

High Precision Web Extraction using Site Knowledge

Meghana Kshirsagar
Rajeev Rastogi
Sandeepkumar Satpal
Srinivasan H Sengamedu
Venu Satuluri



Outline

- Motivation
- Problem Definition
- Proposed Approach
 - Site-Knowledge
 - Segmentation
 - Segment Label Selection
 - Node Label Correction
 - Extensions
- Experimental Results
- Conclusions




Information Extraction: What & Why?

Shopping > Electronics > Digital Cameras > All Digital Cameras > Canon EOS 5D Digital Camera

Canon EOS 5D Digital Camera

Overview | Compare Prices | Specifications | Reviews



\$2,399.99 ~~\$2,499.99~~

★★★★★ 140 Ratings (21 Reviews)

Write a Review

Featured Merchants

TigerDirect	\$2,399.99	Go to store
amazon.com Marketplace	\$2,499.99	Go to store

See new & used prices (2) »

Product Description: Canon EOS 5D

Widen Your Horizons.

The new Canon EOS 5D offers advanced photographers a lightweight, rugged digital SLR that uses Canon's superlative EF lenses without a conversion factor. Its full-frame 12.8 megapixel CMOS sensor combines with Canon's DIGIC II Image Processor, a high-precision 9-point autofocus system with 6 assist points, and "Picture Style" color control to deliver...

Product Specification: Canon EOS 5D

Megapixels:	12.8 Megapixels
Lens System:	Body only
LCD Screen Size:	2.5 in

Web Images Video Local Shopping More

YAHOO! canon camera digital

View Notes (1)

SearchScan - On

229,000,000 results for canon camera digital:

Also try: [canon camera digital malaysia](#), [More...](#)

Sponsored Results

Shop for Products Online


Free quotes-Multiple dealers, compare prices, & save big.

[www.Alibaba.com](#)

Canon Digital Cameras

- Canon EOS 5D**
From \$2399.99 • 12.8 Megapixels • Body only
★★★★★ (140) [Reviews](#) | [Prices](#) | [Specifications](#)
- Canon PowerShot G7**
From \$799.99 • 10 Megapixels • 6 X
[Reviews](#) | [Prices](#) | [Specifications](#)
- Canon PowerShot SD800 IS Digital ELPH**
From \$199.99 • 7.1 Megapixels • 3.8 X
★★★★★ (141) [Reviews](#) | [Prices](#) | [Specifications](#)

More Results...



Name	Price	Rating	Num Rating	Resolution	Lens
Canon EOS 5D	2399.99	5	140	12.8	Body Only



Approaches to Extraction: Wrapper

Search for (e.g. taco, salon, Max's) Near (Address, City, State or Zip) Long Beach, Ca Search

Welcome About Me Write a Review **Find Reviews** Invite Friends Messaging Talk Events Member Search | Account | Log In

Dean Anthony **5.0** based on 15 reviews Rating Details

Categories: Barbers, Skin Care, Hair Salons [Edit]

5287 E 2nd St
Long Beach, CA 90803
(562) 434-3326
www.deananthonymb.com

Hours: Tue-Fri 9:00 a.m.-9:00 p.m. Sat 9:00 a.m.-6:00 p.m. Price Range: \$\$\$ Accepts Credit Cards: Yes Wheelchair Accessible: Yes Parking: Street By Appointment Only: No

Edit Business Info Is this your business? First to Review Benjamin M.

Send to Friend Bookmark Send to Phone Write a Review Print version

You Might Also Consider Sponsored Result

Splittends "An Open Letter of Apology to my Regular Stylist: It is with great shame and guilt that I write this review. You knew it when we met, I'm a..." read more

136 reviews Costa Mesa, CA

15 Reviews for Dean Anthony Search Reviews

Sort by: Recent + Votes | Date | Rating | Elites

09/05/2008
I recently went to dean anthony and was very impressed i was referred by the cigar shop down the street and i must say that i was very happy with my experience.

Mark N. Irvine, CA
The stylist was super nice and very skilled, everything from the cut to the shampoo to the mini facial was excellent i would have to say i would and have recommended it to friends and family and have hear nothing but a excellent response.

Keep up the good work everyone over there.. you will be seeing me for a clean up very soon :)

People Who Viewed This Also Viewed...

- Syndicate Barber Shop 4.0 44 reviews Long Beach, CA Category: Barbers
- Salon Pop & Barber Shop 4.0 49 reviews Long Beach, CA Category: Skin Care
- D'aversa the Salon 4.0 9 reviews Long Beach, CA Category: Hair Salons
- The Spot 4.0 34 reviews Seal Beach, CA Category: Hair Salons

Search for (e.g. taco, salon, Max's) Near (Address, Neighborhood, City, State or Zip) New York, NY Search

Welcome About Me Write a Review **Find Reviews** Invite Friends Messaging Talk Events Member Search | Account | Log In

Novecento **4.5** based on 39 reviews Rating Details

Category: Argentine [Edit]

Neighborhood: SoHo
343 W Broadway
(between Broome St & Catherine Ln)
New York, NY 10013
(212) 925-4706
www.novecento.com

Nearest Transit: Canal-Church Sts (A, C, E) Canal-Varick Sts (1) Canal Street (J, M, Z, N, Q, R, W, 6)

Hours: Mon-Sun. 12:00 p.m. - 12:00 a.m. Accepts Credit Cards: Yes Attire: Casual Good for Kids: No Waiter Service: Yes Good for: Dinner

Price Range: \$\$\$ Parking: Street Good for Groups: Yes Takes Reservations: Yes Wheelchair Accessible: Yes Alcohol: Full Bar

Edit Business Info First to Review B. K.

Send to Friend Bookmark Send to Phone Write a Review Print version

39 reviews for Novecento Search Reviews

Review Highlights What's this?

- "Try the ham & cheese empanada with the skirt steak." (in 19 reviews)
- "Absolutely juicy, so tasty, with mounds of mashed potatoes." (in 10 reviews)
- "Entree was steak chimichurri, chicken or pesto penne." (in 7 reviews)

Rating Distribution | Trend

5 stars	10
4 stars	15
3 stars	5
2 stars	2
1 star	1

People Who Viewed This Also Viewed...

