Minimally-Supervised Extraction of Entities from Text Advertisements

Sameer Singh Dustin Hillard Chris Leggetter

Department of Computer Science
University of Massachusetts, Amherst MA

Yahoo! Labs, Silicon Valley
Santa Clara CA

Human Language Technologies: North American Chapter of the Association for Computational Linguistics (NAACL HLT)
June 2-4, 2010

Outline

1. **Entity Extraction for Text Advertisements**
2. **Minimally Supervised Learning**
3. **Features**
   - Unsupervised Signal $\{f_k\}$
   - Semi-Supervised Signal $\{f'_k\}$
4. **Experiments**
5. **Conclusions**

Outline

Sponsored Search

- **Problem:** Given a web search query, which ads to display

Sponsored Search

- **Problem:** Given a web search query, which ads to display
- Current solutions consider word- and phrase-based matches
  - *doesn't always work very well:*
  Query: california hotel **Ad:** Hotel California Lyrics . . .

Sponsored Search

- **Problem:** Given a web search query, which ads to display
- Current solutions consider word- and phrase-based matches
  - *doesn't always work very well:*
  Query: california hotel **Ad:** Hotel California Lyrics ...
- There is a need to understand the *intent*
  `Hotel California` is a MediaTitle, not Lodging
- In our work, *intent* takes the form of "entity recognition"

Objective

Input: Bradley International Airport Hotel
Marriott Hartford,CT Airport hotel-free shuttle service & parking.

Objective

Input: Bradley International Airport Hotel
Marriott Hartford,CT Airport hotel-free shuttle service & parking.

Output: Bradley International Airport Hotel
Marriott Hartford, CT Airport hotel free shuttle service & parking.

Labels: airport , travel , lodging_name , product , city , state

Objective

Input: Bradley International Airport Hotel
Marriott Hartford,CT Airport hotel-free shuttle service & parking.

Output: Bradley International Airport Hotel
Marriott Hartford, CT Airport hotel free shuttle service & parking.

Labels: airport , travel , lodging_name , product , city , state

Combined Segmentation and Tagging

Label Taxonomy

place	person	**org_name**	**product**
airport	media_title	sports_team	tech_prod
city	manufacturer	media_org	auto_prod
state	prod_name	apparel_org	media_prod
country	event	tech_org	travel
continent	**business**	airline	apparel
zipcode	tech_business	restaurant	education_prod
occasion	media_business	lodging	other
...	...	...	...

45 such labels

Data

- Lots of unlabeled data available (millions of ads!)

Data

- Lots of unlabeled data available (millions of ads!)
- Labeling a small subset manually is not ideal:
  1. Expensive and time-consuming (domain knowledge required)
  2. Error-prone (editors disagree and make mistakes)
  3. Overfitting

Data

- Lots of unlabeled data available (millions of ads!)
- Labeling a small subset manually is not ideal:
  1. Expensive and time-consuming (domain knowledge required)
  2. Error-prone (editors disagree and make mistakes)
  3. Overfitting
- Partially and noisily labeling lots of data is easy!

Data

- Lots of unlabeled data available (millions of ads!)
- Labeling a small subset manually is not ideal:
    1. Expensive and time-consuming (domain knowledge required)
    2. Error-prone (editors disagree and make mistakes)
    3. Overfitting
- Partially and noisily labeling lots of data is easy!
    - `New Delhi` is a CITY most of the time
    - Token that ends with `.com` is almost always a URL
    - `for`, `and` and `buy` are almost never AIRPORTS
    - Most tokens are useless, don't tag them

Data

- Lots of unlabeled data available (millions of ads!)
- Labeling a small subset manually is not ideal:
  1. Expensive and time-consuming (domain knowledge required)
  2. Error-prone (editors disagree and make mistakes)
  3. Overfitting
- Partially and noisily labeling lots of data is easy!
  - `New Delhi` is a CITY most of the time
  - Token that ends with `.com` is almost always a URL
  - `for`, `and` and `buy` are almost never AIRPORTS
  - Most tokens are useless, don't tag them
- In this work, we rely only on such *partial* and *probabilistic* labels

