Hybrid Syntactic Category Induction

Bryan Jurish

jurish@ling.uni-potsdam.de

Universität Potsdam, Institut für Linguistik,

Potsdam, Germany

July 26, 2005

Outline

The Big Picture

- Motivation
- Evaluation

Clustering Phase — Token → Type

- Target & Bound Selection
- Monotonic Bernoulli Entropy
- Simulated Melting
- Bootstrapping

Ambiguity Resolution Phase — Type → Token

- 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 θ defining $\tau_\theta : \mathcal{A}^* \rightarrow C^*$

The Question, or "What the bejeebers is a `tag29`?"

- How can we judge the quality of θ?

One Answer, or "Beats the heck outta me, bro…"

- Evaluate wrt. "gold standard" $\tau_G : S \rightarrow C_G^{|S|}$
- Train θ' (supervised), defining $\tau_{\theta'} : C^* \rightarrow C_G^*$
- Total meta-model precision is then:

$$\frac{|\{i \ : \ 1 \leq i \leq |S| \ \ \& \ \ [\tau_{\theta'}(\tau_\theta(S))]_i = [\tau_G(S)]_i\}|}{|S|}$$

Clustering Phase

A Snappy Quote

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, $1 \leq k \leq K$
  - Roberts (2002), Finch & Chater (1993)
- Bootstrap classification using earlier solutions
  - Cutting et al. (1992b)
- Use fuzzy membership heuristic
  - Pereira et al. (1993), Lee (1997)
  - “Simulated melting”

Clustering: Algorithm

for $k = 1$ to K **do**

$T_k = \text{select}(\mathcal{A}, k, K, r_1)$ /* Target selection */

if k == 1 **then**

$B_k = T_k$ /* Prototyping stage */

$M_k = [\varphi(f_k \restriction B_k \times \{w\})]_{w \in T_k}$ /* Target profile */

$C_k = \text{cluster}(M_k, d)$ /* Clustering */

else

$B_k = C_{<k}$ /* Attachment stages */

$M_k = [\varphi(f_k \restriction B_k \times \{w\})]_{w \in T_k}$ /* Targets */

$\hat{M}_k = [\varphi(f_k \restriction B_k \times \{c\})]_{c \in C_{<k}}$ /* Centroids */

$C_k = \text{attach}(M_k, \hat{M}_k, d)$ /* Attachment */

end if

$\hat{p}_k(C_k|T_k) = \text{fuzzy}(d(M_k, C_k), \beta_k, m)$ /* Membership */

end for

Clustering: Data

Base Data (bigram frequencies):

$$f_0 : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{N}$$

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

$$\begin{aligned} f_{\ell,1}(w,b) &= f_0(b,w) \\ f_{r,1}(w,b) &= f_0(w,b) \end{aligned}$$

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

$$\begin{aligned} f_{z,k}(w) &= \sum_{b \in B_k} f_{z,k}(w,b) \\ f_{z,k}(b) &= \sum_{w \in T_k} f_{z,k}(w,b) \\ N_{z,k} &= \sum_{w \in T_k} f_{z,k}(w) \end{aligned}$$

Target & Bound Selection

Stage 1:

$$\begin{aligned} T_1 &= \{w \in \mathcal{A} \mid \mathrm{rank}_{f_0}(w) < r_1\} \\ B_1 &= T_1 \end{aligned}$$

Stage $k > 1$:

$$\begin{aligned} B_k &= C_{<k} \\ T_k &= \arg\max_{w \in \mathcal{A} - T_{<k}}^{r_k - r_{k-1}} \mathrm{rank}_{f_k}(w) \\ \log r_k &= \log r_{k-1} + \frac{\log(|\mathcal{A}|) - \log(|r_1|)}{K-1} \end{aligned}$$

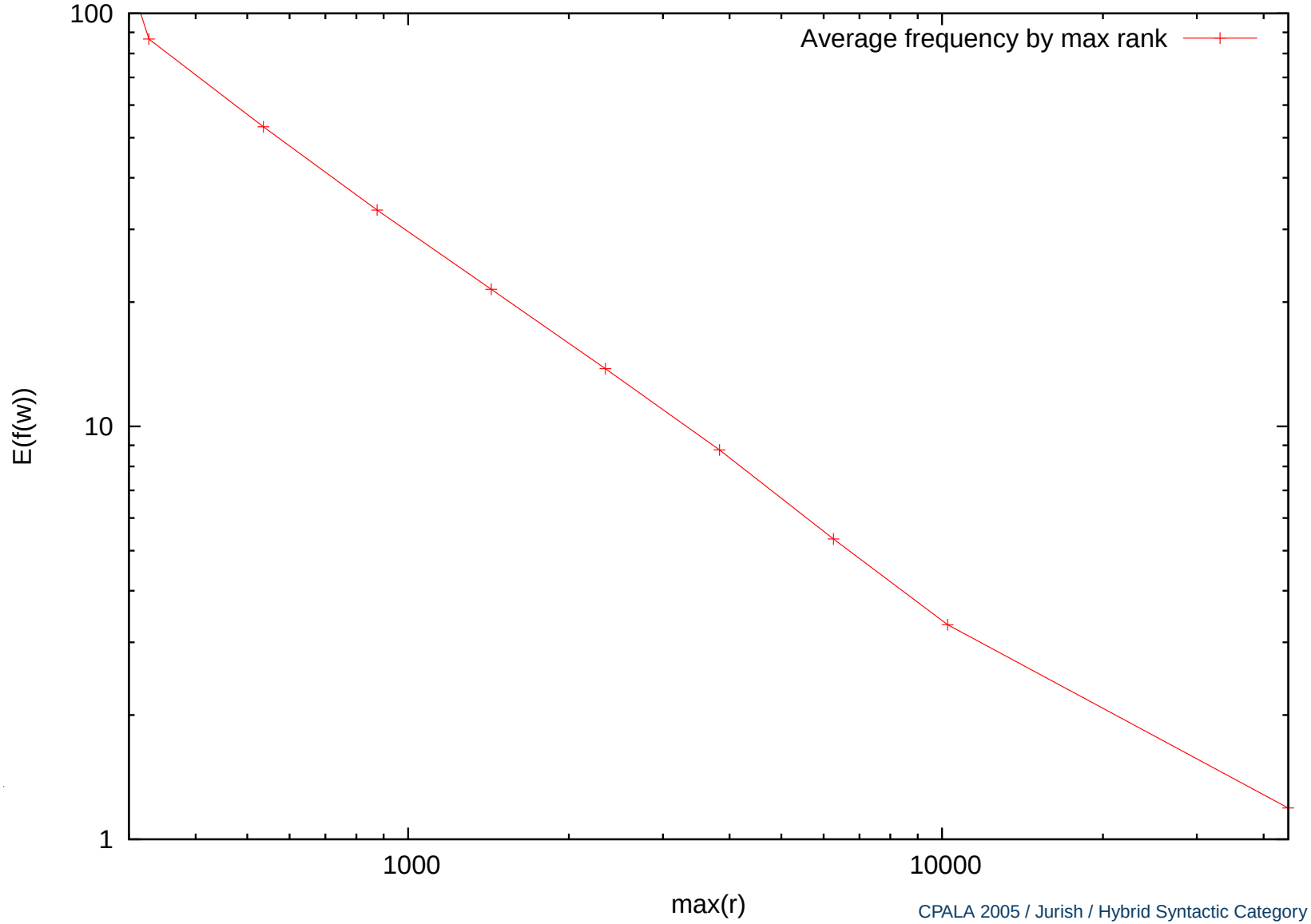
Properties:

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

$$\mathrm{avg}_w \log f_k(w) \approx ak + b \quad \text{(Zipf)}$$

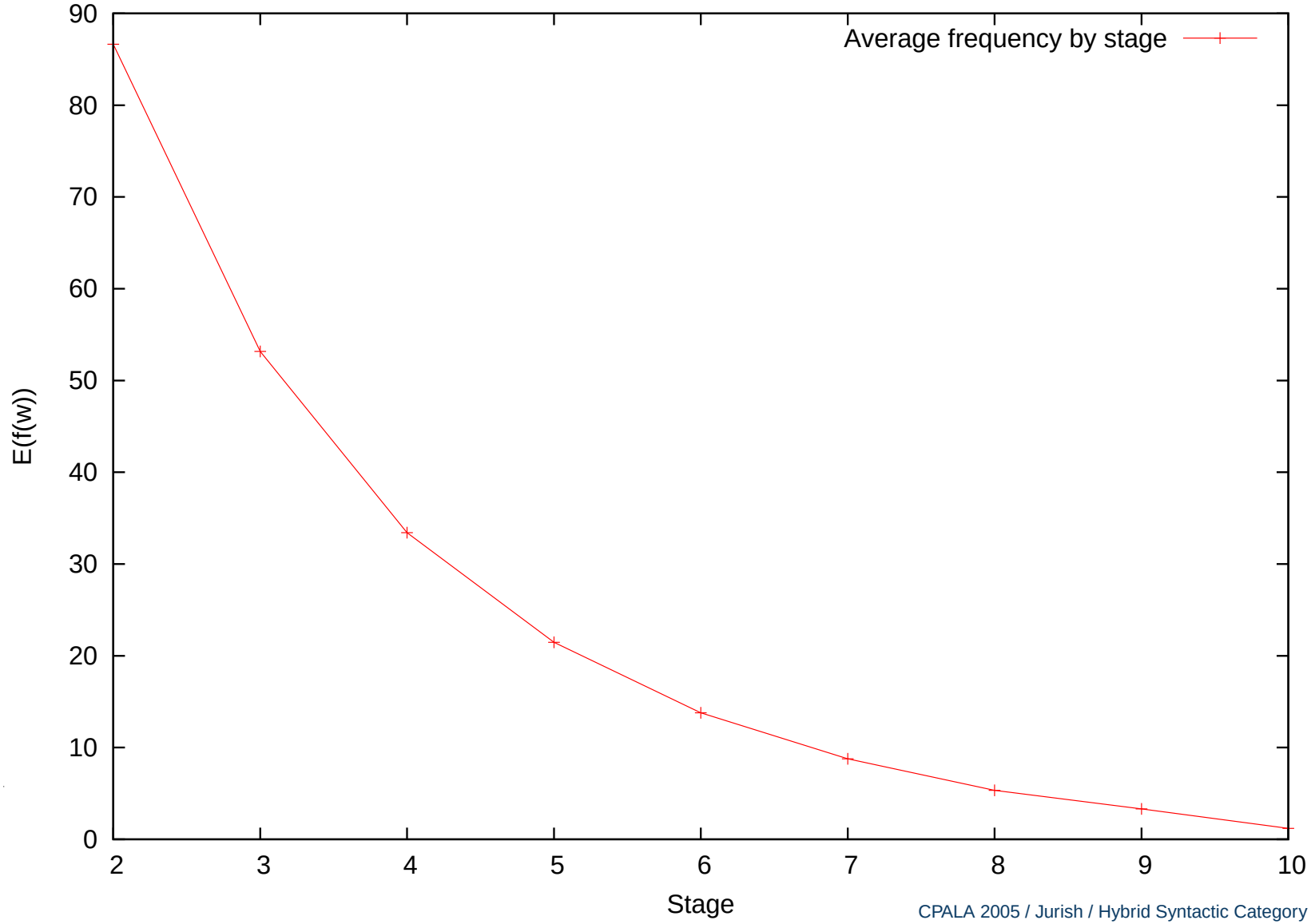
Target Selection

Average Frequency by Rank (log scale)



Target Selection

Average Frequency by Stage (linear scale)



Vector Assembly

ML probability estimartion:

$$\mathrm{P}_{z,k}(w,b) = \frac{f_{z,k}(w,b)}{N_{z,k}}$$

$$\mathrm{P}_{z,k}(w) = \frac{f_{z,k}(w)}{N_{z,k}}$$

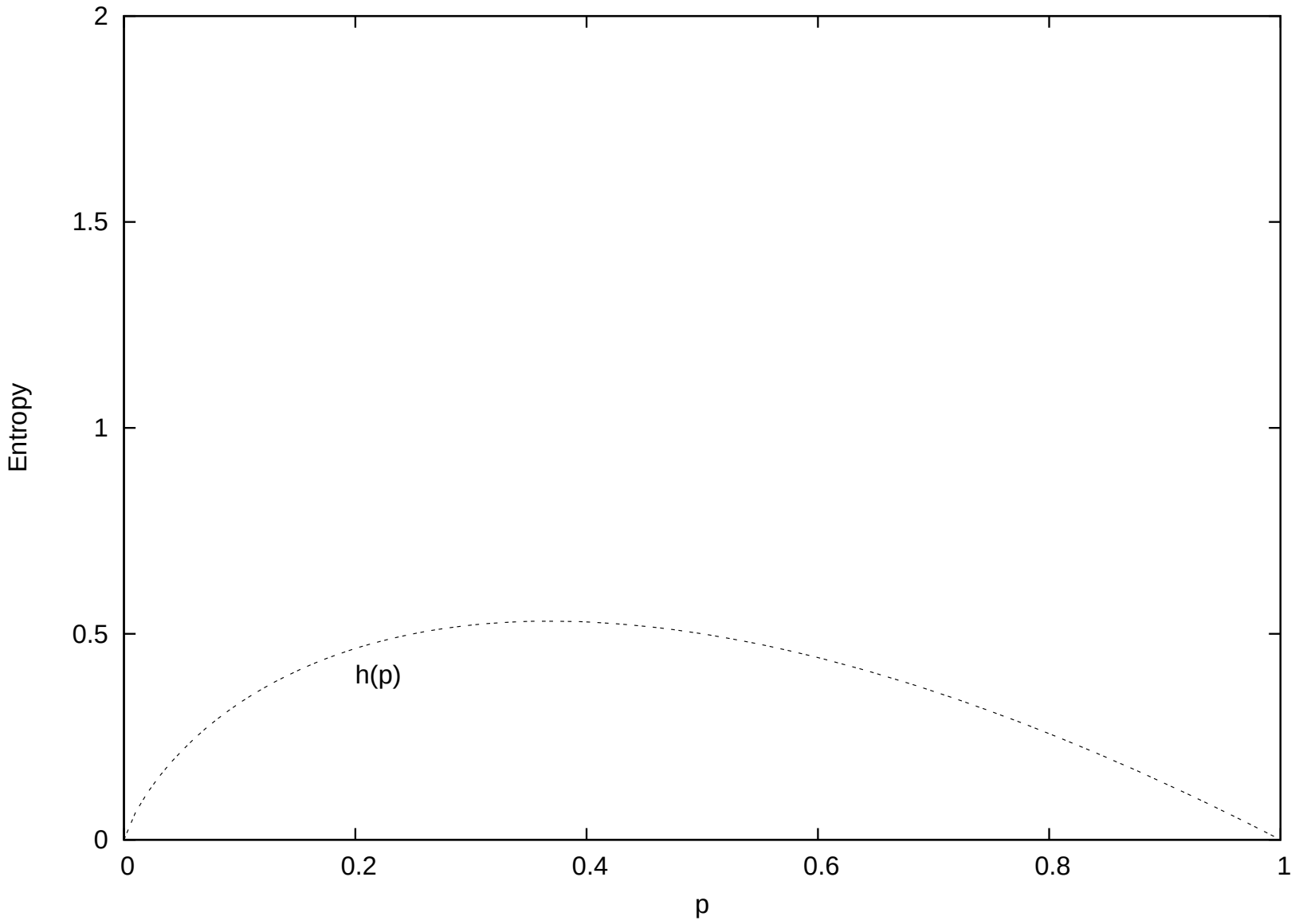
Target vector construction:

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

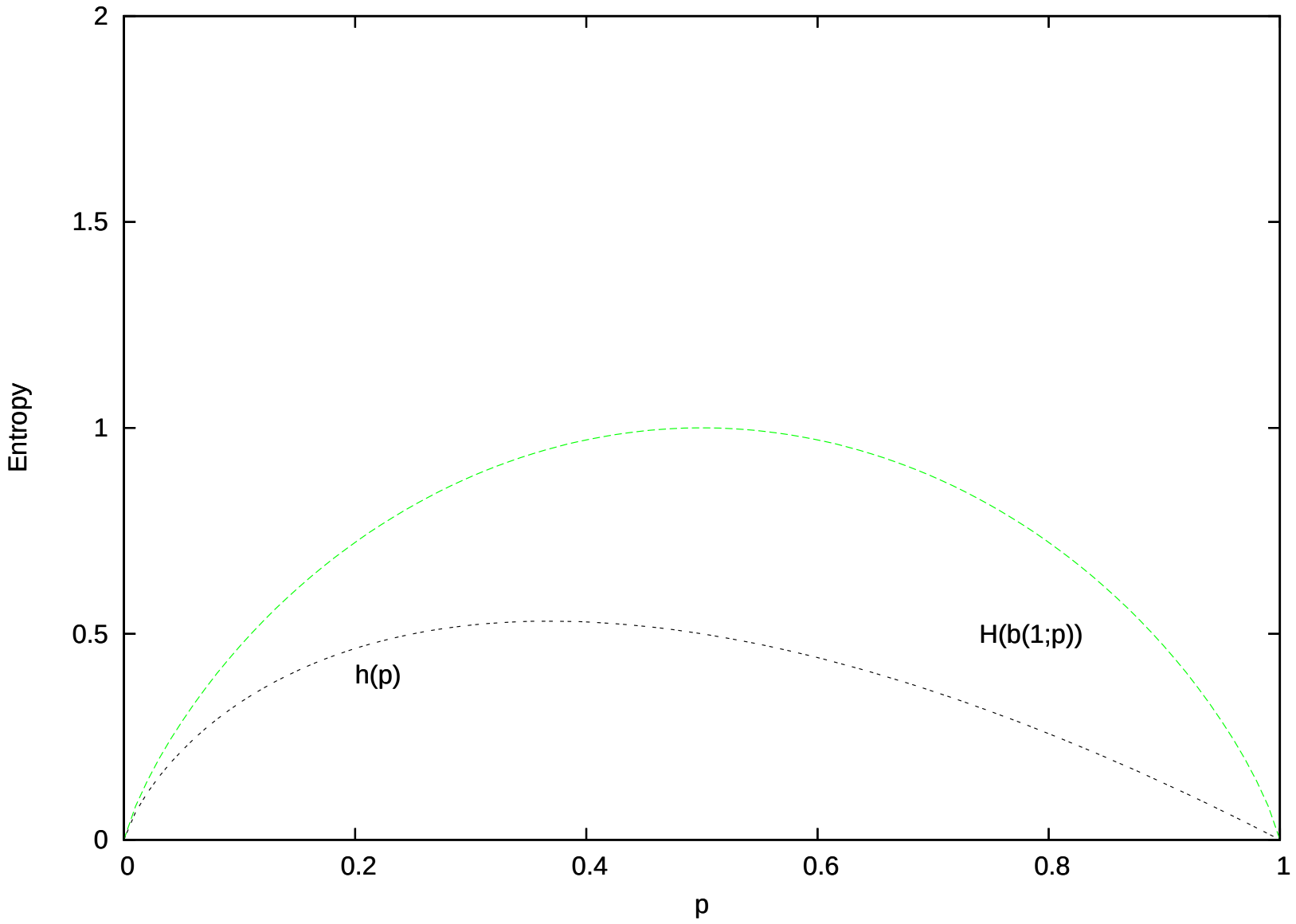
Conditional bigram vectors:

$$\vec{w}_{z,k}(i) = \mathrm{P}_{z,k}(b_i|w) = \frac{\mathrm{P}_{z,k}(w,b_i)}{\mathrm{P}_{z,k}(w)}$$ or . . .

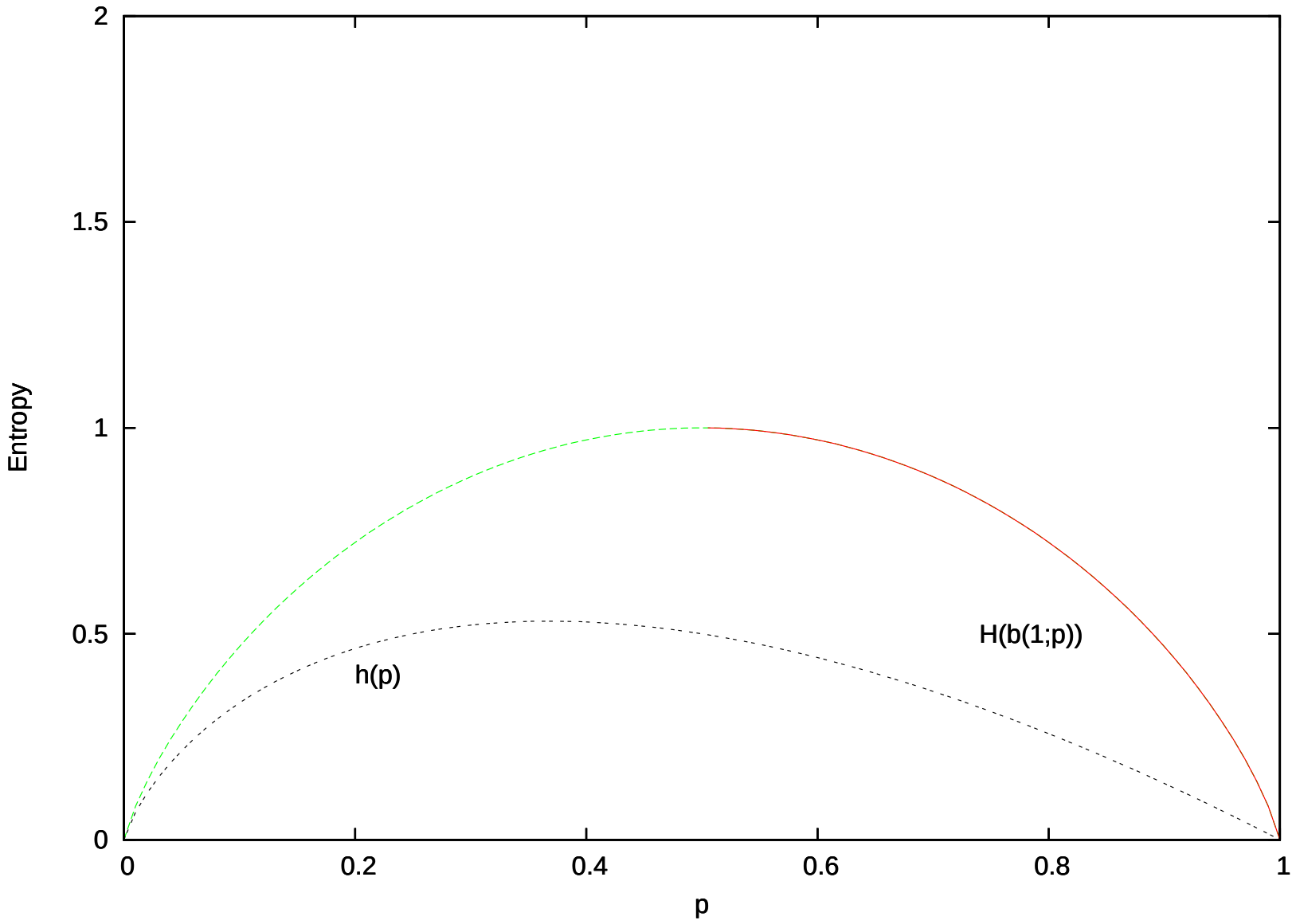
Pointwise Entropy



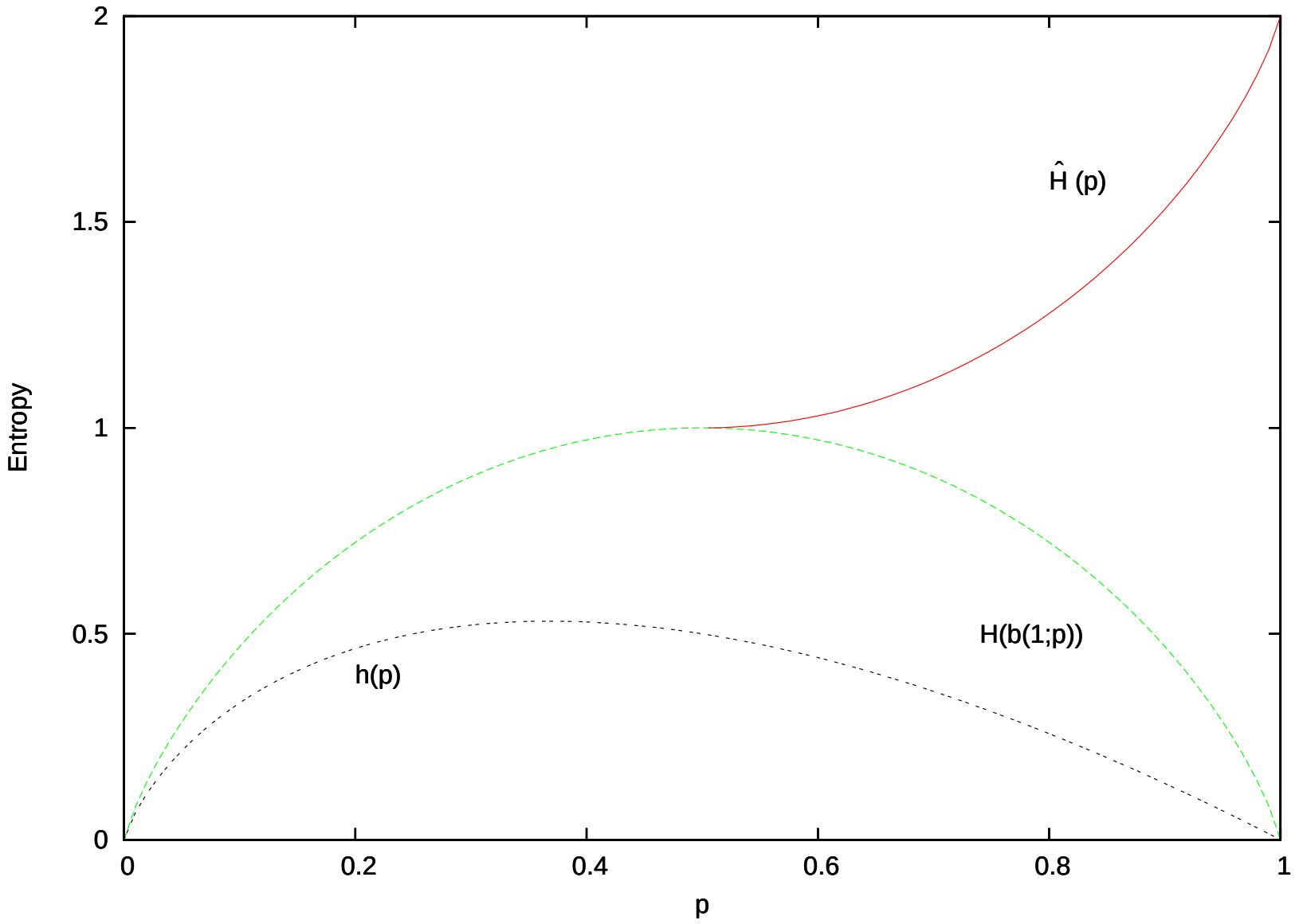
Bernoulli Entropy



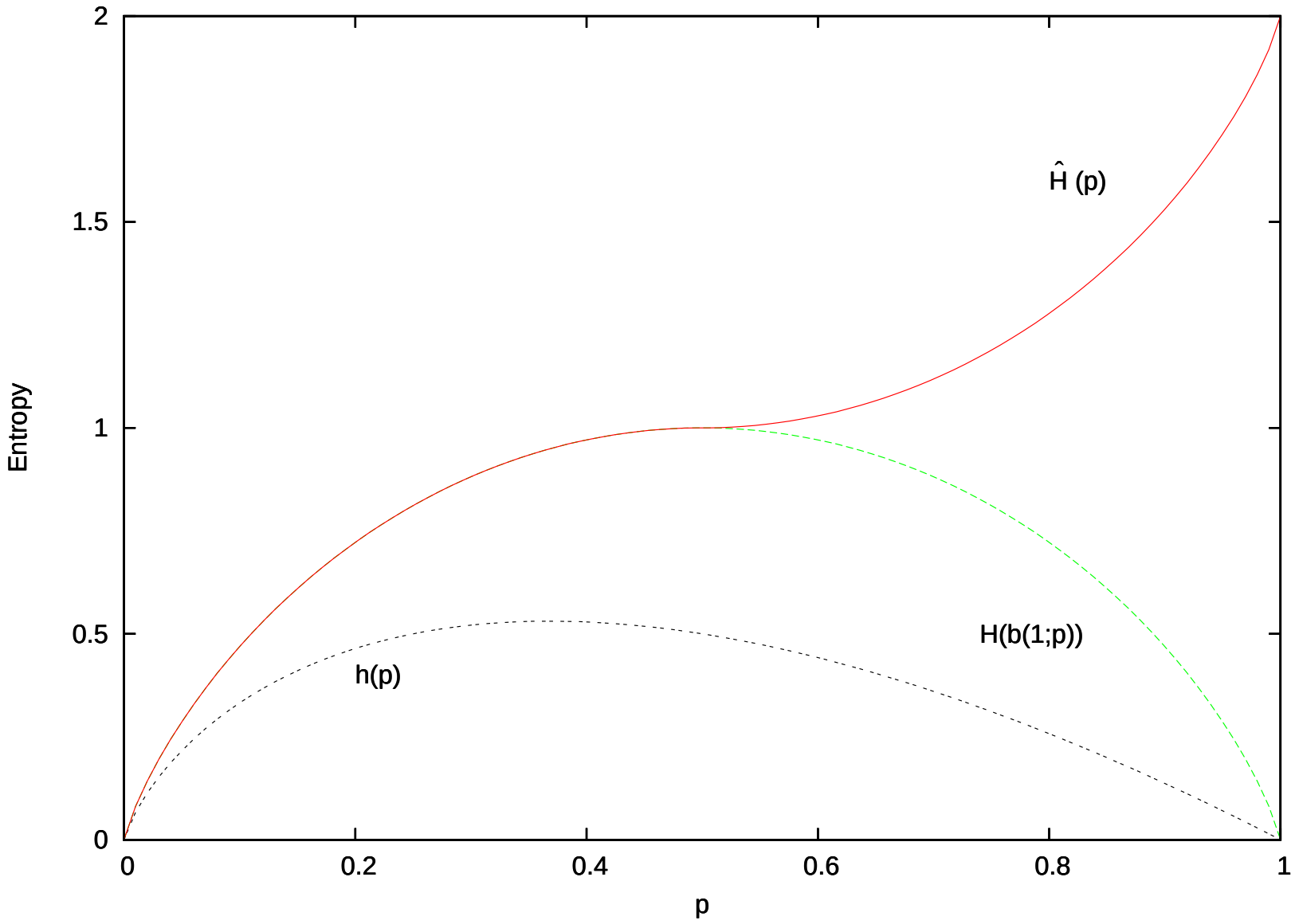
Bernoulli Entropy



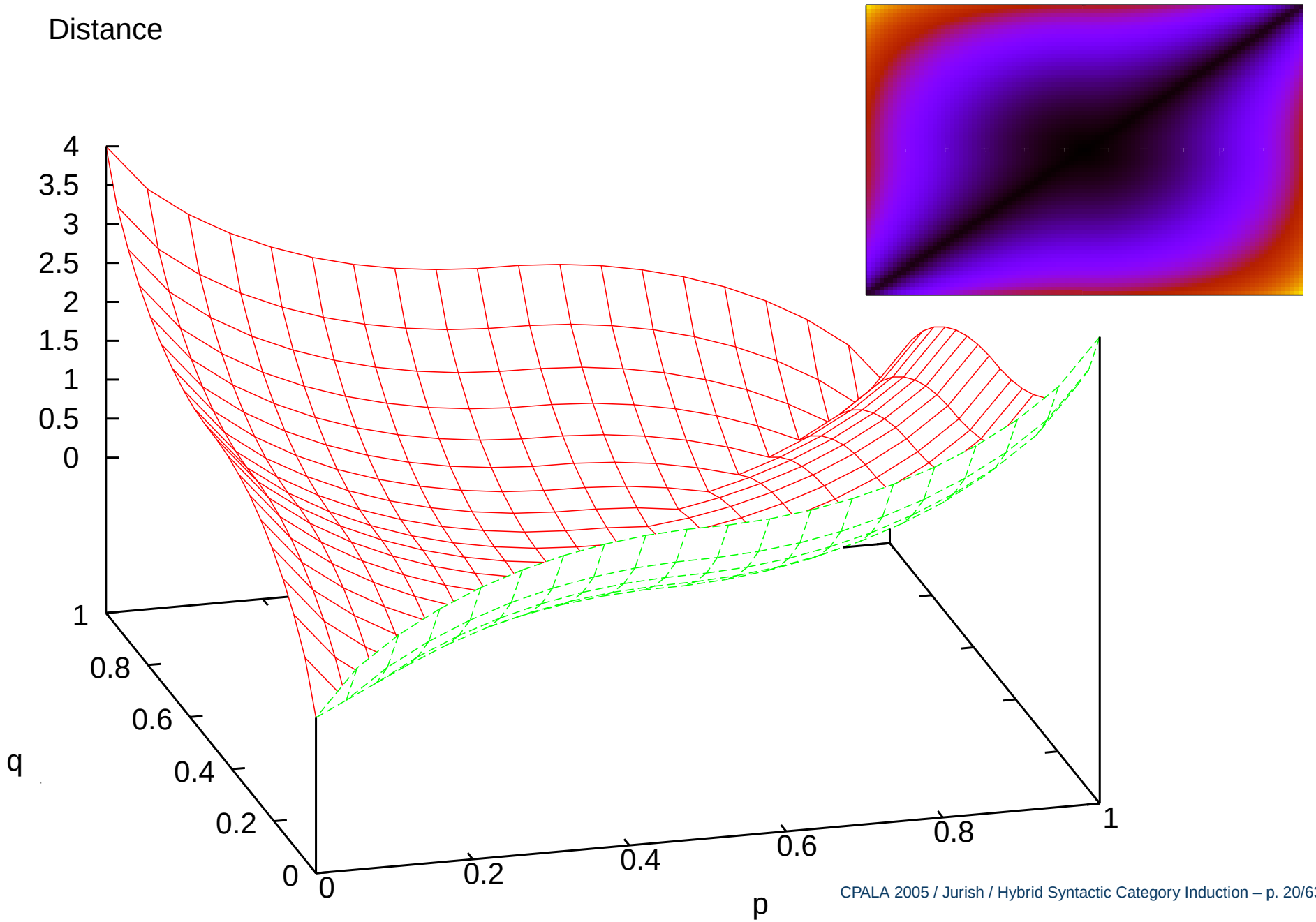
Monotonic Bernoulli Entropy



Monotonic Bernoulli Entropy



Distance: $d_{L1}(\hat{\mathrm{H}}(p), \hat{\mathrm{H}}(q))$



Entropy Distance: Variants

$$\mathrm{H}(Y|X=x) \equiv \mathrm{I}(Y;X=x) \equiv \mathrm{D}(Y|X=x\|Y)$$

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

$\mathrm{H}(Y|X = x) \equiv I(X = x; Y)$

$$
\begin{aligned}
&\forall x_1, x_2 \in \Omega_X : \\
&|I(X = x_1; Y) - I(X = x_2; Y)| \\
&\quad = |\,(H(Y) - H(Y|X = x_1)) - (H(Y) - H(Y|X = x_2))\,| \\
&\quad = |H(Y) - H(Y) - H(Y|X = x_1) + H(Y|X = x_2)| \\
&\quad = |H(Y|X = x_2) - H(Y|X = x_1)| \\
&\quad = |H(Y|X = x_1) - H(Y|X = x_2)|
\end{aligned}
$$

$\square$

$\mathrm{D}(Y|X=x\|Y) \equiv \mathrm{H}(Y|X=x)$

$$
\begin{aligned}
&|D(Y|X=x_1\|Y) - D(Y|X=x_2\|Y)| \\
&= \left|\left[\sum_y p(y|x_1)\log\frac{p(y|x_1)}{p(y)}\right] - \left[\sum_y p(y|x_2)\log\frac{p(y|x_2)}{p(y)}\right]\right| \\
&= \left|\begin{array}{l}\left[\sum_y p(y|x_1)\left(\log p(y|x_1) - \log p(y)\right)\right] \\ -\left[\sum_y p(y|x_2)\left(\log p(y|x_2) - \log p(y)\right)\right]\end{array}\right| \\
&= \left|\begin{array}{l}\left[\sum_y p(y|x_1)\log p(y|x_1)\right] - \left[\sum_y p(y|x_1)\log p(y)\right] \\ -\left[\sum_y p(y|x_2)\log p(y|x_2)\right] + \left[\sum_y p(y|x_2)\log p(y)\right]\end{array}\right| \\
&= \left|H(Y|x_1) - H(Y|x_2) + \sum_y \left(p(y|x_1) - p(y|x_2)\right)\log p(y)\right| \\
&= \left|H(Y|x_1) - H(Y|x_2) + H_{[p(Y|x_1)-p(Y|x_2)]}(Y)\right| \\
&\equiv |H(Y|x_1) - H(Y|x_2)|
\end{aligned}
$$

$\square$

$\hat{\mathrm{H}}$ Clustering

$\hat{\mathrm{H}}$ Target-Vector Features: $\vec{w}_{z,k}(i) = \hat{\mathrm{H}}(\mathrm{P}_{z,k}(b_i|w))$

Distance Functions

- Spearman's rank correlation coefficient, used by Finch & Chater (1993)
- $L1$ ("city-block") distance, used by Korkmaz & Üçoluk (1997), Roberts (2002)
- Vector cosine, used by Schütze (1993,1995)

Link Methods

- Maximum link: $\hat{d}_{max}(W,V) = \max_{\vec{w}\in W,\vec{v}\in V} d(\vec{w},\vec{v})$
- Average link: $\hat{d}_{avg}(W,V) = \mathrm{avg}_{\vec{w}\in W,\vec{v}\in V} d(\vec{w},\vec{v})$

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: $\hat{\mathrm{s}}_k(c, w) = \exp(-\beta_k d_k(c, \vec{w}_k))$
- Simulated melting: $\beta_k = \frac{1}{k}$
- Membership heuristic: $\hat{p}_k(c|w) = \frac{\hat{\mathrm{s}}_k(c,w)}{Z_{k,w} = \sum_{c' \in C_k} \hat{\mathrm{s}}_k(c',w)}$
- Useful restriction: m-best, $m \approx \frac{|C|}{12}$

Bootstrapping

Cluster-based Profiling

for $w \in T_k, b, c \in B_k = C_{<k}, z \in \{\ell, r\}$,

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

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

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: $O \in \mathcal{O}(\mathcal{C}^2|\mathcal{A}|)$

Corpora

German: NEGRA (Skut et al. 1997)

- 355,096 tokens ; 48,924 types ; reduced tagset

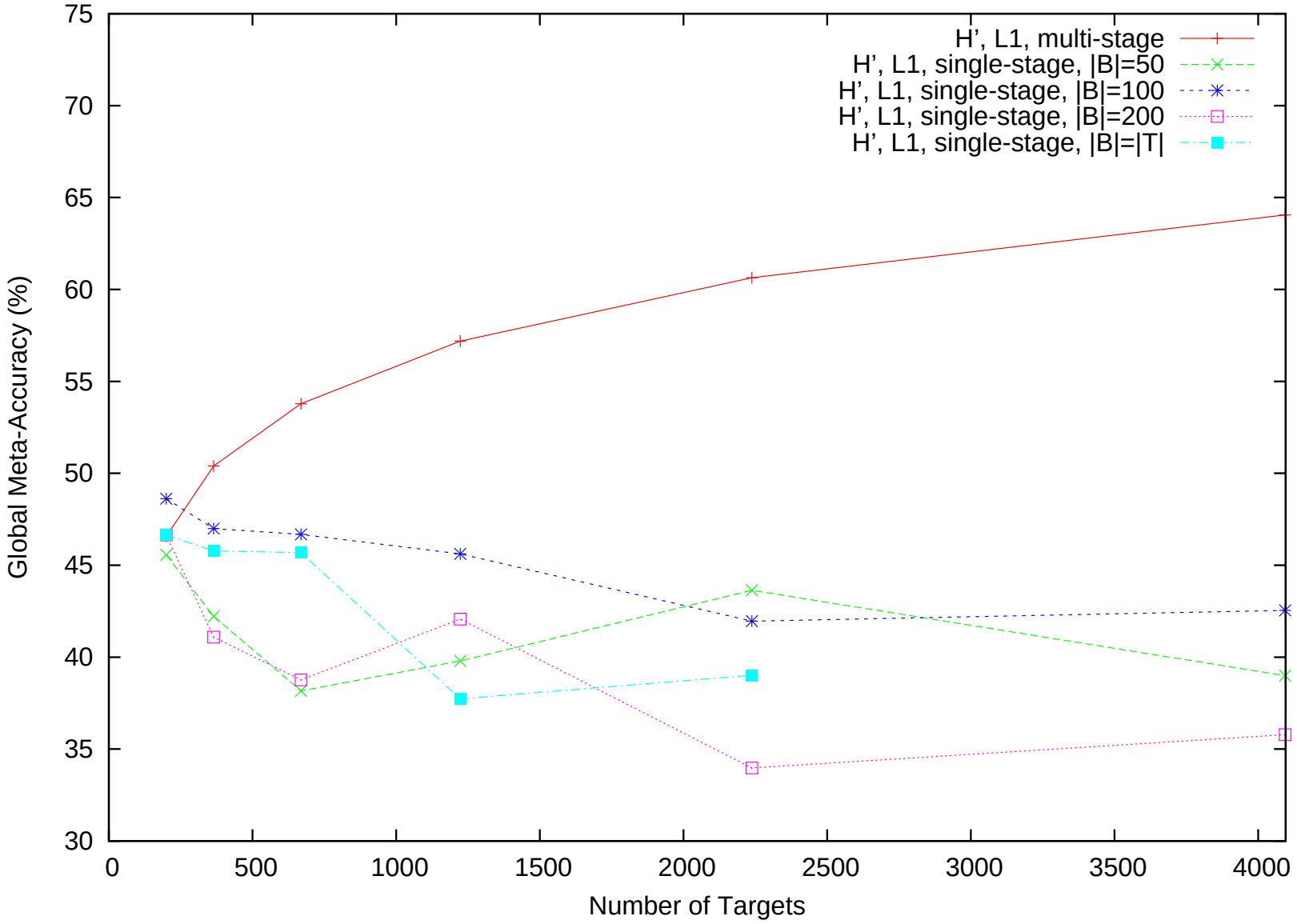
English: SUSANNE (Sampson, 1995)
and *Great Expectations* (Dickens, 1861)

- 374,640 tokens ; 20,600 types ; reduced tagset

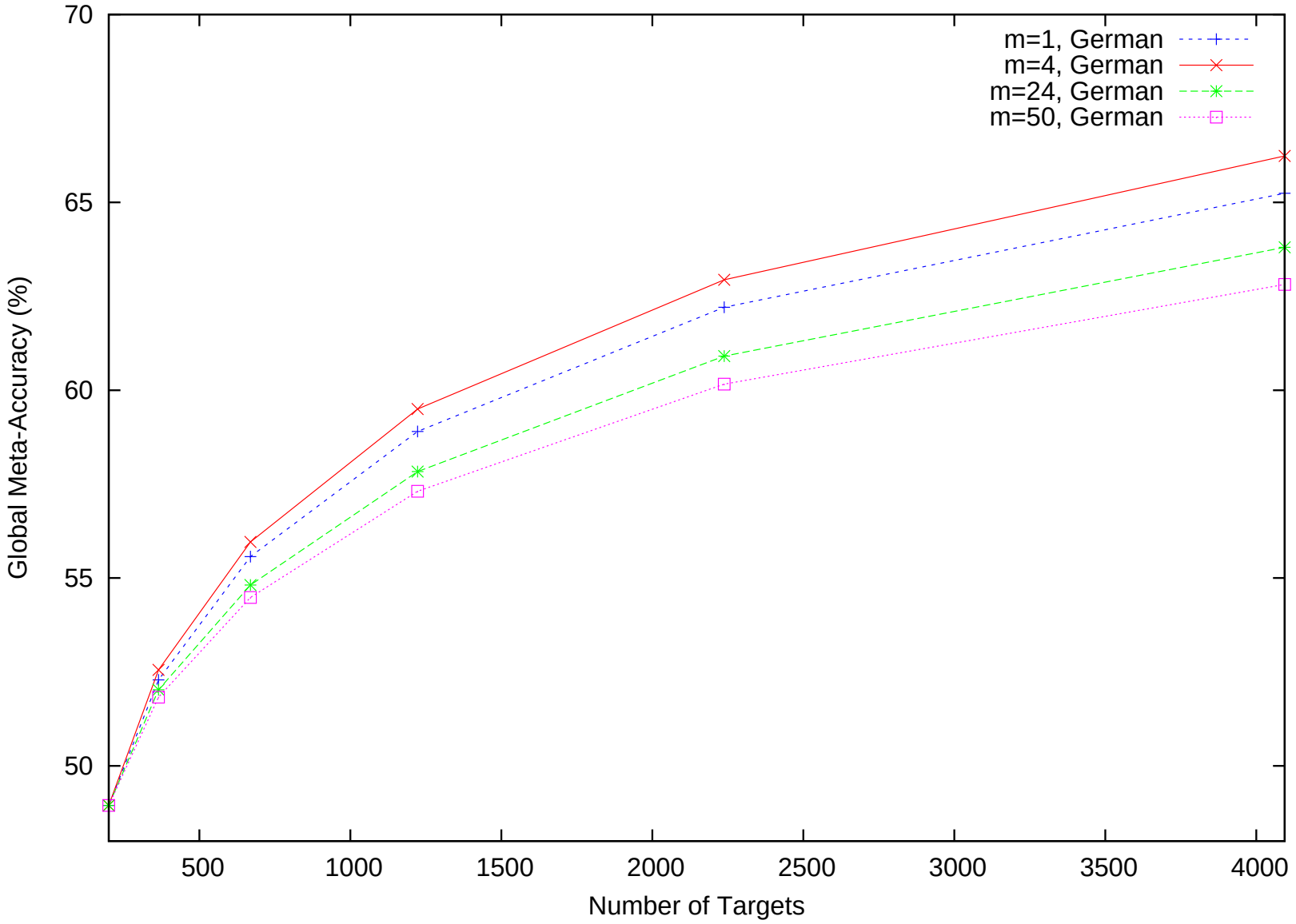
Preprocessing

- 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

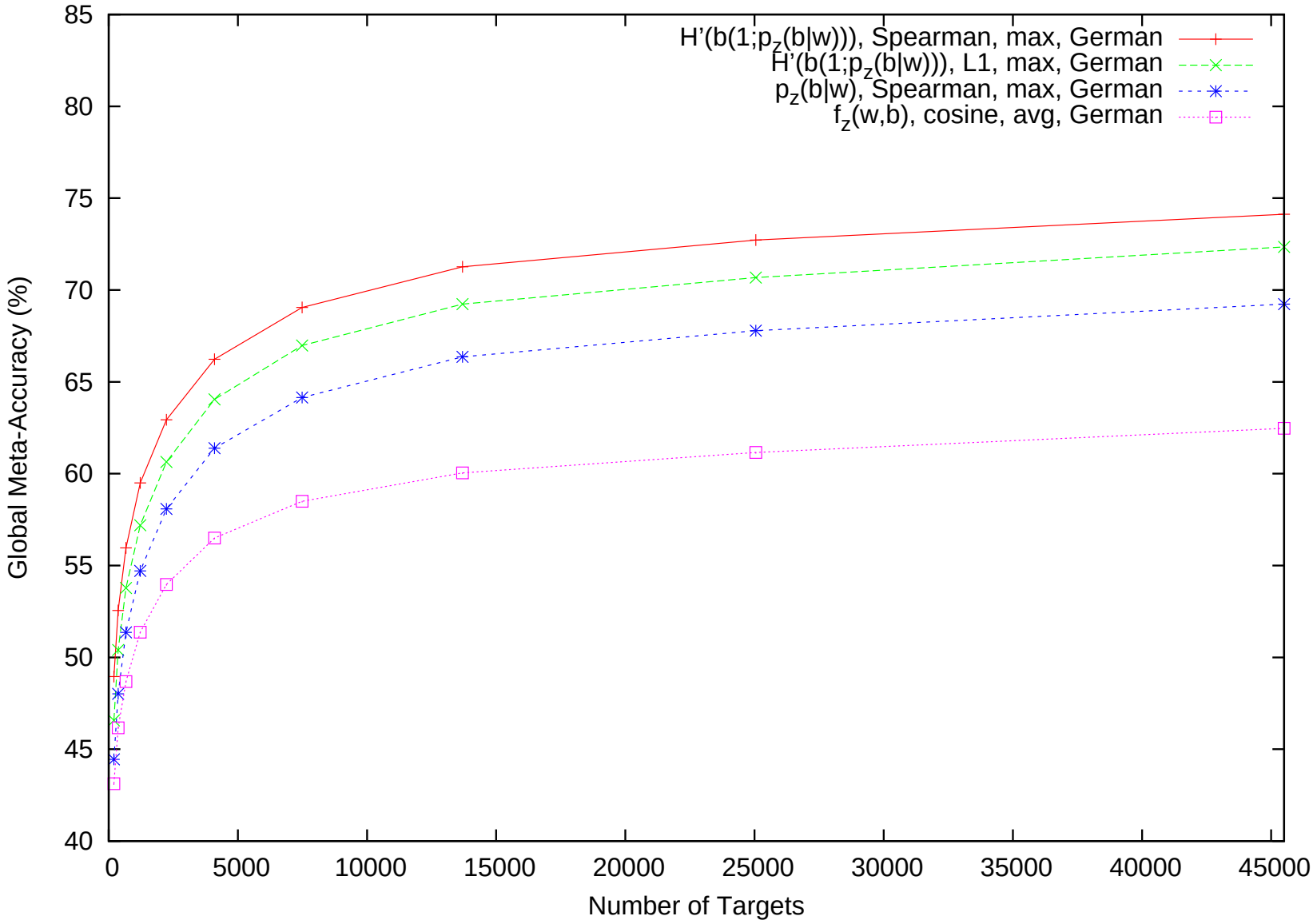
Multi- *vs.* Single-Stage



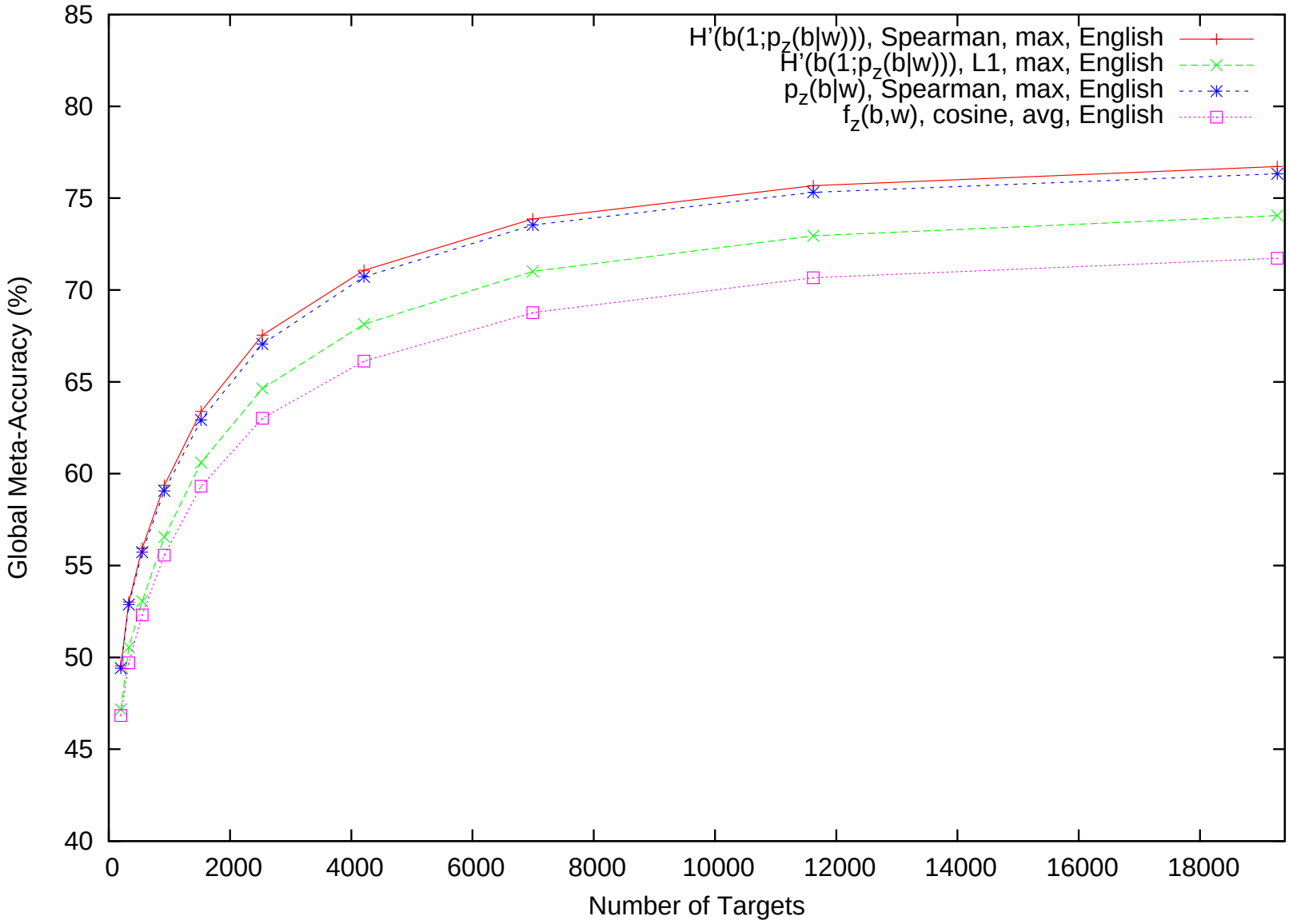
Fuzzy Clusters



Global Precision, German

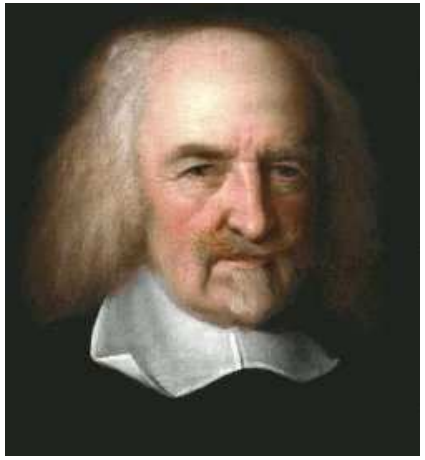


Global Precision, English



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

$$\pi(q) = P(Q_1 = q)$$ (Initial probabilties)

$$A(q_i, q_j) = P(Q_{t+1} = q_j | Q_t = q_i)$$ (Arc probabilities)

$$B(w, q) = P(W_t = w | Q_t = q)$$ (Emission probabilities)

Parameter Initialization (B)

$$\hat{p}_{\leq K}(w|c) = \frac{\hat{p}_{\leq K}(c|w)\hat{p}_{\leq K}(w)}{\hat{p}_{\leq K}(c)}$$ (Bayes)

where:

$$\hat{p}_{\leq K}(w) = \frac{\mathrm{P}_{\ell,K}(w) + \mathrm{P}_{r,K}(w)}{2}$$ (MLE)

$$\hat{p}_{\leq K}(c) = \sum_{w \in T_{\leq k}} \hat{p}_{\leq K}(w, c)$$ (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

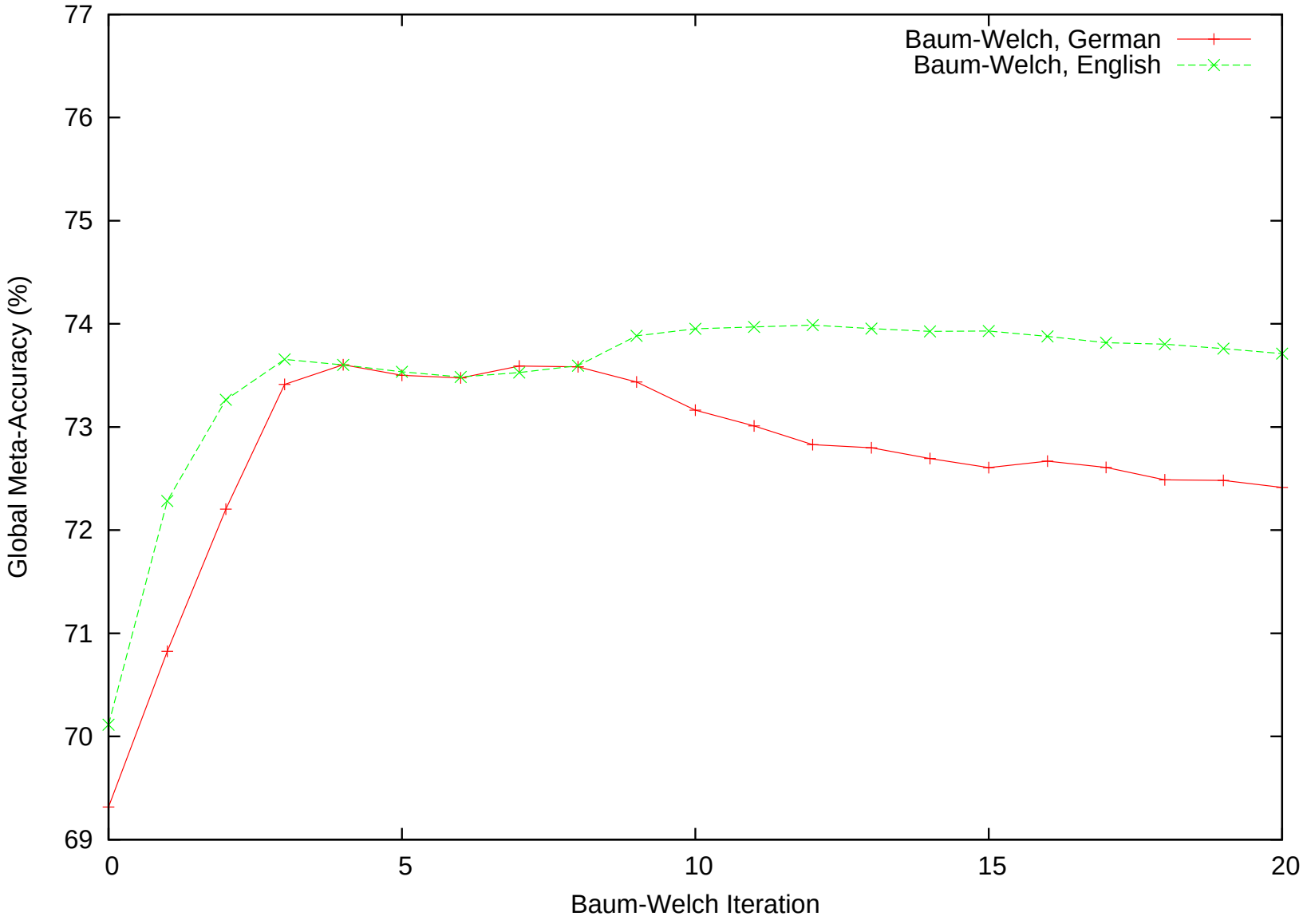
Results

- Hard clusters from π_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)



Trigram Clustering

Idea

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

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

$$\overrightarrow{\langle w_1, w_2, w_3 \rangle} = \overrightarrow{(w_1)}_r \circ \overrightarrow{(w_2)}_\ell \circ \overrightarrow{(w_2)}_r \circ \overrightarrow{(w_3)}_\ell$$

But wait, there's more!

- Cluster-based component profiling
- Frequency-based prototype selection
- Attachment of remaining trigrams to centroids

Trigram Clustering: Results

Clustering Stage	German		English	
	Amb. Rate	Meta-Acc.	Amb. Rate	Meta-Acc.
π_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

i he she we they who
said made saw took done found asked told . . . seen
were been are am
the a his my an their your our its every
going being looking having coming taking getting ...
hands head heart eyes face mouth arm hair water word ...
see hear get put give take want tell speak mean love ...
that which what how where whom
did thought knew felt hope wanted does doubt need
not n’t only rather
then now still why yet perhaps thus sometimes therefore
herbert biddy estella pumblechook provis compeyson ...

Suspicious Clusters

of in with at for on by from into about ... to
, ; : ( ) – ago
this one such these another those many each town
all more ... nothing something ... saying thinking
as if when while though because since whether ... afraid
old young long short same certain strange ... present
gentleman lady fellow convict ... havisham jaggers wopsle
it you me him them us yourself drummle
out up down back off ... near himself care ready known
few two three four five ... great general kitchen state
there here none indeed sir uncle

Tagsets

Reduced STTS Tagset

Tag	STTS Tag(s)	Description
ADJ	ADJA ADJD PIDAT	Adjective
ADV	ADV APPO APZR *AV PTK* (except PTKZU)	Adverb
CARD	CARD	Cardinal number
CCONJ	KON	Coord. conjunction
DET	ART PDAT PIAT PPOSAT PRELAT PWAT	Determiner
MISC	FM ITJ XY	Miscellaneous
NOUN	NE NN TRUNC	Nominal
PREP	APPRART APPR	Preposition
PRON	PDS PIS PPER PPOSS PRELS PRF PWS	Pronominal
SCONJ	KOKOM KOUI KOUS	Subord. conjunction
TO	PTKZU	Infi nitival *zu*
VFIN	VAFIN VAIMP VMFIN VVFIN VVIMP	Finite verb
VINF	V* (except V*INF V*ZU V*PP)	Infi nitive, participle
$,	$,	Comma
$.	$.	Sentence-fi nal punct.
$(	$(	Sentence-internal punct.

Reduced SUSANNE Tagset

Tag	SUSANNE Tag(s)	Description
ADV	FA* FB* LE* XX	Adverb
DET	A* D*	Determiner
CARD	M*	Cardinal number
CCONJ	CC*	Coordinating conjunction
POS	G*	Genitive marker
MISC	FO* FU* FW* UH ZZ*	Miscellaneous
NOUN	N* BTO22	Nominal
PREP	I* BTO21	Preposition
PRON	P* EX	Pronominal
PUNC	Y*	Punctuation
SCONJ	CS*	Subordinating conjunction
TO	TO	Infi nitival *to*
VFIN	V* (except V*0, V*G*)	Finite verb form
VINF	VB0 VD0 VH0	Infi nitive verb form
VING	VBG VDG VHG VVG*	-ing verb form

Monotonic Bernoulli Entropy

Features: Shannon Entropy

- Shannon Entropy (Shannon and Weaver, 1949):

$$\mathrm{H}(p) = -\sum_{x \in \mathrm{dom}(p)} p(x) \log_2 p(x)$$

- Measure of the (un)predictability of p
- Average length of a message that an event from $dom(p)$ has occurred under an optimal encoding
- Properties:
  - $\mathrm{H}(p) \geq 0$
  - Measurement unit: *bits*
- Pointwise entropy: $h_p(x) = -p(x) \log p(x)$
  - Unfortunately asymmetric

Bernoulli Entropy

- Assume a Bernoulli distribution $X_x \sim b(1; p = \mathrm{P}(x))$ for each relevant point x, then:

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

- $\mathrm{H}(b(1;p))$ 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 $\hat{\mathrm{H}}$

Definition

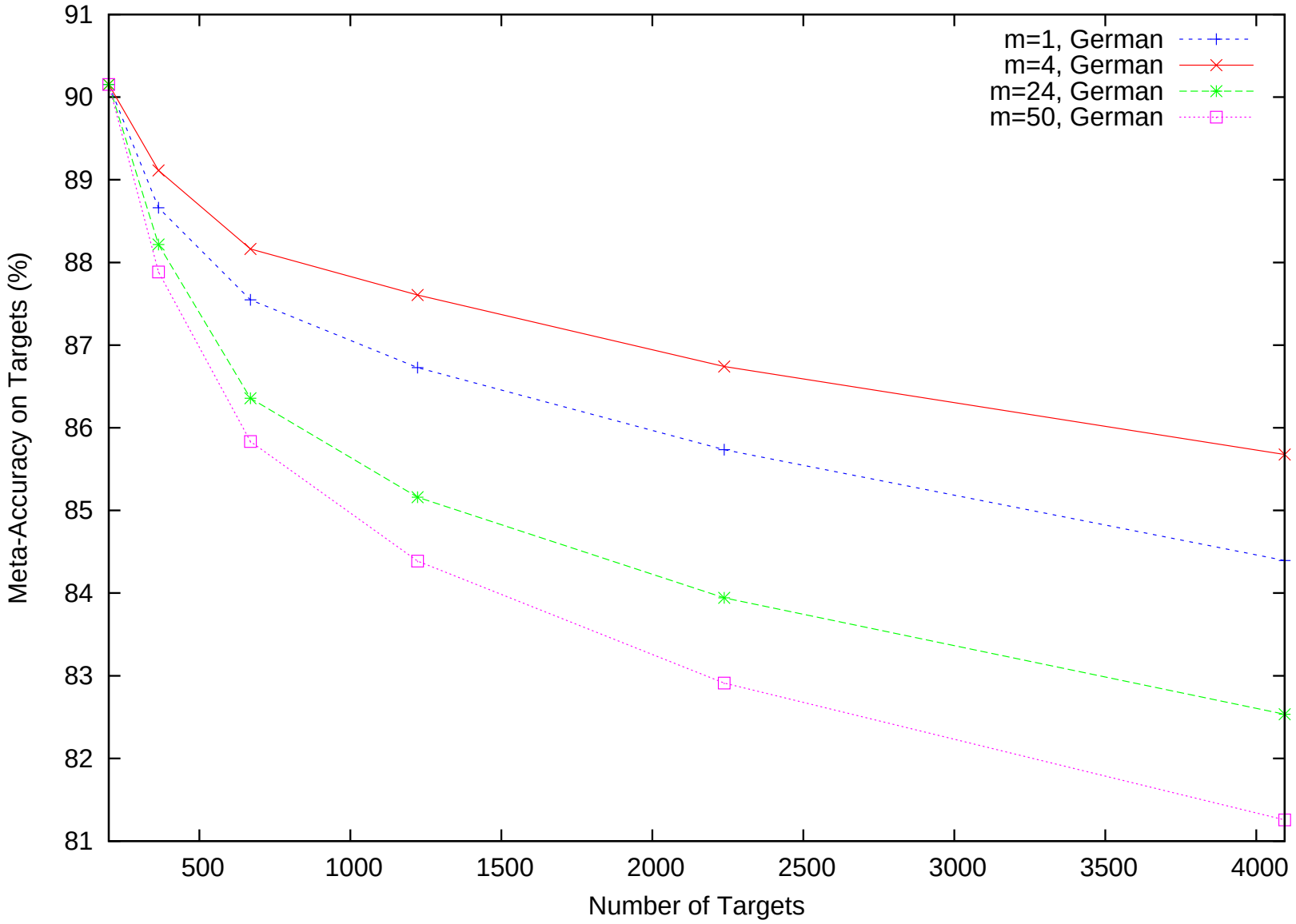
$$\hat{\mathrm{H}}(p) = \begin{cases} \mathrm{H}(\mathrm{b}(1;p)) & \text{if } p \leq \frac{1}{2} \\ 2 - \mathrm{H}(\mathrm{b}(1;p)) & \text{otherwise} \end{cases}$$

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

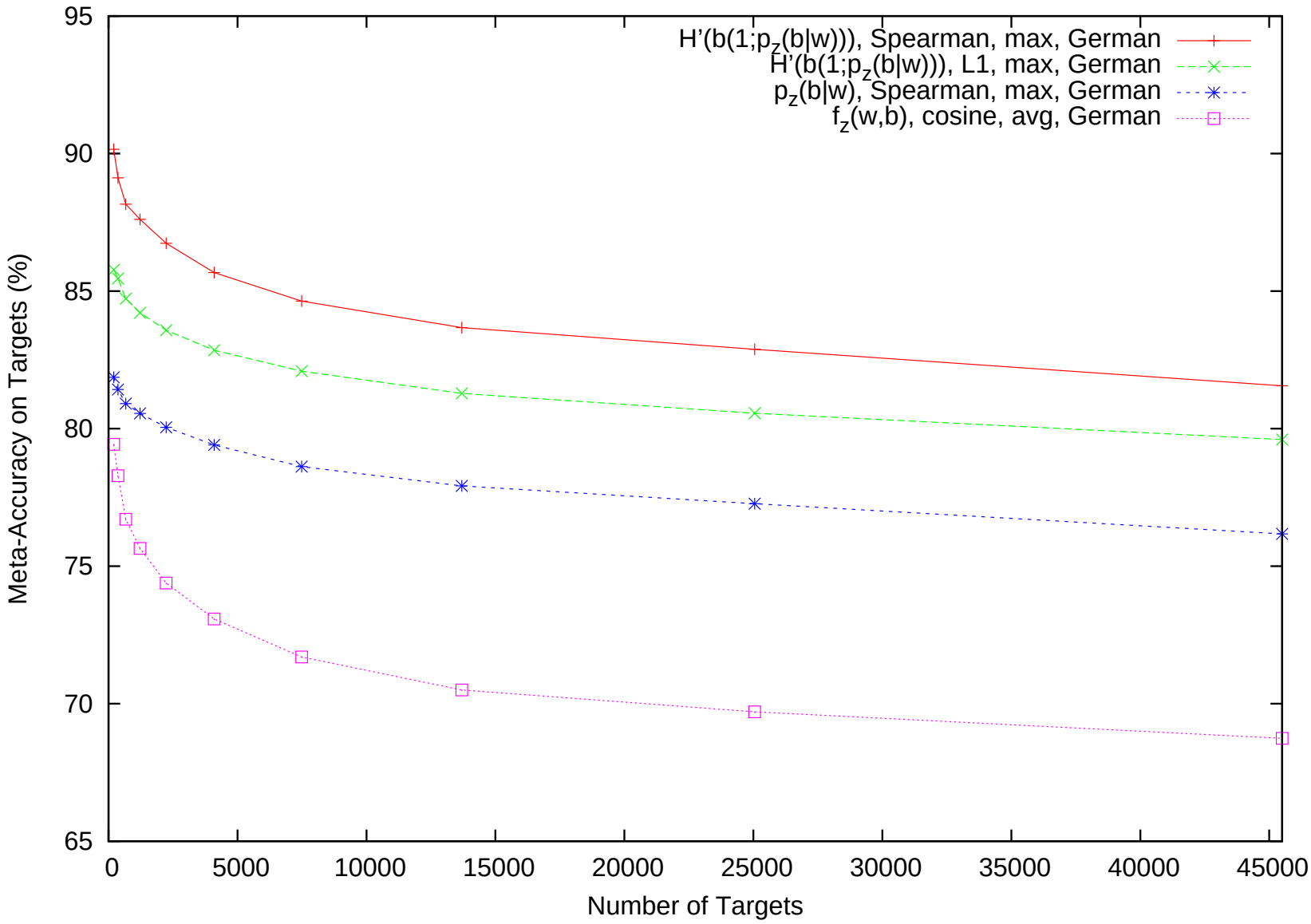
- Mnemonic utility of chunking a boundary event b into a target event w.

Target Data

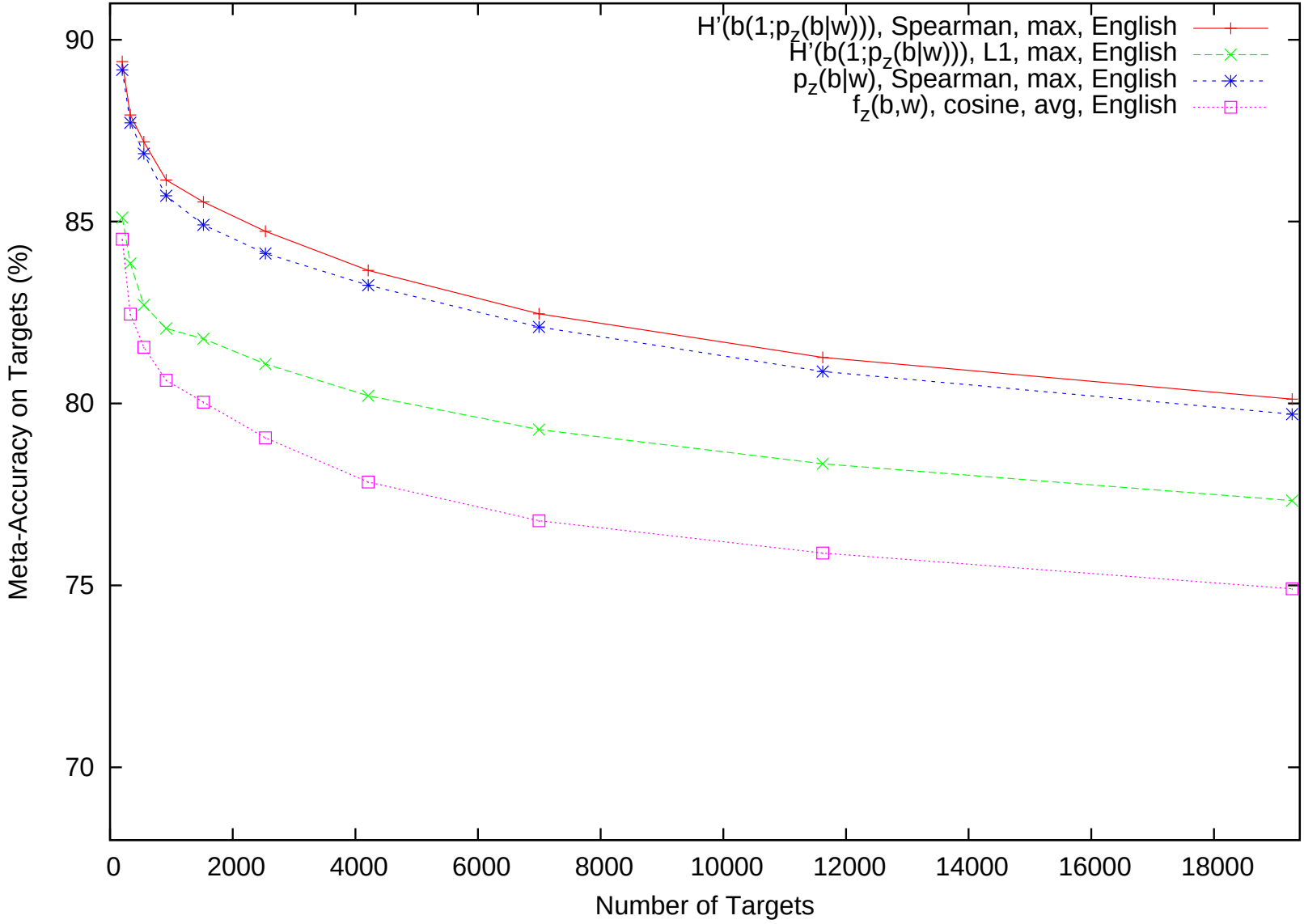
Fuzzy Clusters



Target Precision, German



Target Precision, English



HMM Data

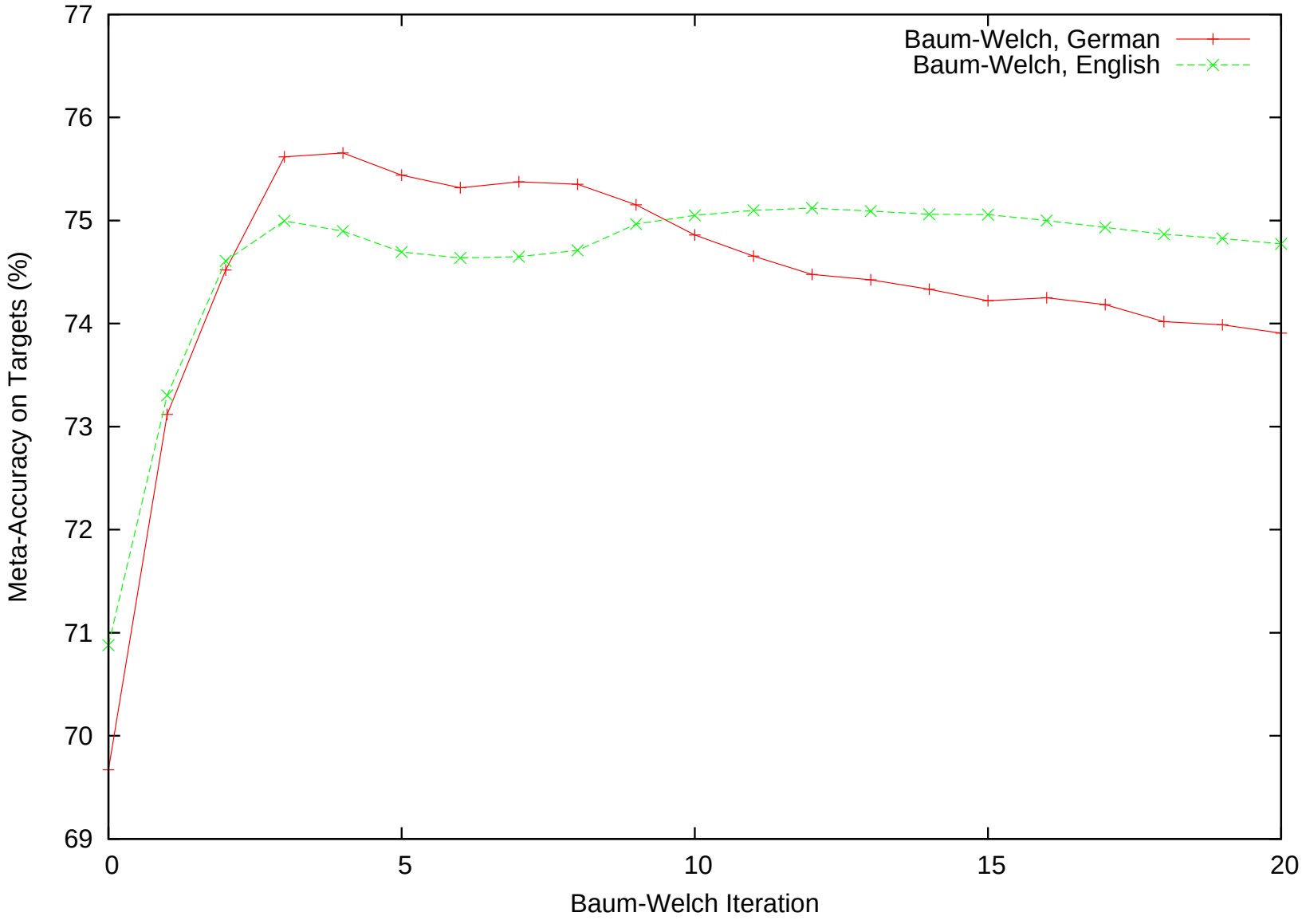
HMM EM: Results: German

Baum-Welch Iteration	Global		Targets	
	Amb. Rate	Meta-Acc.	Amb. Rate	Meta-Acc.
π_K	1.00	74.13 %	1.00	81.56 %
0	1.00	69.32 %	1.00	69.67 %
4	1.22	73.60 %	1.31	75.66 %
8	1.29	73.58 %	1.40	75.35 %
12	1.32	72.83 %	1.46	74.48 %
16	1.34	72.67 %	1.49	74.25 %
20	1.35	72.41 %	1.50	73.91 %

HMM EM: Results: English

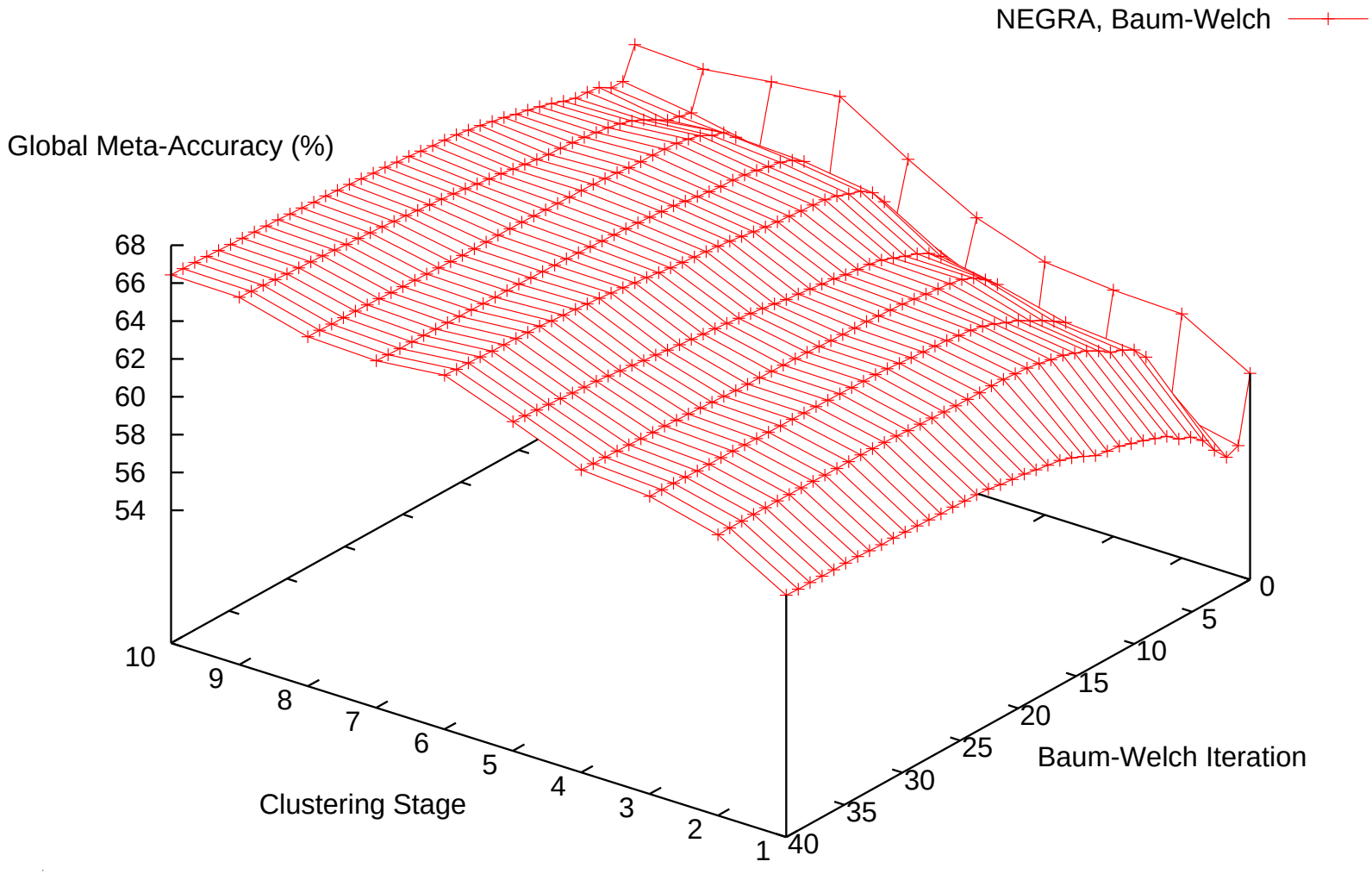
Baum-Welch Iteration	Global		Targets	
	Amb. Rate	Meta-Acc.	Amb. Rate	Meta-Acc.
π_K	1.00	76.72 %	1.00	80.12 %
0	1.00	70.11 %	1.00	70.88 %
4	1.37	73.60 %	1.46	74.90 %
8	1.48	73.60 %	1.59	74.71 %
12	1.56	73.99 %	1.68	75.12 %
16	1.61	73.88 %	1.75	75.00 %
20	1.65	73.71 %	1.80	74.77 %

HMM EM: Results (Targets)



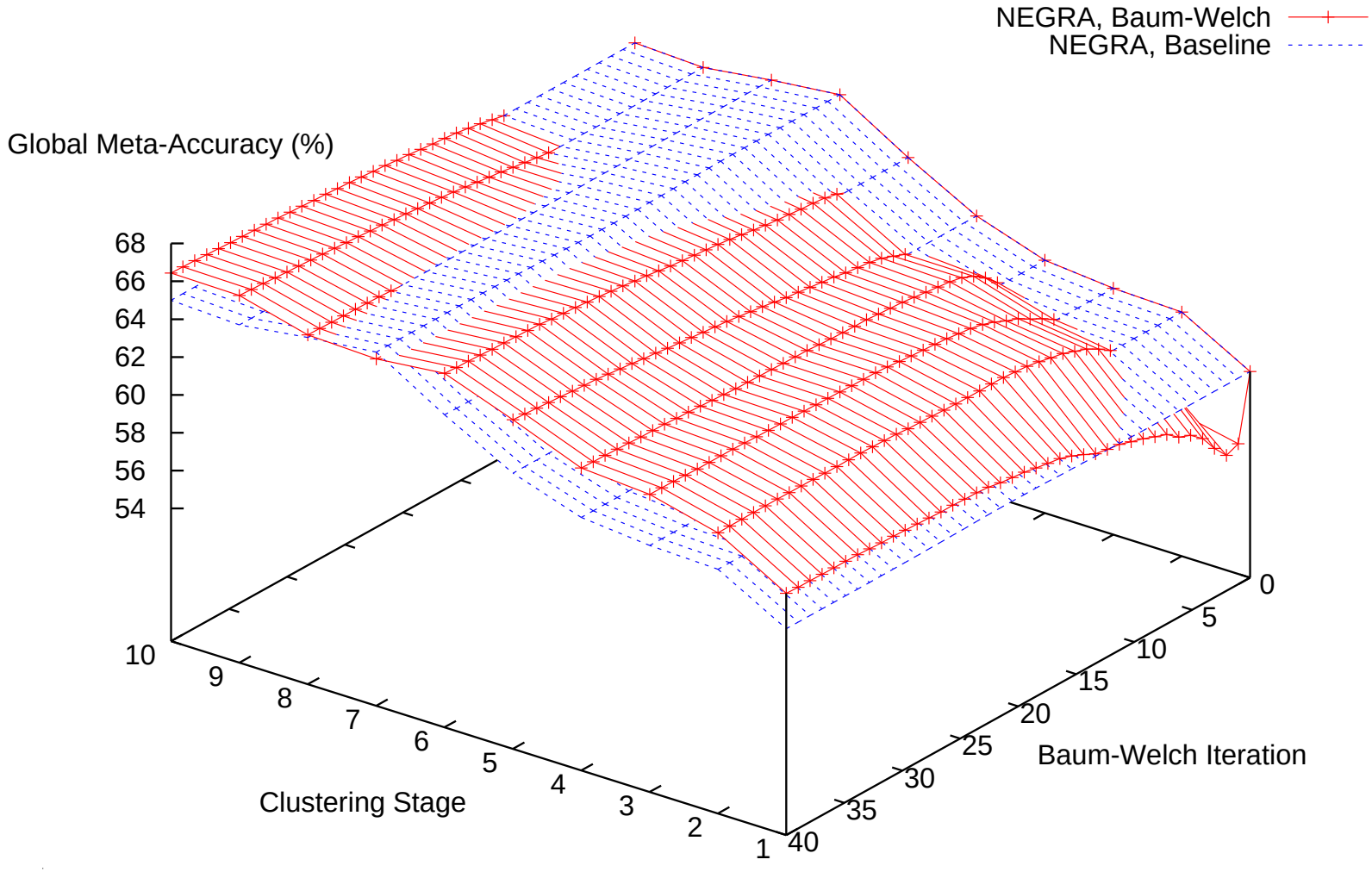
Old Data

HMM Reestimation, Global*



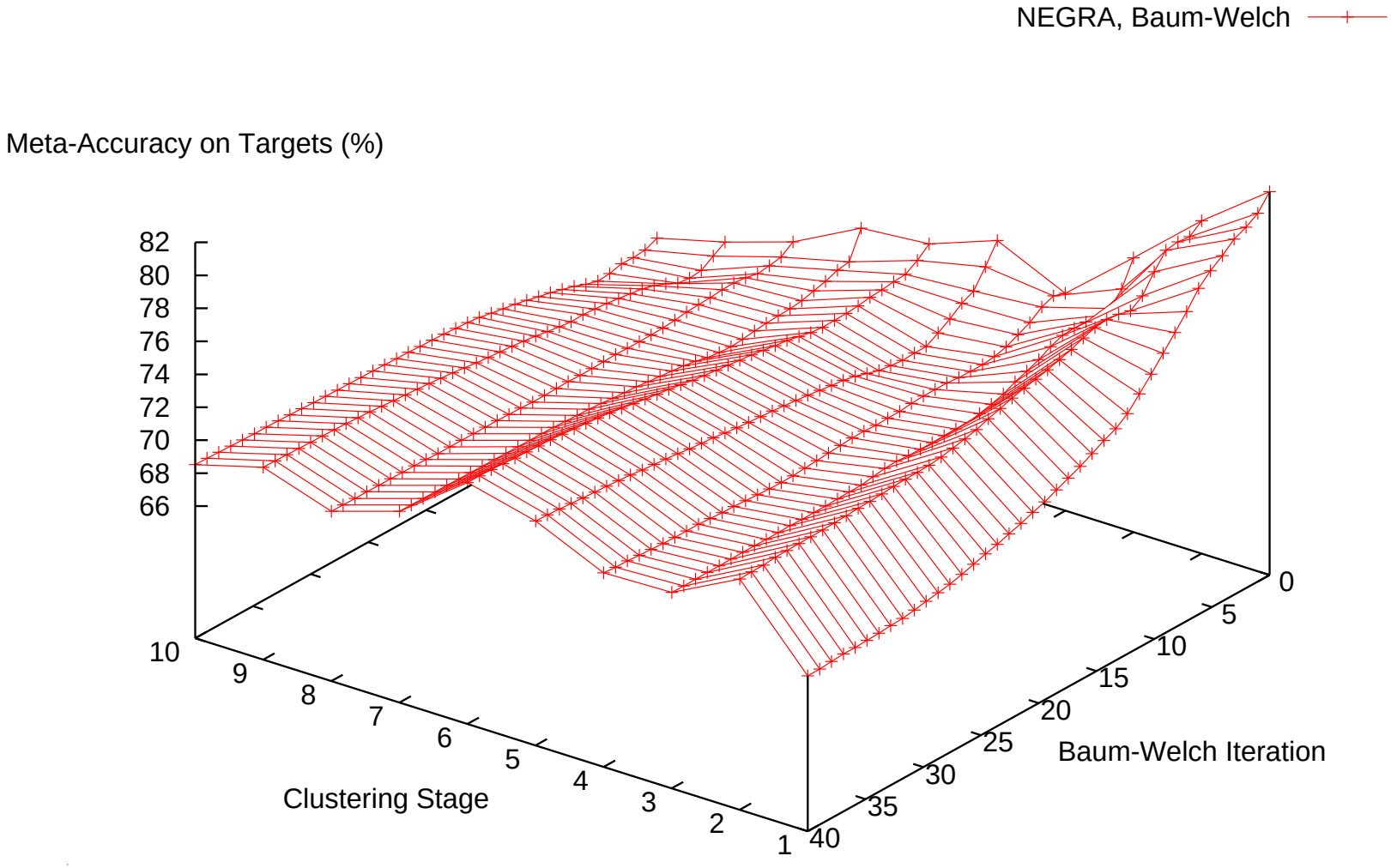
* alternate confi guration

HMM Reestimation, Global*



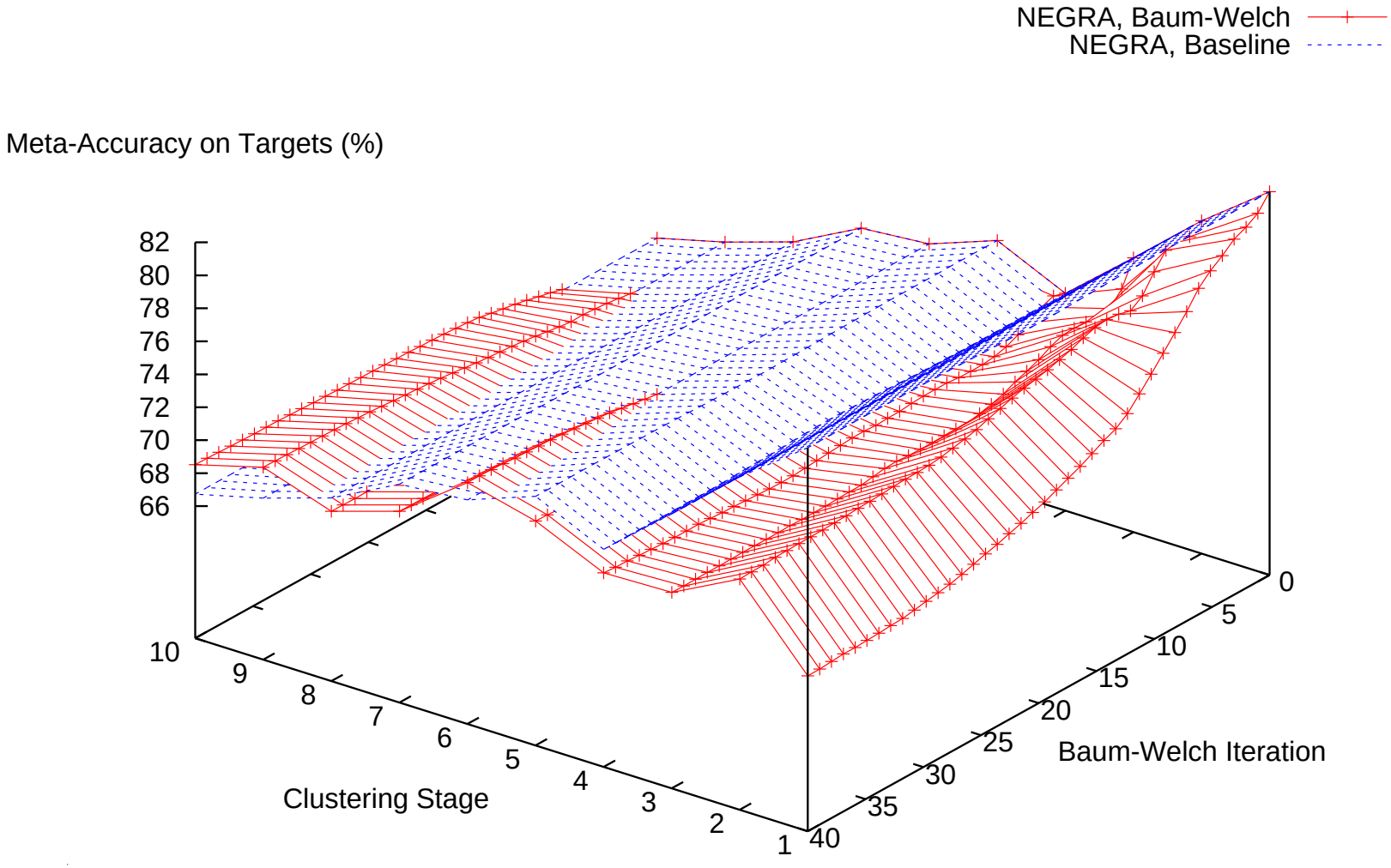
* alternate confi guration

HMM Reestimation, Targets*



* alternate confi guration

HMM Reestimation, Targets*



* alternate confi guration