Introduction to Computational Modeling of Lexical and Grammatical Knowledge Acquisition using Machine Learning Techniques

December 2004

Damir Ćavar

Indiana University

Information Theory

- Mutual Information

$$I(X;Y) = P(XY)\log_2 \frac{P(XY)}{P(X)P(Y)}$$

  - How many bits can we spare by storing <*xy*> together, rather than each separate?
  - How much do we expect *y* given *x*?

Information Theory

- Relative Entropy

$$D(y||x) = p(y)lg\frac{p(y)}{p(y|x)}$$

  - Distance between two distributions:
    - Independent: P(y)
    - Conditional: P(y|x)
  - How many bits more would we need to represent <xy> when we store them together, or when we store them as separate units?

Cue-based Learning

- Rationalist view:
  - Finite set of parameters P = {p1, p2, ..., pn}
  - Each p is a variable for a value from a finite set of values.
  - Specific values for the respective parameter explain language variability as well as language acquisition.

Cue-based Learning

- Parameter values are variable from the outset.
- Most/some parameters are associated with specific lexical properties.
- A set of lexical items has to be learned (i.e. mental lexicon).
- Cues in the input have to be identified to set parameters.
- General idea: E-language properties are mapped to I-language properties.

Cue-based Learning

- Paradox:
  - Cues are innate (Lightfoot, 1999)
  - Cues are identified by a specific innate e-language acquisition device, e.g. Superparser (Fodor & Teller, 2000)

Cue-based Learning

- Conceptual problems:
  - Everything is innate: e- and i-language
  - High-level orientation:
    - Syntax or semantics
- Myths:
  - Chaotic input!

Cue-based Learning

- Central research questions:
  - Can elementary language units and their properties be identified and associated with I-language parameters?
  - What is learnable and what can be learned given what kind of knowledge?

Cue-based Learning

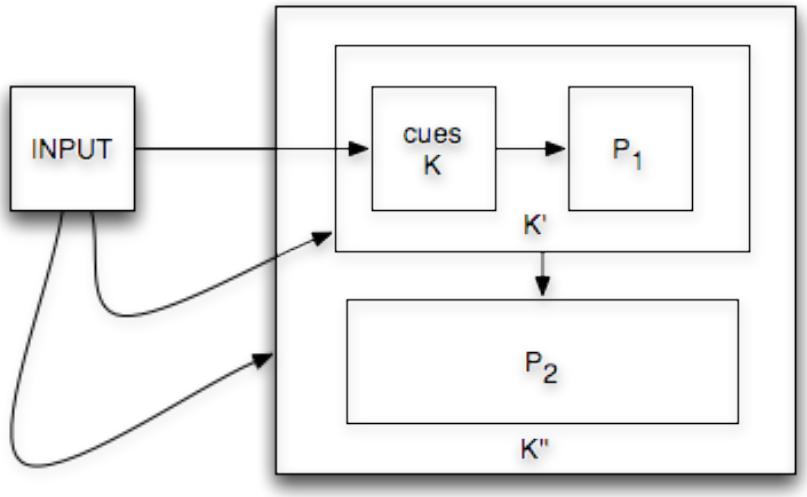
- Favorable outcome:
  - Cues can be identified without additional and language specific machinery.
  - Cues are used to set specific parameters.
  - Snowball effect or chain reaction:
    - Induced parameter values and cues are used to derive further (secondary) cues, which set more parameters, and so on.

Cue-based Learning

- Iterative & Incremental Cue-based Learning
  - Initial Bootstrapping Phase: An initial set of cues K identifies a specific parameter value P1 given some input.
  - Subsequent Bootstrapping Phases: Together with the set of cues K and the knowledge of parameter value P1 a new set of cues K' is derived, and so on.

Cue-based Learning

- Iterative & Incremental Cue-based Learning
  - Initial Bootstrapping Phase: An initial set of cues K identifies a specific parameter value P1 given some input.



Cue Identification

- How can we identify the initial set of cues, i.e. the kernel cues (k-cues)?
- Use observable properties of language input:
  - Intrinsic and Extrinsic properties of elementary language units.

Cue Identification

- Intrinsic properties of elementary language units:
  - Frequency
  - Length
    - Number of X
  - ...

Cue Identification

- Hypothesis 1:
  - Frequency is a cue for lexical information.
  - Length is related to frequency.
- Testing:
  - Frequency (and length) based clustering of words.

Cue Identification

- Weaknesses:
  - Intrinsic features alone are insufficient.
  - Clustering on intrinsic and extrinsic features is more promising.
- Cue identification on the basis of intrinsic and extrinsic features without clustering...

