# EM-Based Clustering

Detlef Prescher

DFKI Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
prescher@dfki.de

March 1, 2002

# Overview of the Talk

- **Part I**:

  Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of ACL-99*.

- **Part II**:

  Detlef Prescher. Novel Properties and Well-Tried Performance of EM-Based Multivariate Clustering. In *Proceedings of RANLP-01*.

# Part I — Overview

- Motivation
- Inducing selection classes via EM-based clustering
  - Theory
  - Clustering examples
- Inducing class labels for subcat-slots via EM-estimation
  - Theory
  - Experiments with English and German data
  - Linguistic evaluation
- Conclusion

# Part I — Motivation

**Aim:**

Induction of *semantic class labels for slots of subcategorization frames* from *unannotated* corpora, i.e., no taxonomy of semantic classes required, rather semantic classes are tailored to unannotated data from arbitrary languages.

**Method:**

- Collect data tuples consisting of verbs annotated with subcat-frames and their nominal fillers by parsing large corpora,
- Induce semantic selection classes for verb-frames and nouns via EM-based clustering,
- Induce class labels for subcategorization slots of given verb-frames by further application of EM to noun-filler data.

# Part I — Inducing Selection Classes via EM-Based Clustering

Given: Data tuples $<v, n>$ of intransitive and transitive verb-frames $v$ and their subjects and objects $n$.

Aim: Smooth class-based probability model on data, together with well-defined and simple method for statistical inference from unannotated data.

**Class-based selection model:**

$$p(v, n) = \sum_{c \in C} p(c, v, n) = \sum_{c \in C} p(c) p(v, n|c) = \sum_{c \in C} p(c) p(v|c) p(n|c)$$

parameterized by $p(c)$, $p(v|c)$, and $p(n|c)$.

# Part I – EM-Estimation for Class-Based Selection Model

Let $y = <v, n>$ and $x = <c, y>$ be fixed, and $f(y)$ be the sample frequency of $y$. The **EM-estimates** are given as:

$$
\begin{aligned}
\hat{p}(v|c) &= \frac{\sum_{y \in \{v\} \times N} f(y) p(x|y)}{\sum_y f(y) p(x|y)}, \\
\hat{p}(n|c) &= \frac{\sum_{y \in V \times \{n\}} f(y) p(x|y)}{\sum_y f(y) p(x|y)}, \\
\hat{p}(c) &= \frac{\sum_y f(y) p(x|y)}{\sum_y f(y)}.
\end{aligned}
$$

According to the EM-theory, each estimation step increases the likelihood $L(p) = \prod_y p(y)^{f(y)}$ of the data, and a sequence of estimates converges (under weak conditions) to a local maximum of $L(p)$.

# Part I — "Communicative Action" (English Data)

| Class 5 PROB 0.0412 | | 0.0148 | 0.0084 | 0.0082 | 0.0078 | 0.0074 | 0.0071 | 0.0054 | 0.0049 | 0.0048 | 0.0047 | 0.0046 | 0.0041 | 0.0040 | 0.0040 | 0.0039 | 0.0039 | 0.0039 | 0.0039 | 0.0038 | 0.0038 | 0.0038 | 0.0037 | 0.0035 | 0.0035 | 0.0035 | 0.0034 | 0.0033 | 0.0033 | 0.0033 | 0.0033 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | man | ruth | corbett | doctor | woman | athelstan | cranston | benjamin | stephen | adam | girl | laura | maggie | voice | john | harry | emily | one | people | boy | rachel | ashley | jane | caroline | jack | burun | juliet | blanche | helen | edward |
| 0.0542 | ask.as:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● |
| 0.0340 | nod.as:s | | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | ● | | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 0.0299 | think.as:s | ● | ● | ● | ● | ● | ● | | | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● |
| 0.0287 | shake.aso:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● |
| 0.0264 | smile.as:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● |
| 0.0213 | laugh.as:s | ● | ● | | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 0.0207 | reply.as:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | ● | | | | ● | ● |
| 0.0167 | shrug.as:s | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | ● | ● | | | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● |
| 0.0148 | wonder.as:s | ● | ● | ● | ● | ● | ● | | | ● | ● | ● | | | | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 0.0141 | feel.aso:s | ● | | ● | ● | ● | | | | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● |
| 0.0133 | take.aso:s | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● |
| 0.0121 | sigh.as:s | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | | | ● | ● | ● | | ● | | ● | ● | ● | ● | ● | ● | | ● | ● | ● |
| 0.0110 | watch.aso:s | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | | ● | | ● | ● | ● | ● | ● | ● | ● | | ● | | ● | ● | ● | | ● | |
| 0.0106 | ask.aso:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | | ● | | ● | | | ● | ● | ● |
| 0.0104 | tell.aso:s | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| 0.0094 | look.as:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | | ● | | ● | ● | ● | ● | ● | | | | ● | ● | ● | | | ● |
| 0.0092 | give.aso:s | ● | ● | | ● | ● | | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |

# Part I — “Scalar Change” (English Data)

| **Class 17** PROB 0.0265 | | 0.0379 | 0.0315 | 0.0313 | 0.0249 | 0.0164 | 0.0143 | 0.0110 | 0.0109 | 0.0105 | 0.0103 | 0.0099 | 0.0091 | 0.0089 | 0.0088 | 0.0082 | 0.0077 | 0.0073 | 0.0071 | 0.0071 | 0.0070 | 0.0068 | 0.0067 | 0.0065 | 0.0065 | 0.0058 | 0.0057 | 0.0057 | 0.0054 | 0.0051 | 0.0050 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | number | rate | price | cost | level | amount | sale | value | interest | demand | chance | standard | share | risk | profit | pressure | income | performance | benefit | size | population | proportion | temperature | tax | fee | time | power | quality | supplely | money |
| 0.0437 | increase.as:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | |
| 0.0392 | increase.aso:o | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | ● | ● | ● | ● | ● |
| 0.0344 | fall.as:s | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | | ● | ● | ● | ● | | | ● | ● | ● | ● | | ● | | ● | ● | |
| 0.0337 | pay.aso:o | ● | ● | ● | | ● | ● | | ● | ● | ● | | | ● | | | | ● | ● | ● | | | ● | | | ● | ● | | | | ● |
| 0.0329 | reduce.aso:o | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | |
| 0.0257 | rise.as:s | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | |
| 0.0196 | exceed.aso:o | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | | | ● | | ● | | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | |
| 0.0177 | exceed.aso:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | | ● | ● |
| 0.0169 | affect.aso:o | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | ● | ● | ● | |
| 0.0156 | grow.as:s | ● | | | | | | ● | ● | ● | ● | | | ● | ● | ● | ● | ● | | ● | ● | ● | ● | | | | | ● | | | ● |
| 0.0134 | include.aso:s | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | | | ● | ● | ● | ● | ● | ● |
| 0.0129 | reach.aso:s | ● | ● | ● | ● | ● | ● | ● | | ● | ● | | | ● | | ● | ● | | ● | ● | | ● | ● | ● | | | ● | ● | ● | ● | ● |
| 0.0120 | decline.as:s | ● | ● | ● | ● | ● | | ● | ● | ● | ● | | ● | ● | ● | ● | | | | | ● | ● | ● | ● | | | | | ● | ● | |
| 0.0102 | lose.aso:o | ● | ● | | ● | | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | | ● | | ● | ● | ● | | ● | ● | ● | ● | | ● |
| 0.0099 | act.aso:s | | ● | ● | | | | ● | | | | | ● | | | | | | | ● | | | | | ● | | | ● | | ● | |
| 0.0099 | improve.aso:o | ● | ● | | | ● | | | ● | | | ● | ● | ● | | ● | | | ● | | | | | | | | ● | ● | ● | | |