- Industria Argentina 4.0 16 reviews Neighborhood: TriBeCa Category: Steakhouses
- Estancia 460 4.0 13 reviews Neighborhood: TriBeCa Category: Argentine
- Chimichurri Grill 4.0 26 reviews Neighborhood: Theater District Category: Steakhouses
- Azul Bistro 4.0 37 reviews



Structural Changes

Nymag.com

Chimichurri Grill

606 Ninth Ave., New York, NY 10036
nr. 43rd St. [See Map](#) | [Subway Directions](#)

212-586-8655 [Send to Phone](#)

- Price Range: \$\$\$\$
- Reader Rating: 9.0 out of 10 2 Reviews | [Write a Review](#)
- Cuisine: South American, Steakhouse



[Map](#) [Official Website](#) [Rate & Review](#)

Yelp.com


Chimichurri Grill

★★★★☆ based on 17 reviews [Rating Details](#)

Categories: Argentine, Steakhouses [Edit]

Neighborhoods: Manhattan/Hell's Kitchen, Manhattan/Theater District

606 9th Ave
(between 43rd St & 44th St)
New York, NY 10036
(212) 586-8655



[Add Photos](#)

8th Ave-42nd St (A, C, E, 1, 2, 3, S, 7, N, Q, R, W) [Mon-Sun 12:00 p.m.-12:00 a.m.](#)

Price Range: \$\$\$
Accepts Credit Cards: Yes
Attire: Casual
Takes Reservations: Yes
Outdoor Seating: No

Good for Groups: No
Take-out: Yes
Good for: Dinner

Parking: Street
Good for Kids: No
Waiter Service: Yes

[Edit Business Info](#) [Is this your business?](#)

[Send to Friend](#) [Bookmark](#) [Send to Phone](#) [Write a Review](#) [Print version](#)

You Might Also Consider Sponsored Result

Gyu-Kaku "Came here on a double date with a couple who would not stop raving about this place. And they were right to rave! Gyu-Kaku is delicious good..." [read](#)

★★★★☆ 130 reviews
Neighborhood: Manhattan/East Village

Site-specific training data is required

Profile

A Ninth Avenue favorite with the pre- and post-theater crowd, this cozy steakhouse specializes in well-seasoned beef and the tangy, eponymous sauce that accompanies it. The Argentinian diet has relied on meat since well before the Atkins craze, so it's no wonder the restaurant imports their prime, marbled Angus and chewy, flavorful short ribs from the homeland. The South American country is also known for its Italian heritage, so vegetarians can opt for homemade raviolis and daily pasta specials. Since prices are a fraction of midtown's classic steakhouses, expect a crowd during peak hours; this small, homey room only has about a dozen tables.

— Carla Spartos

HOURS
Mon-Fri, noon-midnight; Sat, 4pm-midnight; Sun, 4pm-10:30pm

NEARBY SUBWAY STOPS
A, C, E at 42nd St.-Port Authority Bus Terminal

PRICES
\$17-\$28

PAYMENT METHODS
American Express, MasterCard, Visa

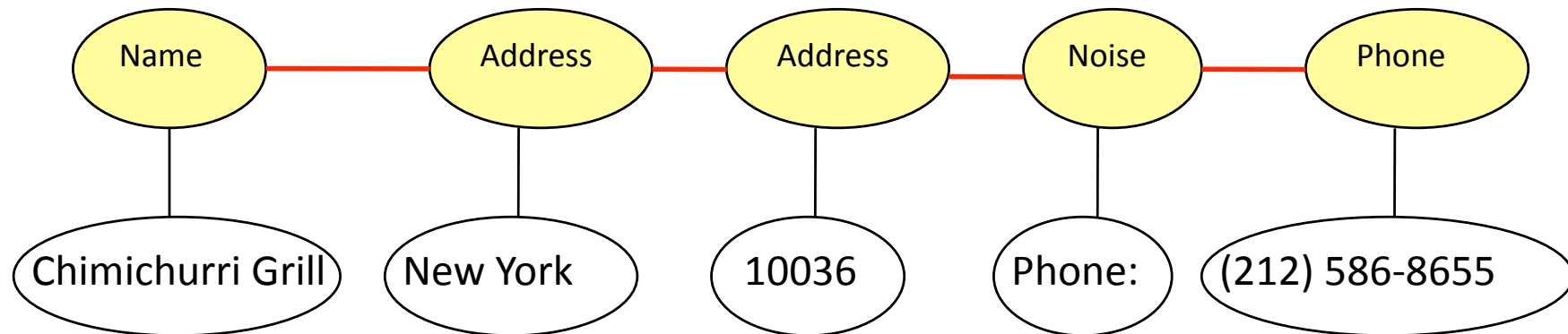


IE as a Labeling Problem

Input: Web Page

Output: Labels for different parts of the page

Labels can be *Restaurant Name, Address, Phone, Rating, Noise, etc.*





Features

Regex features

isAllCapsWord
hasTwoContinuousCaps
isDay
1-2digitNumber
3digitNumber
4digitNumber
5digitNumber
>5digitNumber
dashBetweenDigits
isAlpha
isNumber

