Collective Annotation of Wikipedia Entities in Web Text

Sayali Kulkarni  Amit Singh  Ganesh Ramakrishnan
Soumen Chakrabarti

IIT Bombay
Introduction

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
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 II

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 III

Figure: Possible attachments for a spot

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

Related notations:

\[ \Gamma_s \quad \text{Set of candidate entity labels for spot } s \text{ on a page} \]

\[ \Gamma_0 \quad \bigcup_{s \in S_0} \Gamma_s, \text{ set of all candidate labels for the page} \]
Entity disambiguation

Figure: Clues from local context help in disambiguation

On first looking into the 2009 Jaguar XF, it seems like the ultimate in automotive tech. A red backlight on the engine start button pulses with a heartbeat cadence.

Figure: Disambiguation based on compatibility between spot and label

Related work: SemTag and Seeker[D+03]
Collective entity disambiguation

Figure: Other spots on page help in disambiguation

Related work: Cucerzan [Cuc07] and Milne et al. [MW08]
Relatedness information from entity catalog

- How related are two entities \( \gamma, \gamma' \) in Wikipedia?
- Embed \( \gamma \) in some space using \( g : \Gamma \rightarrow \mathbb{R}^c \)
- 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 \( \gamma \), 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 $\gamma, \gamma'$ in Wikipedia?
- Embed $\gamma$ in some space using $g : \Gamma \rightarrow \mathbb{R}^c$
- 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 $\gamma$, 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 $\gamma, \gamma'$ in Wikipedia?
- Embed $\gamma$ in some space using $g : \Gamma \rightarrow \mathbb{R}^c$
- 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 $\gamma$, 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 $\gamma$.

- Components of $f_s(\gamma)$:
  - Spot feature: Context of spot
  - Wikipedia features:
    - Snippet
    - Full text
    - Anchor text
    - Anchor text with context
  - Similarity measures:
    - Dot product
    - Cosine similarity
    - Jaccard similarity

- 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 $w^T f_s(\gamma)$
- $w$ is learned using a linear adaptation of rankSVM

Clique Score

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

Total objective

$$\frac{1}{|S_0|} \sum_s w^T f_s(y_s) + \frac{1}{\left(\frac{|S_0|}{2}\right)} \sum_{s \neq s'} r(y_s, y_{s'})$$

$y$ 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

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

Node score

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 “\texttt{NA}” to mark a “no attachment”
- Reward a spot for attaching to \texttt{NA} – RNA
- Spots marked \texttt{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 \texttt{NA}
\( A_0 = S_0 \setminus N_0 \) : remaining spots

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

\[
+ \frac{1}{(|S_0| \choose 2)} \sum_{s \neq s' \in A_0} r(y_s, y_{s'}) \quad \text{(Clique Score)}
\]
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

\[ N_0 \subseteq S_0 : \text{spots assigned NA} \]
\[ A_0 = S_0 \setminus N_0 : \text{remaining spots} \]

\[
\max_y \frac{1}{|S_0|} \left( \sum_{s \in N_0} \rho_{NA} + \sum_{s \in A_0} w^T f_s(y_s) \right) \quad \text{(Node Score)}
\]
\[
+ \frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in A_0} r(y_s, y_{s'}) \quad \text{(Clique Score)}
\]
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**
\[
\frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})}
\]
Datasets for evaluation

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

<table>
<thead>
<tr>
<th></th>
<th>IITB</th>
<th>CZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of documents</td>
<td>107</td>
<td>19</td>
</tr>
<tr>
<td>Total number of spots</td>
<td>17,200</td>
<td>288</td>
</tr>
<tr>
<td>Spot per 100 tokens</td>
<td>30</td>
<td>4.48</td>
</tr>
<tr>
<td>Average ambiguity per spot</td>
<td>5.3</td>
<td>18</td>
</tr>
</tbody>
</table>

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

Figure: Recall/precision on CZ data
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:

<table>
<thead>
<tr>
<th></th>
<th>Our system</th>
<th>Cucerzan</th>
<th>Milne et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>70.7%</td>
<td>31.43%</td>
<td>66.1%</td>
</tr>
<tr>
<td>Precision</td>
<td>68.7%</td>
<td>53.41%</td>
<td>19.35%</td>
</tr>
<tr>
<td>F1</td>
<td>69.69%</td>
<td>39.57%</td>
<td>29.94%</td>
</tr>
</tbody>
</table>
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

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

- 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 1

Figure: F1 for Local, Hill1 and LP1 for different RNA values

- Best RNA for Local is lesser than the best RNA for Hill1 and LP1
Effect of tuning RNA II

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

Figure: Precision for different RNA values

Figure: Recall for different RNA values
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%

