  - process data in batch mode
  - categorize word types instead of word tokens
  - pre-define the number of categories
Online category induction

- A few online models of category induction are proposed
  - [Cartwright and Brent, 1997, Parisien et al., 2008]
  - More cognitively motivated
- But may require large amounts of training, and be over-sensitive to context variation
- We propose
  - A simple algorithm which incrementally learns an unbounded number of categories
  - A task-based approach to evaluating human categorization models
1 Lexical category acquisition in humans
2 Online information-theoretic model
3 Task-based evaluation
Informativeness versus parsimony

- A good categorization model partitions words into discrete categories such that:
  - The number and distribution of categories is as simple as possible
  - Categories are highly informative about their members

- In other words trade-off parsimony against informativeness (goodness-of-fit)
Joint entropy criterion

- **Parsimony**

\[ H(Y) = -\sum_{i=1}^{N} P(Y = y_i) \log_2[P(Y = y_i)] \] (1)

- **Informativeness**

\[ H(X|Y) = \sum_{i=1}^{N} P(Y = y_i) H(X|Y = y_i) \] (2)

- Joint entropy minimizes the sum of both

\[ H(X, Y) = H(Y) + H(X|Y) \] (3)
Joint minimization for multiple variables

Optimize simultaneously for all features

\[
\sum_{j=1}^{M} H(X_j, Y) = \sum_{j=1}^{M} \left[ H(X_j|Y) + H(Y) \right] = \sum_{j=1}^{M} H(X_j|Y) + M \times H(Y)
\]
Incremental updates

- At point $t$ find the best assignment $Y = y_i$:

$$\hat{y} = \begin{cases} 
y_{N+1} \\
\arg\min_{y \in \{y\}_{i=1}^{N}} \Delta H^t_y 
\end{cases} \quad \text{if } \forall y_n \left[ \Delta H^t_{y_{N+1}} \leq \Delta H^t_{y_n} \right]$$

otherwise

\begin{equation}
\text{(5)}
\end{equation}

where

$$\Delta H^t_y = \sum_{j=1}^{M} \left[ H^t_y(X_j, Y) - H^{t-1}(X_j, Y) \right] \quad \text{(6)}$$

- $H^t(X_j, Y)$ can be computed incrementally.
1. Lexical category acquisition in humans
2. Online information-theoretic model
3. Task-based evaluation
Data

- Manchester portion of CHILDES, mothers’ turns
- Discard one-word sentences and punctuation

| Data Set    | Sessions | #Sentences  | #Words     |
|-------------|----------|-------------|------------|
| Training    | 26–28    | 22,491      | 125,339    |
| Development | 29–30    | 15,193      | 85,361     |
| Test        | 32–33    | 14,940      | 84,130     |
Labeling with categories

ΔH. Categories induced from the training set
Features: [want, to, try, them, on]

PoS. POS tags from the Manchester corpus

Words. Word types

Parisien. Categories induced by Bayesian model of [Parisien et al., 2008] from the training set.
Example clusters

- playing
- coming
- making
- back
- taking
- doing
- going
- looking
- than
- more
- silly
- harder
- funny
- frightened
- bigger
- dark
How to evaluate induced categories?

- Against gold POS tags
  - Arbitrary choice of granularity and/or criteria for membership
- Task based evaluation
  - Different tasks may call for different category representations
- Proposal: evaluate on a number tasks, simulating key aspects of human language processing
Evaluation against POS labels

- Variation of Information:
  \[ VI(X, X') = H(X) + H(X') - 2I(X, X') \]
- Adjusted Rand Index

\[ \Delta \Delta H \] Parisien Words Gold
\[ VI \] 0 1 2 3 4 5
\[ \Delta \Delta H \] Parisien Words Gold
\[ ARI \] 0 20 40 60 80 100
Task-based evaluation

- **Word prediction**
  - Guess a missing word based on its sentential context

- **Semantic feature prediction**
  - Predict the semantic properties of a novel word based on context

- **Grammaticalitude judgement**
  - Assess the syntactic well-formedness of a sentence based on the category labels assigned to its words
Word prediction

Human subjects are remarkably accurate at guessing words from context, e.g. in Cloze Test:

Petroleum, or crude oil, is one of the world’s (1) —— natural resources. Plastics, synthetic fibres, and (2) —— chemicals are produced from petroleum. It is also used to make lubricants and waxes. (3) ——, its most important use is as a fuel for heating, for (4) —— electricity, and (5) —— for powering vehicles.

