Collective Annotation of Wikipedia Entities in Web Text

Sayali Kulkarni Amit Singh Ganesh Ramakrishnan
Soumen Chakrabarti

IIT Bombay

Our aim

Aggressive open domain annotation of unstructured Web text with uniquely identified entities in a social media (Wikipedia)

The incentive

Use the annotations for search and mining tasks

Outline for today

- Terminologies
- About entity disambiguation
- Our contributions
- Evaluation and results
- Conclusion

Terminologies

Web NASDAQ.com Page: Symbol List: | | | France To Host Meeting Of Afghanistan 's Neighbors -Foreign Min PARIS (AFP)--France has invited Pakistan and Iran to take part in a meeting of Afghanistan 's neighbors to help advance peace in the insurgency -hit country, Foreign Minister Bernard Kouchner said Tuesday . "There will be a meeting , I hope in Paris, of neighboring countries," Kouchner told members of the parliamentary foreign affairs committee . Paris has asked Pakistan , Iran and other neighboring countries to attend because they could play a role in helping Afghanistan reach peace with the Taliban militia , he said. Kouchner didn't give a date for the meeting . The foreign minister reiterated that France supports Afghan President Hamid Karzai 's bid to hold talks with moderates within the Taliban movement , which was ousted from Kabul in 2001 during a U.S. -led invasion

Figure: A plain page from unstructured data source

Terminologies (2)

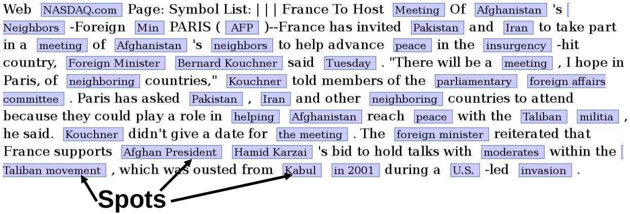


Figure: A spot on a page

Spot is an occurrence of text on a page that can be possibly linked to a Wikipedia article

Related notations:

S_0	All candidate spots in a Web page
$s \in S_0$	One spot, including surrounding context

Terminologies (3)

Figure: Possible attachments for a spot

Attachments are Wikipedia entities that can be possibly linked to a spot

Related notations:

Γ_s	Set of candidate entity labels for spot s on a page
Γ_0	$\bigcup_{s \in S_0} \Gamma_s$, set of all candidate labels for the page

Entity disambiguation

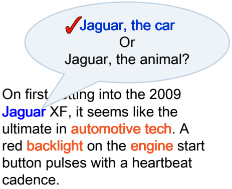


Figure: Clues from local context help in disambiguation

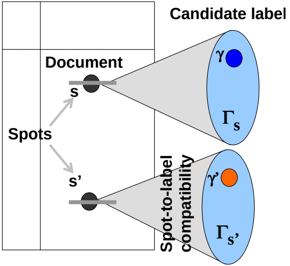


Figure: Disambiguation based on compatibility between spot and label

Related work: SemTag and Seeker[2]

Collective entity disambiguation

Figure: Other spots on page help in disambiguation

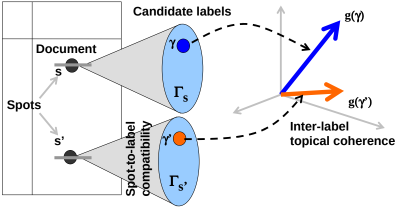


Figure: Disambiguation based on local compatibility and coherence between labels

Related work: Cucerzan[1] and Milne *et al.*[3]

Relatedness information from entity catalog

- How related are two entities γ, γ' in Wikipedia?
- Embed γ in some space using $g : \Gamma \to \mathbb{R}^c$
- Define relatedness $r(\gamma, \gamma') = g(\gamma) \cdot g(\gamma')$ or related
- Cucerzan's proposal: relatedness between entity based on cosine measure
- Milne *et al.* proposal: c = number of Wikipedia pages; $g(\gamma)[p] = 1$ if page p links to page γ, 0 otherwise

$$r(\gamma, \gamma') = \frac{\log |g(\gamma) \cap g(\gamma')| - \log \max\{|g(\gamma)|, |g(\gamma')|\}}{\log c - \log \min\{|g(\gamma)|, |g(\gamma')|\}}$$