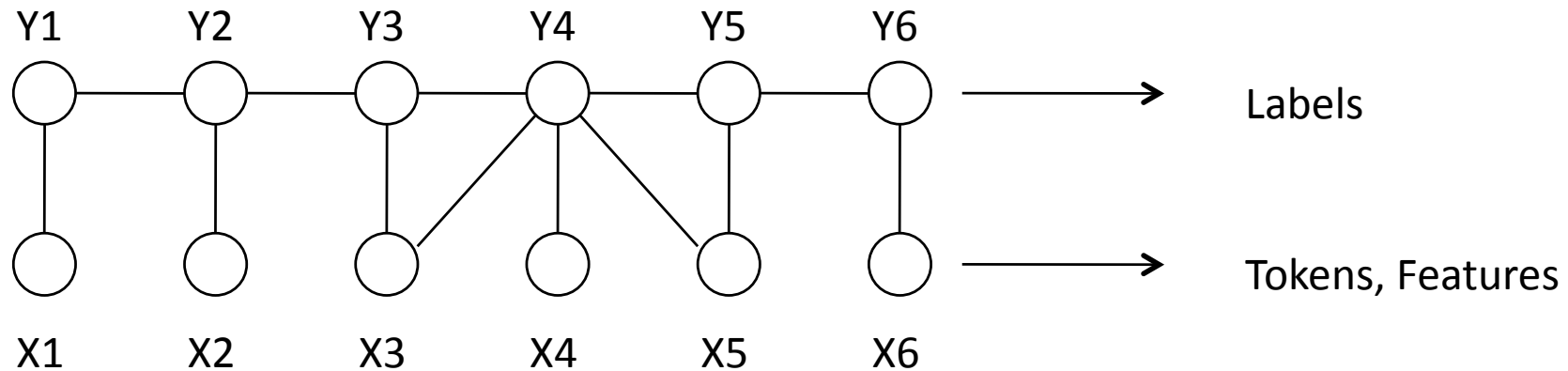
Node-level features

noOfWords>20
noOfWords>50
noOfWords>100
propOfTitleCase<0.2
propOfTitleCase>0.8
overlapWithTitle
prefixOverlapWithTitle



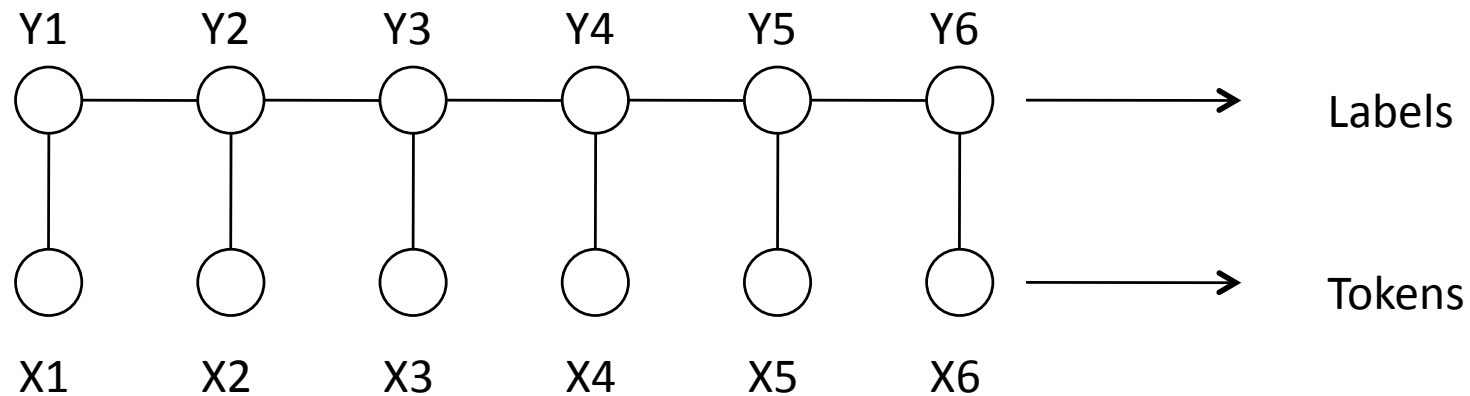
ML Models for Labeling

- Classification
- Sequential Models: HMM, CRF
- FOPC + uncertainty: Markov Logic Networks





Conditional Random Fields



- Features are defined over (x,y) : $f(x,y)$
 - $f([0-9]^*, \text{Phone})$
 - $f(\text{New York}, \text{Address})$
- Conditional random field is a log-linear function over these features

$$\Pr(y|x, w) = \frac{1}{Z(x)} \exp\left(\sum_k w_k f_k(\mathbf{x}, \mathbf{y})\right).$$



Approaches to Extraction: Summary

- Wrapper
 - High Precision (> 99%)
 - Large editorial requirement

- Machine Learning Models
 - Low editorial requirements
 - Low precision due to variable site structure and abundance of noise in web pages

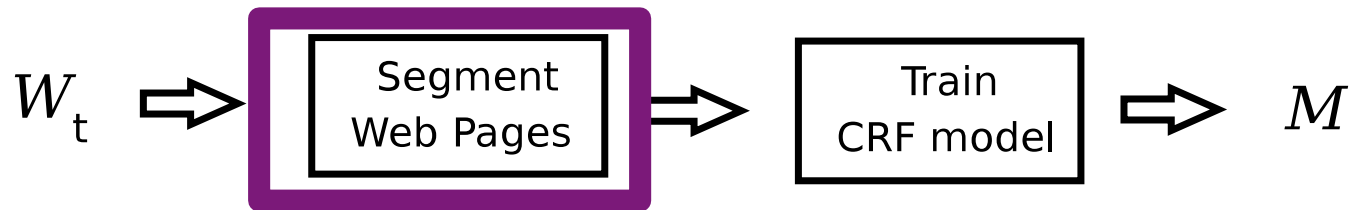


Problem Definition

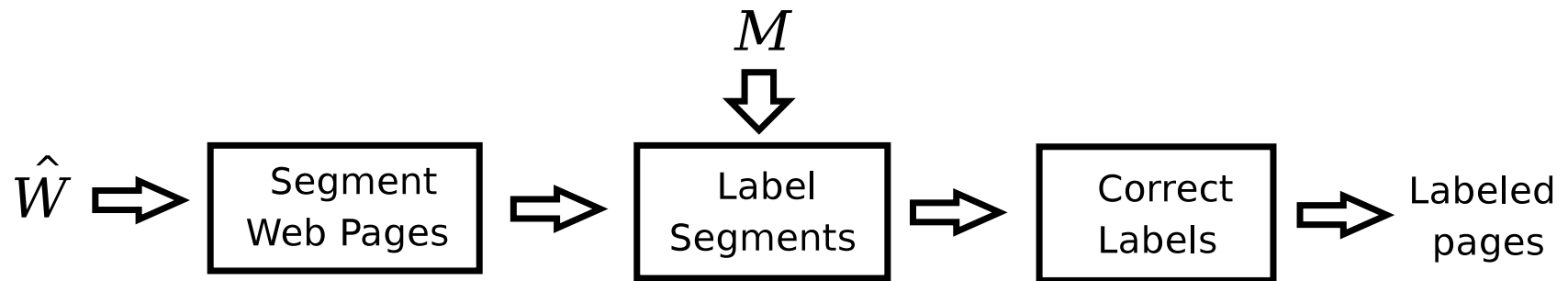
- Problem
 - Extract entities from the Web pages with high precision (> 99%) and very low editorial requirements
- Approach
 - Use CRFs for initial labeling
 - Apply Site Knowledge to improve the precision on a small number of pages
 - Construct Wrappers using these labels
- Site Knowledge
 - Uniqueness: Attributes like *Name*, *Address*, *Hours* are unique per page.
 - Proximity: Attributes describing product/business are close to each other in a page.
 - Sequentiality: Attributes in a site occur in the same sequence in its web pages.



