Hybrid Syntactic Category Induction

Bryan Jurish

jurish@ling.uni-potsdam.de

Universität Potsdam, Institut für Linguistik,
Potsdam, Germany

July 26, 2005
The Big Picture
  • Motivation
  • Evaluation

Clustering Phase
  • Target & Bound Selection
  • Monotonic Bernoulli Entropy
  • Simulated Melting
  • Bootstrapping

Ambiguity Resolution Phase
  • Baum-Welch Hidden Markov Model Reestimation
  • Trigram Clustering
The Big Picture

(1) Category induction ~ PoS identification

(2) Surface modelling ~ Grammar induction

(3) Chunk detection ~ Consituent analysis

(4) Dependency resolution ~ Projection relation

(5) Lexical indexing ~ Lexicon reification
Lexical Category Induction

Motivation

- Both theoretical and empirical results suggest the existence of cognitively salient syntactic categories.
- Provides drastic reduction of the data space
  - sparse data problem workaround

Implementation

- Hybrid iterative-hierarchical fuzzy agglomerative clustering over word-types
  - Clustering features: monotonic Bernoulli entropy
  - Frequency-based target- and bound-selection
  - Zipf’s law used to derive the clustering schedule
- Postpone ambiguity resolution until Phase 2
Categories: Evaluation

The Situation, or “What have we got?”
- Sample text $S \in \mathcal{A}^*$
- Induced model $\theta$ defining $\tau_\theta : \mathcal{A}^* \rightarrow C^*$

The Question, or “What the bejeebers is a tag29?”
- How can we judge the quality of $\theta$?

One Answer, or “Beats the heck outta me, bro…”
- Evaluate wrt. “gold standard” $\tau_G : S \rightarrow C^*_G$
- Train $\theta'$ (supervised), defining $\tau_{\theta'} : C^* \rightarrow C^*_G$
- Total meta-model precision is then:

$$\left\lfloor \left\lfloor \frac{\{i : 1 \leq i \leq |S| \; \&\; [\tau_{\theta'} (\tau_\theta (S))]_i = [\tau_G (S)]_i \}}{|S|} \right\rfloor \right\rfloor$$
Clustering Phase
Eadem sunt, quorum unum potest substitui alteri salva veritate.

“Those things are identical of which one can be substituted for the other with truth preserved.”

Gottfried Wilhelm von Leibniz, ca. 1715
Clustering: The Big Idea

- Break clustering problem down into $K$ stages
  - Brown et al. (1992), Schütze (1993,1995)

- Cluster targets $T_k \subset \mathcal{A}$ wrt. fixed set of bounds $B_k$ into classes $C_k$, $1 \leq k \leq K$

- Bootstrap classification using earlier solutions
  - Cutting et al. (1992b)

- Use fuzzy membership heuristic
  - Pereira et al. (1993), Lee (1997)
  - “Simulated melting”
Clustering: Algorithm

\[
\text{for } k = 1 \text{ to } K \text{ do} \\
T_k = \text{select}(\mathcal{A}, k, K, r_1) \\
\text{if } k == 1 \text{ then} \\
B_k = T_k \\
M_k = [\varphi(f_k \upharpoonright B_k \times \{w\})]_{w \in T_k} \\
C_k = \text{cluster}(M_k, d) \\
\text{else} \\
B_k = C_{<k} \\
M_k = [\varphi(f_k \upharpoonright B_k \times \{c\})]_{c \in C_{<k}} \\
\hat{M}_k = \text{attach}(M_k, \hat{M}_k, d) \\
\text{end if} \\
\hat{p}_k(C_k|T_k) = \text{fuzzy}(d(M_k, C_k), \beta_k, m) \\
\text{end for}
\]
Clustering: Data

Base Data (bigram frequencies):

\[ f_0 : \mathcal{A} \times \mathcal{A} \to \mathbb{N} \]

Stage 1 Data (directed bigrams): for \( w \in T_1, b \in B_1 \),

\[
\begin{align*}
  f_{\ell,1}(w, b) &= f_0(b, w) \\
  f_{r,1}(w, b) &= f_0(w, b)
\end{align*}
\]

