Collective Entity Disambiguation with Structured Gradient Tree Boosting

Yi Yang, Ozan Irsoy, and Kazi Shefaet Rahman

Bloomberg

Bloomberg Engineering

Collective Entity Disambiguation

Entity mentions are ambiguous

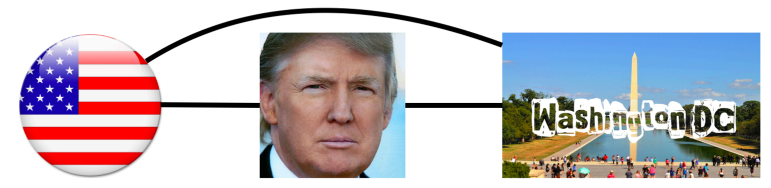
x US president Trump left Washington

y

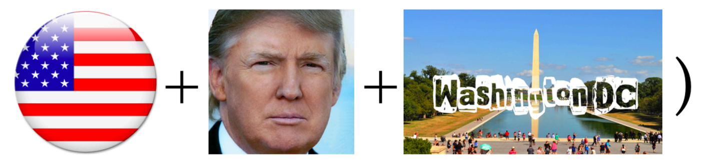
- Local and global context for disambiguation

Structured prediction

- Entity dependencies:

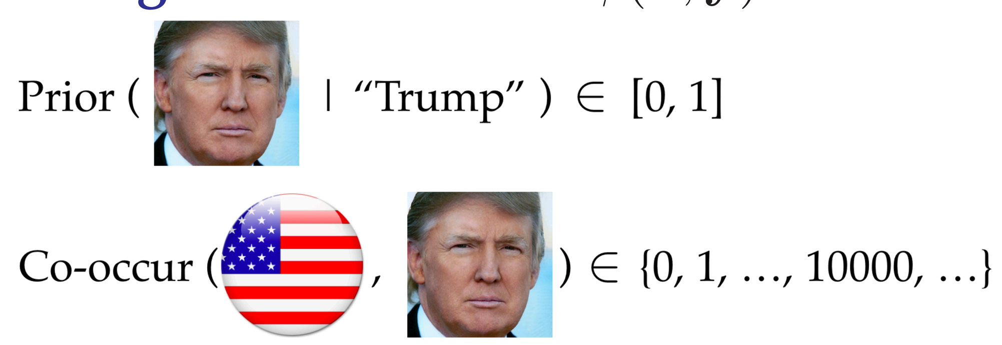


- Inference: $\hat{\mathbf{y}} = \arg\max S(\mathbf{x}, \mathbf{y})$
- Learning:



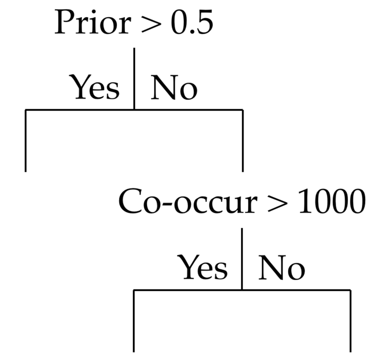
Why Gradient Tree Boosting

Heterogeneous features: $\phi(\mathbf{x}, \mathbf{y})$



- Ideal models can handle:
  - Categorical features and count data
  - Nonlinear relationships between features