# Part I — “Scalar Change” (German Data)

| **Class 26** PROB 0.0223 | | 0.0379 | 0.0226 | 0.0182 | 0.0171 | 0.0137 | 0.0133 | 0.0129 | 0.0086 | 0.0079 | 0.0079 | 0.0073 | 0.0072 | 0.0071 | 0.0067 | 0.0064 | 0.0063 | 0.0061 | 0.0057 | 0.0056 | 0.0053 | 0.0051 | 0.0047 | 0.0046 | 0.0039 | 0.0038 | 0.0037 | 0.0037 | 0.0036 | 0.0036 | 0.0035 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ergebnis | Preis | Menge | Anteil | Stück | Zahl | Gewinn | Kritiker | BürgerMeister | Angst | Umsatz | Einnahme | Zins | Schicksal | Bus | NachFrage | Ertrag | Figur | Verlust | Frucht | ArbeitsLosigkeit | Kost | Last | WahlErgebnis | Temperatur | Solidarität | InflationsRate | Abschluß | Import | unterSchrift |
| 0.0601 | liegen-VPA.np:NP.Nom | • | • | • | • | • | • | • | | • | | | | • | • | | | • | | • | | | • | | | • | | • | | | |
| 0.0351 | tragen-VPA.na:NP.Akk | | | • | • | | | | | | | | | • | | | | | | • | • | | • | • | | | | | | | |
| 0.0230 | steigen-VPA.n:NP.Nom | | • | • | • | | • | | | | | • | • | • | | | • | | | • | | • | | | | • | | • | | | |
| 0.0213 | sagen-VPA.n:NP.Nom | | | | | | | | • | • | | | | | | | | | | | | | | | | | | | | | |
| 0.0182 | gering-ADJ | | • | • | • | | • | | | | | | • | • | | | | • | | • | | • | • | | | • | | • | | | |
| 0.0135 | sinken-VPA.n:NP.Nom | | • | | • | | | • | | | | | • | • | | | • | • | | | | | | | | • | | | | | |
| 0.0135 | steigen-VPA.np:NP.Nom | | | | • | | • | | | | | | | | | | • | • | | • | | • | | | | • | | | | | |
| 0.0119 | erklären-VPA.n:NP.Nom | | | | | | | | | • | | | | | | | | | | | | | | | | | | | | | |
| 0.0100 | positiv-ADJ | • | | | | | | | | | | | | | | | | • | | | | | | | | | | | | | |
| 0.0091 | zurück+gehen-VPA.np:NP.Nom | | | | • | | • | • | | | | • | • | | | | | • | | | | | | | | | | • | | • | |
| 0.0087 | sinken-VPA.np:NP.Nom | | | | • | | • | | | | | • | | • | | | | | | | | | | | | • | • | | | • | |
| 0.0082 | spät-ADJ | | | | | | | | | | | | • | | | | | • | | | | | | | | | | | | | |
| 0.0077 | wirken-VPA.n:NP.Nom | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 0.0071 | ändern-VPA.na:NP.Akk | | | | | | | | | | | | | | • | | | | | | | | | | | | | | | | |

# Part I — Inducing Class Labels for Subcat-Slots using EM-Estimation

Given: Selection Class Model $p_{\text{cluster}}(\cdot)$ on $<v, n>$ data as before; sample $n_1, n_2, \ldots, n_M$ of nominal fillers of subject slot of fixed intransitive verb.

Aim: Tune class-parameters of probability model on noun data to nominal fillers of fixed verbs.

**Class-based slot-labeling model:**

$$p(n) = \sum_{c \in C} p(c, n) = \sum_{c \in C} p(c) \cdot p_{\text{cluster}}(n|c)$$

parameterized by $p(c)$.

# Part I — EM-Mapping for Class-Based Slot Labeling Model

**EM-Mapping:**

$$\hat{p}(n) = \frac{\sum_{n \in N} f(n) p(c|n)}{\sum_{n \in N} f(n)}$$

According to EM-theory, a sequence of such estimates converges (under weak conditions) to a local maximum of the likelihood $L(p) = \prod_n p(n)^{f(n)}$ of the sample of subject fillers.

**Output:** List of slot labels for a fixed verb frame sorted according to $p(c)$; List of nominal fillers for slots of a fixed verb frame sorted according to $f(n)p(c|n)$.