Cue Identification

- Distributional cues:
  - Collocations inform us about the type of a lexical element:
    - *the...*: followed by N, i.e. a syntactic type
    - *of...*: followed by Art, i.e. a syntactic type
    - *John speaks...*: followed by a referent/name to/of a language, i.e. a concept
    - *my name is...*: followed by a proper name, i.e. syntactic type and concept

Cue Identification

- Hypothesis 2:
  - Function words (as well as vowels, derivational and inflectional morphemes etc.) = highly frequent units are the structural landmarks.
- Testing:
  - Distributional properties of function words and substantives (and the relation between them).

Cue Identification

- Elghamry (2004), Ćavar & Elghamry (2004):
  - Language input is highly structured!
  - Distributional regularities in the input provide efficient bootstraps into the grammar of the input language.
  - There is a set of elements in the input (cues) that are learnable and that make language acquisition possible.

Cue Identification

- Language Learnability can be reduced to Cue Learnability.
- Categories and frames are cue-learnable.
- Parameter setting is cue-feasible.

Cues as Features in VS

- Mapping of lexical properties on a vector space:
  - rows: single words (types or tokens)
  - columns: numerical (real number) representation of single cues
    - inherent properties: length, frequency, ...
    - distributional properties: collocation with frequency, position in clause

Cues as Features in VS

- Similarity measure for rows
  - single feature based
  - multiple feature based
  - over whole vector space
- Clustering of lexical items
  - Cluster represent types, i.e. can be replaced with a symbol, i.e. single words can be tagged or replaced with a symbol
  - Distributional regularities can be extracted from symbol collocations: rule formation

Cues as Features in VS

- Dimension problem:
  - Distributional properties
    - high number of columns = complex similarity measures
    - identification of relevant distributional cues
      - but how?

Cue Identification

- The phases of Cue-Based Learning
  - Identification of cues
  - Cue-based induction (e.g. categorization)
  - Generation of new cues:
    - Using cue-based categories in for frame identification (subcategorization)

Cue Identification

- Cue Extraction Criterion (K)
  - The set of Cues, K, is the smallest subset of the elements $\{k_1, ..., k_m\}$ in a corpus R such that all elements in R occurs at least once with at least one member in K.

Cue Identification

- Approximate Cue Extraction Procedure
  - Build decreasing frequency profile for all the words in corpus R.
  - The set of cues K = $\{w_1, \ldots, w_m\}$ is the list of most frequent words, such that the number of words it co-occurs with corresponds to the number of words in the corpus.

Cue Identification

- k-cue identification:
  - From a decreasing frequency profile of types include all the types that co-occur with all the other word types in the corpus.
  - Stop, if no improvement in coverage: stagnation of k-cue - type ratio
  - Coverage:
    - the: 33.0 %, a: 44.0 %, you: 52.0 %, it: 57.0 %, that: 59.0 %, your: 62.0 %, and: 64.0 %, in: 66.0 %, to: 68.0 %, on: 69.0 %, not: 70.0 %
    - [the, a, you, it, that, your, and, in, to, ... w43] = 80%

Cue Identification

- k-cues (Peter corpus):
  - 43 k-cues for 3037 types with 80% coverage
  - 145846 tokens
  - k-cues: ['the', 'a', 'you', 'it', 'that', 'your', 'and', 'in', 'to', 'on', 'not', 'is', 'this', 'i', 'one', 'for', 'its', 'just', 'of', 'what', 'all', 'out', 'now', 'too', 'gonna', 'thats', 'with', 'are', 'peter', 'up', 'some', 'there', 'youre', 'my', 'her', 'right', 'go', 'have', 'we', 'so', 'he', 'can', 'little', 'over']

Cue Identification

Rank	Word	Frequency	Co-occ.
1	the	8705	4010
2	of	4220	3000
3	and	3055	3110
4	to	3049	2233
		Sum[1-4]	**12353** **1st order**
5	in	2629	2030
6	a	2190	1610
7	that	1394	1105
8	is	1160	883
9	was	1090	913
10	it	969	476
11	for	967	943
12	as	847	705
13	on	818	741
14	this	675	456
15	by	664	714
16	with	613	646
17	not	586	377
18	be	548	405
19	but	522	283
20	he	522	314
		Sum[1-20]	**24954** **2nd order**