Regression-tree-based models



Challenges

- Long-term dependencies between entities
- Approximate inference algorithms

Model

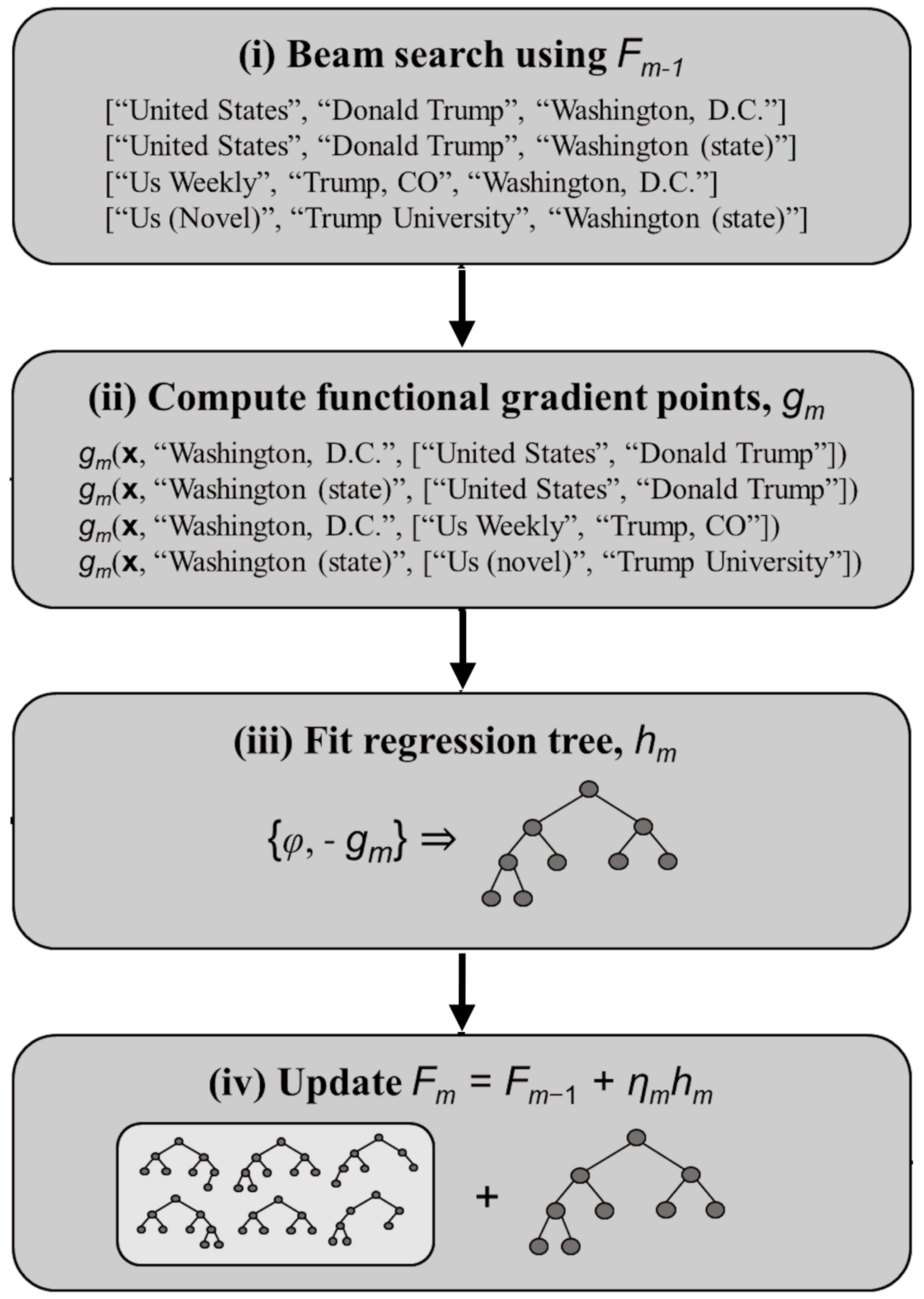
Structured Gradient Tree Boosting (SGTB)

$$S(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} F(\mathbf{x}, y_t, \mathbf{y}_{1:t-1})$$

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\{\sum_{t=1}^{T} F(\mathbf{x}, y_t, \mathbf{y}_{1:t-1})\}}{Z(\mathbf{x})}$$

- Model F using Gradient Tree Boosting.
- Optimize with Functional gradient descent:

$$F_m(\mathbf{x}, y_t, \mathbf{y}_{1:t-1}) = F_{m-1}(\mathbf{x}, y_t, \mathbf{y}_{1:t-1}) - \eta_m g_m(\mathbf{x}, y_t, \mathbf{y}_{1:t-1})$$



