Language Technology and Language Acquisition: an introduction with learning segmentation

Çağrı Çöltekin

c.coltekin@rug.nl

Center for Language and Cognition
University of Groningen

September 21, 2009

An example learning task

```
ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj
```

An example learning task

```
ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj
```

Children need to:

- segment the input to linguistic units (words, morphemes etc).
- assign meanings to these units.
- figure out which combinations of these units are acceptable in the language.
- ...

Overview

- The problem of language acquisition.
- Formal approaches to language learnability.
- How can the computational models help?
- An example: segmentation.

The Problem of Language Acquisition

- Human languages are complex (recursion, ambiguity).
- Children do not receive explicit instruction during language acquisition.
- Language acquisition by children is (arguably) fast and robust.
- The input to children is not enough for learning (*Poverty of Stimulus Argument*).
  - Children do not receive input critical for learning certain phenomena.
  - Human languages are not learnable from positive input (claimed to be formally supported by Gold, 1967). Negative input is not available to children.

Two views on human language acquisition

- **Nativism**
  The nativist theories of language acquisition assume that human language acquisition is guided by an innate *Language Acquisition Device*, or *Universal Grammar* (UG).
  The emphasis is on domain specific rich innate knowledge.
  Role of the input is secondary.
- **Empiricism**
  Empiricist theories claim that language acquisition is possible with general purpose learning systems.
  Emphasis is on the input.

Models of Language Acquisition

- **Principles and Parameters**
  Language acquisition is guided by a UG, consisting of principles and parameters. Learning is achieved by setting a small number of (binary) parameters.
- **Connectionist systems**
  Learning is achieved by general purpose learning algorithms, e.g. backpropagation.

Language acquisition debate: summary

Ground rules:

- *There must be some innate component*:
  - The child born in the same household learns the language, but the kitten does not.
  - No free lunch theorem: we know from the machine learning theory that there is no universal learning algorithm.
- *Learning is a part of the language acquisition*: children learn the language(s) spoken in their environment, not a universal language.

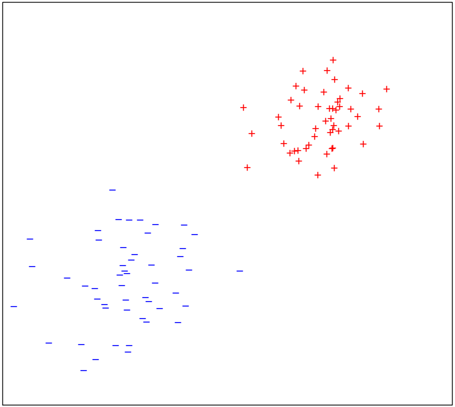
The main dispute is on the nature of the innate component and the learning mechanisms, either they are language specific, or general cognitive mechanisms.

Overview

- The problem of language acquisition.
- Formal approaches to language learnability.
- How can the computational models help?
- An example: segmentation.

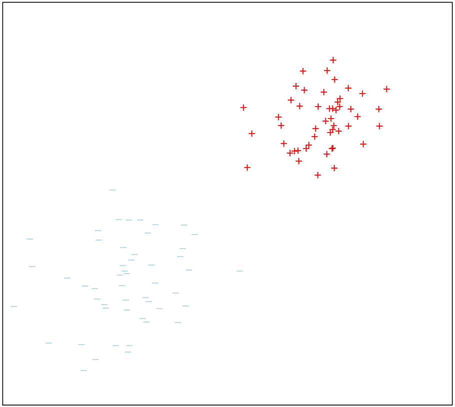
A simple description of the learning task

- Input is a set of positive and (possibly) negative sentences.



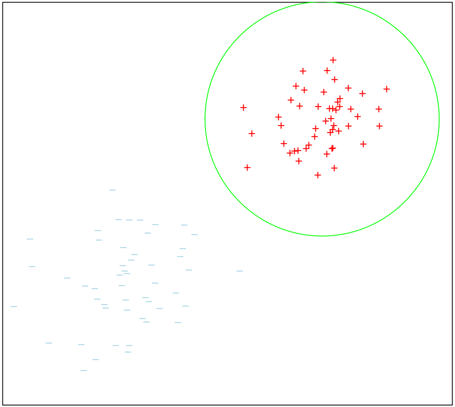
A simple description of the learning task

- Input is a set of positive and (possibly) negative sentences.
- It is common to assume that the learner is not exposed to negative examples.