Cue Use

- Distributional properties of cues and other words with cues:
  - Bi-gram models
  - Mutual Information (MI) calculation (Fano, 1961: 27-28; Jelinek, 1968: 120) :
    - Left and right point-wise MI for each kCue (Elghamry, 2004; Ćavar & Elghamry, 2004)

Cue Use

- Left and right point-wise Mutual Information

$$I(x;y) = lg\frac{P(< x,y >)}{P(x)P(y)}$$

- Average over left and right proportion pw-MI:

0.158840	you	0.841160
0.311925	the	0.688075
0.062290	it	0.937710
0.269833	a	0.730167

Cue Use

- Guessing of Headedness
  - Max sidewise point-wise MI marks syntactic or structural selection direction

Experiments and Results

- CHILDES database: Peter in Bloom (1970)
- Child-oriented speech used.
- Number of utterances: 25148
- Number of tokens: 156646
- Number of types: 3086

Experiments and Results

K	*Co-occ. Set*
the	1093
you	810
a	806
it	673
Total	3382

Implications

- Problems with MI:
  - Sparse data high MI value
  - Token-token ratio is counterintuitive:
    - High MI = high variation but selection implies low variation
    - Reason: edge-effect
      - Solution: include the edge effect, maximize the edge effect, but how?

Implications

- Input driven learning is possible:
  - MI-based selection preferences
  - Frame preferences or structural chunking (cf. treelets)
  - Lexical categorization or classification

Cue Identification II

- Higher level cue-identification:
  - Sparse data problem on the token level.
  - Syntactic Structure
    - Syntactic boundaries for *n*-grams and distributional properties
  - Morphological cues:
    - Additional information
      - Affixes are cues, i.e. they identify types
    - Statistics of types resolves sparse data problem with tokens

MI-based parsing

- Mutual Information
  - How many bits can we spare by storing two words together?
  - How much is Y expected given that X occurred (given the bi-gram <X,Y>)?

$$I(X;Y) = \sum_{xy} P(x,y) lg \frac{P(x,y)}{P(x)P(y)}$$

MI-based parsing

- Looking at MI of each bi-gram in a sentence:

$$I(x;y) = P(x,y) lg \frac{P(x,y)}{P(x)P(y)}$$

- Cutting the sentence/chunk at the lowest MI value into two pieces, and each resulting chunk again into two pieces, and so on
- → binary branching

MI-based parsing

- Looking at MI of each bi-gram in a sentence:

$$I(x;y) = P(x,y) lg \frac{P(x,y)}{P(x)P(y)}$$

- Merging elements in an *n*-gram with the highest MI value into one constituent, and each resulting chunk again into the next constituent, and so on
- → binary branching, *n*-branching

MI and RE

- MI:
  - Frequency sensitive
  - Symmetric
- RE (Kullback-Leibler Divergence)
  - Frequency sensitive
  - Asymmetric

- Syntactic structures are asymmetric!

Motivation

- Learning and Processing in parallel model
- Unsupervised
- Incremental

Previous approaches

- Magerman ea. (1990)
  - Assumptions and algorithm:
    - MI for POS *n*-grams
    - Cut sentence at local MI minima within a specific window
    - Variable window size and Generalized MI (= sum over MI for each possible sequence)
    - Combining constituents again based on *n*-gram MI
    - Supervision via grammar, i.e. a list of *distituents*, e.g. “Noun Preposition”

Previous approaches

- Magerman ea. (1990)
  - Results:
    - Good at parsing short sentences
      - average: one error per sentence
      - with conjunctions: two errors per sentence
    - Long sentences (16-30 words)
      - average 5 and 6 errors
    - What type of errors?

Our approach

- Mutual Information (MI) and Relative Entropy (RE)
- RE for <x,y>:

$$P(y) lg \frac{P(y)}{P(y|x)}$$

- With unknown types or tokens:

$$lg(len(alphabet)) * len(< x, y >)$$

Our approach

- Procedure: Cut sentence/chunk recursively into two sub-constituents
  - at minimum MI
  - at maximum RE
  - for each bi-gram type

Our approach

- Comparison of accuracy on bi-grams only:
  - token - token
  - token - type
  - type - token
  - type - type