Inference

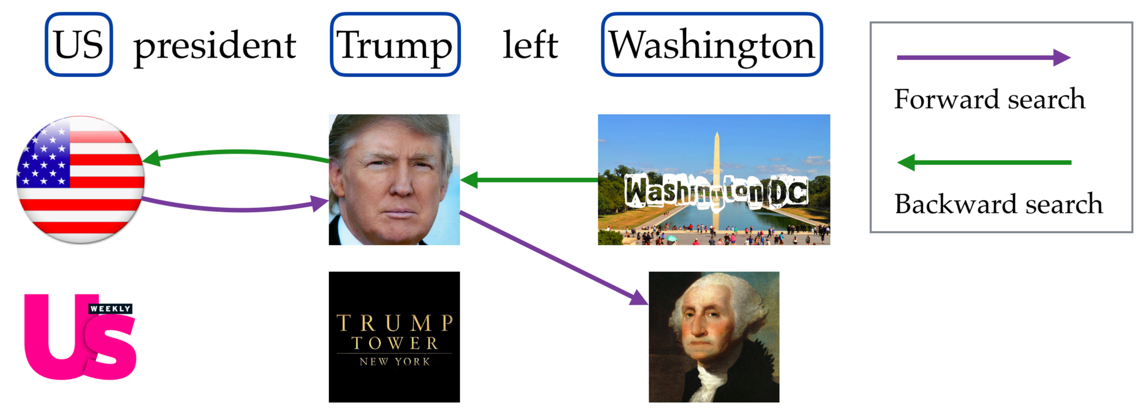
Computing g_m requires inference

$$g_m(\mathbf{x}, y_t, \mathbf{y}_{1:t-1}) = \frac{\partial L(\mathbf{y}^*, S(\mathbf{x}, \mathbf{y}))}{\partial F(\mathbf{x}, y_t, \mathbf{y}_{1:t-1})} = p(\mathbf{y}_{1:t}|\mathbf{x}) - \mathbf{1}[\mathbf{y}_{1:t} = \mathbf{y}^*_{1:t}]$$

$$L(\mathbf{y}^*, S(\mathbf{x}, \mathbf{y})) = \log Z(\mathbf{x}) - S(\mathbf{x}, \mathbf{y}^*)$$

- Exact inference is intractable.
- Beam search is used to approximate $p(\mathbf{y}_{1:t}|\mathbf{x})$.

Bi-directional beam search



- New scoring function: $S(\mathbf{x}, \mathbf{y}) = (\overrightarrow{S} + \overleftarrow{S})/2$

Input : input document $\mathbf{x}$, candidate sequences $\{\mathbf{y}\}$, joint scoring function $S(\mathbf{x}, \mathbf{y}_{t_1:t_2})$
Output: beam sequence set C
$C \leftarrow \emptyset$
while *not converged* **do**
  // forward beam search
  for $t = 1, \cdots, T$ **do**
    $C^{(F)} \leftarrow \text{top-B}_{\mathbf{y}_{1:t}}[S(\mathbf{x}, \mathbf{y}_{1:t}) + S(\mathbf{x}, \mathbf{y}_{T:t})]$
    // add gold subsequence
    $C^{(F)} \leftarrow C^{(F)} \cup \{\mathbf{y}^*_{1:t}\}$
    $C \leftarrow C \cup C^{(F)}$
  end
  // backward beam search
  for $t = T, \cdots, 1$ **do**
    $C^{(B)} \leftarrow \text{top-B}_{\mathbf{y}_{T:t}}[S(\mathbf{x}, \mathbf{y}_{T:t}) + S(\mathbf{x}, \mathbf{y}_{1:t})]$
    // add gold subsequence
    $C^{(B)} \leftarrow C^{(B)} \cup \{\mathbf{y}^*_{T:t}\}$
    $C \leftarrow C \cup C^{(B)}$
  end
end

Experiments

Data

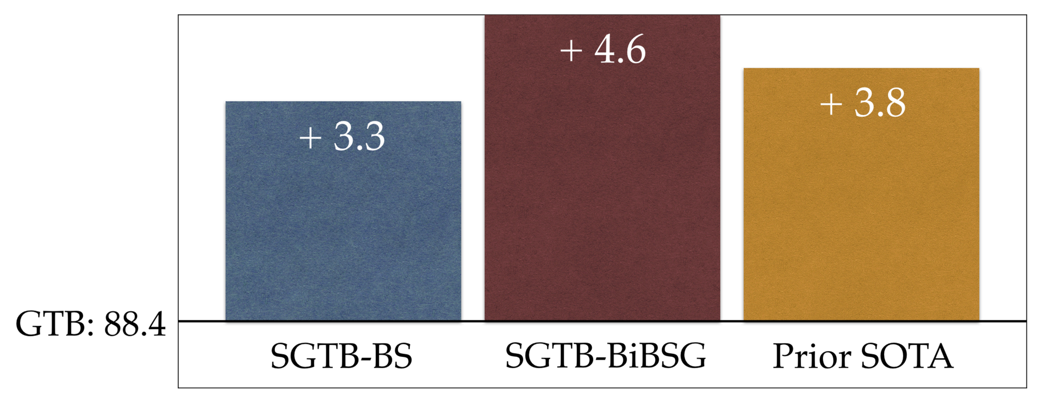
Class	Dataset	# mention	# document	# mention per document
Training	AIDA-train	18,448	946	19.5
Validation	AIDA-dev	4,791	216	22.1
In-domain testing	AIDA-test	4,485	231	19.4
Cross-domain testing	AQUAINT	727	50	14.5
	MSNBC	656	20	32.8
	ACE	257	36	7.1
	CWEB	11,154	320	34.8
	WIKI	6,821	320	21.3

