The contribution of abstract linguistically-motivated learning biases to statistical learning of lexical categories

Erwin Chan

Linguistics colloquium,
University of Arizona
November 7, 2008
Outline

• Introduction: statistical learning and lexical category induction
• Previous work
• Computational modeling of language acquisition
• A new approach
• Conclusion
Lexical categories

• Lexical / open-class categories: “Noun”, “Verb”, “Adjective” in English

• Assumed in generative linguistics

• (Some typological problems)
Origin of lexical categories?

- Innate

- Reflects cognitive predisposition
  - Objects = nouns, properties = adjectives, actions = verbs

- Induced from experience
  - Statistical regularities in input data
  - General-purpose, domain-independent learning procedures

- Induced but mapped onto innate/cognitive categories
Develop computational model

• How do children acquire lexical categories?
  – Language acquisition problem

• Abstract computational problem:
  – Given a corpus of linguistic data, group together words into their lexical categories

• Develop a computational model
  – Try to replicate problem in artificial setting
  – Formulate precise, step-by-step procedures to learn categories from a corpus

• Solution to computational problem is suggestive about human learning
Human language acquisition

Linguistic environment

Universal Grammar + sensorimotor systems + pattern recognition

Language production

Colorless green ideas sleep furiously.
Learning from a corpus

Corpus → Data structures + algorithms → Language production

Colorless green ideas sleep furiously.
Computation = Cognition

Data structures + algorithms

(Description of)
Universal Grammar + sensorimotor systems + pattern recognition
Outline

• Introduction: statistical learning and lexical category induction
• Previous work
• Computational modeling of language acquisition
• A new approach
• Conclusion
Previous work in syn. category induction

• Distributional learning hypothesis
  – to ___ → Verb
  – the ___ is → Noun

• Previous work
  – Structuralist linguistics: Harris 1946/1954, Fries 1952
  – Psycholinguistics: Redington, Chater, & Finch 1998; Mintz 2002

• General claims
  – Distributional learning works
  – Demonstrates utility of statistics
  – Undermines nativism
How it works

• Demonstrate for English

• Input format: orthographic text corpus
  – Child-directed speech (psycholinguistics)
  – Texts written for adults (NLP)
  – Redington et. al. 1998: very similar results

• Linguistic assumptions
  – Split between open- and closed-class words
Examine distributional contexts

- Count occurrences of function words to left/right of content words

- Function words = highest frequency words in a language
- Content words = lower frequency words
Form matrix of co-occurrence counts

<table>
<thead>
<tr>
<th>Words to be clustered</th>
<th>Left context features</th>
<th>Right context features</th>
</tr>
</thead>
<tbody>
<tr>
<td>box</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Each row describes a word’s distributional context in the entire corpus
Apply clustering algorithm

• Group together rows that are similar into clusters
  – Some clustering algorithm
    • Hierarchical clustering
    • Dimensionality reduction approaches
    • Probabilistic models
    • etc.

• Words with similar distributional environments are placed in the same cluster

• If distributional learning hypothesis is correct, the words in a cluster will belong to same syntactic category
Goal for lexical category induction

• One-to-one correspondence between clusters and lexical categories
  – cluster 1 = all nouns
  – cluster 2 = all adjectives
  – cluster 3 = all verbs
Problem: no algorithm discovers the exact set of lexical categories

• Clusters are too fine-grained
  – Singular nouns
  – Plural nouns
  – Verb participles

• Clusters are too coarse-grained
  – Adjective and Noun conflated
  – Open- and closed-class categories conflated
Example: hierarchical clustering

• Algorithm
  – Each item begins in its own cluster
  – Find two most similar clusters, merge them into a new cluster
  – Produces a tree (dendrogram) describing the merging process

• Previous work
  – Redington, Chater, & Finch 1998
  – Mintz, Newport, & Bever 2002
from Redington et. al. 1998
Figure 1. The discrete clusters at a similarity level of 0.8 from the analysis of the CHILDES corpus. The clusters have been labelled by hand with the syntactic categories to which they correspond. The number of items in each cluster is shown in parentheses. Only clusters with 10 or more members are shown here.
Problem: need a discrete set of clusters