Our Approach



(a) Training Phase



(b) Labeling Phase



Static Text in Scripted Pages

Search for (e.g. taco, salon, Max's) Near (Address, City, State or Zip)

Welcome About Me Write a Review **Find Reviews** Invite Friends Messaging Talk Events [Member Search](#) [Account](#) | [Log In](#)

Dean Anthony

★★★★★ based on 15 reviews [Rating Details](#)

Categories: [Barbers](#), [Skin Care](#), [Hair Salons](#) [Edit]

5287 E 2nd St
Long Beach, CA 90803
(562) 434-3326
www.deananthonymb.com

Hours: Tue-Fri 9:00 a.m.-9:00 p.m.
Sat 9:00 a.m.-6:00 p.m.

Price Range: \$\$\$
Parking: Street
By Appointment Only: No

Accepts Credit Cards: Yes
Wheelchair Accessible: Yes

[Edit Business Info](#) [Is this your business?](#) [First to Review](#) 

[Send to Friend](#) [Bookmark](#) [Send to Phone](#) [Write a Review](#) [Print version](#)

You Might Also Consider *Sponsored Result*

Splittens
★★★★★ 136 reviews
Costa Mesa, CA
"An Open Letter of Apology to my Regular Stylist: It is with great shame and guilt that I write this review. You knew it when we met, I'm a..." [read more](#)

15 Reviews for Dean Anthony

Sort by: **Recent + Votes** | [Date](#) | [Rating](#) | [Elites](#)

 0  1 **★★★★★** 09/05/2008

I recently went to dean anthony and was very impressed i was referred by the cigar shop down the street and i must say that i was very happy with my experience.

The stylist was super nice and very skilled, everything from the cut to the shampoo to the mini facial was excellent i would have to say i would and have recommended it to friends and family and have hear nothing but a excellent response.

Keep up the good work everyone over there.. you will be seeing me for a clean up very soon :)

People Who Viewed This Also Viewed...

- Syndicate Barber Shop**
★★★★★ 44 reviews
Long Beach, CA
Category: Barbers
- Salon Pop & Barber Shop**
★★★★★ 39 reviews
Long Beach, CA
Category: Skin Care
- D'aversa the Salon**
★★★★★ 9 reviews
Long Beach, CA
Category: Hair Salons
- The Spot**
★★★★★ 34 reviews
Seal Beach, CA
Category: Hair Salons

Static Text

Search for (e.g. taco, salon, Max's) Near (Address, Neighborhood, City, State or Zip) Now in the UK!

Welcome About Me Write a Review **Find Reviews** Invite Friends Messaging Talk Events [Member Search](#) [Account](#) | [Log In](#)

Novecento

★★★★★ based on 39 reviews [Rating Details](#)

Category: [Argentine](#) [Edit]


Neighborhood: SoHo
343 W Broadway
(between Broome St & Catherine Ln)
New York, NY 10013
(212) 925-4706
www.novecento.com

Nearest Transit: Canal-Church Sts (A, C, E)
Canal-Varick Sts (1)
Canal Street (J, M, Z, N, Q, R, W, 6)

Hours: Mon-Sun: 12:00 p.m. - 12:00 a.m.

Accepts Credit Cards: Yes
Attire: Casual
Good for Kids: No
Waiter Service: Yes
Good for: Dinner




Price Range: \$\$\$
Parking: Street
Good for Groups: Yes
Takes Reservations: Yes
Wheelchair Accessible: Yes
Alcohol: Full Bar

[Edit Business Info](#) [First to Review](#) 

[Send to Friend](#) [Bookmark](#) [Send to Phone](#) [Write a Review](#) [Print version](#)

39 reviews for Novecento

Review Highlights [What's this?](#)

-  "Try the ham & cheese empanada with the **skirt steak**." (in 19 reviews)
-  "Absolutely juicy, so tasty, with mounds of **mashed potatoes**." (in 10 reviews)
-  "Entree was steak **chimichurri**, chicken or pesto penne." (in 7 reviews)

Rating Distribution | Trend

5 stars	10
4 stars	15
3 stars	10
2 stars	5
1 star	0

People Who Viewed This Also Viewed...

- Industria Argentina**
★★★★★ 16 reviews
Neighborhood: TriBeCa
Category: Steakhouses
- Estancia 460**
★★★★★ 13 reviews
Neighborhood: TriBeCa
Category: Argentine
- Chimichurri Grill**
★★★★★ 26 reviews
Neighborhood: Theater District
Category: Steakhouses
- Azul Bistro**
★★★★★ 37 reviews



Segmentation

Chimichurri Grill

★★★★☆ based on 17 reviews [Rating Details](#) »

Categories: [Argentine](#), [Steakhouses](#) [Edit]

Neighborhoods: [Manhattan/Hell's Kitchen](#), [Manhattan/Theater District](#)

606 9th Ave
(between 43rd St & 44th St)
New York, NY 10036
(212) 586-8655

www.chimichurri.grill.com



Add Photos

Nearest Transit:

8th Ave-42nd St (A, C, E, 1, 2, 3, S, 7, N, Q, R, W)

Good for Groups: No

Take-out: Yes

Good for: Dinner

Hours:

Mon-Sun 12:00 p.m.-12:00 a.m.

Parking: Street

Good for Kids: No

Waiter Service: Yes

Price Range: \$\$\$

Accepts Credit Cards: Yes

Attire: Casual

Takes Reservations: Yes

Outdoor Seating: No

[Edit Business Info](#) [Is this your business?](#)

[First to Review](#) Danielle c.

[Send to Friend](#)

[Bookmark](#)

[Send to Phone](#)

[Write a Review](#)

[Print version](#)

You Might Also Consider

Sponsored Result

Gyu-Kaku

★★★★☆ 130 reviews

Neighborhood: [Manhattan/East Village](#)

"Came here on a double date with a couple who would not stop raving about this place. And they were right to rave! Gyu-Kaku is delicious good..." [read](#)

- Static Node
 - Same (text,xpath) in majority of pages
- Segmenting Web page
 - Partition Web page into Segments using Static nodes

Segmented Sequence

- [Chimichurri Grill]
- [based on 17 reviews]
- { Rating details }
- { Categories }
- [Steakhouses, Argentine]
- { Neighbourhoods }
- [Theatre district, Kitchen]
- [603 9th Ave]
- [.....]
- [(212) 586-8655]
- [www.chimichurri.grill.com]
- { Nearest Transit: }
- [8th Ave]
- [.....]

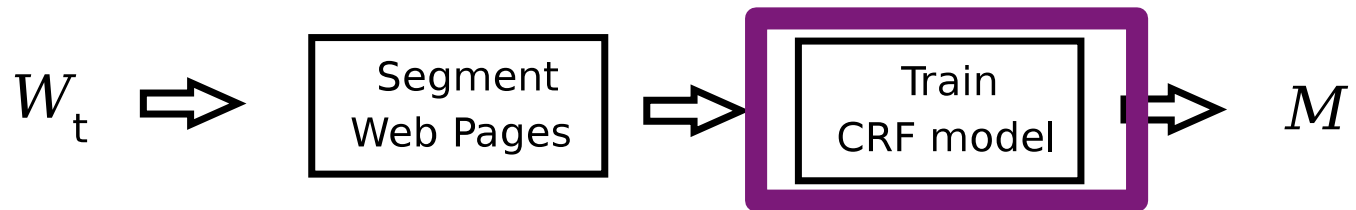


Benefits of Static Text and Segmentation

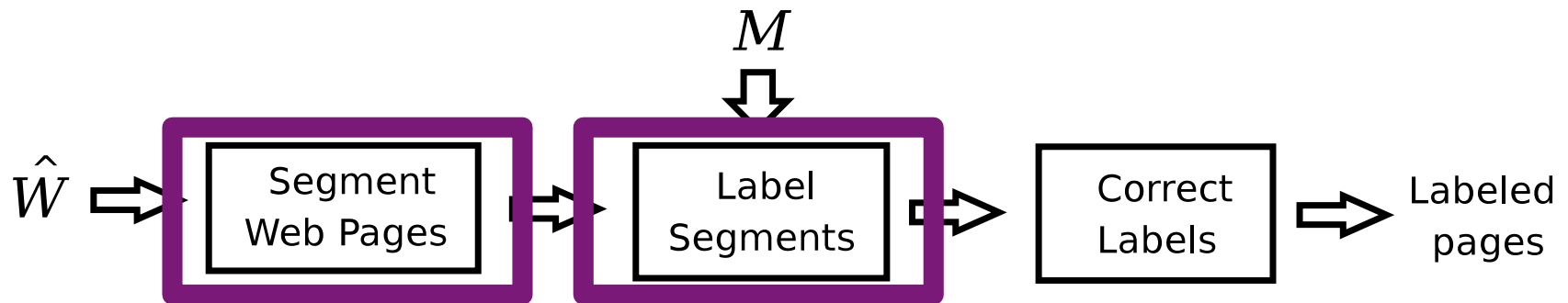
- Noise removal (40%)
- Time requires to train a model is less due to small Instances
- Better control on Precision and Recall by controlling number of Noisy segments (10%)
- Very useful to define context