Setup

- Metrics
  - In-KB accuracy
  - Bag-of-Title (BoT) F1 score
- Competing systems
  - Gradient Tree Boosting (GTB)
  - SGTB with Beam search (SGTB-BS)
  - Bidir. BS using Gold path (SGTB-BiBSG)
  - Previous state-of-the-art (SOTA) systems

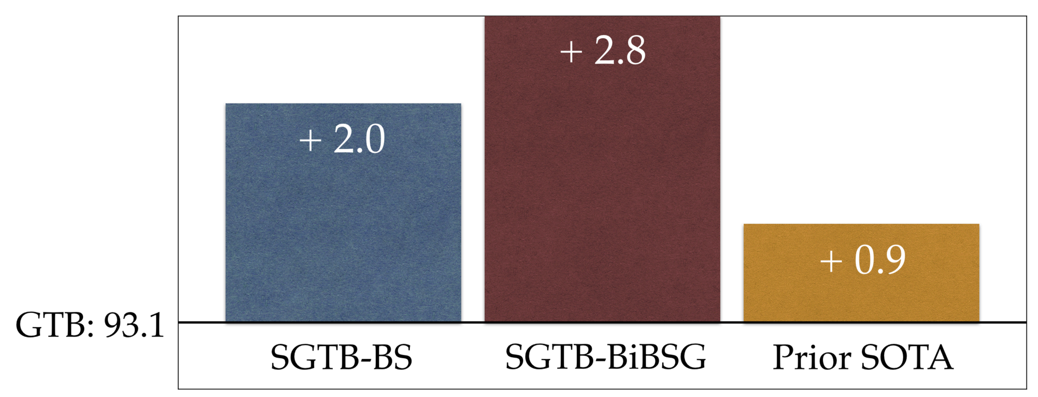
In-domain results

In-KB accuracy gains over GTB



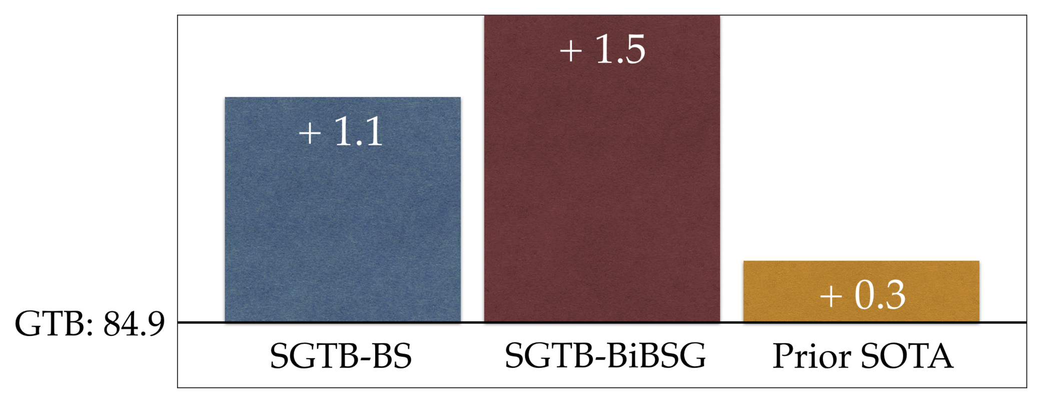
- Generate candidates with PPRforNED

In-KB accuracy gains over GTB



Cross-domain results

Average Bag-of-Title F1 gains over GTB



Summary

- We present a novel Structured Gradient Tree Boosting (SGTB) model for collectively disambiguating entities in a document.
- SGTB combines structured learning with Gradient Tree Boosting to produce globally optimal entity assignments for all the mentions.
- We present Bidirectional Beam Search with Gold path (BiBSG), an efficient approximate inference algorithm tailored for SGTB.
- SGTB achieves state-of-the-art (SOTA) results on popular entity disambiguation datasets of different domains.