Relatedness information from entity catalog

- How related are two entities γ, γ' in Wikipedia?
- Embed γ in some space using $g : \Gamma \to \mathbb{R}^c$
- Define relatedness $r(\gamma, \gamma') = g(\gamma) \cdot g(\gamma')$ or related
- Cucerzan's proposal: relatedness between entity based on cosine measure
- Milne *et al.* proposal: c = number of Wikipedia pages; $g(\gamma)[p] = 1$ if page p links to page γ, 0 otherwise

$$r(\gamma, \gamma') = \frac{\log |g(\gamma) \cap g(\gamma')| - \log \max\{|g(\gamma)|, |g(\gamma')|\}}{\log c - \log \min\{|g(\gamma)|, |g(\gamma')|\}}$$

Relatedness information from entity catalog

- How related are two entities γ, γ' in Wikipedia?
- Embed γ in some space using $g : \Gamma \to \mathbb{R}^c$
- Define relatedness $r(\gamma, \gamma') = g(\gamma) \cdot g(\gamma')$ or related
- Cucerzan's proposal: relatedness between entity based on cosine measure
- Milne *et al.* proposal: c = number of Wikipedia pages; $g(\gamma)[p] = 1$ if page p links to page γ, 0 otherwise

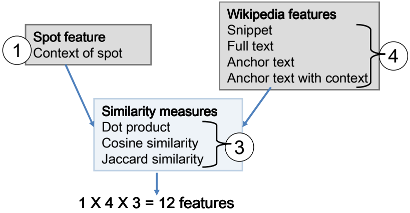
$$r(\gamma, \gamma') = \frac{\log |g(\gamma) \cap g(\gamma')| - \log \max\{|g(\gamma)|, |g(\gamma')|\}}{\log c - \log \min\{|g(\gamma)|, |g(\gamma')|\}}$$

Our contributions

- Posing entity disambiguation as an optimization problem
- Single optimization objective
  - Using integer linear programs (NP Hard)
  - Heuristics for approximate solutions
- Rich node features with systematic learning
- Back off strategy for controlled annotations

Modeling local compatibility

- Feature vector $f_s(\gamma) \in \mathbb{R}^d$ expresses local textual compatibility between (context of) spot s and candidate label γ
- Components of $f_s(\gamma)$



- Sense probability prior: probability that a Wikipedia entity can be associated with a spot ($Pr(\gamma|s)$)

Components of our objective

Node score

- Node scoring model $w \in \mathbb{R}^d$
- Node score defined as $w^\top f_s(\gamma)$
- w is learned using a linear adaptation of rankSVM

Clique Score

- Use relatedness measure (r) as described by Milne *et. al.*

Total objective

$$\underbrace{\frac{1}{|S_0|}\sum_s w^\top f_s(y_s)}_{\text{Node score}} + \underbrace{\frac{1}{\binom{|S_0|}{2}}\sum_{s \neq s'} r(y_s, y_{s'})}_{\text{Clique score}}$$

y is the final set of assignments on a page

Components of our objective

Node score

- Node scoring model $w \in \mathbb{R}^d$
- Node score defined as $w^\top f_s(\gamma)$
- w is learned using a linear adaptation of rankSVM

Clique Score

- Use relatedness measure (r) as described by Milne *et. al.*

Total objective

$$\underbrace{\frac{1}{|S_0|} \sum_{s} w^\top f_s(y_s)}_{\text{Node score}} + \underbrace{\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s'} r(y_s, y_{s'})}_{\text{Clique score}}$$

y is the final set of assignments on a page

Backoff strategy

- Not all spots may be tagged. Allow backoff from tagging
- Assign a special label "NA" to mark a "no attachment"
- Reward a spot for attaching to NA – RNA
- Spots marked NA do not contribute to clique potential
- Smaller the value of RNA, more aggressive is the tagging

Modified Objective

$N_0 \subseteq S_0$: spots assigned NA

$A_0 = S_0 \setminus N_0$: remaining spots

$$\max_y \frac{1}{|S_0|}\left(\sum_{s \in N_0} \rho_{\mathrm{NA}} + \sum_{s \in A_0} w^\top f_s(y_s)\right) \qquad \text{(Node Score)}$$

$$+ \frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in A_0} r(y_s, y_{s'}) \qquad \text{(Clique Score)}$$

