Quantitative and qualitative computational analysis of language and text similarities, clustering and classification

Damir Ćavar

CLS 2010, September 2010

University of Zadar

Agenda

- Fuzzy geometrical approaches:
  - Clustering
- Comparing probability distributions
- Grammar induction

Fuzzy Clustering

Fuzzy Clustering

- In K-Means:
  - Assign every individual vector (representing a document with term frequencies or any other measure) to every centroid/cluster
  - Take the proportion of the distance to any centroid as representing some relative assignment likelihood

Fuzzy Clustering

- See also Expectation Maximization (EM)

Comparing Frequency Profiles

Kullback–Leibler divergence

- We can calculate the number of bits that we need to encode some strings with individually specific distributional probabilities (extracted from a corpus)
- We can compare two distributions wrt. the amount of memory they require (the closer the distributions, the smaller the difference of the encoding in bits)

Kullback–Leibler divergence

- Definition:

$$D_{KL}(P \parallel Q) = \sum_i P(i) log_2 \frac{P(i)}{Q(i)}$$

- We can compare the distance between two distributions (e.g. frequency profiles of N-grams)
- The smaller D_{KL}, the more similar two distributions are.

Kullback–Leibler divergence

- See code example: kld.py

Kullback–Leibler divergence

- Grammar = Compression
  - Symbolic
  - Probabilistic
- Example:
  - Grammar Induction or Language Learning Models
- Minimum Description Length Principle (MDL), Kolmogorov Complexity

Research Examples

Lexical Induction Example

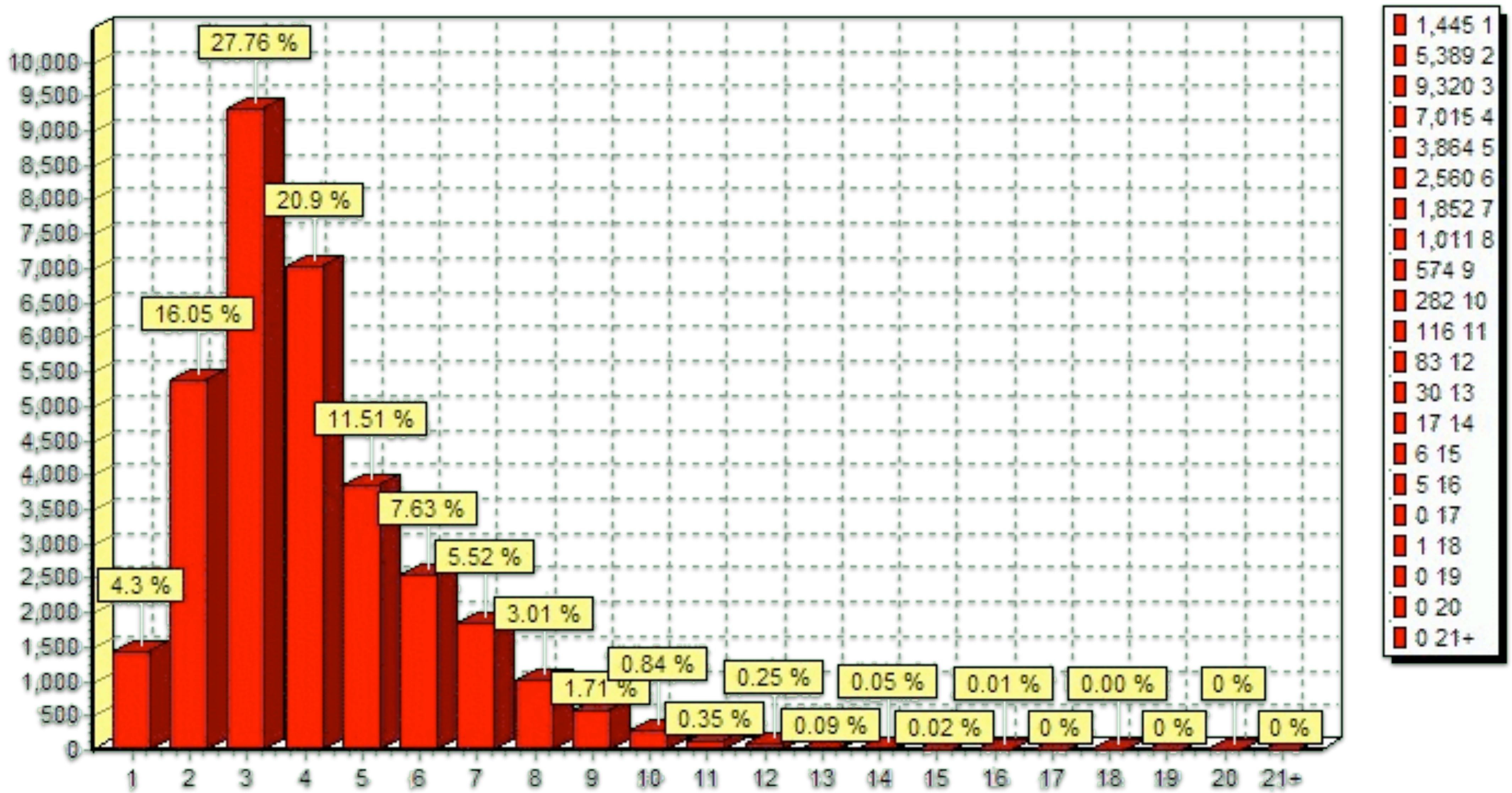
- Distributional properties of lexical items:
  - Expectations for:
    - X the Y
    - X on Y
    - X say Y
    - ...
  - Expectations are usually directional, i.e. X or Y is expected to be a certain token, category etc.