Outline

1. Entity Extraction for Text Advertisements
2. **Minimally Supervised Learning**
3. Features
   - Unsupervised Signal $\{f_k\}$
   - Semi-Supervised Signal $\{f'_k\}$
4. Experiments
5. Conclusions

Semi-Markov CRF Model

- **Input:** Each ad is a sequence $\mathbf{x}$ of tokens

Semi-Markov CRF Model

- **Input:** Each ad is a sequence $\mathbf{x}$ of tokens
- **Output:** Segmentation $\mathbf{s}$ for the input $\mathbf{x}$
  where $\mathbf{s} = \{s_j\}$ and segment $s_j = \langle str_j, end_j, y_j \rangle$

Semi-Markov CRF Model

- **Input:** Each ad is a sequence $\mathbf{x}$ of tokens
- **Output:** Segmentation $\mathbf{s}$ for the input $\mathbf{x}$
  where $\mathbf{s} = \{s_j\}$ and segment $s_j = \langle str_j, end_j, y_j \rangle$
- **Features:** defined over segments, $\{f_k(\mathbf{x}, \mathbf{s}_j)\}_k$
  - Is the segment `New Delhi` and the label CITY
  - the segment length is ≥ 2

Semi-Markov CRF Model

- **Input:** Each ad is a sequence $\mathbf{x}$ of tokens
- **Output:** Segmentation $\mathbf{s}$ for the input $\mathbf{x}$
  where $\mathbf{s} = \{s_j\}$ and segment $s_j = \langle str_j, end_j, y_j \rangle$
- **Features:** defined over segments, $\{f_k(\mathbf{x}, \mathbf{s}_j)\}_k$
  - Is the segment `New Delhi` and the label CITY
  - the segment length is ≥ 2
- **Model** p**:** $Pr_p(\mathbf{s}|\mathbf{x}) = F\left(\{f_k(\mathbf{x}, \mathbf{s})\}_k, \theta_p\right)$
  - If features are Markov, inference can be performed exactly[1]

[1] Sarawagi and Cohen, NIPS 2004

Supervised Learning

Given labeled data $\{\mathbf{x_i}, \mathbf{s_i}\}$:

$$\forall f_k, \sum_{i=1}^{N} E_{p(\mathbf{s}|\mathbf{x}_i)}[f_k(\mathbf{x}_i, \mathbf{s})] = \sum_{i=1}^{N} f_k(\mathbf{x}_i, \mathbf{s}_i)$$

Supervised Learning

Given labeled data $\{\mathbf{x_i}, \mathbf{s_i}\}$:

$$\forall f_k, \sum_{i=1}^{N} E_{p(\mathbf{s}|\mathbf{x}_i)}[f_k(\mathbf{x}_i, \mathbf{s})] = \sum_{i=1}^{N} f_k(\mathbf{x}_i, \mathbf{s}_i)$$

- Unlabeled data do not have targets (RHS) for the expectations

Semi-Supervised Learning

Given labeled data $\{\mathbf{x_i}, \mathbf{s_i}\}$:

$$\forall f_k, \sum_{i=1}^{N} E_{p(\mathbf{s}|\mathbf{x}_i)}[f_k(\mathbf{x}_i, \mathbf{s})] = \sum_{i=1}^{N} f_k(\mathbf{x}_i, \mathbf{s}_i)$$

- Unlabeled data do not have targets (RHS) for the expectations
- For a subset $\{f'_k\}$, provide constraints manually

$$E[f'_k(\mathbf{x}, \mathbf{s})] \geq u_k$$

Semi-Supervised Learning

Given labeled data $\{\mathbf{x_i}, \mathbf{s_i}\}$:

$$\forall f_k, \sum_{i=1}^{N} E_{p(\mathbf{s}|\mathbf{x}_i)}[f_k(\mathbf{x}_i, \mathbf{s})] = \sum_{i=1}^{N} f_k(\mathbf{x}_i, \mathbf{s}_i)$$

- Unlabeled data do not have targets (RHS) for the expectations
- For a subset $\{f'_k\}$, provide constraints manually

$$E[f'_k(\mathbf{x}, \mathbf{s})] \geq u_k$$

$$[[\texttt{Label=City}\ \textit{given}\ \texttt{``New Delhi''}]] \geq 0.5$$

Semi-Supervised Learning

Given labeled data $\{\mathbf{x_i}, \mathbf{s_i}\}$:

$$\forall f_k, \sum_{i=1}^{N} E_{p(\mathbf{s}|\mathbf{x}_i)}[f_k(\mathbf{x}_i, \mathbf{s})] = \sum_{i=1}^{N} f_k(\mathbf{x}_i, \mathbf{s}_i)$$