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#Spots tagged by more than one person</td>
<td>1390</td>
</tr>
<tr>
<td>#<code>NA</code> among these spots</td>
<td>524</td>
</tr>
<tr>
<td>#Spots with disagreement</td>
<td>278</td>
</tr>
<tr>
<td>#Spots with disagreement involving <code>NA</code></td>
<td>218</td>
</tr>
</tbody>
</table>

Figure: Inter-annotator agreement.
System identifies spots and mentions

Shows pull-down list of (subset of) $\Gamma_s$ for each $s$

User selects $\gamma^* \in \Gamma_s \cup \text{NA}$
Integer linear program (ILP) based formulation

Variables:

\[ z_{sg} = \begin{cases} 1 & \text{spot } s \text{ is assigned label } \gamma \in \Gamma_s \\ 0 & \text{otherwise} \end{cases} \]

\[ u_{\gamma\gamma'} = \begin{cases} 1 & \text{both } \gamma, \gamma' \text{ assigned to spots} \\ 0 & \text{otherwise} \end{cases} \]

Figure: Defining the variables for ILP
Integer linear program (ILP) based formulation

Objective:

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

Node potential:

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

Clique potential:

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

Subject to constraints:

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

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

$$\forall s : \sum_{\gamma} z_{s\gamma} = 1. \quad (3)$$
Objective:

$$\max \{ z_{s \gamma}, u_{\gamma \gamma'} \} \quad (NP') + (CP1')$$

Node potential:

$$\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s \gamma} w^T f_s(\gamma) \quad (NP')$$

Clique potential:

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

Subject to constraints:

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

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

$$\forall s : \sum_{\gamma} z_{s \gamma} = 1. \quad (3)$$
Integer linear program (ILP) based formulation

Objective:

\[ \max \{ z_{s\gamma}, u_{\gamma\gamma'} \} \text{ (NP')} + \text{ (CP1')} \]

Node potential:

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

Clique potential:

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

Subject to constraints:

\[ \forall s, \gamma : z_{s\gamma} \in \{0, 1\}, \quad \forall \gamma, \gamma' : u_{\gamma\gamma'} \in \{0, 1\} \quad (1) \]
\[ \forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \quad (2) \]
\[ \forall s : \sum_\gamma z_{s\gamma} = 1. \quad (3) \]
Integer linear program (ILP) based formulation

Objective:
\[
\max \{z_{s\gamma}, u_{\gamma\gamma'}\} \quad (\text{NP}') + (\text{CP1}')
\]

Node potential:
\[
\frac{1}{|S_0|} \sum_{s \in S_0} \sum_{\gamma \in \Gamma_s} z_{s\gamma} w^\top f_s(\gamma) \quad (\text{NP}')
\]

Clique potential:
\[
\frac{1}{\binom{|S_0|}{2}} \sum_{s \neq s' \in S_0} \sum_{\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}} u_{\gamma\gamma'} r(\gamma, \gamma') \quad (\text{CP1}')
\]

Subject to constraints:
\[
\forall s, \gamma : z_{s\gamma} \in \{0, 1\}, \quad \forall \gamma, \gamma' : u_{\gamma\gamma'} \in \{0, 1\} \quad (1)
\]
\[
\forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_{s\gamma} \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \quad (2)
\]
\[
\forall s : \sum_{\gamma} z_{s\gamma} = 1. \quad (3)
\]
LP relaxation for the ILP formulation

- Relax the constraints in the formulation as:

\[ \forall s, \gamma : 0 \leq z_{s\gamma} \leq 1, \quad \forall \gamma, \gamma' : 0 \leq u_{\gamma\gamma'} \leq 1 \]

\[ \forall s, \gamma, \gamma' : u_{\gamma\gamma'} \leq z_s \quad \text{and} \quad u_{\gamma\gamma'} \leq z_{s\gamma'} \]

\[ \forall s : \sum_{\gamma} z_{s\gamma} = 1. \]

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

1: initialize some assignment $y^{(0)}$
2: \textbf{for } $k = 1, 2, \ldots$ \textbf{do}
3: \hspace{1em} select a small spot set $S_\Delta$
4: \hspace{1em} \textbf{for each } $s \in S_\Delta$ \textbf{do}
5: \hspace{2em} find new $\gamma$ that improves objective
6: \hspace{2em} change $y_s^{(k-1)}$ to $y_s^{(k)} = \gamma$ greedily
7: \hspace{1em} \textbf{if} objective could not be improved \textbf{then}
8: \hspace{1em} \textbf{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 $|S_0|$
- Hill1 takes about 2–3 seconds
- LP1 takes around 4–6 seconds
References


References II