Lexical Induction Example

- Distributional properties of lexical items:
  - Expectations for:
    - X dog Y
    - X rains Y
    - X calls Y
    - ...

Distributional properties of terms

- Mapping of term length to frequency



Friday, September 3, 2010

Distributional properties of terms

- 49 most frequent words:
- THE, AND, OF, TO, A, HE, HIS, IN, THAT, WITH, HIM, WAS, IT, I, HER, FOR, IS, ME, HAD, THEY, BUT, ON, AS, AT, SHE, NOT, FROM, THEIR, SAID, THOU, THEM, THEE, WHEN, WHO, WERE, SO, HAVE, LITTLE, OUT, YOUNG, MY, BY, BE, SOUL, THERE, CAME, THIS, WILL, INTO

Lexical Induction Example

- Local distributional properties match with specific lexical properties
  - Map distributional properties on a vector space (left and right context)
  - Prominence of function words: indicating syntactic structure, co-occurring with categories etc.

Function Words

- Invariant part of the mental lexicon
- Highly frequent
- Functioning with placement restrictions and contextual constraints
- Coding fundamental grammatical properties, but being semantically vacuous

Lexical Differences

- Frequency
  - Function words are highly frequent (cross-linguistically)
  - Substantives are less frequent (cross-linguistically)
  - Highly frequent term tend to be shorter (remember the Entropy effect?)

Clustering Lexical Items

- Clustering algorithms:
  - k-means
  - Expectation Maximization (EM)
- Clustering words from child oriented speech in Peter corpus (Bloom, 1970) (CHILDES):
  - binary clustering
  - features: [frequency, length]

Clustering Lexical Items

- Clustering results (k-means, iterative subclustering):
  - 1. ['the', 'it', 'you']
  - 2. ['here', 'me', 'want', 'one', 'do', 'is', 'in', 'right', 'no', 'did', 'can', 'not', 'think', 'that', 'and', 'see', 'gonna', 'on', 'ok', 'oh', 'your', 'to', 'what', 'a', 'its', 'put', 'are', 'go', 'thats', 'this', 'mmhm', 'there', 'have', 'I', 'well']
  - 3. all other tokens

Clustering Lexical Items

- Weaknesses:
  - Intrinsic features alone are insufficient.
- Clustering on intrinsic and extrinsic features is more promising.

Clustering Lexical Items

- 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).