- Unlabeled data do not have targets (RHS) for the expectations
- For a subset $\{f'_k\}$, provide constraints manually

$$E[f'_k(\mathbf{x}, \mathbf{s})] \geq u_k$$
$$[[\texttt{Label=City}\ \textit{given}\ \texttt{``New Delhi''}]] \geq 0.5$$

- Constraints on $\{f'_k\}$ are used to learn θ_p over all features $\{f_k\}$
  - online training algorithm in *Bellare et al., UAI 2009*

Outline

Conventional CRF and semi-CRF Features

- Emission Features
  - Token $\times$ Label
  - WindowTokens $\times$ Label
- Transition Features
  - PrevLabel $\times$ Label
- Segment Features
  - SegLength $==$ L
  - SegLength $\times$ Label

Conventional CRF and semi-CRF Features

- Emission Features
  - Token $\times$ Label
  - WindowTokens $\times$ Label
- Transition Features
  - PrevLabel $\times$ Label
- Segment Features
  - SegLength $==$ L
  - SegLength $\times$ Label

We need more features to propagate the constraints

Segment Clusters

- `London` is similar to `Boston`, but context may not capture that

Segment Clusters

- London is similar to Boston, but context may not capture that
- Cluster segments based on a large corpus[2]
  - take 5.1 billion English sentences from the web
  - use co-occurence of segments as distance
  - cluster using K-Means

[2] Pantel et al., EMNLP 2009

Segment Clusters

- `London` is similar to `Boston`, but context may not capture that
- Cluster segments based on a large corpus[2]
  - take 5.1 billion English sentences from the web
  - use co-occurence of segments as distance
  - cluster using K-Means
- Cluster identity of each segment is added as a feature
  - segments in the same cluster should have the same label

[2] Pantel et al., EMNLP 2009

Topic Models

- Ads of the same domain will have similar label distribution
  - Ads in the travel domain usually have PLACE in it

Topic Models

- Ads of the same domain will have similar label distribution
  - Ads in the travel domain usually have PLACE in it
- The domains of the ads are unknown
  - approximate using unsupervised techniques

Topic Models

- Ads of the same domain will have similar label distribution
  - Ads in the travel domain usually have PLACE in it
- The domains of the ads are unknown
  - approximate using unsupervised techniques
- **Topic Models**: given a corpus of documents, identify the "topics"
  - run LDA to obtain topic distributions over the ads
- The topic distribution of each ad is used as a feature

Semi-Supervised Signal

- Constraints are features with associated target expectations
  - e.g. [[`Label=STATE` *given* `''arizona''`]] ≥ 0.5

Semi-Supervised Signal

- Constraints are features with associated target expectations
  - e.g. [[`Label=State` *given* `‘‘arizona’’`]] ≥ 0.5
- Specifying the targets is not easy:
  1. Use prior knowledge
  2. Evaluate on held-out data
  3. Use predictions to tweak the targets
  4. Use output of previous model
- Robustness to noise in targets has not been studied

Dictionary-Based

- Dictionary is a list of segments for a label
  - *airports, cities, countries, . . .*
- Can be obtained from a number of different sources:
  - *databases, lexicons, manual collections, output of another model*

Dictionary-Based

- Dictionary is a list of segments for a label
  - *airports, cities, countries, ...*
- Can be obtained from a number of different sources:
  - *databases, lexicons, manual collections, output of another model*
- Constraint is added for segment match for each dictionary
  - accurate dictionaries get higher targets

Dictionary-Based

- Dictionary is a list of segments for a label
  - *airports, cities, countries, ...*
- Can be obtained from a number of different sources:
  - *databases, lexicons, manual collections, output of another model*
- Constraint is added for segment match for each dictionary
  - accurate dictionaries get higher targets
- **External Databases**
  - lexicons of airports, cities, countries etc. are easily available
  - for other labels, we use product databases within Yahoo!