# Part I — Experiments with English Data

Class 5 ("communicative action") labeling subject-slot of intransitive *blush* and *snarl* together with expected noun-fillers:

| *blush* 5 | 0.982975 | *snarl* 5 | 0.962094 |
|---|---|---|---|
| constance | 3 | mandeville | 2 |
| christina | 3 | jinkwa | 2 |
| willie | 2.99737 | man | 1.99859 |
| ronni | 2 | scott | 1.99761 |
| claudia | 2 | omalley | 1.99755 |
| gabriel | 2 | shamlou | 1 |
| maggie | 2 | angalo | 1 |
| bathsheba | 2 | corbett | 1 |
| sarah | 2 | southgate | 1 |
| girl | 1.9977 | ace | 1 |

# Part I — Experiments with English Data

Class 17 ("scalar change") labeling subject-slot of intransitive and object-slot of transitive *increase*:

| *increase* $17$ | 0.923698 | | |
|---|---|---|---|
| number | 134.147 | proportion | 23.8699 |
| demand | 30.7322 | size | 22.8108 |
| pressure | 30.5844 | rate | 20.9593 |
| temperature | 25.9691 | level | 20.7651 |
| cost | 23.9431 | price | 17.9996 |

| *increase* $(8, 17)$ | 0.3097650 |
|---|---|
| development - pressure | 2.3055 |
| fat - risk | 2.11807 |
| communication - awareness | 2.04227 |
| supplementation - concentration | 1.98918 |
| increase - number | 1.80559 |

# Part I — Experiments with English Data

Grouping of intransitives taking class 17 as best subject-label:

| | |
|---|---|
| 0.977992 decrease | 0.560727 drop |
| 0.948099 double | 0.476524 grow |
| 0.923698 increase | 0.42842 vary |
| 0.908378 decline | 0.365586 improve |
| 0.877338 rise | 0.365374 climb |
| 0.876083 soar | 0.292716 flow |
| 0.803479 fall | 0.280183 cut |
| 0.672409 slow | 0.238182 mount |
| 0.583314 diminish | |

Comparison to Levin's 48 top-level verb classes shows an agreement with her class of "verbs of change of state", except for last three verbs. Furthermore, most verbs take class 17 as best intransitive subject-label and as best transitive object-label, pointing at causative alternation.

# Part I — Experiments with German Data

Class 26 labeling subject-slot of intransitive verb “steigen” (rise) and object-slot of transitive verb “senken”(lower):

| *steigen* 26 | 0.67273 | (rise) |
|---|---|---|
| Zahl | 23.333 | (number) |
| Preis | 15.895 | (price) |
| ArbeitsLosigkeit | 10.8788 | (unemployment) |
| Lohn | 9.72965 | (wage) |
| NachFrage | 6.83619 | (demand) |

| *senken* $(14, 26)$ | 0.450352 | (lower) |
|---|---|---|
| BundesBank - LeitZins | 5.81457 | (Federal bank - base rate) |
| BundesBank - Zins | 2.97838 | (Federal bank - interest) |
| superMarkt - Preis | 1 | (super market - price) |
| SommerGeschft - Verlust | 1 | (summer business - loss) |
| BundesBank - DiskontSatz | 0.99999 | (Federal bank - minimum lending rate) |

# Part I — Experiments with German Data

Grouping of intransitives taking class 26 as best subject-label:

| | |
|---|---|
| 0.741467 ansteigen | (go up) |
| 0.720221 steigen | (rise) |
| 0.693922 absinken | (sink) |
| 0.656021 sinken | (go down) |
| 0.438486 schrumpfen | (shrink) |
| 0.375039 zurckgehen | (decrease) |
| 0.316081 anwachsen | (increase) |
| 0.215156 stagnieren | (stagnate) |
| 0.160317 wachsen | (grow) |
| 0.154633 hinzukommen | (be added) |

Comparison to Schuhmacher's verb classification shows correspondence with class of "einfache Änderungsverben" (simple verbs of change of state) except for *anwachsen* (increase) and *stagnieren* (stagnate).

# Conclusion

Proposed algorithms for clustering and slot labeling from unannotated data are

- well-founded in the theory of parametrized statistical inference,
- efficient enough to map very large corpora of observations of frames and their complements to structured lexical entries,
- applicable to any natural language where text samples of sufficient size together with a robust parser are available.

# Part II — Overview

- Motivation
- Theory
  - Probabilistic Multivariate Clustering
  - Simplified EM-Based Re-Estimation Formulas
  - Invariance Property
- Experiments
  - Pruning
  - Smoothing
- Well-Tried Performance
- Discussion and Conclusion

# Part II — Motivation

- **Aim:** Unsupervised induction of probabilistic classes for arbitrary linguistic data, e.g. semantic labels for verb-argument pairs or phonological labels for onset-nucleus-coda triples, to
  - automatically reveal the **hidden structural information** in a given linguistic corpus,
  - obtain a **smooth model** on both the data types **in** the corpus and the reasonable data types **not in** the corpus.

- **Constraints:** In detecting and generalizing information, induced probabilistic clustering models should
  - preserve the main characteristics of a given **real-world corpus**, and
  - have a good performance when utilized for **real-world applications**.

# Probabilistic Multivariate Clustering

The probability of a $d$-dimensional data type $y = (y_1, .., y_d) \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_d$ is defined as:

$$p(y) = \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) \cdot p(y|c) = \sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c) \ .$$

The **independence assumption**

$$p(y|c) = \prod_{i=1}^{d} p(y_i|c)$$

for each class $c$ makes clustering feasible since an independence assumption is not adequate for the whole sample, and therefore **forces the sample to break into clusters**.

# Multivariate Clustering via Regular Grammars

Given the simple **probabilistic regular grammar**

$$\begin{array}{lll} S \rightarrow Y_1^c \ldots Y_i^c \ldots Y_d^c & \qquad & p(c) \\ Y_i^c \rightarrow y_i & & p(y_i|c) \end{array}$$

with starting symbol $S$, non-terminals $Y_i^c$ (with $c \in C$ and $i = 1 \ldots d$) and terminals $y_i \in \mathcal{Y}_i$, a "sentence" $y_1 \ldots y_i \ldots y_d$ has exactly $|C|$ "syntax trees":

$$\begin{array}{c} S \\ \begin{array}{ccccc} Y_1^c & \ldots & Y_i^c & \ldots & Y_d^c \\ y_1 & \ldots & y_i & \ldots & y_d \end{array} \end{array}$$

Each "syntax tree" has the probability of $p(c) \prod_{i=1}^{d} p(y_i|c)$ yielding a "sentence probability" of $\sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c)$ which is equal to the probabilty of the type $y_1 \ldots y_d$ given the multivariate clustering model.

# EM-Based Multivariate Clustering

Using the well-known EM-based re-estimation formulas for probabilistic context-free grammars, simple **EM-based re-estimation formulas** for probabilistic multivariate clustering can be derived:

$$\begin{aligned}\hat{p}(c) &= |f|^{-1} \cdot |f_c|, \\ \hat{p}(y_i|c) &= |f_c|^{-1} \cdot f_c(y_i),\end{aligned}$$

where the total frequencies $|f| := \sum_y f(y)$ and $|f_c| := \sum_y f_c(y)$ are computed using the sample frequencies $f(y)$ and **class-based frequencies**

$$f_c(y) = f(y) \cdot p(c|y),$$

and $f_c(y_i) := \sum_{y \in \mathcal{H}(y_i)} f_c(y)$ is a **marginal class-based frequency** characterized by the **hyper plane** $\mathcal{H}(y_i)$ defined as the set of data types $y$ with fixed coordinate $y_i$.