Clustering Lexical Items

- 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 input cues is learnable and that make language acquisition possible (distributional properties and individual tokens)

Clustering Lexical Items

- The set of cues is K = $\{w_1, \ldots, w_m\}$, such that if we add up the number of words X_1, that co-occur with w_1 and the number of words X_2, that co-occur with w_2, until the m-most frequent word, w_m, the number of words

$$\sum_{i=1}^{m} X_i$$

- converges to an order α (= 1, 2, 3 ...), of n, where n is the number of word types in corpus R.

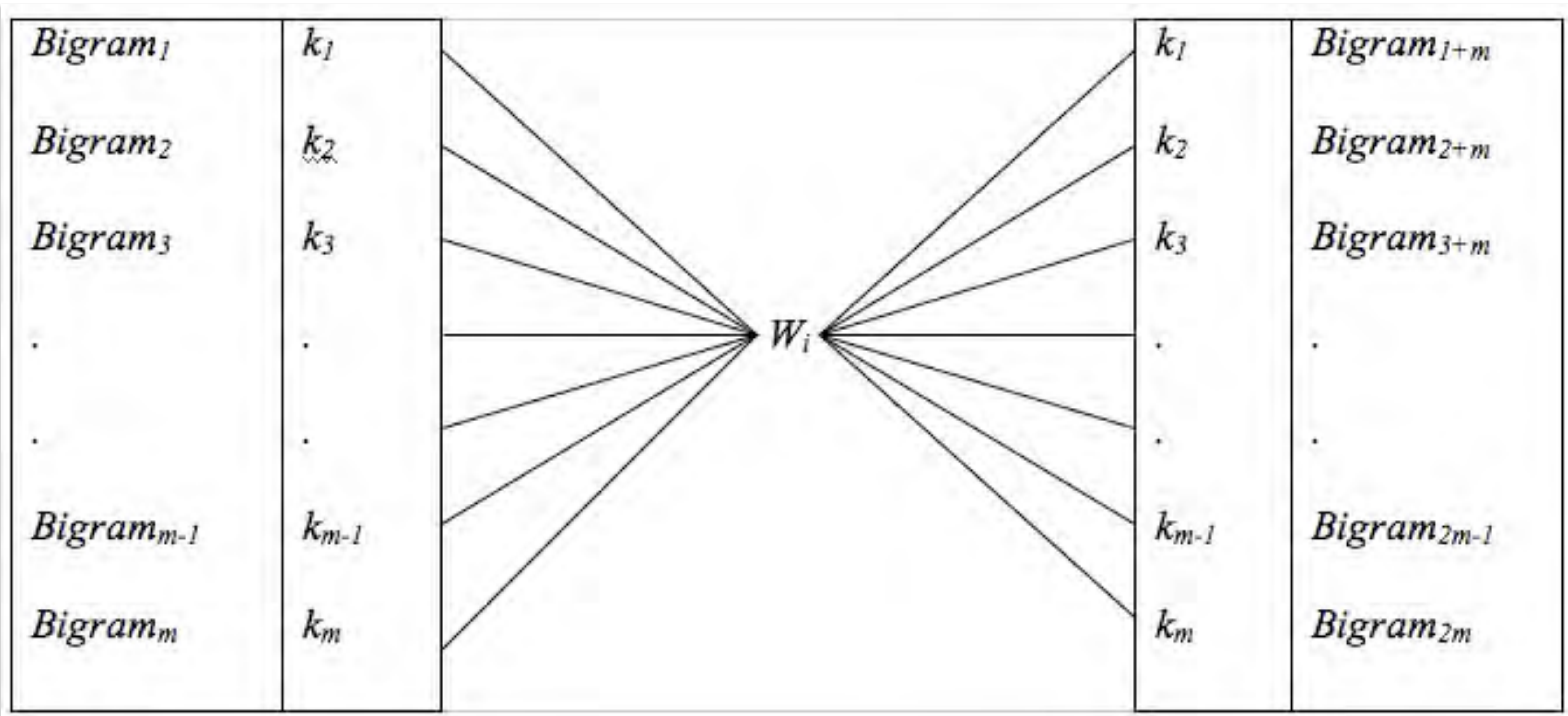
Clustering Lexical Items

- Variant 2 of 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, ... w_{43}] = 80%

Clustering Lexical Items

- Variant 2 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']

Lexical Vector Space



Clustering Experiments with Child-oriented Speech

- Mintz ea. (2002) The distributional structure of grammatical categories in speech to young children. *Cognitive Science* 26, 393-424.