Backoff strategy

- Not all spots may be tagged. Allow backoff from tagging
- Assign a special label "NA" to mark a "no attachment"
- Reward a spot for attaching to NA – RNA
- Spots marked NA do not contribute to clique potential
- Smaller the value of RNA, more aggressive is the tagging

Modified Objective

$N_0 \subseteq S_0$: spots assigned NA
$A_0 = S_0 \setminus N_0$: remaining spots

$$\max_{y} \frac{1}{|S_0|} \left(\sum_{s \in N_0} \rho_{\mathrm{NA}} + \sum_{s \in A_0} w^\top f_s(y_s) \right) \qquad \text{(Node Score)}$$

$$+ \frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in A_0} r(y_s, y_{s'}) \qquad \text{(Clique Score)}$$

Methodologies for solving the objective

Integer linear program (ILP) based formulation

- Casting as $0/1$ integer linear program
- Using up to $|\Gamma_0| + |\Gamma_0|^2$ variables
- Relaxing it to an LP

Simpler heuristics

- Hill climbing for optimization

Evaluation of the annotation system

Evaluation measures:

Precision

Number of spots tagged correctly out of total number of spots tagged

Recall

Number of spots tagged correctly out of total number of spots in ground truth

F1

$$\frac{2 \times Recall \times Precision}{(Recall + Precision)}$$

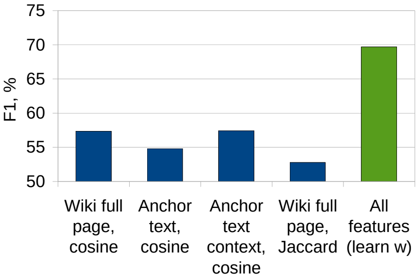
Datasets for evaluation

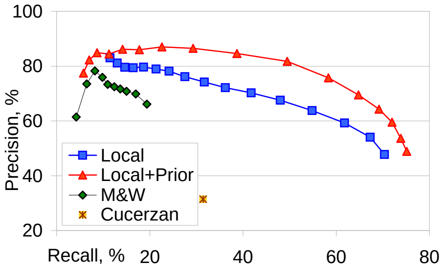
- Documents(IITB) crawled from popular sites
- Publicly available data from Cucerzan's experiments (CZ)

	IITB	**CZ**
Number of documents	107	19
Total number of spots	17,200	288
Spot per 100 tokens	30	4.48
Average ambiguity per spot	5.3	18

Figure: Corpus statistics.

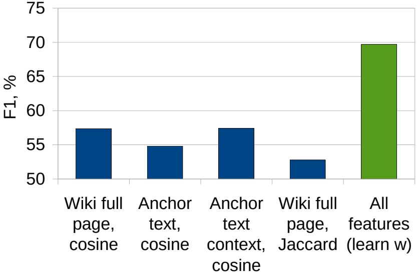
Effect of learning in node score calculation

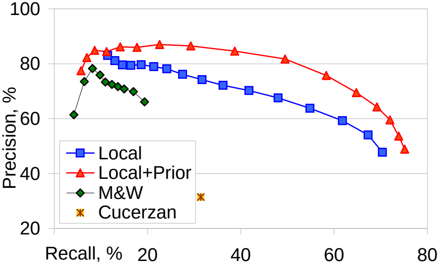




- Using w is better than using individual node features in isolation
- Enough to outperform other baseline systems

Effect of learning in node score calculation





- Using w is better than using individual node features in isolation
- Enough to outperform other baseline systems

Benefits of collective annotation

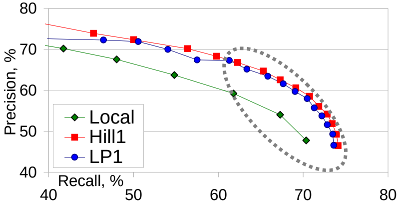
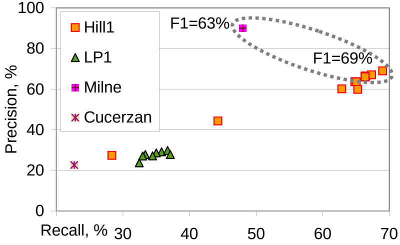


