High Precision Web Extraction using Site Knowledge

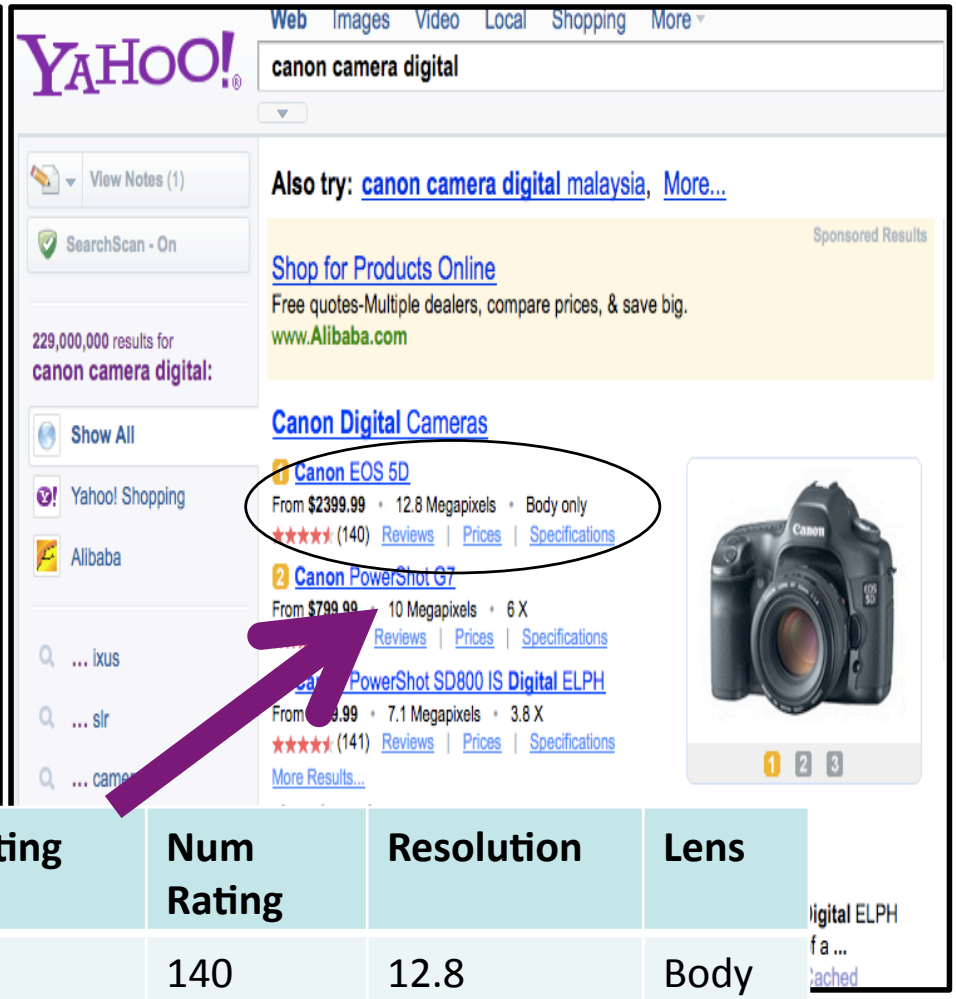
Meghana Kshirsagar
Rajeev Rastogi
Sandeepkumar Satpal
Srinivasan H Sengamedu
Venu Satuluri