- Dendrogram describes how words are gradually clustered, until there is just one cluster containing all words

- Obtain discrete set of clusters by choosing a cutoff value for cluster similarity

- Resulting clusters are the set of syntactic categories
Two categories

Pronouns, Pronouns + Aux, Aux, Aux + Negation (49)

WH-, WH- + Aux, Pronoun + Aux (53)

Verb (105)

Verb (62)

Verb, Present Part. (50)

Determiner, Possessive Pronoun (29)

Conjunction, Interjection, Proper Noun (91)

Proper Noun (19)

Preposition (33)

Noun (317)

Adjective (92)

Proper Noun (10)
all Verbs

Pronouns, Pronouns + Aux, Aux, Aux + Negation (49)
WH-, WH- + Aux, Pronoun + Aux (53)

Verb (105)

Verb (62)

Verb, Present Part. (50)

Determiner, Possessive Pronoun (29)

Conjunction, Interjection, Proper Noun (91)

Proper Noun (19)

Preposition (33)

Noun (317)

Adjective (92)

Proper Noun (10)

Nouns, Adjectives conflated
Three different Verb clusters

Nouns, Adjectives separate
Mintz et. al. 2002: same issue
Finding separate lexical categories

- Want exactly one cluster for each lexical category: Noun, Adjective, Verb
- No algorithm finds exact set of lexical categories
  - One Verb cluster → Noun, Adj conflated together
  - Noun, Adjective separate → multiple Verb clusters
- Distributional learning doesn’t work…
  - Despite large corpora
  - Despite accumulation of lots of statistics
Nouns and adjectives in English are hard to separate.
(Need an algorithm!)

- Mintz 2003: “frequent frames”
  - the ___ is $\rightarrow$ noun
  - to ___ the $\rightarrow$ verb

- In English, a frequent frame is a good indicator of words that are homogeneous w.r.t. category
  - e.g., most words in frame the ___ is are nouns

- However, no algorithm for grouping together all the frequent frames for one category
Outline

• Introduction: statistical learning and lexical category induction
• Previous work
• Computational modeling of language acquisition
• A new approach
• Conclusion
How to design a computational model of language acquisition?

- Chomsky 2005: 3 factors
  - 1. Genetic endowment / Universal Grammar (**main emphasis**)  
  - 2. Linguistic experience  
  - 3. Principles not specific to language, including computation

- Computational interpretation:
  - 1. Linguistic representation  
  - 2. Input data statistics  
  - 3. Learning algorithm

- Computational modeling addresses all 3 factors
  - Interaction between factors currently not well understood
Component 1: Data representation

• What structures must be built into the model in order for it to succeed?

• What should the result look like?
  – Extensional definition
    • Verbs = { swim, give, talk, laugh, …, eat }
    • Words considered to be “Verbs” appear in same cluster
  – Intensional definition
    • Infinite set of Verbs
    • Define a Verb by its properties: “ to ___ “, “ will ___ “

• What input data representation is necessary to obtain the desired output?
Component 2: statistics of environment

• View #1:
  – Poverty of Stimulus
  – Believed that not possible to account for from linguistic experience,
  – Especially given time constraints
Statistics: view #2

- Learner encounters lots of data
- Sufficient linguistic experience to learn and generalize
- Demonstrations that child learners are capable of calculating statistics
- Demonstrations of statistical learning from corpora through computational models
Statistics: view #3

• There are statistical properties of naturally occurring linguistic data that are invariant

• Can be investigated by looking at a corpus
  – Must be naturally occurring corpus
  – Not artificial

• Consequences for learning…
Zipf’s law: distribution of inflections

- # words per inflection, Czech Verbs
- One inflection has very high frequency
- A few inflections have high or moderately high frequency
- Most inflections are infrequent or rare
Catalan verbs