A. as important  
B. most important  
C. so importantly  
D. less importantly  
E. too important
Word prediction

Reciprocal rank

want to | put | them on
## Word prediction

### Reciprocal rank

| want to put them on $y_{123}$ | make take put get sit eat let |
|--------------------------------|--------------------------------|

$rank^{-1} = \frac{1}{3}$
Word prediction: variants

- $\Delta H_{\text{max}}$

\[
P(w|h) = P(w|\arg\max_i R(y_i|h)^{-1})
\]

- $\Delta H_{\Sigma}$

\[
P(w|h) = \sum_{i=1}^{N} P(w|y_i) \frac{R(y_i|h)^{-1}}{\sum_{i=1}^{N} R(y_i|h)^{-1}}
\]
Word prediction: Results

- Gold POS
- Parisien
- $\Delta H_{\text{max}}$
- $\Delta H_{\Sigma}$

MRR

Chrupala and Alishahi (UdS)
Online Category Acquisition
CoNLL 2010
Comparison to n-gram language models

Gold
LM₁
LM₂
LM₃
LM₄
LM₅
ΔH₅

MRR

0 5 10 15 20 25 30 35
[Gelman and Taylor, 1984]: 2-year-olds treat words preceded by a determiner ("the zav") as common nouns, and interpret them as category members (block-like toy).
[Gelman and Taylor, 1984]: 2-year-olds treat words not preceded by a determiner ("Zav") as proper nouns, and interpret them as individuals (animal-like toy).
Semantic features from WordNet and VerbNet

Semantic profile for each category is the multiset union of the semantic sets of its members.
Semantic feature prediction task

I had | cake | for lunch
Semantic feature prediction task

I had cake for lunch

\[ y_{123} \]

\[
\begin{pmatrix}
    y_{123} & \text{entity} & \text{cake} \\
    \text{substance} & \text{baked goods} \\
    \text{matter} & \text{food} \\
    \text{food} & \text{solid} \\
    \text{edible} & \text{substance} \\
    \ldots
\end{pmatrix}
\]

\[
\text{AP}(F, R) = \frac{1}{|R|} \sum_{r=1}^{|F|} P(r) \times 1_R(F_r)
\] (7)
Predicting semantic properties: Results

![Graph showing MAP results for Gold POS, Parisien, and ΔH]
Grammaticality judgement

Both children and adults have a reliable concept of what is grammatical [Theakston, 2004]:

“She gave the book me”
Is it ok, or is it a bit silly?

Silly

“She gave me the book”
Is it ok, or is it a bit silly?

OK
Grammaticality task

\[ \text{score}(y) = \min_{i=1}^{n} P(y_i | y_{i-2}, y_{i-1}) \]

want to put them on
Grammaticality task

\[ \text{score}(y) = \min_{i=1}^{n} P(y_i|y_{i-2}, y_{i-1}) \]

want to put them on

\( y_{41} \) \( y_{21} \) \( y_{123} \) \( y_2 \) \( y_3 \)
Grammaticality task

\[ \text{score}(y) = \min_{i=1}^{n} P(y_i|y_{i-2}, y_{i-1}) \]

want to put them on

\[ y_{41} \quad y_{21} \quad y_{123} \quad y_2 \quad y_3 \]

\[ 0.02 \quad 0.1 \quad 0.05 \quad 0.01 \quad 0.03 \quad = 0.0100 \]
Grammaticality task

\[ score(y) = \min_{i=1}^{n} P(y_{i} | y_{i-2}, y_{i-1}) \]