YAHOO! Labs Bangalore

Outline

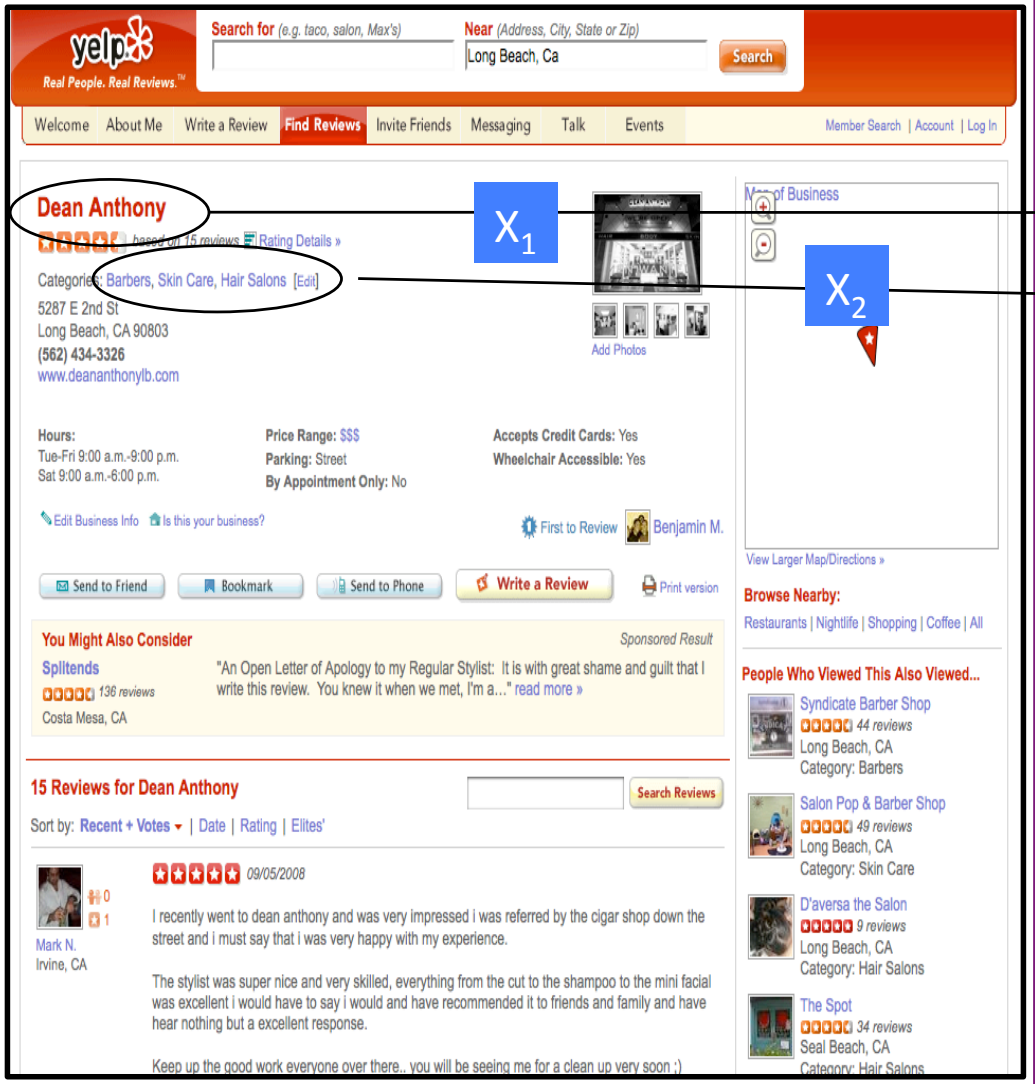
- Motivation
- Problem Definition
- Proposed Approach
  - Site-Knowledge
  - Segmentation
  - Segment Label Selection
  - Node Label Correction
  - Extensions
- Experimental Results
- Conclusions

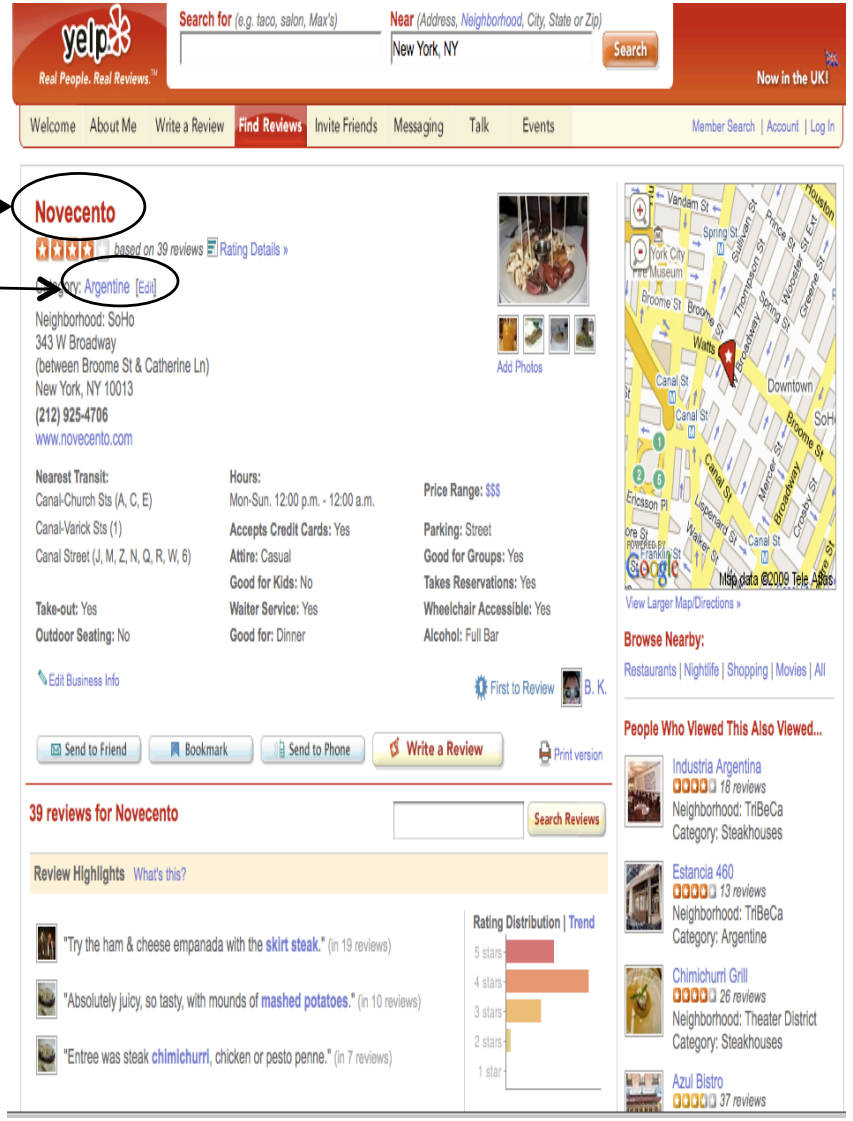
Information Extraction: What & Why?