General Data: for \( z \in \{\ell, r\}, k \in K \),

\[
\begin{align*}
  f_{z,k}(w) &= \sum_{b \in B_k} f_{z,k}(w, b) \\
  f_{z,k}(b) &= \sum_{w \in T_k} f_{z,k}(w, b) \\
  N_{z,k} &= \sum_{w \in T_k} f_{z,k}(w)
\end{align*}
\]
Target & Bound Selection

Stage 1:

\[ T_1 = \{ w \in A \mid \text{rank}_{f_0}(w) < r_1 \} \]

\[ B_1 = T_1 \]

Stage \( k > 1 \):

\[ B_k = C_{<k} \]

\[ T_k = \arg \max_{w \in A - T_{<k}} \text{rank}_{f_k}(w) \]

\[ \log r_k = \log r_{k-1} + \frac{\log(|A|) - \log(|r_1|)}{K-1} \]

Properties:

\[ i \neq j \implies T_i \cap T_j = \emptyset \quad \text{(OaOO)} \]

\[ \text{avg}_w \log f_k(w) \approx ak + b \quad \text{(Zipf)} \]
Target Selection

Average Frequency by Rank (log scale)

Average frequency by max rank

$E(f(w))$ vs. $\max(r)$
Target Selection

Average Frequency by Stage (linear scale)

Average frequency by stage

E(f(w))

Stage
Vector Assembly

ML probability estimation:

\[ P_{z,k}(w, b) = \frac{f_{z,k}(w, b)}{N_{z,k}} \]
\[ P_{z,k}(w) = \frac{f_{z,k}(w)}{N_{z,k}} \]

Target vector construction:

\[ \vec{w}_{z,k} = [\vec{w}_{z,k}(1), \ldots, \vec{w}_{z,k}(|B_k|)] \]
\[ \vec{w}_k = \vec{w}_{\ell,k} \circ \vec{w}_{r,k} \]
\[ = [\vec{w}_{\ell,k}(1), \ldots, \vec{w}_{\ell,k}(|B_k|), \vec{w}_{r,k}(1), \ldots, \vec{w}_{r,k}(|B_k|)] \]

Conditional bigram vectors:

\[ \vec{w}_{z,k}(i) = P_{z,k}(b_i|w) = \frac{P_{z,k}(w, b_i)}{P_{z,k}(w)} \text{ or } \ldots \]
Pointwise Entropy

\[ h(p) \]
Bernoulli Entropy

![Graph of Bernoulli Entropy](image)
Bernoulli Entropy

The diagram illustrates the relationship between the probability of a Bernoulli trial, \( p \), and the entropy function \( H(b(1;p)) \). The entropy function measures the uncertainty or randomness associated with the outcomes of a Bernoulli trial. The graph shows how entropy decreases as the probability \( p \) increases, indicating that the more certain the outcome (i.e., \( p \) approaching 0 or 1), the less uncertain the system is.
Monotonic Bernoulli Entropy

The graph shows the functions $H(p)$ and $H(b(1;p))$. The y-axis represents entropy, and the x-axis represents the probability $p$. The function $H(p)$ is shown in red, and $H(b(1;p))$ is shown in green. The functions are plotted for $p$ values ranging from 0 to 1.
Monotonic Bernoulli Entropy

The graph shows the relationship between entropy and the probability of an event, $p$. The functions $H(p)$ and $h(p)$ are plotted, with $H(p)$ representing the entropy of a Bernoulli distribution with parameter $p$, and $h(p)$ being a decreasing function. The graph illustrates how entropy increases monotonically with $p$. The curve $H(b(1:p))$ is also plotted, indicating a different behavior compared to $H(p)$. The x-axis represents the probability $p$, while the y-axis represents the entropy.
Distance: $d_{L1}(\hat{H}(p), \hat{H}(q))$
Entropy Distance: Variants

\[ H(Y|X = x) \equiv I(Y; X = x) \equiv D(Y|X = x||Y) \]

- \(X\) is a target random variable, \(\Omega_X = T_k\)
- \(Y\) is a boundary random variable, \(\Omega_Y = B_k\)
- \(H(Y|X = x)\) is the conditional entropy of a boundary \(Y\) given the target event \(x\)
- \(I(Y; X = x)\) is the semi-pointwise MI between a boundary \(Y\) and the target event \(x\)
- \(D(Y|X = x||Y)\) is the KL divergence between:
  - the conditional distribution \(p(Y|X = x)\) of a boundary \(Y\) given the word event \(x\), and
  - the global distribution \(p(Y)\) of \(Y\).
\[ H(Y \mid X = x) \equiv I(X = x; Y) \]

\[ \forall x_1, x_2 \in \Omega_X : \]
\[ |I(X = x_1; Y) - I(X = x_2; Y)| \]
\[ = |H(Y) - H(Y \mid X = x_1)) - (H(Y) - H(Y \mid X = x_2))| \]
\[ = |H(Y \mid X = x_2) - H(Y \mid X = x_1)| \]
\[ = |H(Y \mid X = x_1) - H(Y \mid X = x_2)| \]

\[ \square \]
\[ D(Y \mid X = x_1 \parallel Y) - D(Y \mid X = x_2 \parallel Y) \]

\[
= \left| \left[ \sum_y p(y \mid x_1) \log \frac{p(y \mid x_1)}{p(y)} \right] - \left[ \sum_y p(y \mid x_2) \log \frac{p(y \mid x_2)}{p(y)} \right] \right|
\]

\[
= \left| \left[ \sum_y p(y \mid x_1) (\log p(y \mid x_1) - \log p(y)) \right] - \left[ \sum_y p(y \mid x_2) (\log p(y \mid x_2) - \log p(y)) \right] \right|
\]

\[
= |H(Y \mid x_1) - H(Y \mid x_2) + \sum_y (p(y \mid x_1) - p(y \mid x_2)) \log p(y)|
\]

\[ \equiv |H(Y \mid x_1) - H(Y \mid x_2)| \]
Clustering

Target-Vector Features:
\[ \hat{w}_{z,k}(i) = \hat{H}(P_{z,k}(b_i | w)) \]

Distance Functions
- **Spearman’s rank correlation coefficient**, used by Finch & Chater (1993)

Link Methods
- **Maximum link**: \[ d_{max}(W, V) = \max_{\vec{w} \in W, \vec{v} \in V} d(\vec{w}, \vec{v}) \]
- **Average link**: \[ d_{avg}(W, V) = \text{avg}_{\vec{w} \in W, \vec{v} \in V} d(\vec{w}, \vec{v}) \]

Tree Pruning: 50 output clusters
Fuzzy Cluster Membership

Desideratum:

- Membership distribution: $\hat{p}_k(C_k|T_k)$

Available Source Data:

- Hard partitioning: $\pi_k : T_k \rightarrow C_k$
- Distance function: $d_k : \mathcal{P}(\mathbb{R}^2|B_k|) \times \mathcal{P}(\mathbb{R}^2|B_k|) \rightarrow \mathbb{R}$

Heuristic: (Jaynes, 1983)

- Similarity function: $\hat{s}_k(c, w) = \exp(-\beta_k d_k(c, \bar{w}_k))$
- Simulated melting: $\beta_k = \frac{1}{k}$
- Membership heuristic: $\hat{p}_k(c|w) = \frac{\hat{s}_k(c, w)}{Z_{k,w}}$
- Useful restriction: $m$-best, $m \approx \frac{|C|}{12}$
Bootstrapping

Cluster-based Profiling

\[ f_{z,k}(w, b) = \sum_{v \in T_{<k}} \hat{p}_{<k}(b|v) f_{z,0}(w, v) \quad \text{(Clusters as bounds)} \]

\[ f_{z,k}(c, b) = \sum_{w \in \pi_{<k}^{-1}(c)} f_{z,k}(w, b) \quad \text{(Clusters as targets)} \]

Underlying Assumptions (mostly harmless)

\[ f_{z,k}(w, b) = p_{z,k}(w, b) N_{z,k} \quad \text{(MLE)} \]

\[ f_{z,0}(v, w) = p_{z,k}(w, v) N_{z,k} \quad \text{(MLE)} \]

\[ p_{z,k}(w, b) = \sum_{v \in T_{<k}} p_{z,k}(v, w, b) \quad \text{(Marginal)} \]

\[ p_{z,k}(b|v, w) = \hat{p}_{<k}(b|v) \quad \text{(Independence)} \]

Computational Complexity: \( \mathcal{O}(C^2 |A|) \)
Corpora

**German**: NEGRA (Skut et al. 1997)
- 355,096 tokens ; 48,924 types ; reduced tagset

**English**: SUSANNE (Sampson, 1995) and *Great Expectations* (Dickens, 1861)
- 374,640 tokens ; 20,600 types ; reduced tagset

**Preprocessing**
- Token- and sentence-boundaries were marked
- All words were converted to lower-case
- Punctuation was preserved as separate tokens
- 10% of each corpus were reserved for testing and ambiguity resolution
Multi- vs. Single-Stage

![Graph showing the comparison between Multi-stage and Single-stage](image)

- H', L1, multi-stage
- H', L1, single-stage, |B|=50
- H', L1, single-stage, |B|=100
- H', L1, single-stage, |B|=200
- H', L1, single-stage, |B|=|T|

Global Meta-Accuracy (%) vs. Number of Targets
Ambiguity Resolution Phase
Another Snappy Quote

The Light of humane minds is Perspicuous Words, but by exact definitions first snuffed, and purged from ambiguity

Thomas Hobbes, *Leviathan* (1651)
HMM Initialization

Hidden Markov Model Parameters

\[
\begin{align*}
\pi(q) &= P(Q_1 = q) \quad \text{(Initial probabilities)} \\
A(q_i, q_j) &= P(Q_{t+1} = q_j | Q_t = q_i) \quad \text{(Arc probabilities)} \\
B(w, q) &= P(W_t = w | Q_t = q) \quad \text{(Emission probabilities)}
\end{align*}
\]

Parameter Initialization (B)

\[
\hat{p}_{\leq k}(w|c) = \frac{\hat{p}_{\leq k}(c|w)\hat{p}_{\leq k}(w)}{\hat{p}_{\leq k}(c)} \quad \text{(Bayes)}
\]

where:

\[
\hat{p}_{\leq k}(w) = \frac{P_{\ell,K}(w)+P_{r,K}(w)}{2} \quad \text{(MLE)}
\]

\[
\hat{p}_{\leq k}(c) = \sum_{w \in T_{\leq k}} \hat{p}_{\leq k}(w, c) \quad \text{(Marginal)}
\]

\[
= \sum_{w \in T_{\leq k}} \hat{p}_{\leq k}(c|w)\hat{p}_{\leq k}(w)
\]
HMM Reestimation

Method

- 20 iterations of the Baum-Welch Algorithm applied to test corpus

Results

- Hard clusters from $\pi_K$ outperformed all HMMs!
- Reestimation showed an early maximum pattern in the sense of Elworthy (1994)

Speculation

- Fuzzy membership heuristic is too lenient “Shock-freezing” $\Rightarrow$ initial maximum
- $m$-best method biases emission estimates
- Independence assumptions are incompatible
HMM EM: Results (Global)

![Graph showing the global meta-accuracy over Baum-Welch iterations for German and English.]

- Baum-Welch, German
- Baum-Welch, English
Trigram Clustering

Idea

• Variant of Schütze’s (1995) method
• Cluster trigram types by reference to their components’ context vectors

Gory Details: for $w_1, w_2, w_3 \in A$,

$$\langle w_1, w_2, w_3 \rangle = \overrightarrow{w_1} \circ \overrightarrow{w_2} \circ \overrightarrow{w_3}$$

But wait, there’s more!

• Cluster-based component profiling
• Frequency-based prototype selection
• Attachment of remaining trigrams to centroids
### Trigram Clustering: Results

| Clustering | German |  | English |  |
|-------------|--------|----------------|----------------|
| Stage       | Amb. Rate | Meta-Acc. | Amb. Rate | Meta-Acc. |
| $\pi_K$     | 1.00   | 74.13 %     | 1.00   | 76.72 %   |
| 7           | 1.25   | 73.57 %     | 1.52   | 73.86 %   |
| 8           | 1.23   | 75.25 %     | 1.53   | 74.57 %   |
| 9           | 1.24   | 77.08 %     | 1.51   | 74.49 %   |
| 10          | 1.25   | 75.59 %     | 1.55   | 74.83 %   |
Summary

Clustering Phase

- Monotonic Bernoulli entropy is a useful discriminator for syntactic category
- Multi-stage clustering outperforms many single-stage methods
- Possible improvement: add morphology

Ambiguity Resolution Phase

- Needs work
- Use trigram clustering output to initialize HMM?
- Interpolate reestimated HMMs with (estimated) cluster unigram model?
The End
Examples
## Example Clusters

<table>
<thead>
<tr>
<th>i</th>
<th>he</th>
<th>she</th>
<th>we</th>
<th>they</th>
<th>who</th>
</tr>
</thead>
<tbody>
<tr>
<td>said</td>
<td>made</td>
<td>saw</td>
<td>took</td>
<td>done</td>
<td>found</td>
</tr>
<tr>
<td>were</td>
<td>been</td>
<td>are</td>
<td>am</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>a</td>
<td>his</td>
<td>my</td>
<td>an</td>
<td>your</td>
</tr>
<tr>
<td>going</td>
<td>being</td>
<td>having</td>
<td>taking</td>
<td>getting</td>
<td></td>
</tr>
<tr>
<td>hands</td>
<td>head</td>
<td>heart</td>
<td>eyes</td>
<td>face</td>
<td>mouth</td>
</tr>
<tr>
<td>see</td>
<td>hear</td>
<td>get</td>
<td>put</td>
<td>give</td>
<td>take</td>
</tr>
<tr>
<td>that</td>
<td>which</td>
<td>what</td>
<td>how</td>
<td>where</td>
<td>whom</td>
</tr>
<tr>
<td>did</td>
<td>thought</td>
<td>knew</td>
<td>hope</td>
<td>wanted</td>
<td>does</td>
</tr>
<tr>
<td>not</td>
<td>n’t</td>
<td>only</td>
<td>rather</td>
<td></td>
<td></td>
</tr>
<tr>
<td>then</td>
<td>now</td>
<td>still</td>
<td>why</td>
<td>yet</td>
<td>perhaps</td>
</tr>
<tr>
<td>herbert</td>
<td>biddy</td>
<td>estella</td>
<td>pumblechook</td>
<td>provis</td>
<td>compeyson</td>
</tr>
<tr>
<td>SUSPICIOUS CLUSTERS</td>
<td>COHESION FEATURES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of in with at for on by from into about ... to</td>
<td>, ; : ( ) – ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>this one such these another those many each town</td>
<td>all more ... nothing something ... saying thinking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>as if when though because since whether ... afraid</td>
<td>old young long short certain strange ... present</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gentleman lady fellow convict ... havisham jaggers wopsle</td>
<td>it you me him them us yourself drummle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>out up down back off ... near himself care ready known</td>
<td>few two three four five ... great general kitchen state</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>there here none indeed sir uncle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CPALA 2005 / Jurish / Hybrid Syntactic Category Induction – p. 43/63
Tagsets
<table>
<thead>
<tr>
<th>Tag</th>
<th>STTS Tag(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADJA ADJD PIDAT</td>
<td>Adjective</td>
</tr>
<tr>
<td>ADV</td>
<td>ADV APPO APZR <em>AV PTK</em> (except PTKZU)</td>
<td>Adverb</td>
</tr>
<tr>
<td>CARD</td>
<td>CARD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>CCONJ</td>
<td>KON</td>
<td>Coord. conjunction</td>
</tr>
<tr>
<td>DET</td>
<td>ART PDAT PIAT PPOSAT PRELAT PWAT</td>
<td>Determiner</td>
</tr>
<tr>
<td>MISC</td>
<td>FM ITJ XY</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>DET</td>
<td>ART PDAT PIAT PPOSAT PRELAT PWAT</td>
<td>Determiner</td>
</tr>
<tr>
<td>MISC</td>
<td>FM ITJ XY</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>NOUN</td>
<td>NE NN TRUNC</td>
<td>Nominal</td>
</tr>
<tr>
<td>PREP</td>
<td>APPRART APPR</td>
<td>Preposition</td>
</tr>
<tr>
<td>PRON</td>
<td>PDS PIS PPER PPOSSE PRELS PRF PWS</td>
<td>Pronominal</td>
</tr>
<tr>
<td>SCONJ</td>
<td>KOKOM KOUI KOUS</td>
<td>Subord. conjunction</td>
</tr>
<tr>
<td>TO</td>
<td>PTKZU</td>
<td>Infinital zu</td>
</tr>
<tr>
<td>VFIN</td>
<td>VAFIN VAIMP VMFIN VVFIN VVIMP</td>
<td>Finite verb</td>
</tr>
<tr>
<td>VINF</td>
<td>V* (except V<em>INF V</em>ZU V*PP)</td>
<td>Infinitive, participle</td>
</tr>
<tr>
<td>$,</td>
<td>$</td>
<td>Comma</td>
</tr>
<tr>
<td>$,</td>
<td>$</td>
<td>Sentence-final punct.</td>
</tr>
<tr>
<td>$(</td>
<td>$(</td>
<td>Sentence-internal punct.</td>
</tr>
</tbody>
</table>
## Reduced SUSANNE Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>SUSANNE Tag(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV</td>
<td>FA* FB* LE* XX</td>
<td>Adverb</td>
</tr>
<tr>
<td>DET</td>
<td>A* D*</td>
<td>Determiner</td>
</tr>
<tr>
<td>CARD</td>
<td>M*</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>CCONJ</td>
<td>CC*</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>POS</td>
<td>G*</td>
<td>Genitive marker</td>
</tr>
<tr>
<td>MISC</td>
<td>FO* FU* FW* UH ZZ*</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>NOUN</td>
<td>N* BTO22</td>
<td>Nominal</td>
</tr>
<tr>
<td>PREP</td>
<td>I* BTO21</td>
<td>Preposition</td>
</tr>
<tr>
<td>PRON</td>
<td>P* EX</td>
<td>Pronominal</td>
</tr>
<tr>
<td>PUNC</td>
<td>Y*</td>
<td>Punctuation</td>
</tr>
<tr>
<td>SCONJ</td>
<td>CS*</td>
<td>Subordinating conjunction</td>
</tr>
<tr>
<td>TO</td>
<td>TO</td>
<td>Infinitival to</td>
</tr>
<tr>
<td>VFIN</td>
<td>V* (except V<em>0, V</em>G*)</td>
<td>Finite verb form</td>
</tr>
<tr>
<td>VINF</td>
<td>VB0 VD0 VH0</td>
<td>Infinitive verb form</td>
</tr>
<tr>
<td>VING</td>
<td>VBG VDG VHG VVG*</td>
<td>-ing verb form</td>
</tr>
</tbody>
</table>
Monotonic Bernoulli Entropy
Features: Shannon Entropy

- Shannon Entropy (Shannon and Weaver, 1949):
  \[ H(p) = - \sum_{x \in \text{dom}(p)} p(x) \log_2 p(x) \]

- Measure of the (un)predictability of \( p \)

- Average length of a message that an event from \( \text{dom}(p) \) has occurred under an optimal encoding

- Properties:
  - \( H(p) \geq 0 \)
  - Measurement unit: \( \text{bits} \)
  - Pointwise entropy: \( h_p(x) = -p(x) \log p(x) \)
  - Unfortunately asymmetric
Assume a Bernoulli distribution $X_x \sim b(1; p = P(x))$ for each relevant point $x$, then:

$$H(X \sim b(1; p)) = \sum_{x \in \{0, 1\}} h(x) = -p \log p - (1 - p) \log(1 - p)$$

- $H(b(1; p))$ is **symmetric**
- ... but unfortunately **non-monotonic**: No difference is drawn between high- and low-probability points.
Monotonic Bernoulli Entropy

Idea

- Use Bernoulli assumption for symmetry
- Modify $H$ to produce a monotonically growing function $\hat{H}$

Definition

$$\hat{H}(p) = \begin{cases} 
H(b(1; p)) & \text{if } p \leq \frac{1}{2} \\
2 - H(b(1; p)) & \text{otherwise}
\end{cases}$$

Intuitive Interpretation of $\hat{H}(P(B = b|T = w))$:

- Mnemonic utility of chunking a boundary event $b$ into a target event $w$. 
Target Data
Fuzzy Clusters

The graph shows the meta-accuracy on targets (%) for different numbers of targets, varying the number of clusters. The lines represent different numbers of clusters:
- m=1, German
- m=4, German
- m=24, German
- m=50, German

The meta-accuracy decreases as the number of targets increases for all configurations.
Target Precision, German

Meta-Accuracy on Targets (%)

Number of Targets

H'(b(1;p_z(b|w))), Spearman, max, German
H'(b(1;p_z(b|w))), L1, max, German
p_z(b|w), Spearman, max, German
f_z(w,b), cosine, avg, German
Target Precision, English

Graph showing the meta-accuracy on targets (%): 
- $H'(b(1:p_z(b|w)))$, Spearman, max, English
- $H'(b(1:p_z(b|w)))$, L1, max, English
- $p_z(b|w)$, Spearman, max, English
- $f_z(b,w)$, cosine, avg, English

Y-axis: Meta-Accuracy on Targets (%)
X-axis: Number of Targets

CPALA 2005 / Jurish / Hybrid Syntactic Category Induction – p. 54/63
HMM Data
HMM EM: Results: German

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi K$</td>
<td>1.00</td>
<td>74.13 %</td>
<td>1.00</td>
<td>81.56 %</td>
</tr>
<tr>
<td>0</td>
<td>1.00</td>
<td>69.32 %</td>
<td>1.00</td>
<td>69.67 %</td>
</tr>
<tr>
<td>4</td>
<td>1.22</td>
<td>73.60 %</td>
<td>1.31</td>
<td>75.66 %</td>
</tr>
<tr>
<td>8</td>
<td>1.29</td>
<td>73.58 %</td>
<td>1.40</td>
<td>75.35 %</td>
</tr>
<tr>
<td>12</td>
<td>1.32</td>
<td>72.83 %</td>
<td>1.46</td>
<td>74.48 %</td>
</tr>
<tr>
<td>16</td>
<td>1.34</td>
<td>72.67 %</td>
<td>1.49</td>
<td>74.25 %</td>
</tr>
<tr>
<td>20</td>
<td>1.35</td>
<td>72.41 %</td>
<td>1.50</td>
<td>73.91 %</td>
</tr>
</tbody>
</table>
## Baum-Welch Results: English

| Iteration | Global | | | Targets | | |
|---|---|---|---|---|---|
| $\pi K$ | 1.00 | 76.72 % | | 1.00 | 80.12 % |
| 0 | 1.00 | 70.11 % | | 1.00 | 70.88 % |
| 4 | 1.37 | 73.60 % | | 1.46 | 74.90 % |
| 8 | 1.48 | 73.60 % | | 1.59 | 74.71 % |
| 12 | 1.56 | 73.99 % | | 1.68 | 75.12 % |
| 16 | 1.61 | 73.88 % | | 1.75 | 75.00 % |
| 20 | 1.65 | 73.71 % | | 1.80 | 74.77 % |
HMM EM: Results (Targets)

Meta-Accuracy on Targets (%)

Baum-Welch, German
Baum-Welch, English

Baum-Welch Iteration
Old Data
HMM Reestimation, Global

* alternate configuration
HMM Reestimation, Global

* alternate configuration
HMM Reestimation, Targets

Meta-Accuracy on Targets (%)

* alternate configuration
HMM Reestimation, Targets

Meta-Accuracy on Targets (%)

* alternate configuration

NEGRA, Baum-Welch
NEGRA, Baseline