Experiment

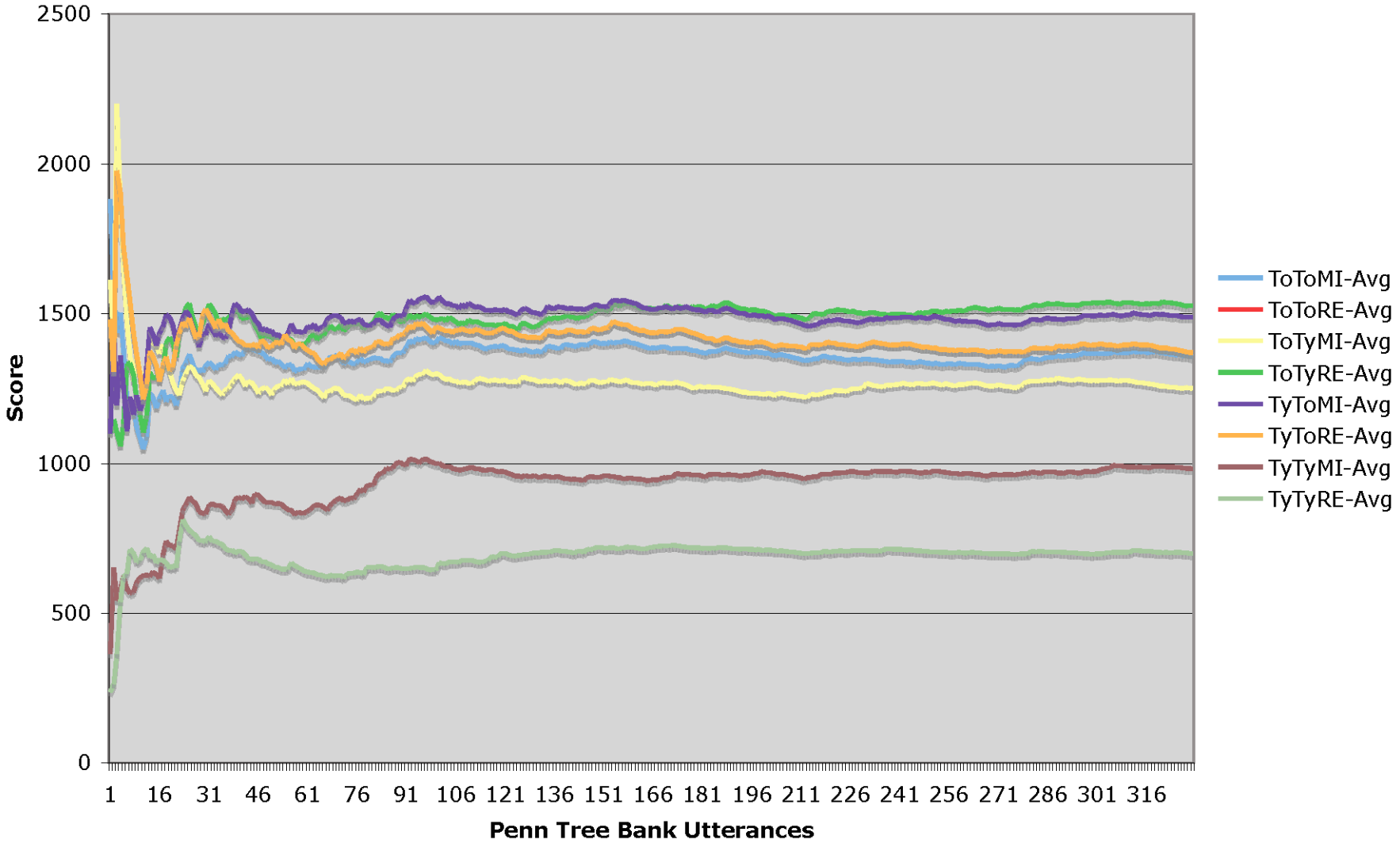
- Brown corpus with reduced tag-set (only basic category, i.e. N, V etc., ca. 14 tags) for training

- 5% Penn Treebank for evaluation

Evaluation

- Online evaluation:
  - For every input sentence and every output parse, compare brackets with Penn Treebank parse.
  - Incremental learning and parsing in parallel
    - Evaluation of the incremental effect of learning more words, types and bi-grams.

Evaluation


2500
2000
1500
1000
500
0
Score
1 16 31 46 61 76 91 106 121 136 151 166 181 196 211 226 241 256 271 286 301 316
Penn Tree Bank Utterances
ToToMI-Avg
ToToRE-Avg
ToTyMI-Avg
ToTyRE-Avg
TyToMI-Avg
TyToRE-Avg
TyTyMI-Avg
TyTyRE-Avg

Conclusions

- Type-type relations give best results
  - Syntactic relations are determined on the type level
- RE outperforms MI
  - Syntactic relations are asymmetric

Conclusions

- Chunks as domains for *n*-gram statistics
- Self-supervision:
  - Online update of distributional information for lexical properties
    - bias for directional properties
    - inclusion of only relevant collocation elements in vector space for clustering or classification
      - dynamic vector space for incremental learning

Self-supervision

- Selection of cues:
  - The smallest set of tokens that co-occurs with all other tokens (cf. Elghamry & Ćavar, 2004)
    - On the inverse frequency profile include all the most frequent tokens that co-occur with all other tokens.