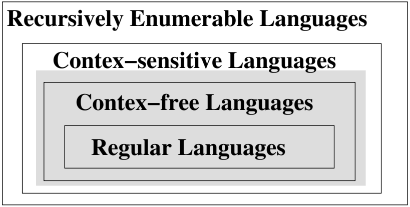


A simple description of the learning task

- Input is a set of positive and (possibly) negative sentences.
- It is common to assume that the learner is not exposed to negative examples.
- Task is learning a grammar that separates grammatical and ungrammatical sentences.



Chomsky hierarchy and Language Acquisition



- Human language syntax seems to require slightly more expressive power than context-free languages.
- Gold's theorem states that the languages in none of these classes are **identifiable in the limit** using positive examples.
- All are identifiable in the limit from positive and negative examples.

Is it innate then?

Theoretical results especially by Gold (1967), frequently (mis)used as a support for nativist theories. However,

- Other learning paradigms, e.g. PAC learning (Valiant 1984), are more suitable for modeling human learning.
- Different classification of grammars may allow learning in Gold's framework (e.g. Angluin, 1980; Shinohara, 1990; Kanazawa, 1998; Clark etal. 2008).
- Distribution of input may have a significance in learning.
- Input may contain negative data.

A cautionary note: identifiability in the limit does guarantee practical learnability.

Overview

- The problem of language acquisition.
- Formal approaches to language learnability.
- How can the computational models help?
- An example: segmentation.

Computational Models of Language Acquisition

- Answers to the questions on language acquisition should eventually come from neuroscience. But we seem to be far from this yet.
- Formal learnability results are useful by identifying learnable/unlearnable well defined (formal) languages. It seems to be difficult to formally characterize
  - The class of human languages.
  - The input to human learner.
- Computational models provide other (complementary) means of investigating these questions.

Computational Models of Language Acquisition

- Computational models can help us test claims of learnability directly: we can use real the data (e.g. CHILDES database) and empirical experiments with the models of language acquisition.
- Computational models can help identify the innate knowledge necessary (or not) for learning languages.
- Computational models require theories to be described explicitly.

A short divergence: levels of processing

Theories or models provide explanations at different levels. One attempt to formalize this notion of levels of processing/representation is due to Marr (1982).

- Computational level: *What* does the system do, and *why*.
- Algorithmic level: *How* does the system carry out the computations, and how is the input/output represented.
- Implementation level: How the system is physically realized.

The classification is not always clear-cut, but while evaluating the computational models of cognitive phenomena, one should always keep in mind at which level the model tries to answer the questions.

Computational Models for Language Acquisition

Computational models of human language acquisition has to meet some criteria that is not always applicable for engineering oriented CL applications.

- Modes should use realistic input, such as naturally occurring child directed speech.
- Any additional source of information, or heuristics should be justifiable.
- Learning should proceed on-line: models should not require all the input data available at once.
- Models should not pose unrealistic bounds on memory and computation resources.
- The assumptions and predictions of the model should match (at least should not conflict with) psycholinguistic evidence.

Overview

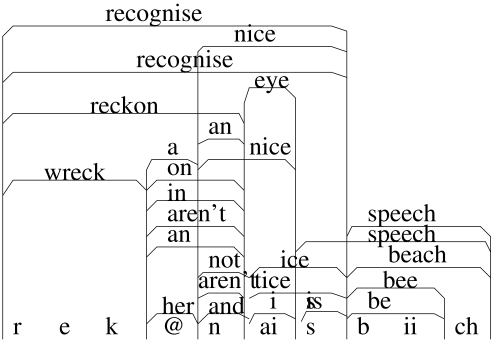
- The problem of language acquisition.
- Formal approaches to language learnability.
- How can the computational models help?
- An example: segmentation.

Segmentation: introduction

- Spoken language does come with blanks: there is no reliable cue for spotting boundaries of linguistic units (words, morphemes etc.).
- Children need to segment continuous speech into useful units.

Segmentation: introduction

- Spoken language does come with blanks: there is no reliable cue for spotting boundaries of linguistic units (words, morphemes etc.).
- Children need to segment continuous speech into useful units.



*Example re-produced from: ?