Rare words
CHILDES Catalan verbs
Zipf’s law: not unique to language usage

- Magnitude of earthquakes
- Death toll in wars
- Population of cities
- Number of adherents to religions
- Net worth of individuals
- Popularity of websites
- Number of species per genus
Consequences of Zipf’s law and sparse data for learning

- Statistical invariants places constraints on cognitively plausible representations and algorithms

- Zipf’s law
  - Causes sparse data problem / Poverty of the stimulus

- Learners cannot assume abundant statistical evidence

- Must exist alternative means for efficient statistical learning given sparse data
Component 3: algorithms

- No Free Lunch theorem
  - Not possible to formulate an algorithm that works equally well for all problems
  - Any algorithm must be tuned towards specific problem, to some extent

- Any algorithm works best for:
  - Some particular representation of data
  - Some particular statistical distribution of data

- Learning bias: set of assumptions made
What kind of learning bias?

- Strong representation, weak statistics
  - Nativist position

- Weak representation, strong statistics
  - Statistical / distributional learning

- Intermediate in both
  - My proposed solution
Summary: must balance 3 factors

• Representation
  – Want to learn intensional definitions

• Statistics
  – Naturalistic data is sparse

• Algorithms
  – Balance learning bias: linguistic and statistical
Outline

• Introduction: statistical learning and lexical category induction
• Previous work
• Computational modeling of language acquisition
• A new approach
• Conclusion
Key problems to address

• Fine-grained clusters

• Categories conflated into same cluster

• Previous work
  – Cluster all frequent content words
  – Look at immediately adjacent words only
  – Does not consider data statistics
Key aspects of solution: stronger linguistic bias

• Use abstract distributional contexts
  – Instead of just immediate words

• Cluster morphological base forms
  – Instead of clustering words of all inflections

• Concept of open-class clusters
  – Instead of arbitrary merging of clusters
Input data

• Brown corpus + Penn Treebank
  – Text corpus
  – 2 million words
  – Written for adults

• (Redington et. al. 1998: no differences in cluster analysis between child-directed speech and adult-directed text)
Solution part 1: Too many categories

- Problem: fine-grained clusters often correspond to different morphological inflections of same lexical category

- Solution: cluster one inflection per category
  - instead of words of all inflections
  - One inflection per underlying lexical category → one cluster per underlying lexical category

- Lexicon + rules model of morphophonology
  - Cluster “morphological base form”
Rule-based model of morphology and lexical categories

Lexicon

base1: A
base2: A
base3: B
base4: B
base5: A
...

POS categories

A

B
Morphological base form

• English “base” or “bare form”

• Dog
• Bake
• Sell

• Not: dogs, dog’s
  baking, baked
  sells, sold
Freq. distribution of inflections

• Zipfian dist. of inflections
  – # word types per inflection

• Most type-frequent inflections in corpora
  – Nouns, Adjectives:
    nominative singular
  – Verbs: Infinitive or
    3rd person present singular
  – “canonical base”
Learning rules from sparse data

• Sketch of morphology acquisition algorithm:
  – For a particular lexical category:
  – 1. Identify base through high frequency
  – 2. Relate other, less-frequent inflections to base through morphophonological rules
Discover rule-based grammar
English top 1000 base forms

Green = Verbs

Blue = Nouns

Red = Adjectives
Solution part 2: conflated categories

• Previous work:
  – Large clusters get merged together
    (nouns, adjs)

• Solution:
  – Identify *open-class* clusters
  – Let each open-class cluster gain more members
    – *But don’t merge open-class clusters*
Clustering algorithm

• Let each data point be in a cluster.