- MI for such cues is reliable
- RE for such cues is more reliable!

Self-supervision

- Calculate the average left- and right-side MI for each cue X:
  - Average over all bi-grams with X in left position is the right-side MI for X.
  - Average over all bi-grams with X in right position is the left-side MI for X.

Self-supervision

- Average over left- and right-side MI for all cues X:
  - preferred selection direction

- However...
  - things are not that simple...

Self-supervision

- For elements like *the*, MI will be higher left of the token, lower right of it.
  - Reason: *the* occurs sentence initially!

- On the type level, MI values give the expected results, i.e. number of types right of *Art* is lower, MI increases.

Self-supervision

- Splitting of chunks is excluded if a type has a lower side-wise point-wise MI value on the side of the possible split.

- In the sense of Magerman ea. (1990):
  - Generation of "*distituents*" can be unsupervised!

Morphological Cues

- Morphological Structure
  - affixes behave like function words
    - structural and type landmarks
      - co-occur only with specific lexical types
    - they are highly frequent
    - they are more stable (closed class)
  - inclusion in VS
    - similarity analysis
    - clustering or classification

Morphological Cue Identification

- Hypothesis generation for morphological structure:
  - Random
  - Statistical:
    - Transitional probabilities (Harris, 1955)
    - EM-based (Brent, et al.)
  - Alignment based (ABL) (van Zaanen, 2001)

Morphological Cue Identification

- Hypothesis generation: MI

$$\sum_{y \in \{<xY>\}} p(<xy>|x) lg \frac{p(<xy>)}{p(x)p(y)}$$

- Pairwise summation of left MI of x and right MI of y.
- Cutting morpheme boundaries at local MI-minima.

Morphological Cue Identification

- Hypothesis generation: ABL
  - Substitutability and Complementarity
    - Given two words, the edges of matching substrings mark morphological boundaries.
  - Advantage:
    - Learning from previous knowledge.

Morphological Cue Identification

- Hypothesis evaluation:
  - Minimum Description Length
  - Relative Entropy
  - Mutual Information
  - Other quantitative measures

Morphological Cue Identification

- Grammar size
  - Minimum Description Length Principle (MDL)
    - From n grammars that describe the same corpus, chose the grammar with the smallest size (e.g. number of symbols, length of terminals)
  - Size for the morphology grammar:
    - morphemes + signature
      - dog = [NULL, -s]
      - -s = [dog, cat, car, ...]
      - morphological net: morpheme to signature, signature to morphemes

Morphological Cue Identification

- Grammar size *S*
  - For all morphemes *m*, with *n* the number of morphemes, sum length of *m*, and length of the pointer *p* to a corresponding signature
  - For all signatures *σ*, with *m* the number of signatures, sum the product of the length of the signature with the length of a pointer *p* to a morpheme

$$S = \left(\sum_{i=1}^{n} len(w_i) + len(p) \right) + \left(\sum_{i=1}^{m} len(\sigma_i) * len(p) \right)$$

Morphological Cue Identification

- MDL Grammar size
  - Independent of frequency of elements and signatures and elements in the signatures

Morphological Cue Identification

- Grammar size
  - Relative Entropy
    - From a set of hypotheses about the structure of an input *i*, add the hypothesis *h* to the set of grammar rules/hypotheses that results in lowest divergence from the original grammar.

Morphological Cue Identification

- Grammar size
  - Relative Entropy
    - We calculate RE as a variant of the Kullback-Leibler Divergence
    - Given grammars G1 and G2, choose the grammar that has the smallest divergence from the initial grammar G0.