# Toy-Example: EM-Based 2-dimensional Clustering

**Training corpus**: given via the so-called **empirical distribution**

$$\tilde{p}(y_1, y_2) := |f|^{-1} \cdot f(y_1, y_2).$$

| $\tilde{p}(y_1, y_2)$ | smile | laugh | increase | fall |
|---|---|---|---|---|
| man | 0.2 | | | |
| woman | 0.2 | 0.1 | | |
| number | | | 0.1 | |
| price | | | 0.3 | 0.1 |

**Parameters** of trained model (randomly initialized, 10 re-estimation steps):

| | | | | | |
|---|---|---|---|---|---|
| class 0 | 0.5 | woman 0.6 | man 0.4 | smile 0.8 | laugh 0.2 |
| class 1 | 0.5 | price 0.8 | number 0.2 | increase 0.8 | fall 0.2 |

# Toy Example: Empirical versus Model Probabilties

| $\tilde{p}(y_1, y_2)$ | smile | laugh | increase | fall | $\tilde{p}(y_1)$ |
|---|---|---|---|---|---|
| man | 0.2 | | | | 0.2 |
| woman | 0.2 | 0.1 | | | 0.3 |
| number | | | 0.1 | | 0.1 |
| price | | | 0.3 | 0.1 | 0.4 |
| $\tilde{p}(y_2)$ | 0.4 | 0.1 | 0.4 | 0.1 | |

| $p(y_1, y_2)$ | smile | laugh | increase | fall | $p(y_1)$ |
|---|---|---|---|---|---|
| man | 0.16 | 0.04 | | | 0.2 |
| woman | 0.24 | 0.06 | | | 0.3 |
| number | | | 0.08 | 0.02 | 0.1 |
| price | | | 0.32 | 0.08 | 0.4 |
| $p(y_2)$ | 0.4 | 0.1 | 0.4 | 0.1 | |

Comparing the empirical distribution (left) with the model distribution (right), two events attract attention:

- **smoothing takes place**: *(man, laugh)* and *(number,fall)* but not for e.g. *(man, increase)* or *(number,smile)*,

- **invariant marginal distributions** occure, i.e. $\tilde{p}(y_i) = p(y_1)$ $(i = 1, 2)$. **EM-based multivariate clustering preserves** the marginal distributions and thus some **characteristical properties of the given corpus**.

# Invariant Marginal Distributions

**Theorem:** EM-based multivariate clustering leads to invariant marginals.
**Proof:** Using EM-based re-estimated probabilities $(i = 1 \ldots d)$:

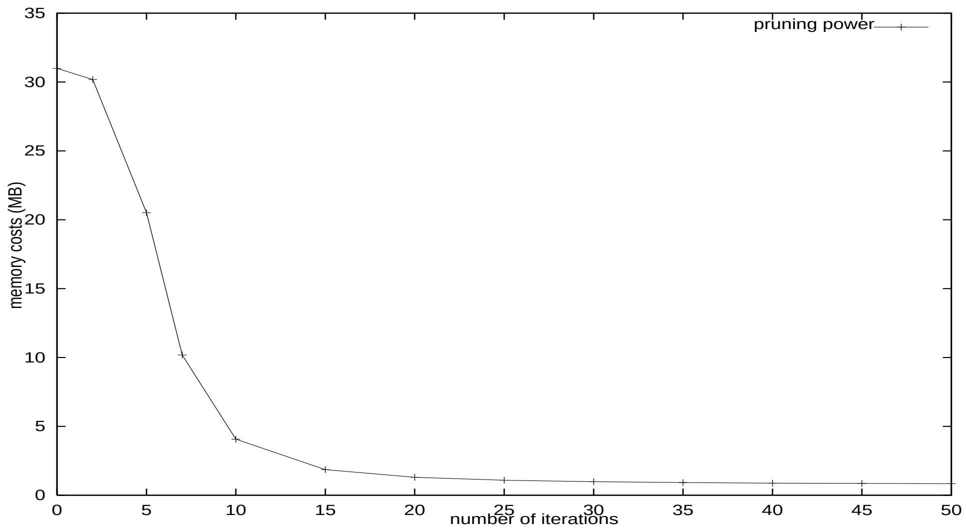
$$
\begin{aligned}
\hat{p}(y_i) &= \sum_{c \in C} \hat{p}(c) \cdot \hat{p}(y_i|c) \\
&= \sum_{c \in C} |f|^{-1} \cdot |f_c| \cdot |f_c|^{-1} \cdot f_c(y_i) \\
&= \sum_{c \in C} |f|^{-1} \sum_{y \in \mathcal{H}(y_i)} f_c(y) \\
&= \sum_{y \in \mathcal{H}(y_i)} |f|^{-1} \cdot f(y) \sum_{c \in C} p(c|y) \\
&= \sum_{y \in \mathcal{H}(y_i)} \tilde{p}(y) \qquad = \tilde{p}(y_i) \qquad \textbf{q.e.d.}
\end{aligned}
$$

# Pruning Experiments

**Pruning:** after each re-estimation step, we deleted all data types with **infrequent class-based frequencies** to decrease the required storage space for EM-based multivariate clustering models:

$$f_c(y) < \epsilon \cdot |C|^{-1} \qquad (\epsilon := 0.001 \text{ in the implementation}).$$

**Results:** for a model with 35 classes of 27 000 verb-object pairs (WSJ)

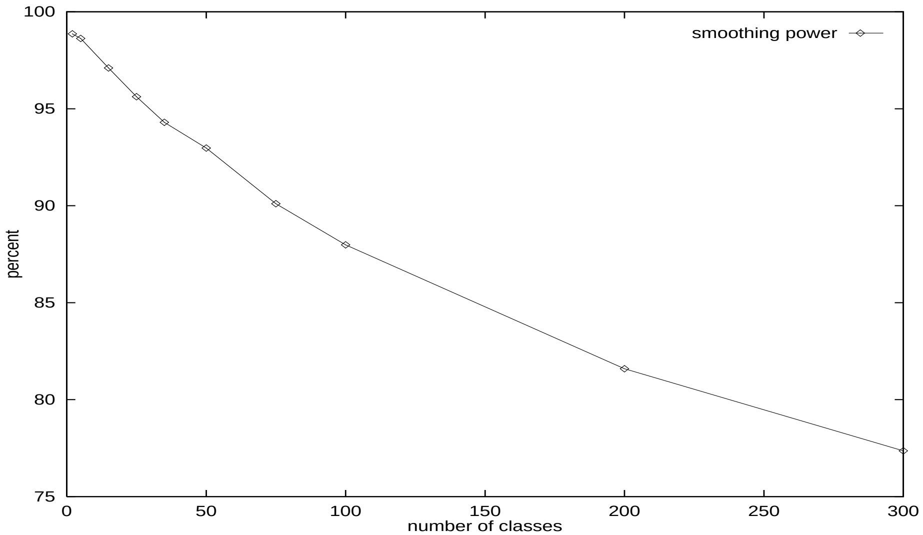


# Smoothing Experiments

**Smoothing power**: percentage of data types having a positive probability.
We expect for **EM-based clustering models**:

- One-class models are **perfect smoothers** since $p(y) = \prod_i \tilde{p}(y_i)$
- In general, **good smoothing behaviour** since $p(y) = \sum_c p(c) \cdot p(y|c)$
- **Pruning decreases the smoothing power** since it delets data

**Results:** for models trained on 1 280 000 verb-noun pairs (BNC) with an initial smoothing power of $0.1\%$



# Well-Tried Performance

**Well-Tried Performance**: EM-Based Multivariate Clustering (EMMC) was evaluated in **real-world applications** on a large number of randomly selected examples of a **real-world corpus**. Here is an overview:

| **NLP application** | **Improvement of performance using EMMC** |
|---|---|
| Induction of semantically annotated lexicons | $\infty$ |
| Identification of collocations/idioms | 0% |
| Grapheme-to-phoneme conversion | 3% (30%) |
| Machine translation | 7 % |
| Stochastic lexicalized parsing | 13-16 % |

# Unsupervised Induction of Semantically annotated Lexicons

**Reference**: Rooth, Riezler, Prescher, Carroll, Beil (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL-99*.

**Application/Method**: Unsupervised induction of slot annotations for subcategorization frames, where annotations consist of the classes of an **EM-based 2-dimensional clustering model** trained on more than 1 000 000 verb-noun pairs gathered by stochastically parsing the British National Corpus (a free text corpus of about 1 00 000 000 words).

**Gain ($\infty$)**: We attribute an infinite gain of performance to EM-based clustering models, since **unsupervised induction** of semantically annotated lexicons completely relies on EM-based clustering models which can be automatically built using a statistical parser and free text.

# Identification of Collocations/Idioms

**Reference**: Prescher, Heid (1999). Probabilistisches Clustering zur Identifikation von Verb-Nomen-Kollokationen. Presented at *DGfS-00*.

**Application/Method**: Identification of verb-noun collocations/idioms based on a simple comparison of the **empirical distribution** with the distribution of an **EM-based 2-dimensional clustering model** trained on about 500 000 verb-noun pairs gathered by parsing a German corpus of about 4 000 000 words.

**Gain (0%)**: Recent evaluation on 400 items shows that this method yields as good results as current state-of-the-art methods.

# Grapheme-to-Phoneme Conversion

**Reference**: Müller, Möbius, Prescher (2000). Inducing probabilistic syllable classes using multivariate clustering. In *Proceedings of ACL-00*.

**Application/Method**: Grapheme-to-phoneme conversion (G2P) was performed (i) using a context-free grammar to produce all phonemic correspondences of a given grapheme string, (ii) applying an **EM-based 3- or 5-dimensional clustering model** trained on syllables, and (iii) predicting pronunciation by choosing the most probable analysis.

**Gain (3% resp. 30%)**: On about 2 000 test items, the G2P system using EM-based 5-dimensional clustering models achieves an increase over the baseline system using the **empirical distribution**. Moreover, it achieves an increase of about 30%, if compared to the **standard probability model** for context-free grammars.

# Machine Translation

**Reference**: Prescher, Riezler, Rooth (2000). Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of COLING-00*.

**Application/Method**: In machine translation, one problem to be solved is to find a correct translation of a source word using only minimal contextual information. The method is, to use the **class-based frequencies** of an EM-based 2-dimensional clustering model of verb-noun data to disambiguate alternative target nouns given their governing verb.

**Gain (7%)**: Evaluation on a corpus with about 800 bilingual sentences yields an increase of 7% precision over the baseline system using the **empirical verb-noun distribution** as disambiguator.

# Stochastic Lexicalized Parsing

**Reference**: Riezler, Prescher, Kuhn, Johnson (2000). Lexicalized stochastic modeling of constrait-based grammars using log-linear measures and EM training. In *Proceedings of ACL-00*.

**Application/Method**: Stochastic modeling of an unification grammar was performed using free text and EM training for log-linear models. With an EM-based 2-dimensional clustering model of verb-noun data, all parses of a sentence were pre-disambiguated according to **class-based frequencies** of the lexical heads in certain grammatical relations.

**Gain (13%-16%)**: Evaluation on a corpus of about 500 items shows that incorporation of the pre-disambiguation routine into the log-linear model improves precision of the stochastic model by about 13%. Moreover, it improves the model by 16%, if simple treebank training is used instead of EM-training.

# Discussion

- In contrast to similiar approaches (Rooth 1995; Pereira, Tishby, Lee 1995; Hofmann, Puzicha 1998), EM-based multivariate clustering is formalized clearly as pure EM algorithm, which plays a major role in the derivation of the invariance property.

- The approaches of (Schütze 1992, Yarowsky 1995) to word sense disambiguation use large amounts of information in terms of large context windows, which is easily obtainable in IR applications, but often unavailable in parsing or translation.

- Previous work on smoothing has shown that most of the developed methods are highly effective. However, EM-based multivariate clustering models utilizes the hidden structural information of given data and this is a feature lying beyond the capability of a pure smoother.

## Conclusion

- **Simplified re-estimation formulas** on the basis of **class-based frequencies** lead to detection of a **new invariance property** of EM-based multivariate clustering,
- a **new pruning technique** relying on **class-based frequencies** makes EM-based multivariate clustering feasible and results in models of only twice the size of the original data,
- we motivated theoretically and showed experimentally that EM-based multivariate clustering models are **good smoother**,
- **EM-based multivariate clustering models** improve several NLP applications, where other clustering methods are not applicable or competing smoothing methods are useles.