Figure: Recall/precision on IITB data



- Adding collective inference adds to the accuracy of the annotations

Results summary

- Selection of features for defining the node score is important
- Collective inference improves accuracy further
- Able to gain high recall without sacrificing much on precision

Evaluation:

	Our system	Cucerzan	Milne *et al.*
Recall	**70.7%**	31.43%	66.1%
Precision	**68.7%**	53.41%	19.35%
F1	**69.69%**	39.57%	29.94%

Future work

- Extending collective inference beyond page-level boundaries
- Associating confidence with annotations
- Reducing cognitive load during the process manual annotations
- Building an entity search system over annotations

KDD Demo: 30 June '09, 17:30 onwards

Future work

- Extending collective inference beyond page-level boundaries
- Associating confidence with annotations
- Reducing cognitive load during the process manual annotations
- Building an entity search system over annotations

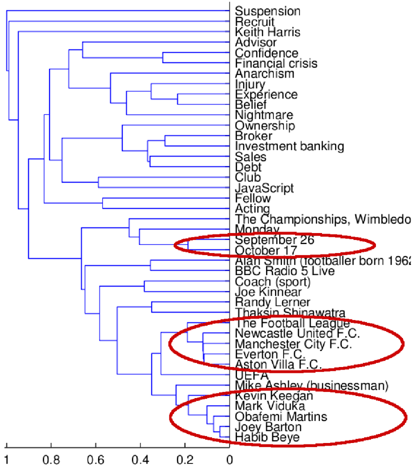
KDD Demo: 30 June '09, 17:30 onwards

Questions?

Additional slides

- Multitopic models
- Belief about the objective
- Tuning RNA
- More about data sets
- Human Supervision
- ILP in detail
- Hill climbing algorithm
- Timing graphs
- References

Dendrogram with multitopic model



Multi-topic model

- Current clique potentials encourages a single cluster model
- The single cluster hypothesis is not always true
- Partition the set of possible attachments as $C = \Gamma^1, \ldots, \Gamma^K$
- Refined clique potential for supporting multitopic model

$$\frac{1}{|C|} \sum_{\Gamma^k \in C} \frac{1}{\binom{\Gamma^k}{2}} \sum_{s,s' :\, y_s, y_{s'} \in \Gamma^k} r(y_s, y_{s'}). \qquad \text{(CPK)}$$

- Using $\binom{\Gamma^k}{2}$ instead of $\binom{S_0}{2}$ to reward smaller coherent clusters
- Node score is not disturbed

Is our belief about the objective correct?

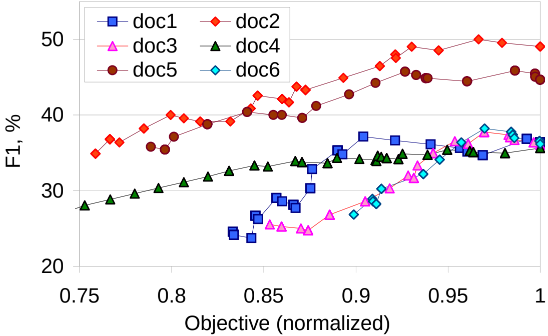


Figure: F1 versus Objective

- As the objective value increases, the F1 increases
- Validates our belief about the objective

Effect of tuning RNA

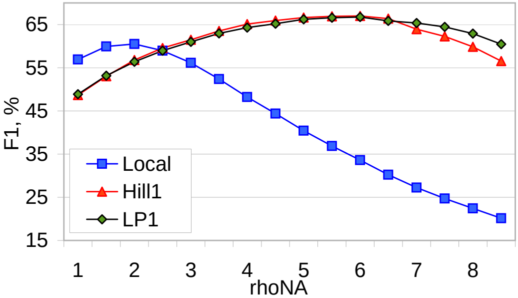


Figure: F1 for Local, Hill and LP for different RNA values

- Best RNA for LOCAL is lesser than the best RNA for HILL1 and LP1

Effect of tuning RNA (2)