Morphological Cue Identification

- Grammar size - Relative Entropy
  - Kullback-Leibler Divergence

$$\sum_{x \in X} P(x) lg \frac{P(x)}{Q(x)}$$

$$\sum_{x \in X} P(x) lg \frac{1}{P(x)}$$

Morphological Cue Identification

- Grammar size - Relative Entropy
  - simple calculation:
    - grammar = the string sequences it can generate
    - string sequences = *n*-grams
    - probability of *n*-grams = their relative frequency
    - Q(x) = *n*-gram statistics of existing string model
    - P(x) = *n*-gram statistics of new string model
    - Simplification of calculation:
      - If P1(x) and P2(x), integrate the new strings from the one, that has the smallest RE score given Q(x)

Morphological Cue Identification

- Grammar size - Relative Entropy
  - complex calculation:
    - for all morphemes and signatures sum of RE
      - probability of signature: relative number of pointers to it
      - probability of morphemes: relative frequency
    - if morpheme/signature not in old grammar, take its entropy

Morphological Cue Identification

- Problems:
  - Statistics over all do not work out
    - generation of nonsense morphemes
- Solution:
  - memory split:
    - long-term (LTM) and short-term memory (STM)
    - short-term memory window of *n*-utterances (remark on Zipf and cognition), implemented as a stack (FIFO)
    - significance check on morphemes and signatures cyclically within STM and copy of significant hypotheses to LTM

Morphological Cue Identification

- Usability related criteria:
  - Frequency of Morpheme Boundaries
  - Number of Morpheme Boundaries
  - Length of Morphemes

Morphological Cue Identification

- Voting-based architecture:
  - Every component votes for a hypothesis (= grammar)
  - The hypotheses with the highest votes win.

Morphological Cue Identification

- Weighting of votes:
  - Every voter can be weighted (0-1)
  - Means of self-supervision.

Morphological Cue Identification

- Results:
  - English: 100% precision, 80% recall
  - Latin: 99% precision, 35% recall
  - No self-supervision system so far, manual choice of values for constraint weights, which turns out to be quite irrelevant.
    - Memory division might be taken to be some sort of self-supervision instrument.
    - If hypotheses enter STM but get eliminated or never integrated into LTM, punish constraints that voted highest for these hypotheses (voting history)

Morphological Cue Identification

- Phases:
  - Acquisition phases for morphology observed in English and in the induction algorithm:
    - verbal inflection first
    - derivational morphology on main categories second
    - prefixes and infixes last
  - Why?
    - Because this mirrors the frequency patterns found in the corpora!

Conclusions

- Supervision suggestions:
  - Goldsmith (2002): every stem has at least one vowel
    - a stem is at least a syllable
  - We don't need this information:
    - stems: morphemes with a small signature and highly frequent elements
    - affixes: morphemes with a large signature and low frequent elements

Conclusions

- Gain:
  - extremely high precision suggests:
    - induced morphological structures are reliable cues
    - integration of morphological properties in the VS
  - “lemmatization” or rather “stemming” as a side effect improves the token statistics
- Expectation:
  - improved cluster purity!

Target Question

- Is the language input chaotic?
  - No!
- Can cues be identified as elementary properties of the speech signal?
  - Of course, and only these are the fundamental cues for bootstrapping!
- Can these cues serve as learning cues in an **incremental** algorithm?
  - In fact, the incremental nature is the big advantage of this system, because of the complexity reduction.

Target Question

- Do the observable statistical properties correlate with language properties?
  - Sure!
- Is there a plausible, formal model of bootstrapping?
  - I think so! It is the cue-based approach described here and argued for empirically!
  - And: P&P is not one!

Target Question

- Are children sensitive to frequency and do they really use frequency and entropy for induction?
  - I guess so, at least many experiments show that children do this in all kinds of domains, not just in language tasks, and adults do it as well.
- Does probability play a role in reasoning?
  - Of course!
- Do we use it in reasoning (plausibility vs. deductive reasoning)?
  - Sure!

Vector Space

- Representing elements in a vector space:
  - x = [2.0, 4.9, 12.4, ...]
  - Matrix:
    - row = elements
    - column = features
  - Representation in an *n*-dimensional space
  - Linear Algebra for analysis of vector similarity
  - Vector similarity for clustering, grouping, association

Clustering

- Data analysis:
  - Exploratory
    - Hypothesis creation
  - Confirmatory
    - Decision-making
- Grouping:
  - Is there a correlation between data patterns?
  - Which data patterns are similar?
    - Which words are similar?
    - What kind of constructions are similar?