Clustering Experiments with Child-oriented Speech

- Linguistic environment of the language learner
- Properties of her computational and representational system.
- Lexical acquisition is related to input.
- Other aspects are internal.

Clustering Experiments with Child-oriented Speech

- Acquisition of major categories
- Verbs, Nouns
  - Universal and fundamental primitives for grammar
- Two models:
  - Semantic
  - Innate

Clustering Experiments with Child-oriented Speech

- Semantic:
- From world observation of referent
  - Concrete object: noun
  - Action or event: verb
- Problem:
  - Abstract concepts and events

Clustering Experiments with Child-oriented Speech

- Some solution suggestions:
  - Generalization from non-prototypical nouns and verbs based on overlapping semantic features shared with prototypical ones.
- Alternative:
  - Distributional similarities

Clustering Experiments with Child-oriented Speech

- Innate lexical specification:
  - Lexical categories as atomic grammar elements are specified
  - Lexical items have to be classified on the basis of this innate taxonomy
  - Various bootstrapping approaches

Clustering Experiments with Child-oriented Speech

- Semantic bootstrapping (innateness)
  - using semantic-syntactic correspondence
  - augmented by distributional properties
- Prosodic bootstrapping
  - using phonological-syntactic correspondence

Clustering Experiments with Child-oriented Speech

- Alternative:
  - Distributional properties
  - Similarities of patterns are mapped on lexical similarity
  - Categories are derived from such similarities

Clustering Experiments with Child-oriented Speech

- Classical criticism:
  - Pinker & Chomsky:
    - Distributional properties in the sense of substitutability might over- and under-generalize
- Consequences:
  - Abandoned: distributional approaches

Clustering Experiments with Child-oriented Speech

- Mintz' approach:
  - Distribution with one context word left and right only
  - Expanding the window to two words and eight words left and right
  - That is: a matrix of n = number of words (rows) times 2 * n (columns)

Clustering Experiments with Child-oriented Speech

- Purpose:
  - Investigate the effect of the context size on categorization
- Evaluation:
  - Compare the categorization with same categorization on randomly generated corpora given the extracted tokens

Clustering Experiments with Child-oriented Speech

- Further settings:
  - Restrict the context to syntactic structure (phrases and phrase boundaries)
  - Reduce representations of elements in the input (assuming that young children do not process this)

Clustering Experiments with Child-oriented Speech

- Child-oriented speech from the CHILDES database
- Utterances directed to children below 2.5
- with 2.5 children already produce utterances that display syntax and lexical knowledge
- Testset: 14,167 utterances

Clustering Experiments with Child-oriented Speech

- Selection of words for the analysis:
  - 200 most frequent (actually less than 200, see footnote)
  - Argument:
    - these words represent 80% of the tokens
    - less frequent words are too low frequent

Clustering Experiments with Child-oriented Speech

- Counting:
  - for every word
  - for every other word
  - how many times does it occur left and right
- Example: *John likes port*

Clustering Experiments with Child-oriented Speech

- Matrix size:
  - for one neighbor context 200 x 400
  - 200 x 800, 200 x 3200
- Example: w1
  - left: [w2:fr, w3:fr, w4:fr, ...]
  - right: [w2:fr, w3:fr, w4:fr, ...]

Clustering Experiments with Child-oriented Speech

- Cosine similarity:
  - Used in document similarity measure
  - Extraction of keywords and their frequencies
  - Each document is represented as a vector with the frequencies of all extracted keywords

Clustering Experiments with Child-oriented Speech

- Measure Cluster purity (based on a given tagged corpus, e.g. Childes)
- Results are very good, even for larger sets of tokens in a corpus, even for other corpus types (not just Child oriented speech)

References

- see some work by Lillian Lee (also her PhD), Sabine Schulte im Walde, Mintz and Newport etc.