Dictionary-Based

- Dictionary is a list of segments for a label
  - *airports, cities, countries, . . .*
- Can be obtained from a number of different sources:
  - *databases, lexicons, manual collections, output of another model*
- Constraint is added for segment match for each dictionary
  - accurate dictionaries get higher targets
- **External Databases**
  - lexicons of airports, cities, countries etc. are easily available
  - for other labels, we use product databases within Yahoo!
- **Query Entity-Extraction Model**
  - similar task of tagging web search queries (similar set of labels)
  - predictions are not good, but provide a weak signal

Pattern-Based

- Dictionaries don't utilize the context

Pattern-Based

- Dictionaries don't utilize the context
- Introduce *patterns* that provide additional signal
- Examples:
  - `Flights to` PLACE
  - `city of` CITY
  - `Looking for` PRODUCT `find it here`
- can also use pattern-discovery algorithms

Domain-Based

- Guide model predictions to avoid degenerate solutions

Domain-Based

- Guide model predictions to avoid degenerate solutions
- Priors of segmentation (independent of the labels)
  - $\Pr(\texttt{SegLength} \leq 2) \geq 0.8$
  - $\Pr(\texttt{SegLength > 6}) \leq \epsilon$
  - Every dictionary also informs the segmentation

Domain-Based

- Guide model predictions to avoid degenerate solutions
- Priors of segmentation (independent of the labels)
  - $\Pr(\texttt{SegLength} \leq \texttt{2}) \geq 0.8$
  - $\Pr(\texttt{SegLength > 6}) \leq \epsilon$
  - Every dictionary also informs the segmentation
- Priors on labels
  - $\Pr(\texttt{label == OTHER}) \geq 0.5$

Outline

Setup

- **Data**
  - Two datasets: $14k$ and $42k$ randomly sampled ads from Yahoo!
  - Training Time: $\sim$90 minutes and $\sim$120 minutes
  - Inference Time: 8 minutes and 32 minutes
  - $2,157$ ads labeled for evaluation (@$20-25$ ads per hour)

Setup

- **Data**
  - Two datasets: $14k$ and $42k$ randomly sampled ads from Yahoo!
  - Training Time: $\sim$90 minutes and $\sim$120 minutes
  - Inference Time: 8 minutes and 32 minutes
  - $2,157$ ads labeled for evaluation (@$20-25$ ads per hour)
- **Methods**
  1. *Bootstrapped:* Dictionary-based predictions
  2. *QSup:* Supervised model using labeled web queries
  3. *Our Method* has 14k and 42k variations

  - Only using labeled ads data gave extremely poor results

Tokenwise Accuracy (w/ partial credit)

Metric	Dictionary	14k	42k	QSup
Overall Accuracy	*46.6*	*62.7*	*64.9*	*68.5*
non-OTHER Recall	*20.5*	*41.2*	*32.5*	*34.2*
non-OTHER Precision	*16.3*	*33.3*	*35.7*	*46.9*
F1-score	*18.2*	*36.8*	*34.0*	*39.5*
F2-score	*19.5*	*39.3*	*33.1*	*36.1*

Outline

1. Entity Extraction for Text Advertisements
2. Minimally Supervised Learning
3. Features
   - Unsupervised Signal $\{f_k\}$
   - Semi-Supervised Signal $\{f'_k\}$
4. Experiments
5. Conclusions

Summary

Contributions

- Entity Recognition for advertisements without labeled data

Summary

Contributions

- Entity Recognition for advertisements without labeled data
- Real-world application of semi-supervised learning

Summary

Contributions

- Entity Recognition for advertisements without labeled data
- Real-world application of semi-supervised learning
- Not having any labeled data is not the end of the world
  - use existing resources as noisy supervision

Summary

Contributions

- Entity Recognition for advertisements without labeled data
- Real-world application of semi-supervised learning
- Not having any labeled data is not the end of the world
  - use existing resources as noisy supervision

Future Work

- Use in downstream applications (*click prediction*, *ad retrieval*, ...)
- Robustness to target expectations
- Add constraints that use other sources

Thanks!

Sameer Singh, Dustin Hillard and Chris Leggetter

University of Massachusetts, Amherst
Yahoo! Labs, Santa Clara