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

Approaches to Extraction: Wrapper

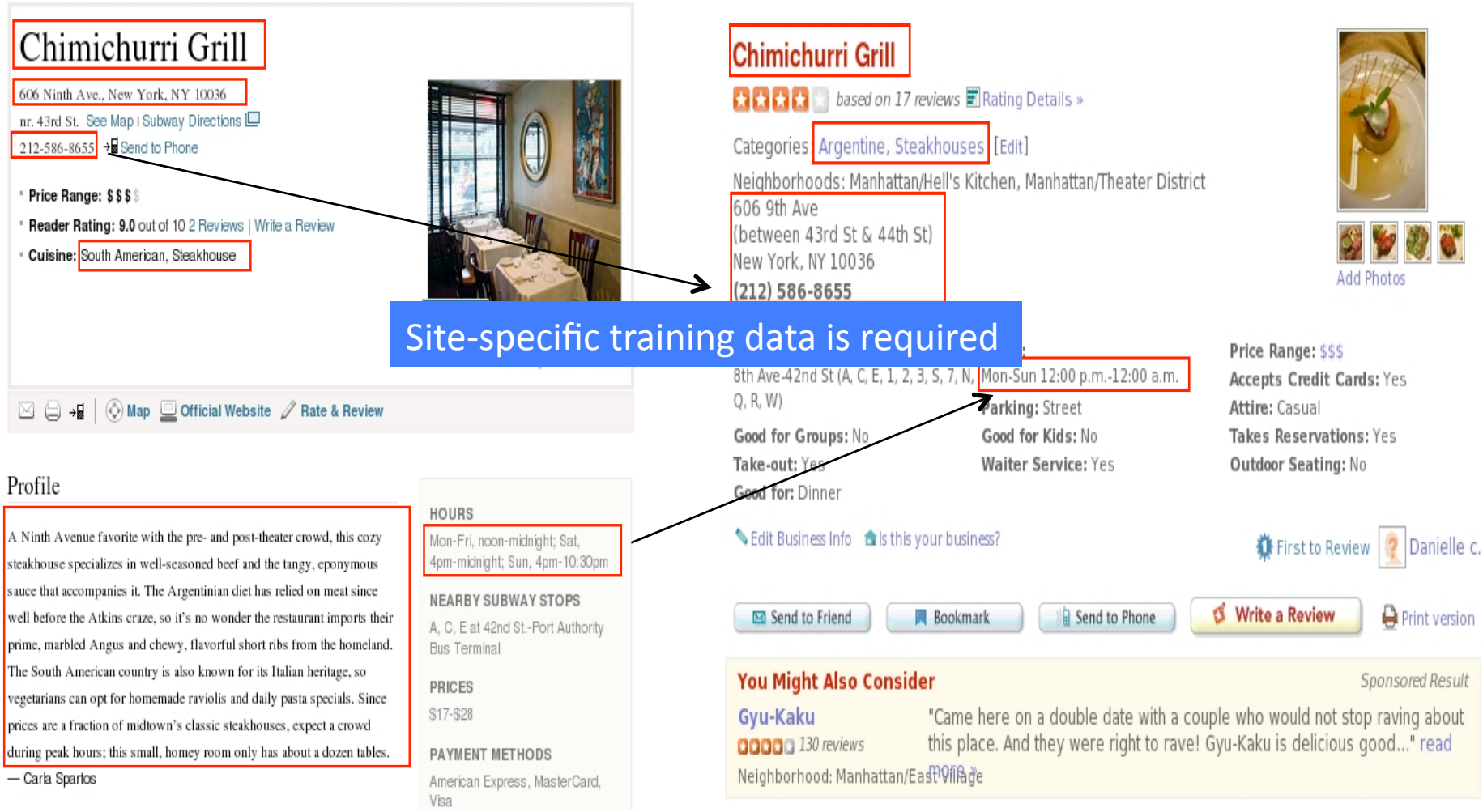




Structural Changes

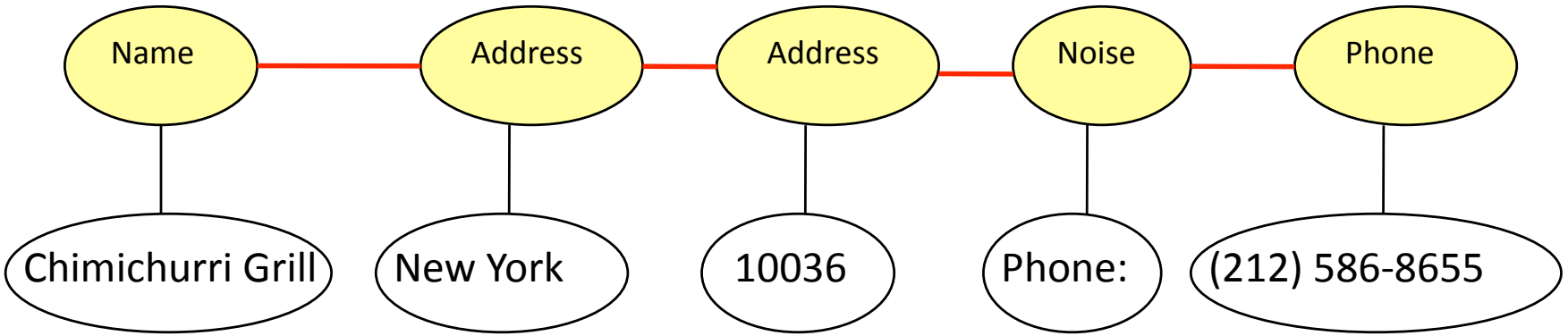
Nymag.com

Yelp.com