| want | to | put | them | on      |
|------|----|-----|------|---------|
| \( y_{41} \) | \( y_{21} \) | \( y_{123} \) | \( y_{2} \) | \( y_{3} \) |
| 0.02 | 0.1 | 0.05 | 0.01 | 0.03 | = 0.0100 |

| want | to | them | put | on      |
|------|----|------|-----|---------|
| \( y_{41} \) | \( y_{21} \) | \( y_{124} \) | \( y_{4} \) | \( y_{3} \) |
| 0.02 | 0.1 | 0.001 | 0.0005 | 0.005 | = 0.0005 |
Grammaticality task

\[
\text{score}(y) = \min_{i=1}^n P(y_i|y_{i-2}, y_{i-1})
\]

| want | to | put | them | on |
|------|----|-----|------|----|
| \(y_{41}\) | \(y_{21}\) | \(y_{123}\) | \(y_2\) | \(y_3\) |
| 0.02 | 0.1 | 0.05 | 0.01 | 0.03 |

\[= 0.0100\]

| want | to | them | put | on |
|------|----|------|-----|----|
| \(y_{41}\) | \(y_{21}\) | \(y_{124}\) | \(y_4\) | \(y_3\) |
| 0.02 | 0.1 | 0.001 | 0.0005 | 0.005 |

\[= 0.0005\]

\[
correct = \begin{cases} 
1 & \text{if } \text{score}(y^{ok}) > \text{score}(y^*) \\
0 & \text{otherwise}
\end{cases}
\]
Grammaticality judgement: Results

![Chart showing the accuracy rates for Gold POS, Words, Parisien, and ΔH categories.](chart.png)
## Summary of results

|        | Gold   | Words | Parisien | $\Delta H_{\text{max}}$ | $\Delta H_{\Sigma}$ |
|--------|--------|-------|----------|--------------------------|---------------------|
| Pred   | 0.354  | -     | 0.212    | 0.309                    | **0.359**           |
| Sem    | 0.351  | -     | 0.213    | **0.366**                | -                   |
| Gram   | **0.728** | 0.685 | 0.683    | 0.715                    | -                   |
Conclusion

- **Learning categories**
  - Categories can be learned from usage data incrementally
  - A simple online information-theoretic approach works well in this scenario

- **Evaluation**
  - Automatically induced categories can work better than PoS tags in language tasks
  - Evaluation of unsupervised category induction models should not rely exclusively on gold POS labels

- **Future directions**
  - Compare the performance of the model to humans
  - Develop a wider range of tasks
Aslin, R., Saffran, J., and Newport, E. (1998).
Computation of conditional probability statistics by 8-month-old infants.
*Psychological Science*, 9(4):321–324.

Brown, P., Mercer, R., Della Pietra, V., and Lai, J. (1992).
Class-based n-gram models of natural language.
*Computational linguistics*, 18(4):467–479.

Cartwright, T. and Brent, M. (1997).
Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis.
*Cognition*, 63(2):121–170.

Clark, A. (2003).
Combining distributional and morphological information for part of speech induction.
In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66.

Gelman, S. and Taylor, M. (1984).
How two-year-old children interpret proper and common names for unfamiliar objects.
*Child Development*, pages 1535–1540.

Goldwater, S. and Griffiths, T. (2007).
A fully Bayesian approach to unsupervised part-of-speech tagging.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 744.

Kemp, N., Lieven, E., and Tomasello, M. (2005).
Young Children’s Knowledge of the” Determiner” and” Adjective” Categories.
*Journal of Speech, Language and Hearing Research*, 48(3):592–609.

Mintz, T. (2002).
Category induction from distributional cues in an artificial language.
*Memory and Cognition*, 30(5):678–686.
Cluster evaluation metrics

- **Variation of information:**
  \[ V I(X; Y) = H(X) + H(Y) - 2I(X, Y) \]

- **Rand Index:**
  \[ R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \]

- **Adjusted Rand Index:**
  \[ AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \]