Syntactic Category Induction

Bryan Jurish

jurish@ling.uni-potsdam.de

Universität Potsdam, Institut für Linguistik,

Potsdam, Germany

2005-06-23

Outline

The Big Picture

- Motivation
- Evaluation

Clustering Phase

- Target & Bound Selection
- Clustering Data
- Fuzzy Clusters
- Bootstrapping

Reestimation Phase

- Parameter Initialization
- Results

The Big Picture

(1) Category induction $\sim$ PoS identification

(2) Surface modelling $\sim$ Grammar induction

(3) Chunk detection $\sim$ Consituent analysis

(4) Dependency resolution $\sim$ Projection relation

(5) Lexical indexing $\sim$ Lexicon reification

Lexical Category Induction

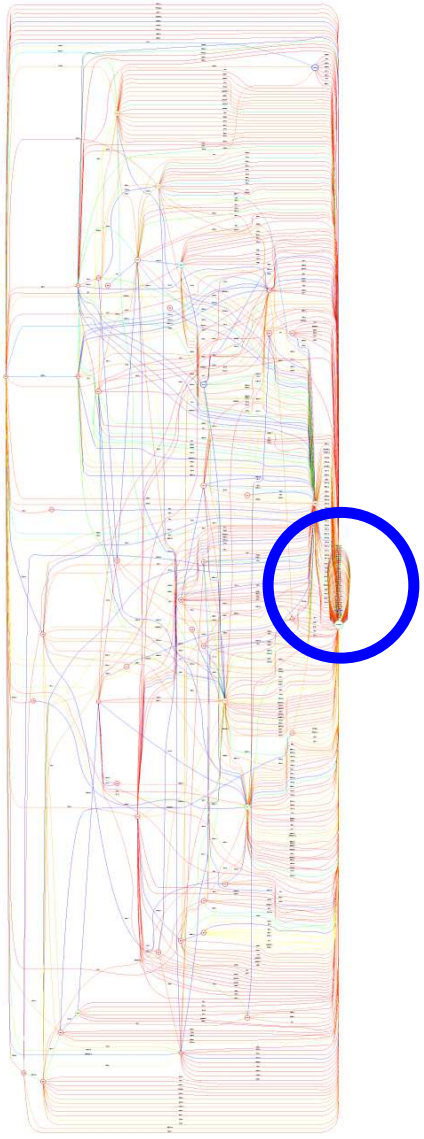
Motivation

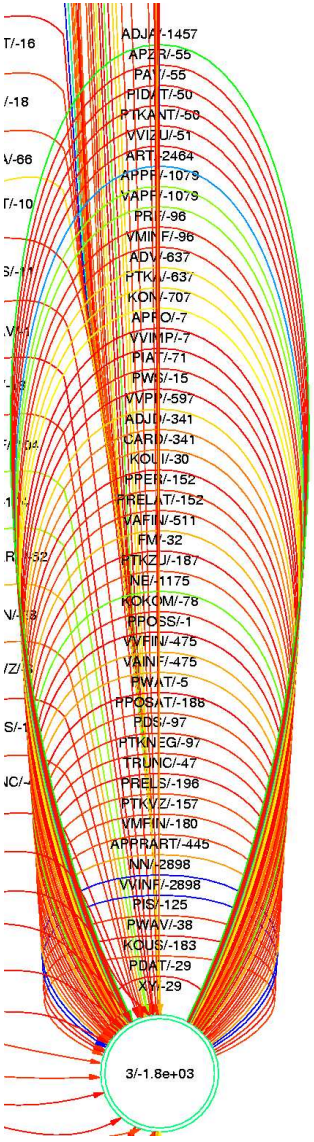
- Both theoretical and empirical results suggest the existence of cognitively salient syntactic categories.
- Provides drastic reduction of the data space
  - sparse data problem workaround

Implementation

- Hybrid iterative-hierarchical fuzzy agglomerative clustering over word-types
  - Clustering features: relative mutual information
  - Frequency-based target- and bound-selection
  - Zipf's law used to derive the clustering schedule
- HMM reestimation: recover token-level ambiguity

Categories: Motivation





Categories: Evaluation

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

- How can the quality of an induced model θ of syntactic category assignment be judged?

The Answers, or "Beats the heck outta me, bro. . ."

- Sample Entropy $\mathrm{H}(\theta|S)$ (Brown et al.)
- Text probability: $P(\mathcal{C}'|\theta)$ (Brown et al., Clark)
- Metamodelling (Clark, Roberts, Schütze)
  - Evaluate wrt. a "gold standard" $\tau_G : \mathcal{C} \rightarrow T_G^{|\mathcal{C}|}$
  - Train θ' (supervised), defining $\tau' : T^* \rightarrow T_G^*$
  - Per-word accuracy of metamodel is then

$$\frac{|\{\langle i,j\rangle \ : \ \tau'(\tau(\mathcal{C}))_{ij} = \tau_G(\mathcal{C})_{ij}\}|}{\sum_{s\in\mathcal{C}}|s|}, 1 \leq i \leq |\mathcal{C}|, 1 \leq j \leq |\mathcal{C}_i|$$

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 Phase

The Big Idea

- Cluster targets $T_k \subset \mathcal{A}$ wrt. fixed set of bounds B_k into classes C_k
  - similar to Roberts (2002), Finch & Chater (1993)
- Break clustering problem down into K stages
  - similar to Brown et al. (1992), Schütze (1993,1995)
- Bootstrap classification using earlier solutions
  - extend traditional agglomerative clustering approach
- Use fuzzy membership heuristic
  - approximate type-level ambiguity

Clustering Phase: Algorithm

for $k = 1$ to K **do**
  Select stage targets $T_1 \subset \mathcal{A}$
  if k==0 **then**
    $B_k = T_k$ /* Bootstrapping stage */
  else
    $B_k = C_{k-1}$ /* Bootstrapped stages */
    $T_k = T_k \cup C_{k-1}$
  end if
  for $w \in T_k$ **do**
    $\vec{w}_1 = \phi(f_k, f_0)$ /* Target vectors */
  end for
  $C_k = \text{cluster}(M_k = [\vec{w}_k]_{w \in T_k})$
  $\hat{p}_k(c \in C_k | w \in T_{\leq k}) = \text{fuzzy}(M_k, C_k)$ /* Heuristic */
end for

Target & Bound Selection

Stage 1:

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

Stage $k > 1$:

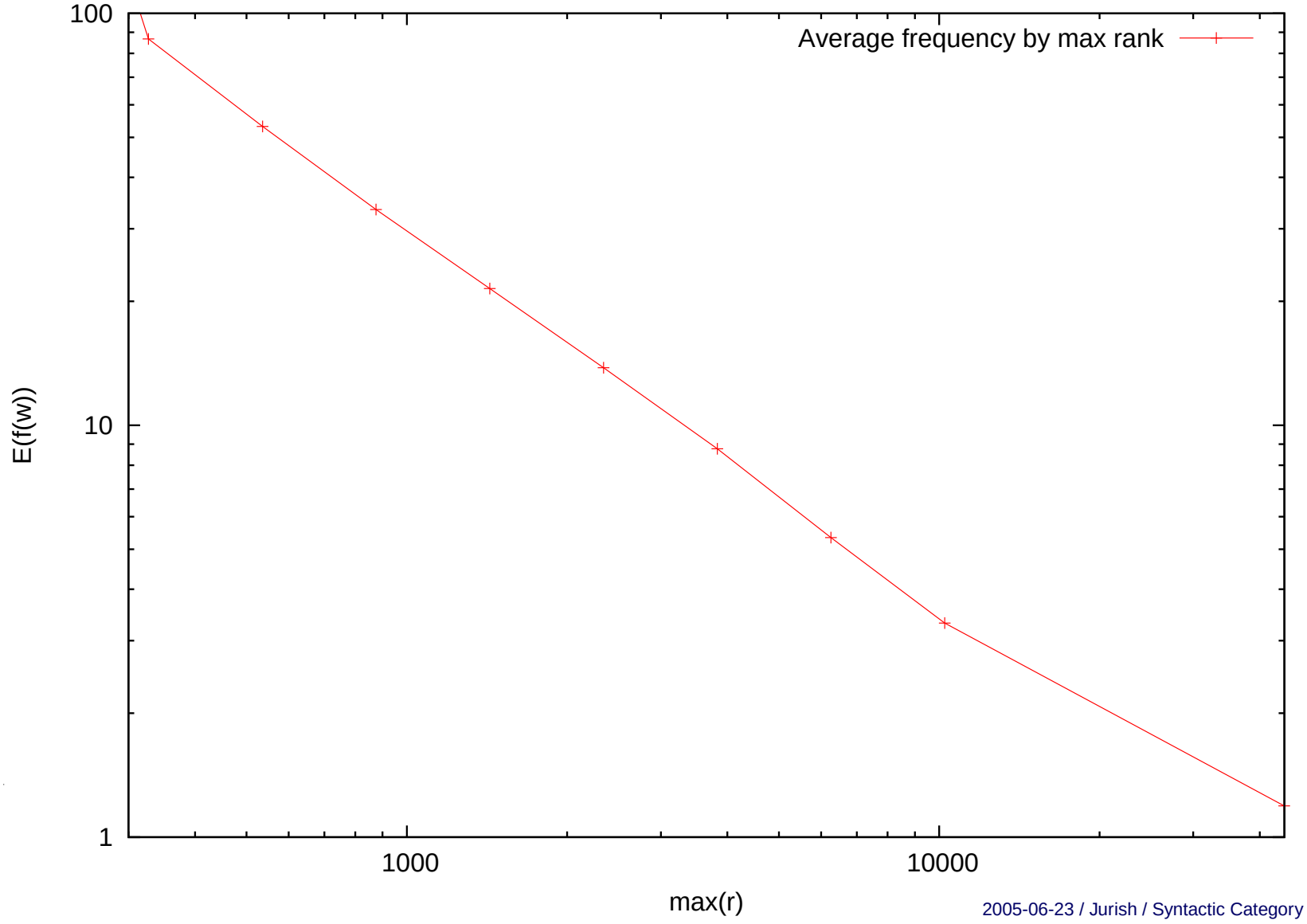
$$\begin{aligned} B_k &= C_{k-1} \\ T_k &= \{w \in \mathcal{A} \mid r_{k-1} \leq r(w) \leq r_k\} \\ \log r_k &= \log r_{k-1} + \left(\frac{\log(|\mathcal{A}|) - \log(|r_1|)}{K-1}\right) \end{aligned}$$

Properties:

$$\begin{aligned} i \neq j &\implies T_i \cap T_j = \emptyset \quad \text{(OaOO)} \\ \operatorname{avg}_w \log f_k(w) &\approx ak + b \quad \text{(Zipf)} \end{aligned}$$

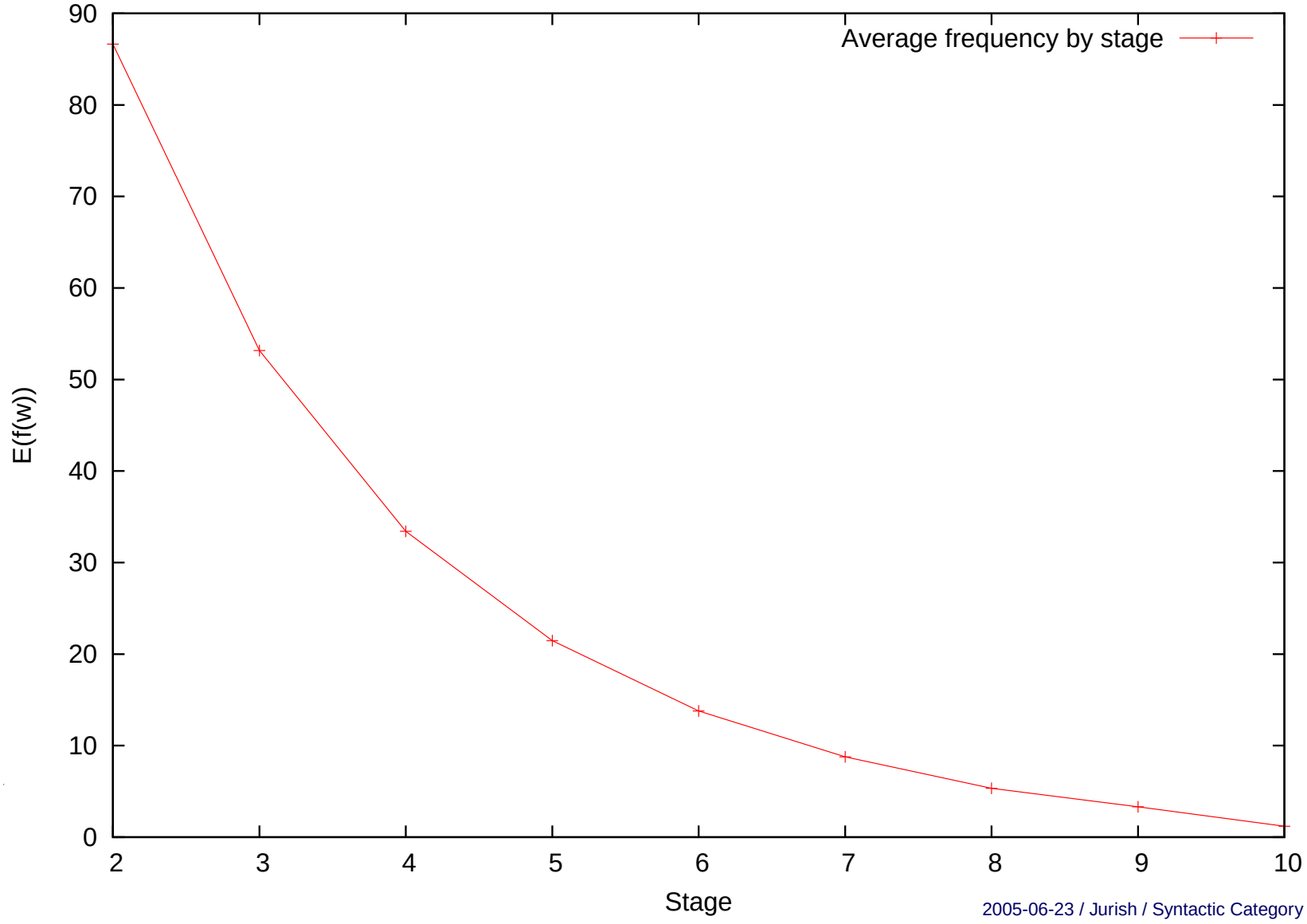
Target Selection: Example

By Rank (log scale)



Target Selection: Example

By Stage



Clustering: Data

Base Data (bigram frequencies):

$$f_0 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$$

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 $z \in \{\ell, r\}, k \in K$,

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

Clustering: More Data

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:

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

Conditional bigram vectors:

$$\vec{w}_{z,1}(i) = P_{z,1}(b_i|w) = \frac{P_{z,1}(w,b_i)}{P_{z,1}(w)}$$

Clustering: Relative MI

Pointwise Mutual Information (PMI)

$$I(x;y) = \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$$

- highly frequency-sensitive

Relative Pointwise Mutual Information (RPMI)

$$\tilde{I}(y|x) = \frac{I(x;y)}{\sum_{y' \in \Omega_Y} I(x;y')}$$

- may (still) be negative

Clustering: Normalized RMI

Normalized Relative Pointwise MI (NRPMI)

$$\hat{I}(x;y) = I(x;y) - \min_{x' \in \Omega_X, y' \in \Omega_Y} I(x';y')$$

$$\hat{I}(y|x) = \frac{\hat{I}(x;y)}{\sum_{y' \in \Omega_Y} \hat{I}(x;y')}$$

- values in the range $[0, 1]$
- Bonus Question: where's the bug?

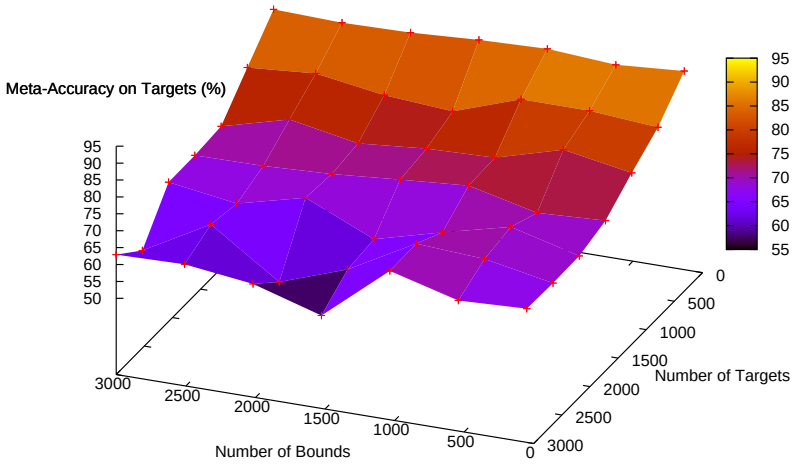
NRPMI target vectors:

$$I_{z,k}(w;b) = \log_2 \frac{P_{z,k}(w,b)}{P_{z,k}(w) P_{z,k}(b)}$$

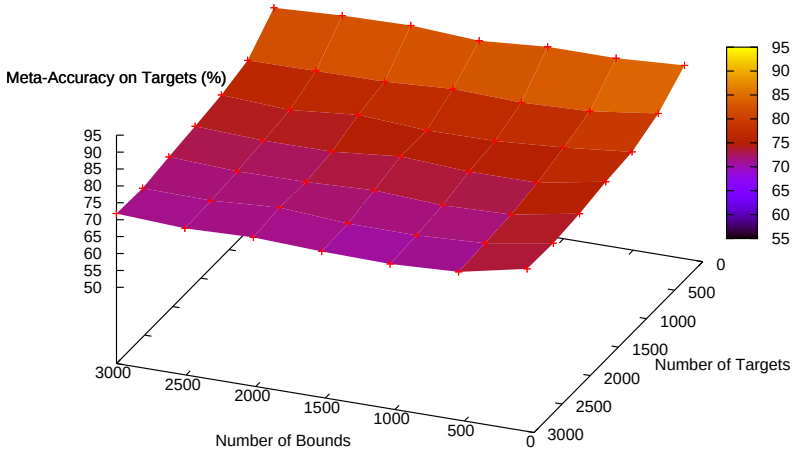
$$\vec{w}_{z,k}(i) = \hat{I}_{z,k}(b_i|w)$$

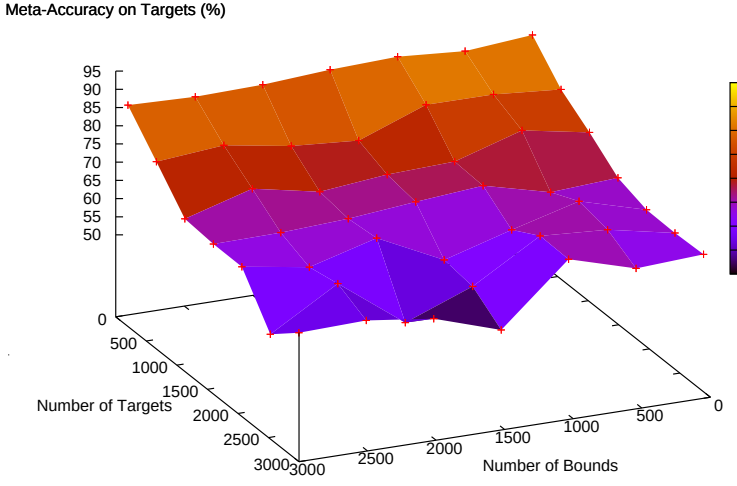
Clustering Data: Comparison

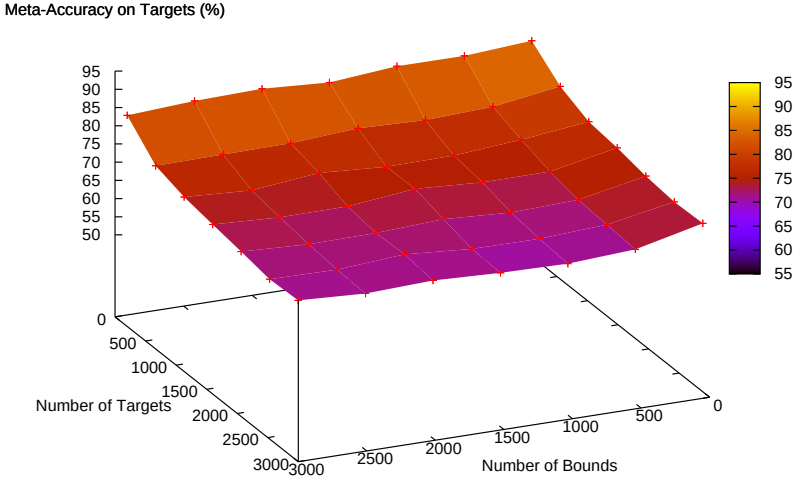
$P_{ML}(b|t)$, $Spearman$



$NRPMI(b|t)$, $L1$







Fuzzy Cluster Membership

Desideratum:

- membership probability distribution:
  $\hat{p}_k(\cdot|\cdot) : T_k \times C_k \rightarrow [0,1]$
- such that $\forall w \in T_k . \sum_{c \in C_k} \hat{p}_k(c|w) = 1$

Available Source Data:

- Clustering distance function: $d : \mathcal{P}(\mathbb{R}^{2|B_k|})^2 \rightarrow \mathbb{R}$

Heuristic:

- let $d_{k,min} = \min_{c \in C_k, w \in T_k : d(c,w) > 0} d_k(c,w)$
- Similarity function: $\hat{s}_k(c,w) = \frac{d_{k,min}}{d_k(c,w) + d_{k,min}}$
- Membership heuristic: $\hat{p}_k(c|w) = \frac{\hat{s}_k(c,w)}{\sum_{c' \in C_k} \hat{s}_k(c',w)}$
- Useful restriction: m-best, $m \approx \frac{|C|}{12}$

Clustering: Bootstrapping

Clusters as Bounds: for $w \in T_k, b \in B_k = C_{k-1}$,

$$f_{\ell,k}(w,b) = \sum_{v \in T_{<k}} \hat{p}_{<k}(b|v) f_0(v,w)$$
$$f_{r,k}(w,b) = \sum_{v \in T_{<k}} \hat{p}_{<k}(b|v) f_0(w,v)$$

Underlying Model Assumptions:

$$f_{\ell,k}(w,b) = p_{\ell,k}(w,b) N_{\ell,k} \quad \text{(MLE)}$$
$$p_{\ell,k}(w,v) = \frac{f_0(v,w)}{N_{\ell,k}} \quad \text{(MLE)}$$
$$p_{\ell,k}(w,b) = \sum_{v \in T_{<k}} p_{\ell,k}(v,w,b) \quad \text{(Marginal)}$$
$$p_{\ell,k}(b|v,w) = \hat{p}_{<k}(b|v) \quad \text{(Independence)}$$

Bonus Question #2

Where's the next bug?

Centroid Reclustering

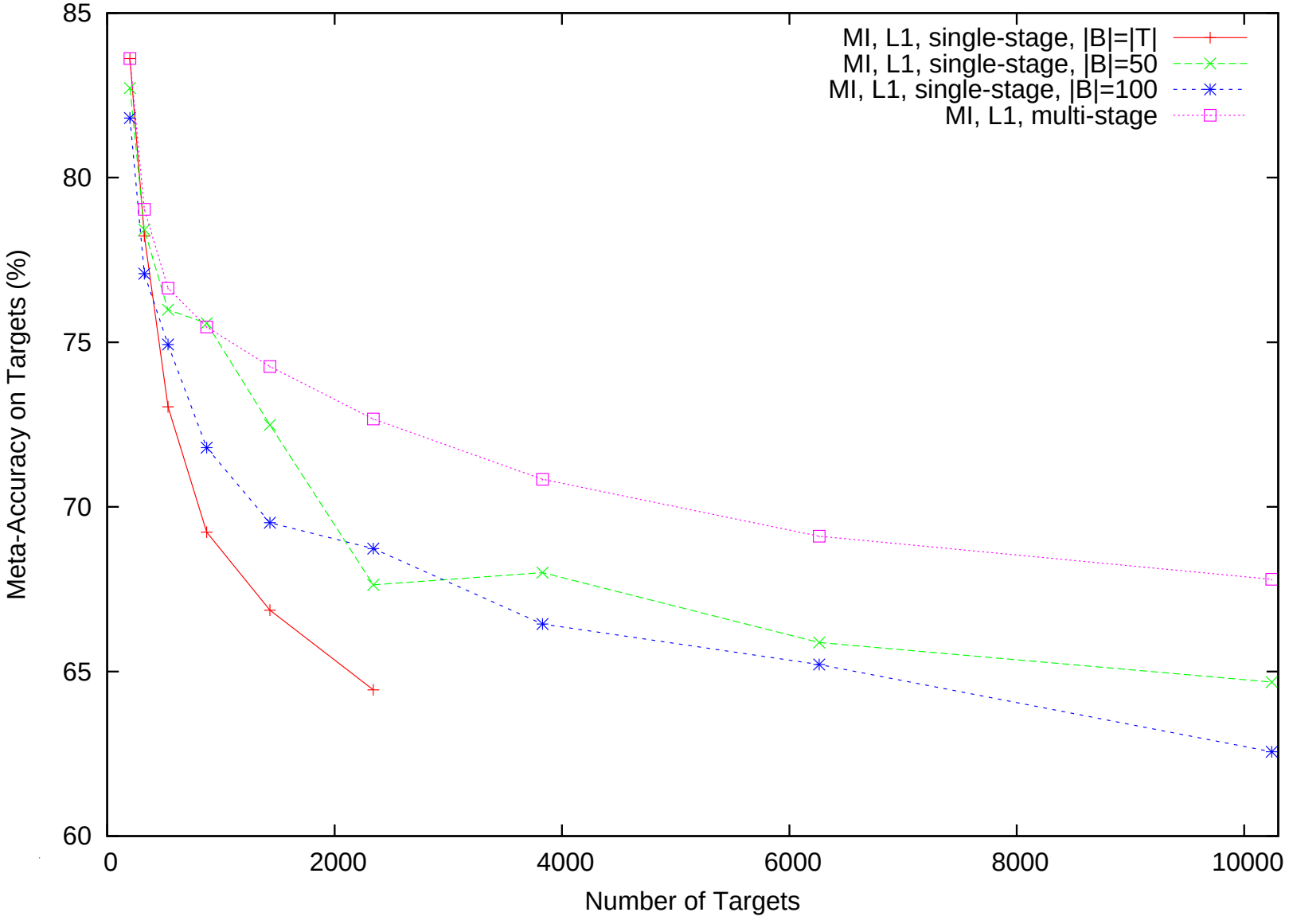
The Problem

- Maintain solution compatibility over mutliple stages
- $i \neq j \implies T_i \cap T_j = \emptyset \implies C_i \cap C_j = \emptyset$

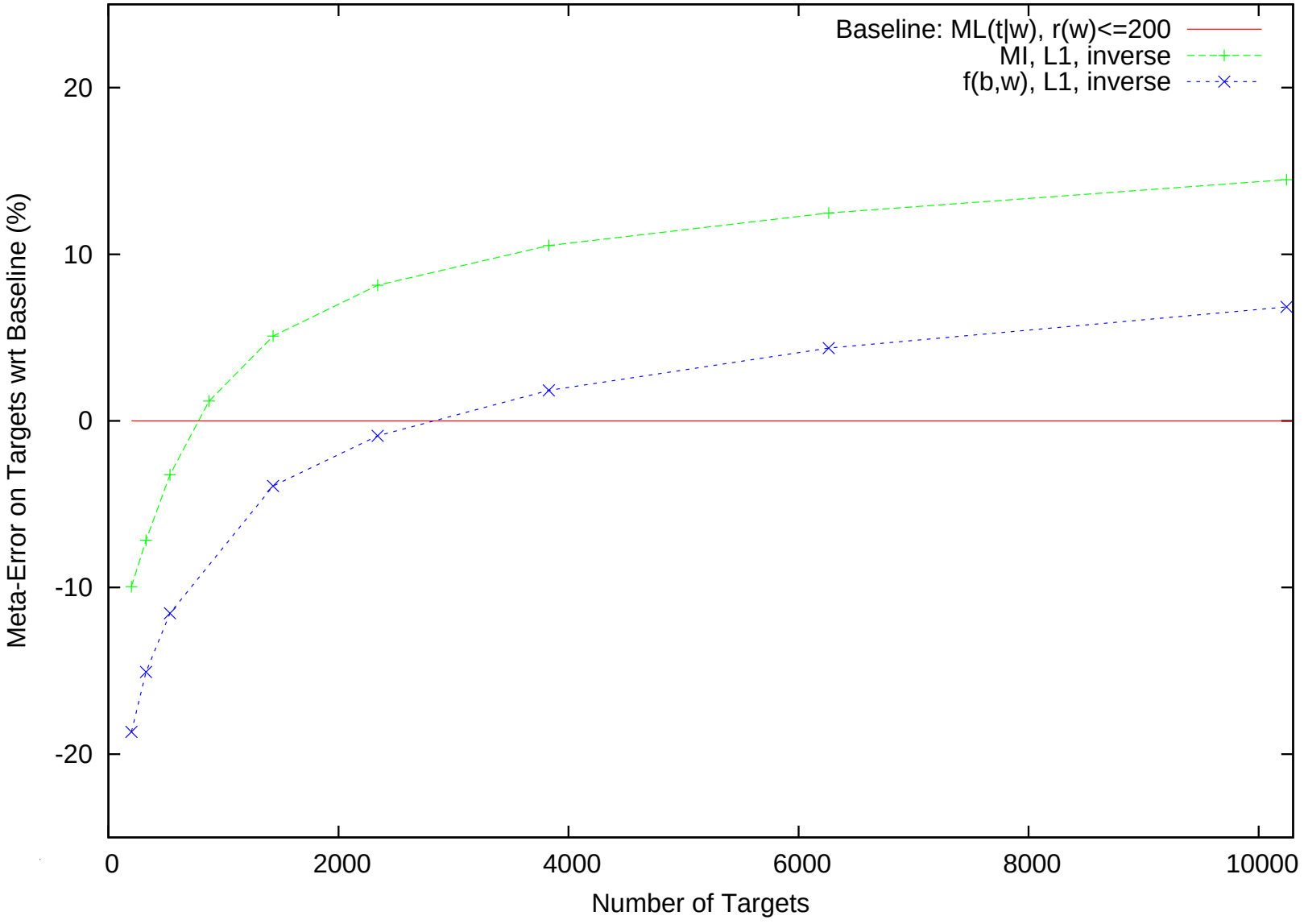
The Approach

- Idea: centroid reclustering – cluster targets
  $\tilde{T}_k = T_k \cup C_{k-1}$
- Categorical cluster assignment function
  $\tilde{c}_{<k}(w) = \arg\max_{c \in C_{k-1}} \hat{p}_{<k}(c|w)$
- Categorical cluster membership function
  $\tilde{c}^{-1}_{<k}(c) = \{w \in T_{<k} \mid \tilde{c}_{<k}(w) = c\}$
- Centroid pseudo-targets, for $c, b \in C_{k-1}$:
  $f_{z,k}(c, b) = \sum_{w \in \tilde{c}^{-1}_{<k}(c)} f_{z,k}(w, b)$

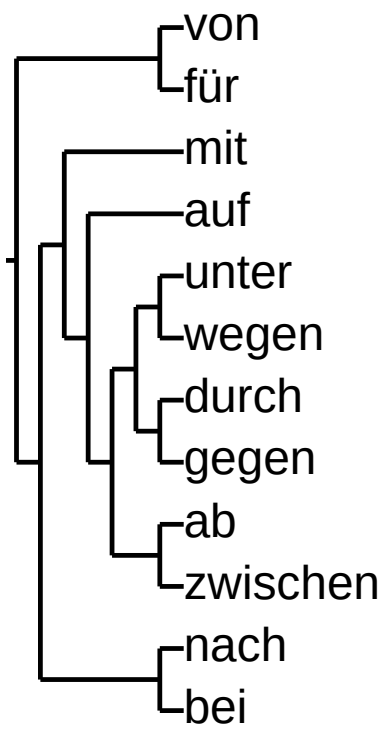
K- vs 1-Stage Clustering

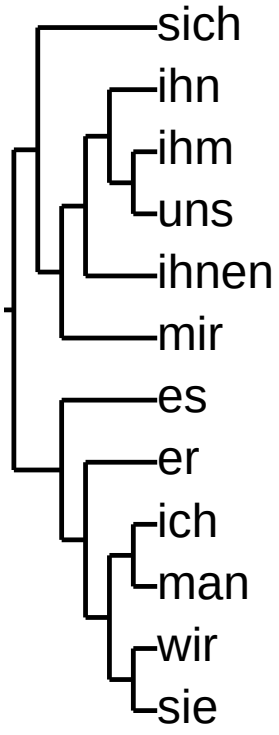


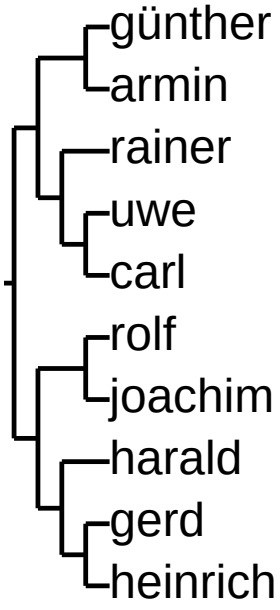
K-stage clustering vs. baseline

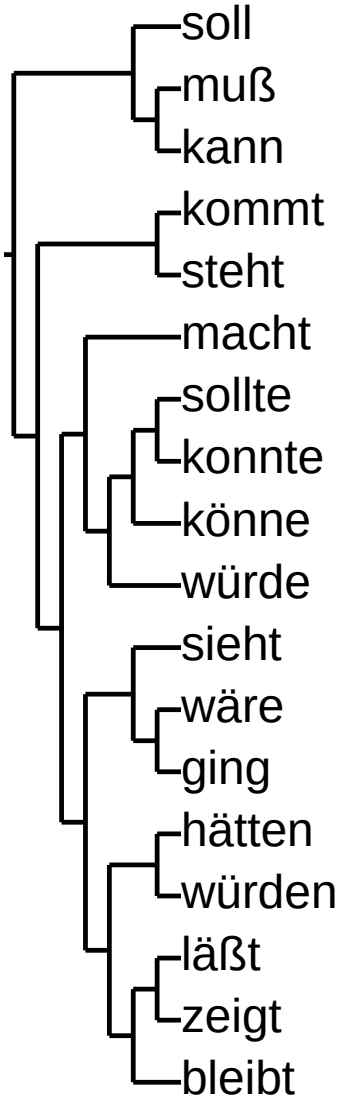


Clustering: Examples

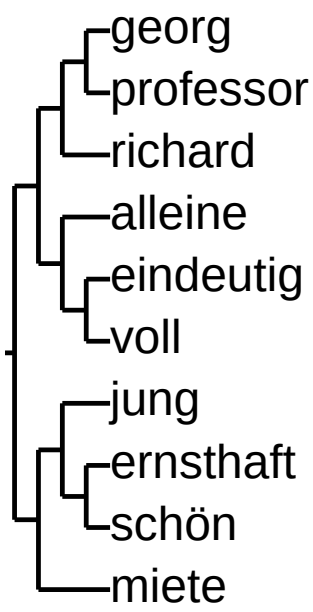


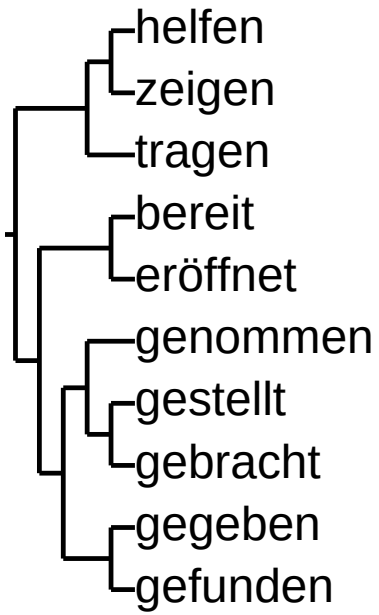


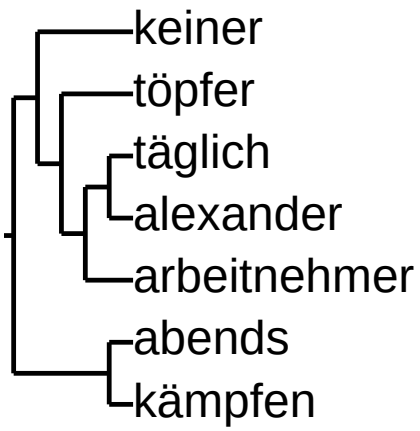




Clustering: Weirder Examples







?

!

Reestimation Phase

Reestimation Phase

Motivation

- Word types may be ambiguous
- Word tokens in context are unambiguous
- Clustering maps types 1-1 to “best” clusters

The Plan

- Use membership probabilities $\hat{p}_{\leq K}$ to initialize a 1^{st}-order HMM
- Reestimate HMM with the Baum-Welch algorithm
- Enable context-dependent ambiguity resolution via the Viterbi algorithm

Parameter Initialization

Hidden Markov Model Parameters

- $A(q_i, q_j) = P(Q_{t+1} = q_j | Q_t = q_i)$
- $B(w, q) = P(W_t = w | Q_t = q)$
- $\pi(q) = P(Q_1 = q)$

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:

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

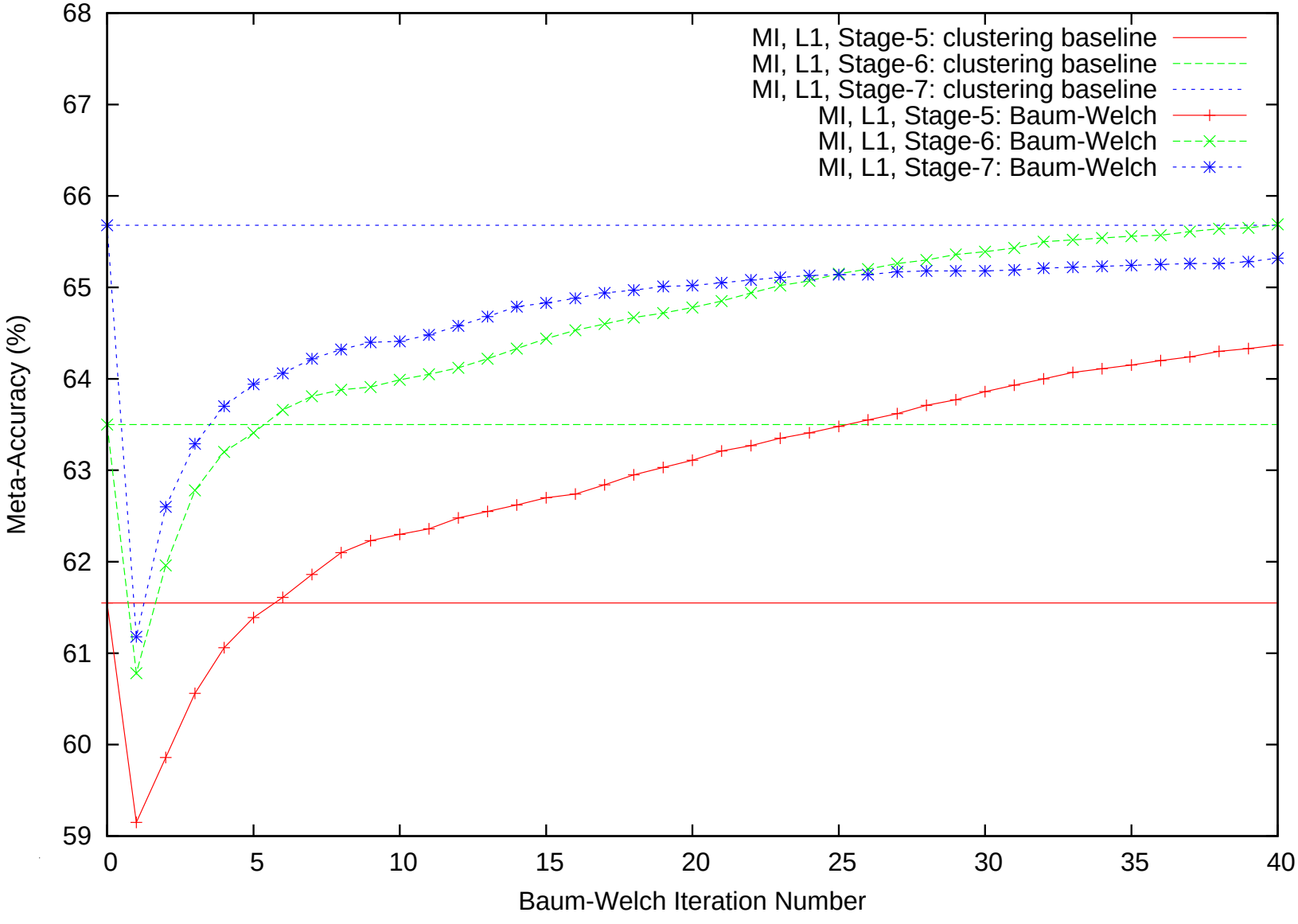
$$\begin{aligned} \hat{p}_{\leq K}(c) &= \sum_{w \in T_{\leq k}} \hat{p}_{\leq K}(w, c) \quad \text{(Marginal)} \\ &= \sum_{w \in T_{\leq k}} \hat{p}_{\leq K}(c|w)\hat{p}_{\leq K}(w) \end{aligned}$$

Ambiguity Rates

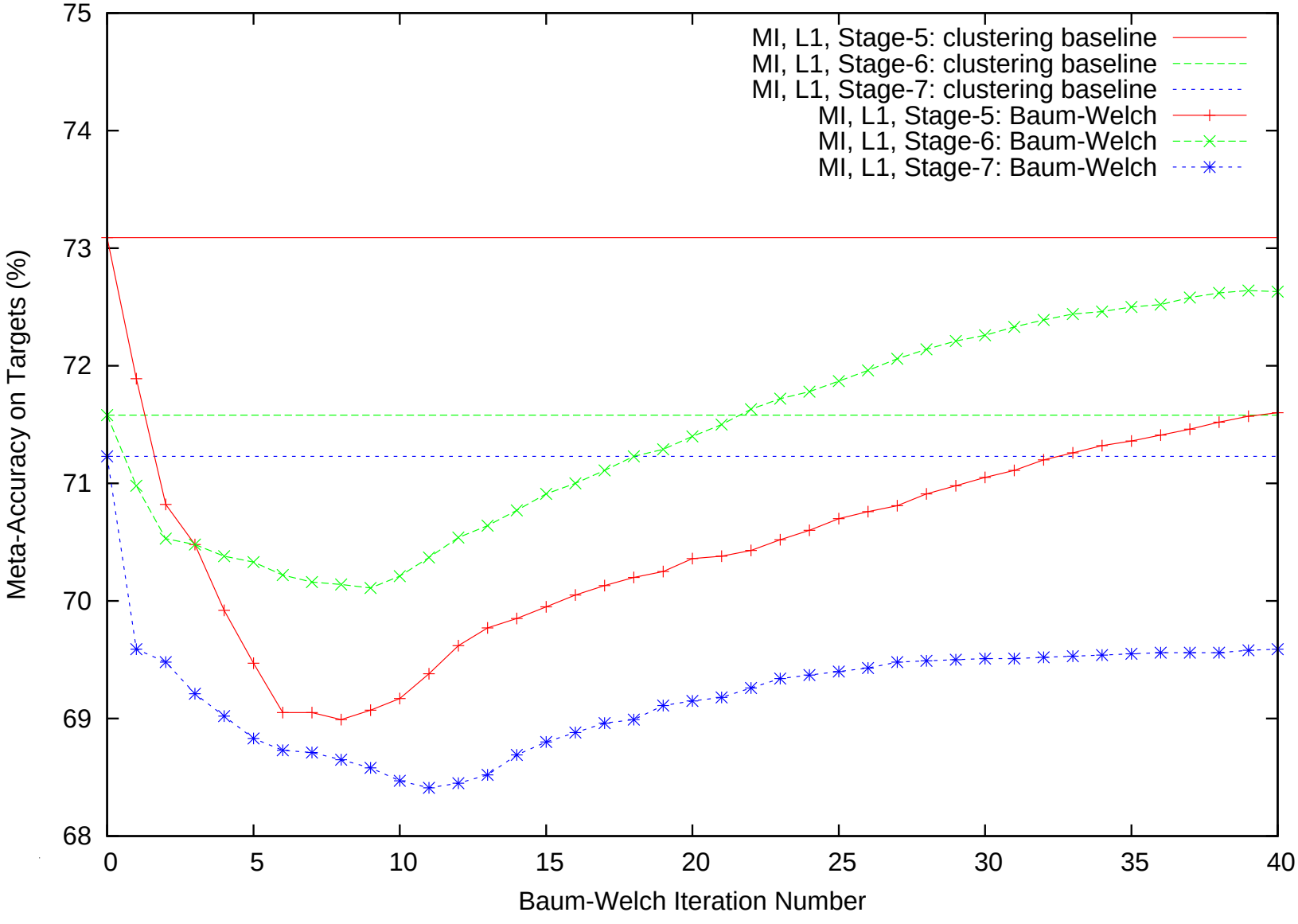
Stage	# Targets	Avg. Ambiguity	Gold-std. Ambiguity
5	1431	1.65	1.40
6	2341	1.57	1.35
7	6262	1.49	1.23

TODO: fill in the blanks...

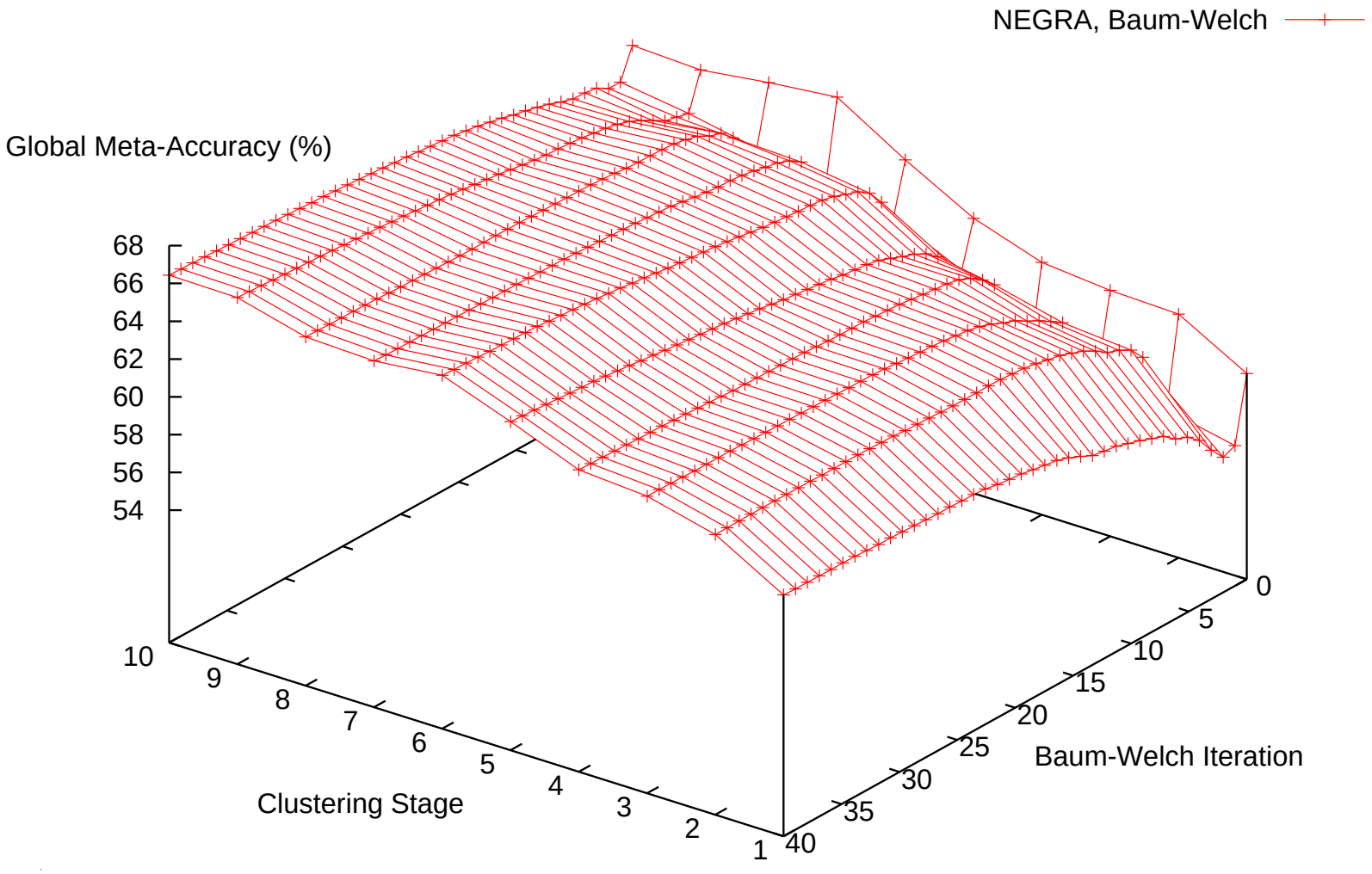
EM Evaluation: Global



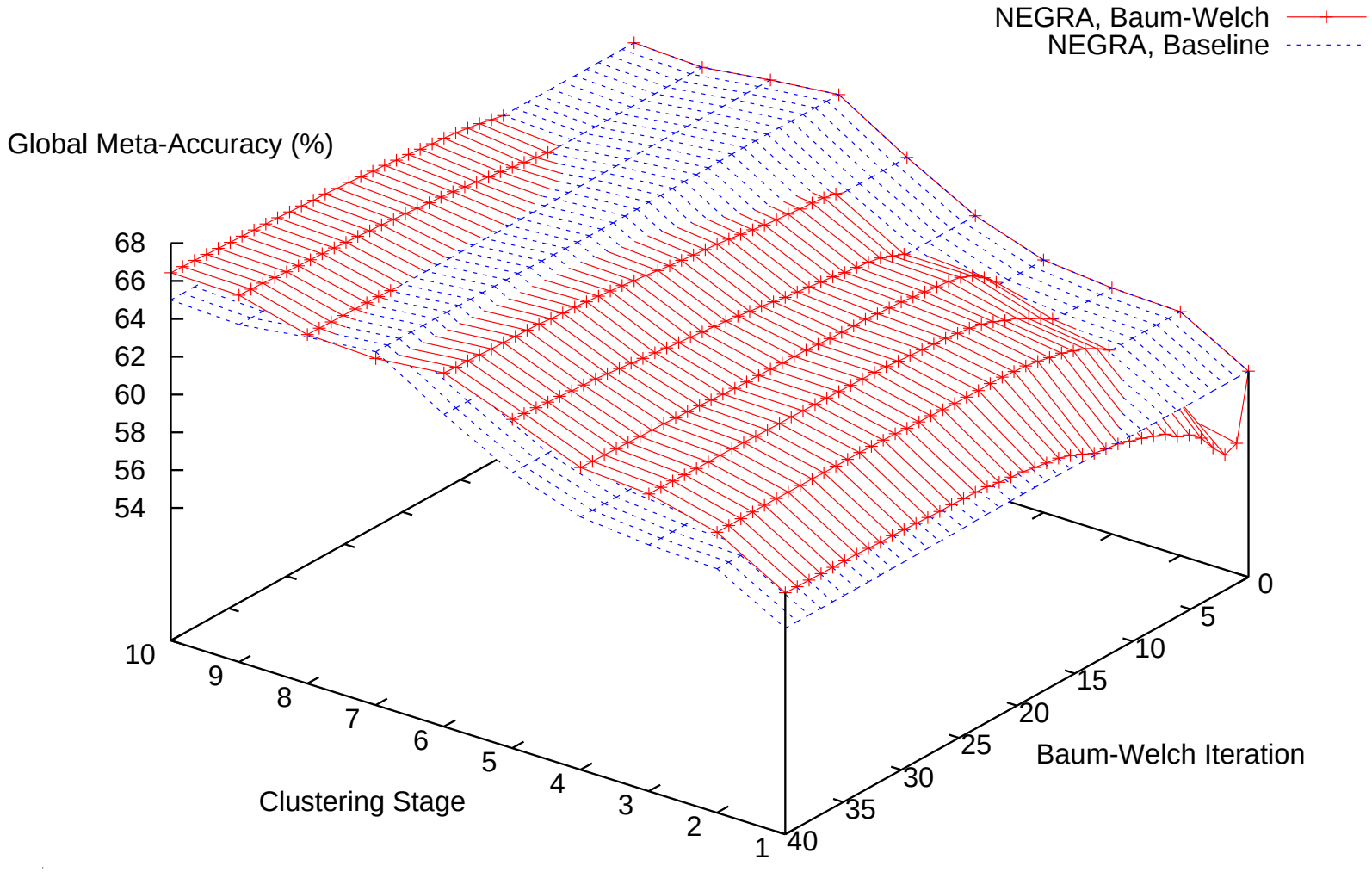
EM Evaluation: Targets Only



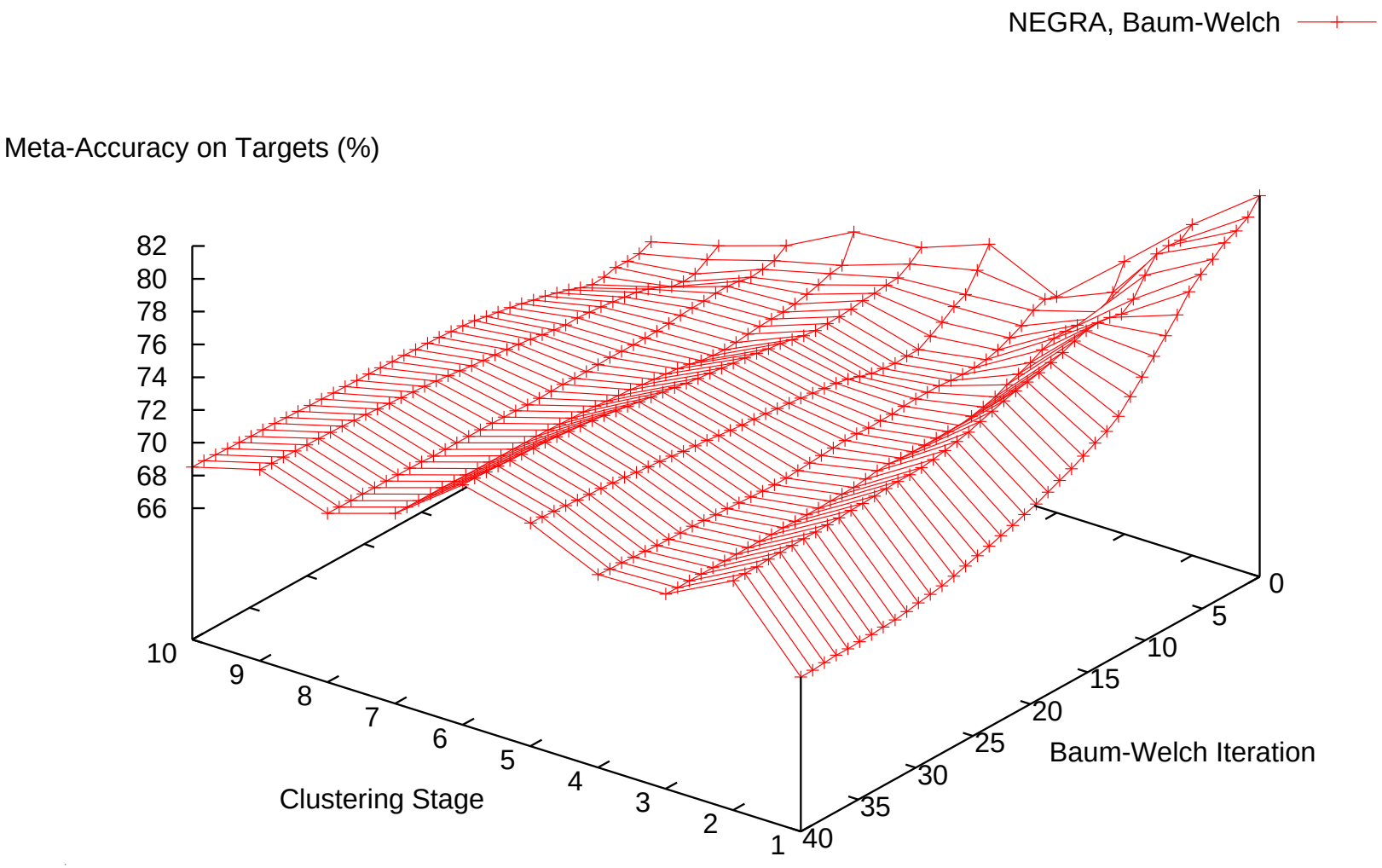
EM: Global: Full



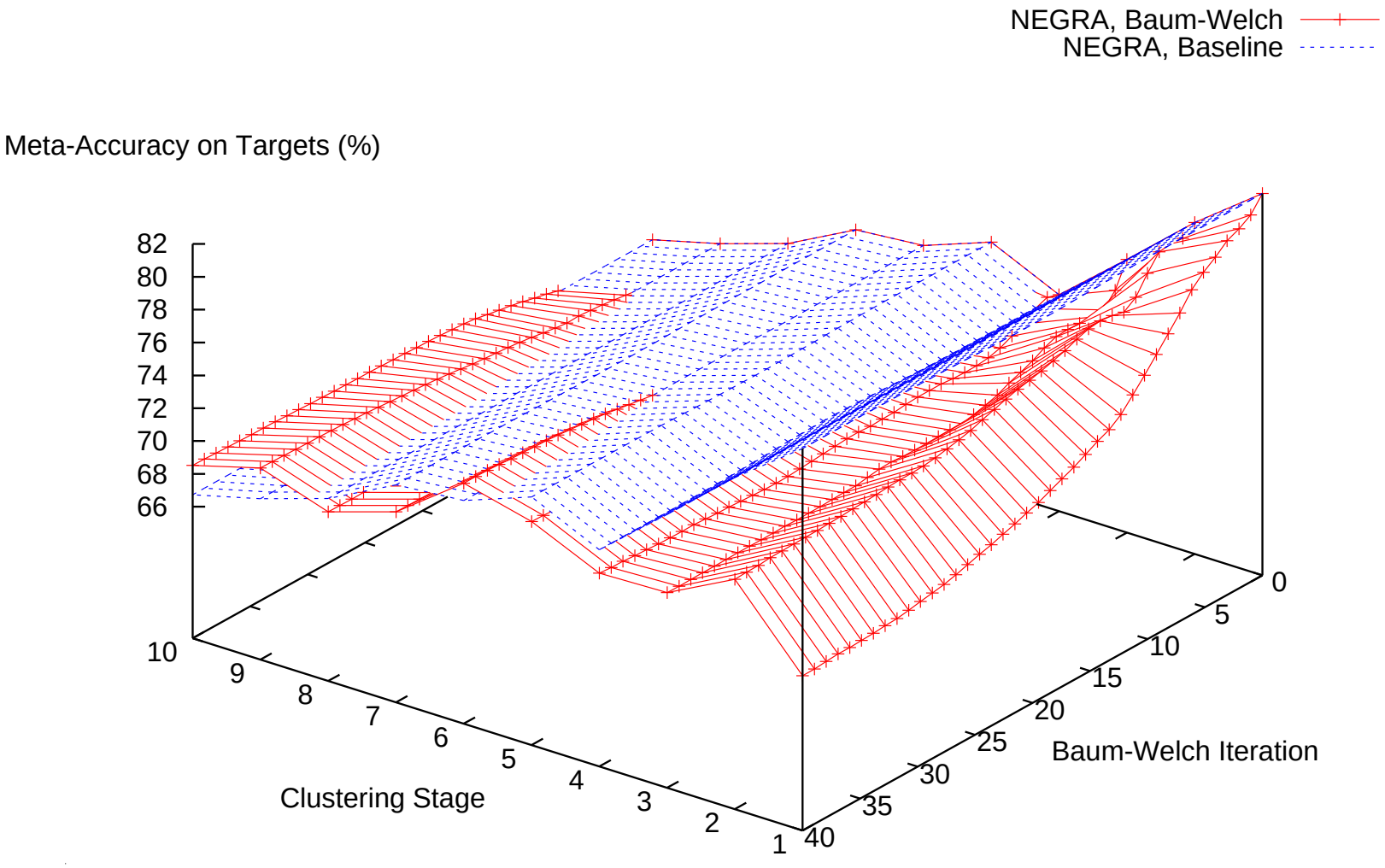
EM: Global: Full



EM: Targets: Full



EM: Targets: Full



The End