Our Approach



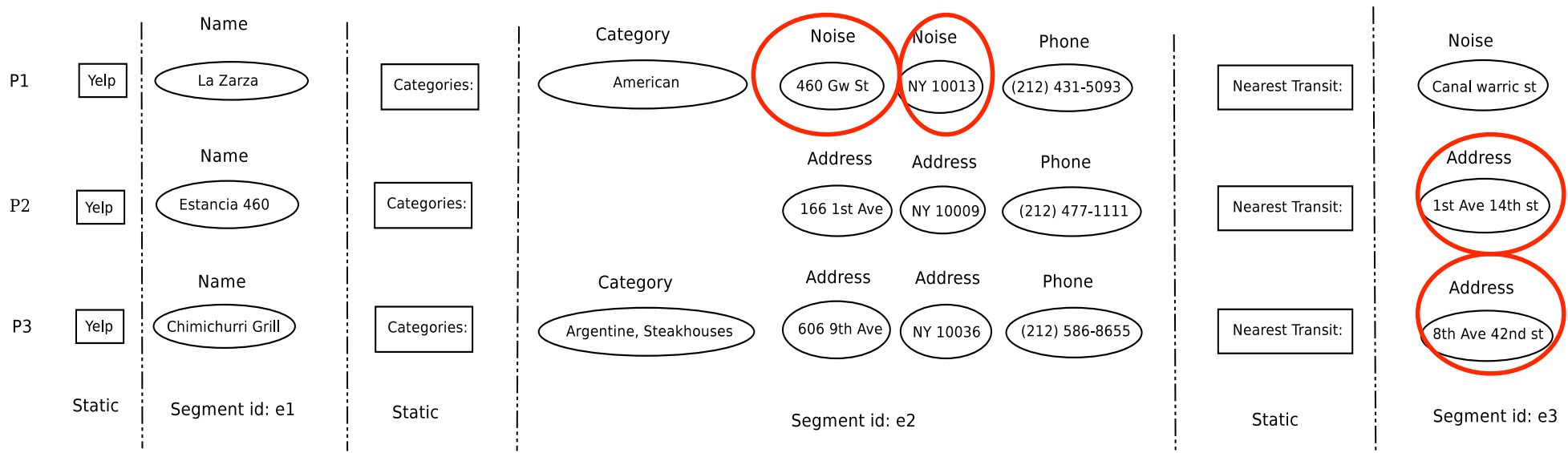
(a) Training Phase



(b) Labeling Phase



CRF Labeling



Identify attribute labels at segment level

seg("address") = e2

Use *Attribute Uniqueness & Proximity*

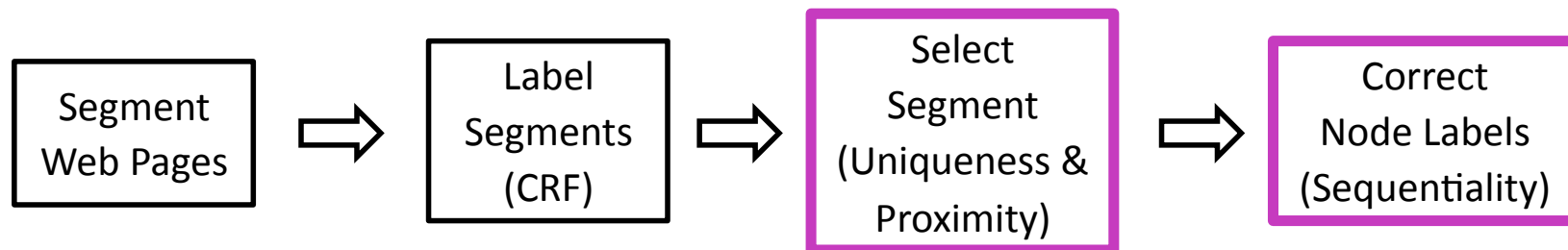
Fix node labels

"Noise" -> "Address" in Segment e2

Use *Sequentiality*



Label Correction



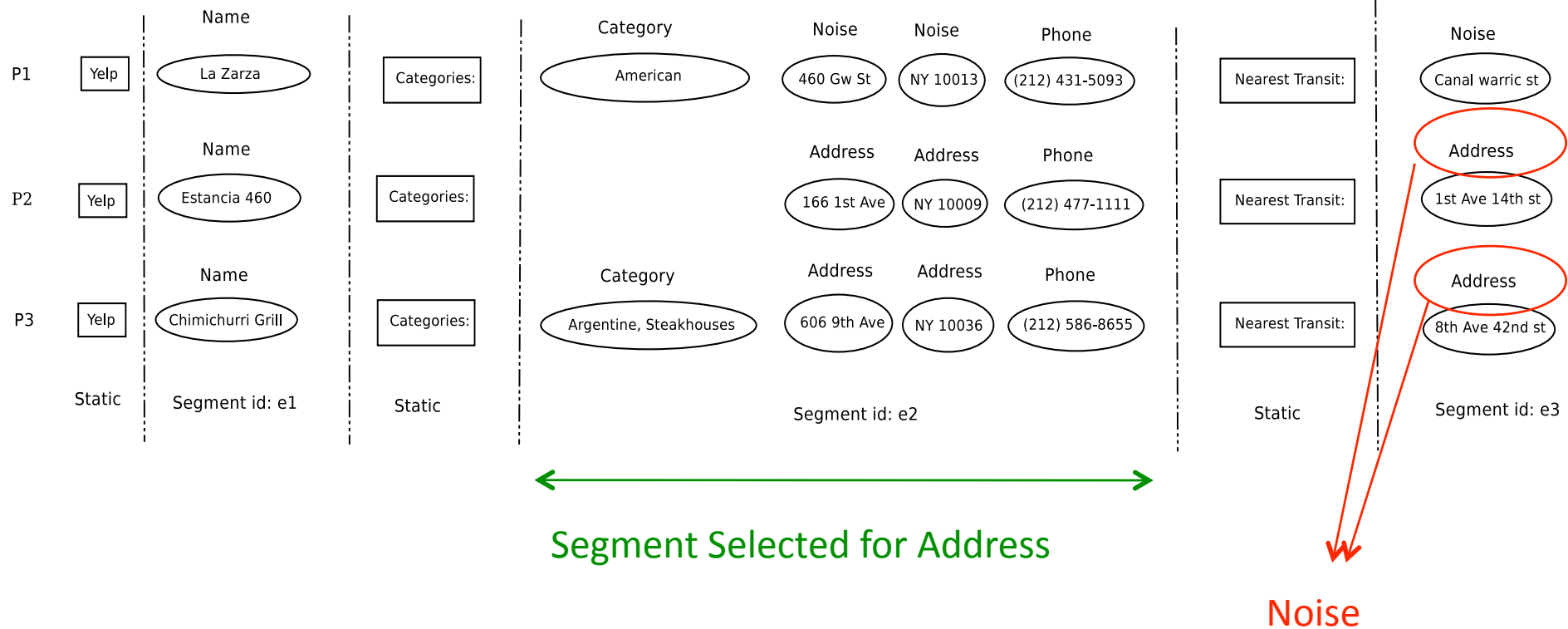


Segment Selection

- Intra-page Constraint (Site Knowledge)
 - Uniqueness Constraint: Attributes like *Name*, *Address*, *Hours* are unique per page
 - Proximity: Attributes describing product/business are close to each other
- Intuition is to select the segments which are in close proximity



Segment Selection





Segment Selection

- For each attribute A , select single segment $seg(A)$ such that

$$\sum_{A, A'} dist(seg(A), seg(A'))$$

is minimum.

- This problem is NP Hard
- Heuristic: for each segment e , define weight w_e as

$$w_e = \sum_f dist(e, f) \cdot |attr(f)|;$$

- For each attribute A , choose the segment with minimum weight.

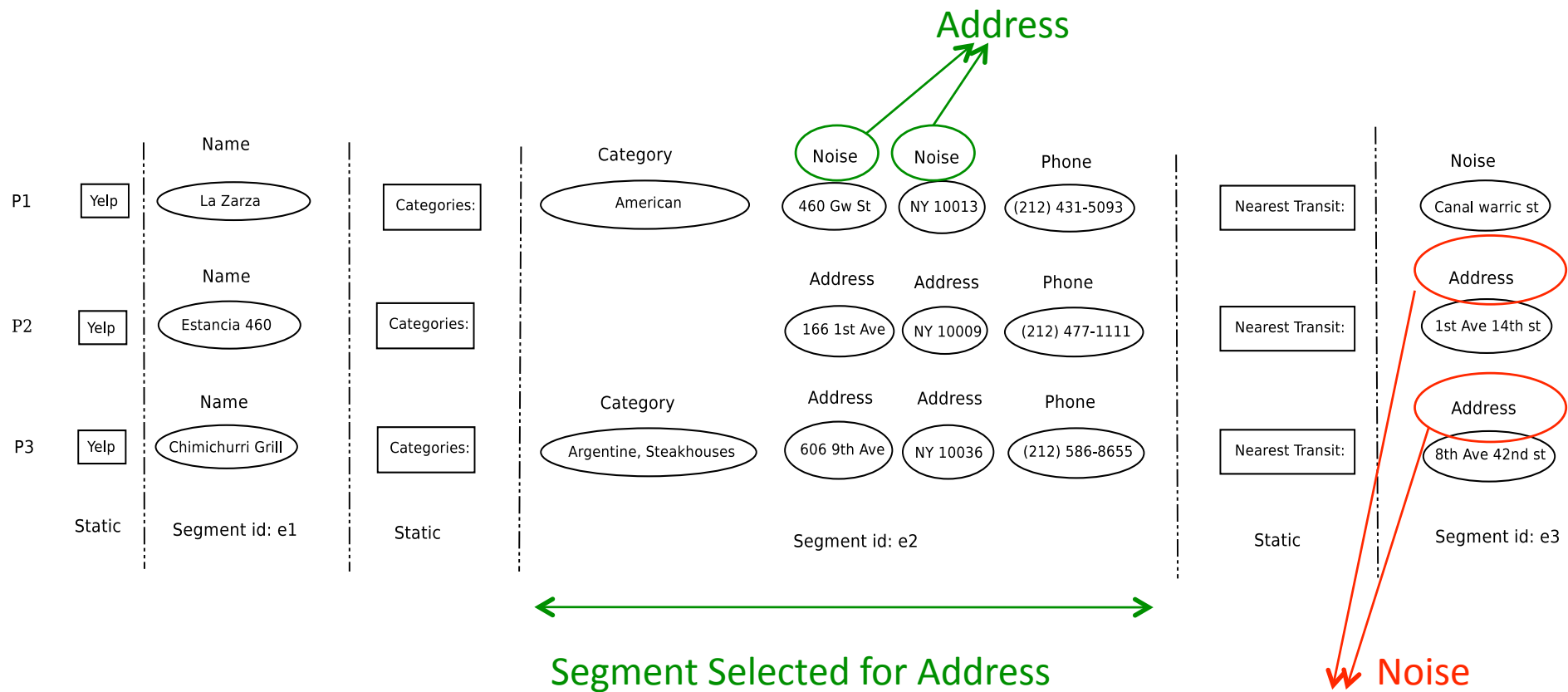


Correct Node Labels

- Inter-page Constraint (Site Knowledge)
 - Same Template: Since pages are script generated, they follow same template
- Label Variations across same segment will be minor and primarily due to
 - Missing or additional nodes in certain segments
 - Incorrectly labeled nodes in some segments
- Intuition: If CRF model assign correct labels in majority of the cases then applying “wisdom of crowd” helps to correct labels



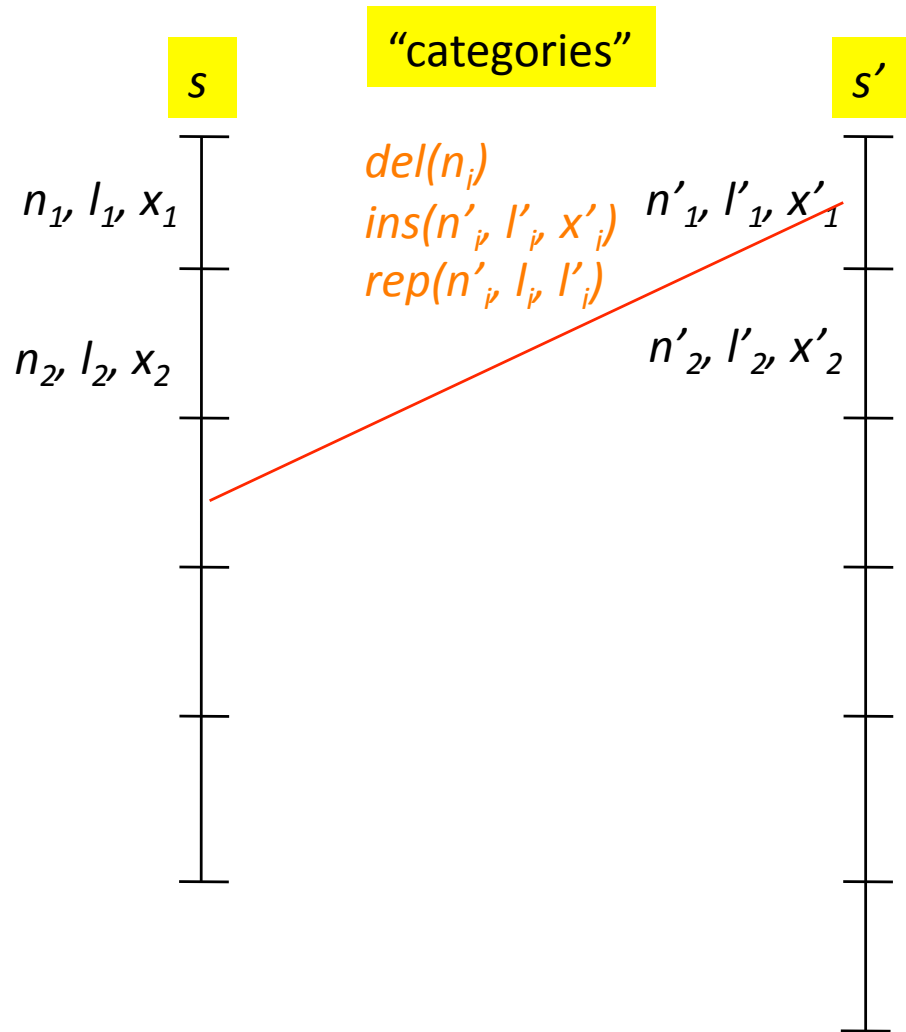
Correct Node Labels



Choose the majority label? Segment alignment is needed.



Correct Node Labels – Node Alignment



1. Find the min cost edit operation sequence with every other sequence with the same id.
2. For each node, choose the majority operation.
3. If the selected operation is *replace*, then the label of the node is changed.



Extensions

- Attributes Spanning Segments
 - Cluster the segments
 - Select cluster whose average weight is minimum
- Missing Static Nodes
 - Insert Static node at appropriate position using Edit Distance



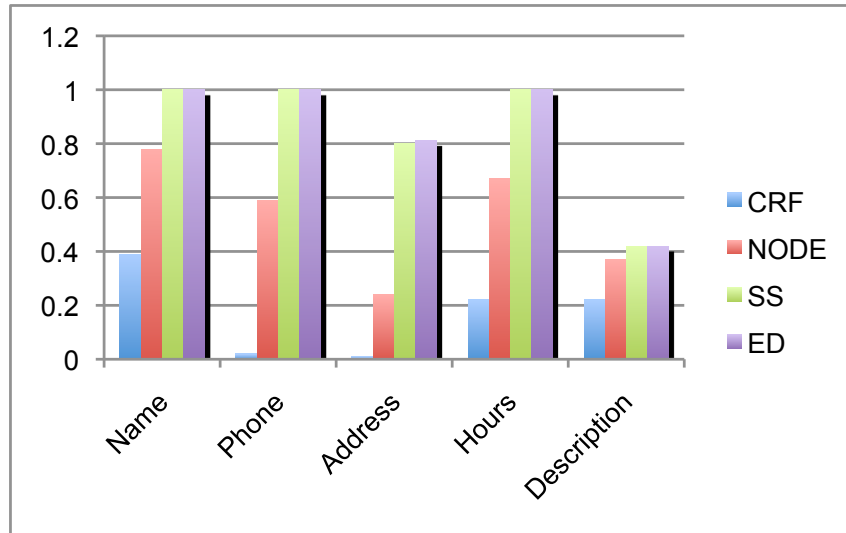
Experiments

- Dataset
 - 5 restaurant sites, ~ 100 pages from each site
 - Attribute: Name, Address, Phone, Hours, Description
 - Attribute order: NAPHD, NHAPD, NAPDH, NAPH
- Features
 - Lexicon, Regex, Node-level
- Experiment
 - Learn on four sites, Label the fifth
- Baselines
 - Full-page CRF
 - HCRF

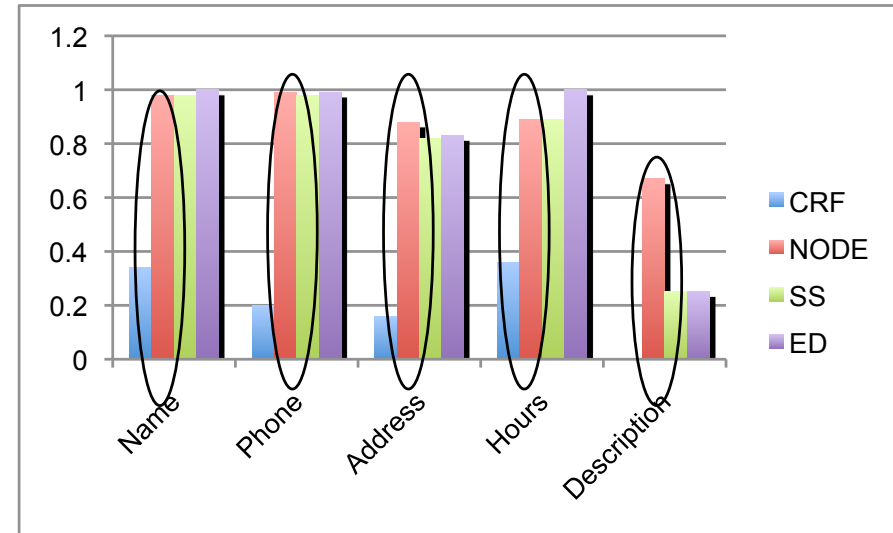


Results

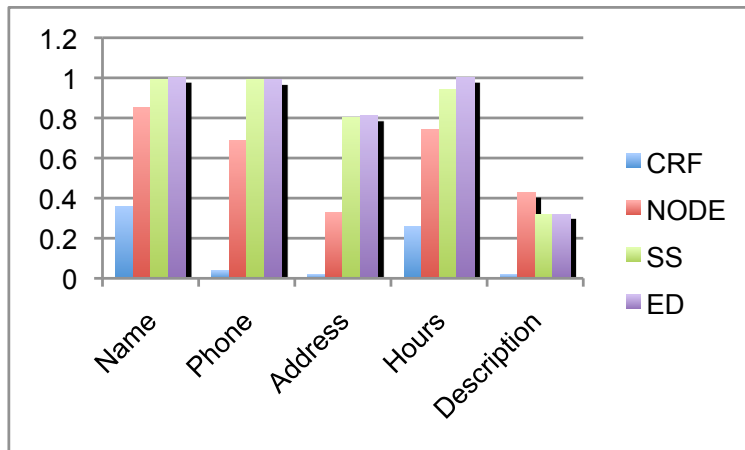
Precision



Recall



F1



HCRF

1. Memory and compute-intensive.
2. On a subset of data, the F1 was 0.262 compared to 0.5271 for the proposed approach.



Conclusions

- Unsupervised extraction is a challenging problem.
- The framework proposed in this paper, leverages site-knowledge to boost the precision of underlying extraction schemes.
- When applied to CRF-based extractors, the proposed method boosts both precision and recall.



Questions?

shs@yahoo-inc.com