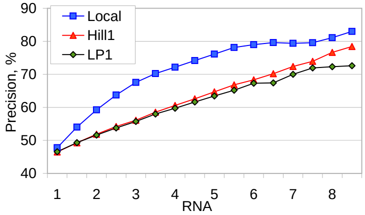


Figure: Precision for different RNA values

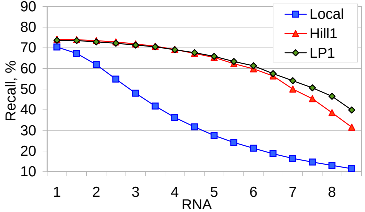


Figure: Recall for different RNA values

- Smaller the value of RNA, more aggressive is the tagging
- Precision increases with increase in RNA value
- Recall decreases with increase in RNA value

More about data sets

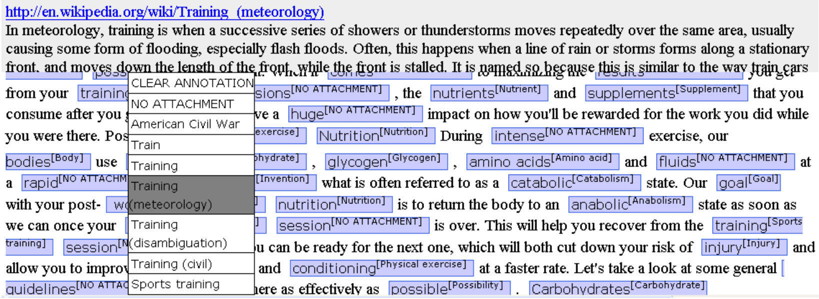
More on IITB dataset

- Collected a total of about **19,000** annotations
- Done by by 6 volunteers
- About 50 man-hours spent in collecting the annotations
- Exhaustive tagging by volunteers
- Spots labeled as NA was about 40%

#Spots tagged by more than one person	1390
#NA among these spots	524
#Spots with disagreement	278
#Spots with disagreement involving NA	218

Figure: Inter-annotator agreement.

Human Supervision



- System identifies spots and mentions
- Shows pull-down list of (subset of) Γ_s for each s
- User selects $\gamma^* \in \Gamma_s \cup \mathrm{NA}$

Integer linear program (ILP) based formulation

Variables:

$$z_{s\gamma} = [\![\text{spot } s \text{ is assigned label } \gamma \in \Gamma_s]\!]$$
$$u_{\gamma\gamma'} = [\![\text{both } \gamma, \gamma' \text{ assigned to spots}]\!]$$

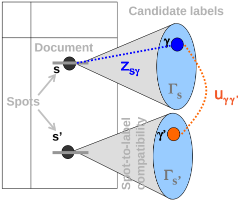


Figure: Defining the variables for ILP

Integer linear program (ILP) based formulation

Objective:

$$\max_{\{z_{s\gamma}, u_{\gamma\gamma'}\}} \; (\text{NP}') + (\text{CP1}')$$

Node potential:

$$\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s\gamma} w^\top f_s(\gamma) \tag{NP$'$}$$

Clique potential:

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} \sum_{\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}} u_{\gamma\gamma'} r(\gamma, \gamma') \tag{CP1$'$}$$

Subject to constraints:

$$\forall s, \gamma : z_{s\gamma} \in \{0,1\}, \quad \forall \gamma, \gamma' : \; u_{\gamma\gamma'} \in \{0,1\} \tag{1}$$

$$\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \tag{2}$$

$$\forall s : \; \textstyle\sum_\gamma z_{s\gamma} = 1. \tag{3}$$

Integer linear program (ILP) based formulation

Objective:

$$\max_{\{z_{s\gamma}, u_{\gamma\gamma'}\}} \text{(NP}'\text{)} + \text{(CP1}'\text{)}$$

Node potential:

$$\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s\gamma} w^\top f_s(\gamma) \qquad (\text{NP}')$$

Clique potential:

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} \sum_{\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}} u_{\gamma\gamma'} r(\gamma, \gamma') \qquad (\text{CP1}')$$

Subject to constraints:

$$\forall s, \gamma : z_{s\gamma} \in \{0, 1\}, \quad \forall \gamma, \gamma' : \; u_{\gamma\gamma'} \in \{0, 1\} \qquad (1)$$

$$\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \qquad (2)$$

$$\forall s : \; \textstyle\sum_\gamma z_{s\gamma} = 1. \qquad (3)$$

Integer linear program (ILP) based formulation

Objective:

$$\max_{\{z_{s\gamma}, u_{\gamma\gamma'}\}} \text{(NP}'\text{)} + \text{(CP1}'\text{)}$$

Node potential:

$$\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s\gamma} w^\top f_s(\gamma) \qquad \text{(NP}'\text{)}$$

Clique potential:

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} \sum_{\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}} u_{\gamma\gamma'} r(\gamma, \gamma') \qquad \text{(CP1}'\text{)}$$

Subject to constraints:

$$\forall s, \gamma : z_{s\gamma} \in \{0, 1\}, \quad \forall \gamma, \gamma' : \; u_{\gamma\gamma'} \in \{0, 1\} \qquad (1)$$

$$\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \qquad (2)$$

$$\forall s : \; \textstyle\sum_\gamma z_{s\gamma} = 1. \qquad (3)$$

Integer linear program (ILP) based formulation

Objective:

$$\max_{\{z_{s\gamma}, u_{\gamma\gamma'}\}} \ (\mathrm{NP}') + (\mathrm{CP1}')$$

Node potential:

$$\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s\gamma} w^\top f_s(\gamma) \tag{NP$'$}$$

Clique potential:

$$\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} \sum_{\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}} u_{\gamma\gamma'} r(\gamma, \gamma') \tag{CP1$'$}$$

Subject to constraints:

$$\forall s, \gamma : z_{s\gamma} \in \{0,1\}, \quad \forall \gamma, \gamma' : \ u_{\gamma\gamma'} \in \{0,1\} \tag{1}$$

$$\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \tag{2}$$

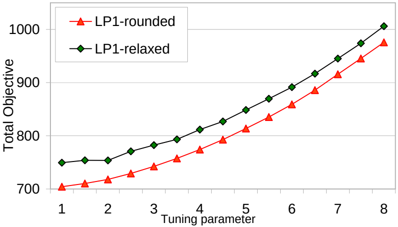
$$\forall s : \ \textstyle\sum_\gamma z_{s\gamma} = 1. \tag{3}$$

LP relaxation for the ILP formulation

- Relax the constraints in the formulation as :

$$\forall s, \gamma : 0 \leq z_{s\gamma} \leq 1, \quad \forall \gamma, \gamma' : \; 0 \leq u_{\gamma\gamma'} \leq 1$$
$$\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'}$$
$$\forall s : \; \sum_{\gamma} z_{s\gamma} = 1.$$

- Margin between objective of relaxed LP and the rounded LP is quite thin



Hill climbing algorithm

```
initialize some assignment y^(0)
for k = 1, 2, ... do
    select a small spot set S_Δ
    for each s ∈ S_Δ do
        find new γ that improves objective
        change y_s^(k−1) to y_s^(k) = γ greedily
    if objective could not be improved then
        return latest solution y^(k)
```

Figure: Outline for hill-climbing algorithm

Scaling and performance measurement

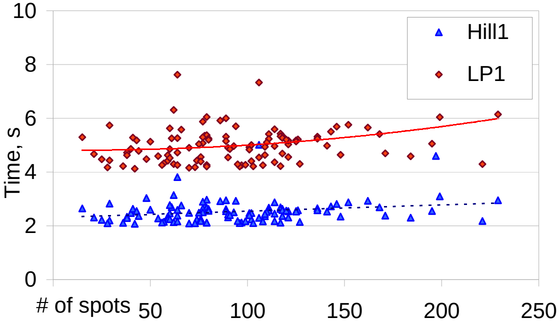


Figure: Scaling the annotation process with number of spots being annotated

- Scaling is mildly quadratically wrt $|S_0|$
- $\mathrm{H\scriptstyle ILL}1$ takes about 2–3 seconds
- LP1 takes around 4–6 seconds

References

[1] S. Cucerzan.
Large-scale named entity disambiguation based on Wikipedia data.
In *EMNLP Conference*, pages 708–716, 2007.

[2] S. Dill et al.
SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation.
In *WWW Conference*, 2003.

[3] D. Milne and I. H. Witten.
Learning to link with Wikipedia.
In *CIKM*, 2008.