• Go through all pairs of data points \((p1, p2)\) in order of decreasing similarity.
  - If \(p1\)’s cluster and \(p2\)’s cluster are both above a threshold size, do not merge their clusters
  - Otherwise merge their clusters
English top 1000 base forms
English top 1000 base forms
English top 1000 base forms
English top 1000 base forms
Intentions of algorithm

• Algorithm discovers cluster centers first

• Clusters continually add new members

• Large cluster \(\rightarrow\) open-class category

• Large clusters never get merged
  – Noun and adj should not be conflated
## Intermediate results

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Adj</th>
<th>Noun</th>
<th>Verb</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster 1</td>
<td>143</td>
<td>302</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>cluster 2</td>
<td>30</td>
<td>365</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>cluster 3</td>
<td>3</td>
<td>18</td>
<td>112</td>
<td>0</td>
</tr>
</tbody>
</table>

 Doesn’t work: Adjectives and Nouns still in same cluster (cluster 1)
Solution part 3: higher-order features

• Definition of adjective?
  – the ___ (ambiguous between adj/noun)
  – “the big dog”
  – “the dog”

• Better definition: ___ Noun
  – Not available when using words as distributional context
Clusters as distributional context

- Annotate the data with the clusters that were learned
  - *The big/1 dog/1 wants to sleep/3.*
- Update co-occurrence matrix
- Cluster again

<table>
<thead>
<tr>
<th>words to be clustered</th>
<th>Left features: words</th>
<th>Right features: words</th>
<th>Left / Right cluster features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cluster 1 “errors”: *american*, *indian*, *soviet*, *african*

<table>
<thead>
<tr>
<th></th>
<th>Adj</th>
<th>Noun</th>
<th>Verb</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster 1</td>
<td>143</td>
<td>302</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>cluster 2</td>
<td>30</td>
<td>365</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>cluster 3</td>
<td>3</td>
<td>18</td>
<td>112</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Adj</th>
<th>Noun</th>
<th>Verb</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster 1</td>
<td>144</td>
<td>75</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>cluster 2</td>
<td>29</td>
<td>604</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>cluster 3</td>
<td>3</td>
<td>6</td>
<td>111</td>
<td>3</td>
</tr>
</tbody>
</table>
Algorithm summary

• Linguistically motivated algorithm:
  – Cluster base forms only, instead of all inflections
  – Concept of open-class categories
    • Do not merge them into a larger category
  – Abstract category formation
    • Available for use in subsequent stages of learning
Solves problems of previous work

• 1. Too many categories
  – *Cluster morphological base forms only*

• 2. Lexical categories merged together
  – *Don’t merge open-class clusters*
  – *Use categories as distributional contextual features*
Current and future work

• Other languages
  – Also examined Spanish: complications of gender classes

• Develop a tagger to extend analysis to rare words

• Syncretism

• Derivational morphology

• Lack of function words

• Other kinds of syntactic relations
Outline

• Introduction: statistical learning and lexical category induction
• Previous work
• Computational modeling of language acquisition
• A new approach
• Conclusion
Summary

• Problem: acquisition of open-class lexical categories

• Treat as abstract computational problem
  – Develop a model that learns from a corpus

• Balance learning bias
  – Linguistic: what representational structures are built into the learner?
  – Statistical: what assumptions about the data distribution does the learner make?
Comparison to previous work

• Previous computational models:
  – Doesn’t attain desired set of lexical categories
  – Distributional learning according to words in immediate context is insufficient

• Current model:
  – Stronger linguistic bias
  – More successfully acquires lexical categories
  – *Strong linguistic bias may be necessary*
Development of proposed solution

• Methodology
  – What is the desired result?
  – Look at statistical distribution of input data
  – What is necessary in the learner in order to produce desired result?

• Build linguistic concepts into architecture of learner
  – Open-class categories
  – Morphological base
  – Formation of abstract categories

• Still a data-driven, statistical / distributional learner
Linguistic and cognitive implications

• A solution to computational problem says something about the language faculty

• Supports:
  – Rule-based model of morphology
  – Formation and further availability of abstract categories

• Against:
  – Highly lexicalist models
  – Item-based learning (no abstractions)
Thank you.
Bibliography

• Zellig Harris. 1954. Distributional structure. Word, 10(2/3), 146-162.