Clustering

- Tryon (1939)
  - Unsupervised classification of observed data into groups (clusters).
    - No *a priori* hypothesis.
    - Grouping of objects or individuals.
    - Grouping of (random) variables.
- Use nowadays:
  - Medicine, Chemistry, Psychiatry, Linguistics, ...
  - Development of taxonomies
  - Dissection of a population
  - Identification of (potential) terrorists :-)

Clustering

- Good overview:
  - Everitt (1974) and Everitt et al (2003)
  - Unsupervised classification of observed data into groups (clusters).
    - No *a priori* hypothesis.
    - Grouping of objects or individuals.
    - Grouping of (random) variables.
- Use nowadays:
  - Medicine, Chemistry, Psychiatry, Linguistics, ...
  - Development of taxonomies
  - Dissection of a population
  - Identification of (potential) terrorists :-)

Clustering

- Objectives today:
  - Typology detection or identification.
  - Model Fitting.
  - Prediction based on groups.
  - Hypothesis testing.
  - Data exploration.
  - Hypothesis generating.
  - Data reduction.

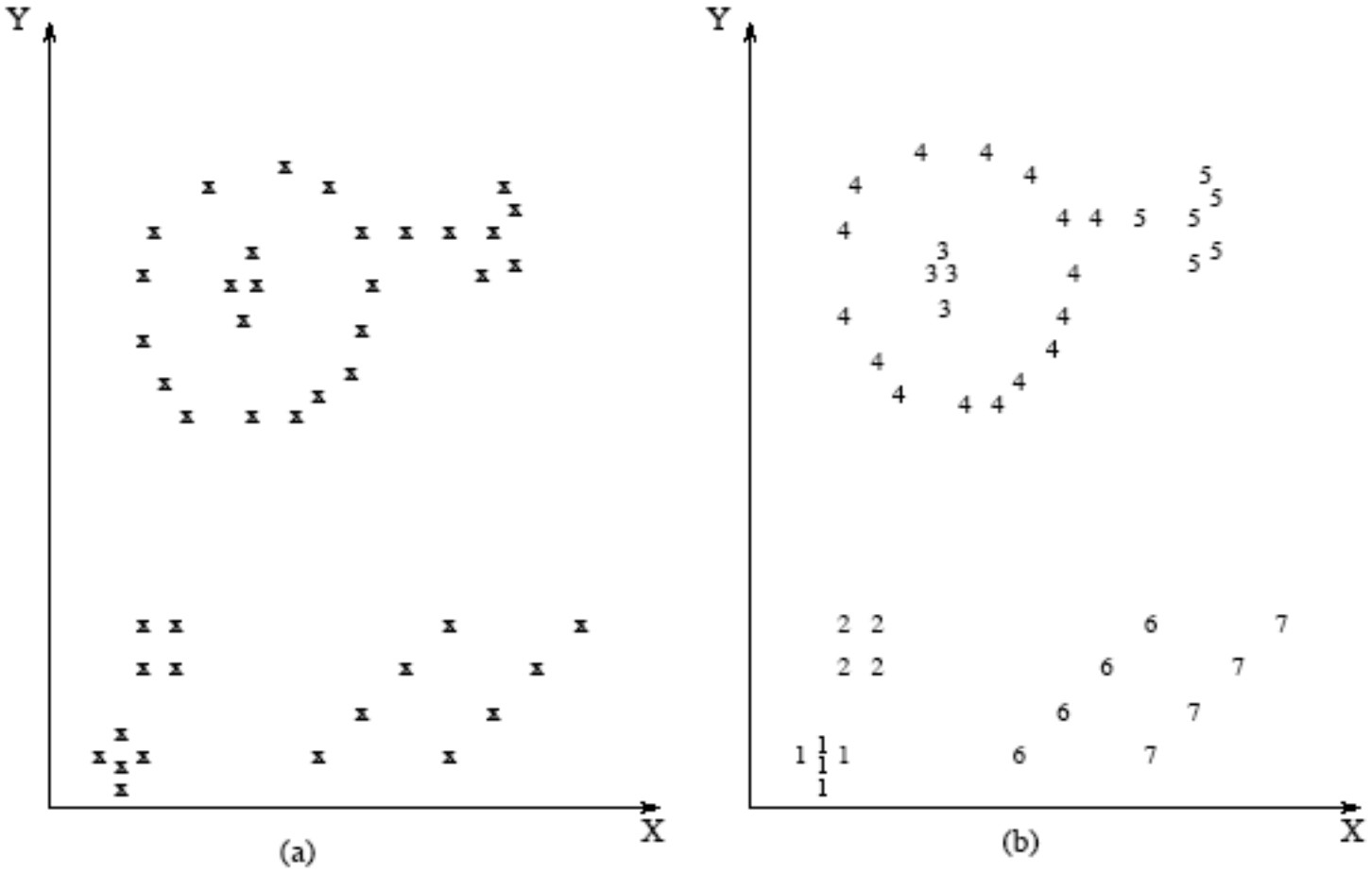
Clustering

- Different names used in the literature:
  - Q-analysis
  - Typology
  - Grouping
  - Clumping
  - Numerical taxonomy
  - Unsupervised pattern recognition