An old algorithm: LSV

- The morpheme boundaries are at the locations where there are more possibilities to follow. (Haris, 1955)
- Try to think about words starting with,

  `compu-`

  probably most words you can think of will continue with `-t`.
- Try to think about words starting with,

  `comput-`

  this time we can find words (at least) with continuations `-e`, `-a`, `-i`.

LSV: example

Consider the following input:

```
READ
READS
READING
READABLE
```

LSV: example

Consider the following input:

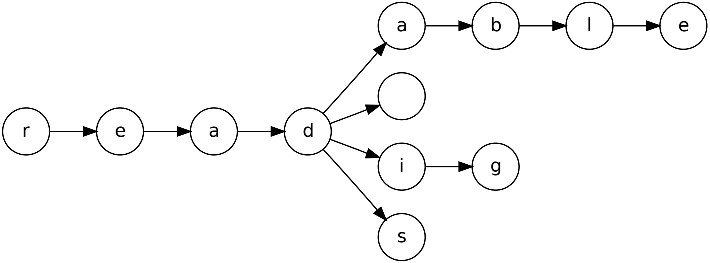
`READ`

`READS`

`READING`

`READABLE`

There is a data structure called *trie* (or prefix tree) that implements the idea efficiently.



Generalization of the idea: entropy/predictability/surprisal

More generally, probability of a segment is higher where next phoneme (or letter) is *not* predictable, lower if predictable.

- $P(I|C)$ is high inside units, low outside the units.
- Entropy is low inside units, high outside units.

$$H(\alpha) = -\sum_{\beta \in succ(\alpha)} P(\alpha\beta)\ log_2 P(\alpha\beta)$$

Segmentation: an example

```
ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj
```

Segmentation: an example

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph

ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

Segmentation: an example

ljuuzuibutsjhiuljuuz

ljuuztbzjubhbjompwfljuuz

xibutuibu

ljuuz

epzpvxbounpsfnjmlipofz

ljuuzljuuzephhjf

opnjxibuepftbljuuztbz

xibuepftbljuuztbz

ephhjfeph

ephhjf

opnjxibuepftuifephhjftbz

xibuepftuifephhjftbz

mjuumfcbczcjsejf

cbczcjsejf

zpvepoumjlfuibupof

plbznpnnzublfuijtpvu

dpx

uifdpxtbztnppnpp

xibuepftuifdpxtbzopnj

Segmentation: an example

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph

ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

$$P(u|j) = \frac{11}{27} = 0.4 \quad P(u|z) = \frac{2}{23} = 0.08$$

Segmentation: an example

```
kitty
thatsright
kitty kitty
sayitagainlove
kitty what
sthat kitty do
youwantmoremi
lkhoney kitty
kitty doggie
nomiwhat
doesakittysay
what does
akitty say do
ggie dog
doggie
nomiwhat does
thedoggiesay
what does
the doggiesay
littlebabybirdie
babybirdie
youdontlikethatone
okaymommytakethisout
cow the cowsay
smoomoo what
does the
cowsaynomi
```

Predictability based models: psychological relevance

Children very early in life (8-months) seem to be sensitive to this type of information in the speech (Saffran, Aslin, Newport 1996)

- Infants are habituated to artificial speech segments built from a simple vocabulary.
- They are tested with non-familiar patterns and familiar patterns.
- On the basis of very short training 8-month-old infants attended familiar examples significantly longer than the unfamiliar ones.

Summary

- Computational models/simulations provide are useful in science, including cognitive sciences, especially when direct methods are not available or feasible.
- Computational models are useful for testing abstract linguistic theories. They, at least, provide more direct answers to questions of learnability.

References

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, *45*, 117–135.

Clark, A., Eyraud, R., & Habrard, A. (2008). A polynomial algorithm for the inference of context free languages. In *Proceedings of International Colloquium on Grammatical Inference*.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190–222.

Kanazawa, M. (1994). *Learnable classes of categorial grammars*. Amsterdam: Institute for Logic, Language and Computation, ILLC dissertation series.

Marr, D. & Vaina, L. (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *214*(1197), 501–524.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926–1928.

Shinohara, T. (1989). Inductive inference from positive data is powerful. Publications in Computer and Information Science 33, Research Institute of Fundamental Information Science, Kyushu University.