Clustering vs. Classification

- Classification:
  - Grouping on the basis of a priori labels
  - Discriminant analysis = supervised classification
  - Given a set of labeled patterns, label an unlabeled pattern
- Clustering:
  - Labeling of unlabeled data sets or patterns
  - Data-driven, not taxonomy driven = unsupervised
  - Labels are related to clusters
  - Cluster labels are obtained solely from data

Clustering



Clustering

- Prerequisites:
  - Representation of data (pattern and features)
  - Data or pattern proximity measure (domain dependent)
  - Clustering algorithm
- Representation of data:
  - pattern and features, graphically or as a vector space
  - Number of classes or clusters
  - Available and expected patterns
  - Features: number, type, scale
  - May partially be opaque or unknown, i.e. can be induced

Clustering

- Feature selection
  - Feature extraction
  - Identification of the subset of features that is most efficient for clustering.
  - Transformation of input features and creation of new salient features.
- Algorithms:
  - Input: Data selection and preparation, Feature selection and/or extraction
  - Evaluation: Proximity measures via clustering algorithm
  - Output: Taxonomy, Grouping, Clusters

Clustering

- The choice of pattern proximity measures is:
  - Domain or data dependent
  - Distance function defined on pairs of patterns
    - e. g. Euclidean distance or cosine similarity of vectors etc.
- Grouping
  - Hierarchical algorithms with nested groups
  - Overlapping groups
  - etc.

Clustering

- Extraction of data sets that are:
  - simple
  - compact
- Machine oriented:
  - efficiency
- Human or cognitively oriented:
  - intuitive and comprehensible

Clustering

- Pre-clustering evaluation:
  - Cluster tendency
- Post-clustering evaluation:
  - Cluster validity or purity
  - Rather subjective
  - Valid: if clusters are not the result of an artifact or randomly chosen.

Clustering

- Evaluation:
  - Cluster validity or purity
  - External assessment:
    - Compare recovered structure to some a priori structure or theory (e.g. lexicological models, psycholinguistic evidence)
    - Automatically compare taxonomies, hierarchical trees, distance of centroids etc.
  - Internal assessment:
    - Are resulting clusters intrinsically appropriate for the data.
  - Relative test:
    - Compare two resulting clusters and measure relative merit.

Clustering

- Clustering algorithms
  - Vast number
  - Selection on the basis of:
    - Way in forming clusters
    - Data-structure
    - Robustness (changes, data types)
    - Computational efficiency
    - Choice of similarity measure
    - Data amount (small, large)
    - Use of domain knowledge or heuristics
    - etc.

Clustering

- Types of algorithms and techniques:
  - Hierarchical
  - Optimization
    - K-means Clustering
    - Expectation Maximization (EM)
  - Density or mode-seeking
  - Clumping

Clustering

- Formalization:
  - Feature Vector, Datum, Pattern:
    - With d measurements: $x = (x_1, x_2, ..., x_d)$
    - x_1, x_2, ..., in general: x_i is a *feature* or *attribute* of x
    - d = *dimension* of pattern or pattern space
  - Pattern set:
    - $X = \{x_1, x_2, ..., x_n\}$
    - The i^{th} pattern in X: $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$
    - or

Vector Space

$$\mathscr{X} = \begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,d} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,d} \\ \vdots & & & \\ \mathbf{x}_{k,1} & \mathbf{x}_{k,2} & \cdots & \mathbf{x}_{k,d} \end{bmatrix}$$

Clustering

- Hard clustering techniques:
  - Assign a label l_i to each pattern x_i identifying its class.
  - For a set of patterns X the set of labels is $L = \{l_1, l_2, ..., l_n\}$ with $l_i \in \{1, ..., k\}$, with k the number of clusters.
- Fuzzy or soft clustering:
  - Assign each pattern x_i a fractional degree of membership f_{ij} in each output cluster j.

Centroid

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- *n* = number of dimensions
- Example:
  - a = (2, 2)
  - b = (3, 4)
  - centroid = ((2+3)/2, (2+4)/2) = (2.5, 3)

Distance

- Euclidean Distance
  - for two dimensions:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

  - for *n* dimensions:

$$d(a,b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Clustering

- Hierarchical clustering
- Optimization clustering
  - K-means
  - Expectation Maximization