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

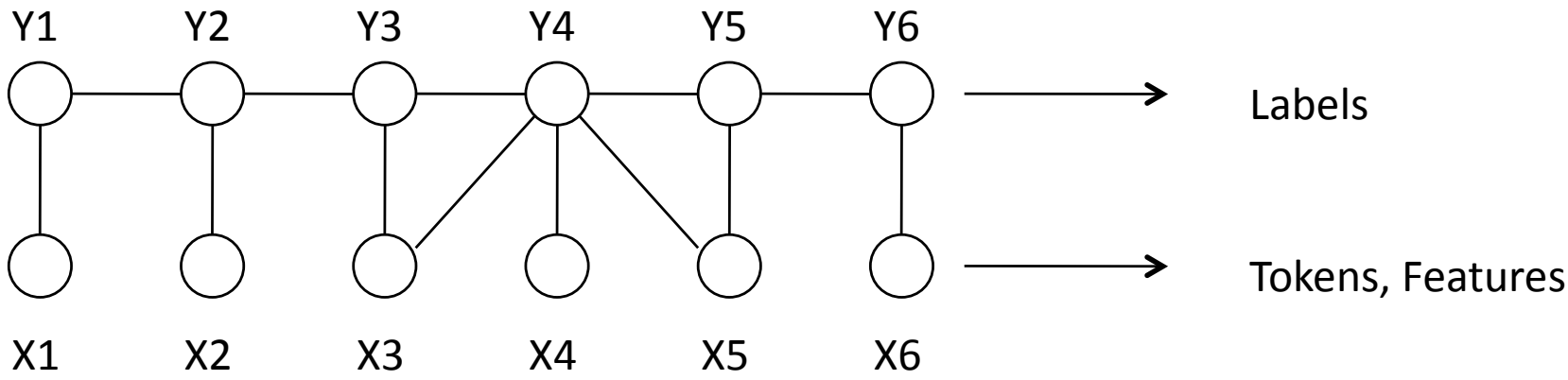
isAllCapsWord
hasTwoContinuousCaps
isDay
1-2digitNumber
3digitNumber
4digitNumber
5digitNumber
>5digitNumber
dashBetweenDigits
isAlpha
isNumber

Node-level features

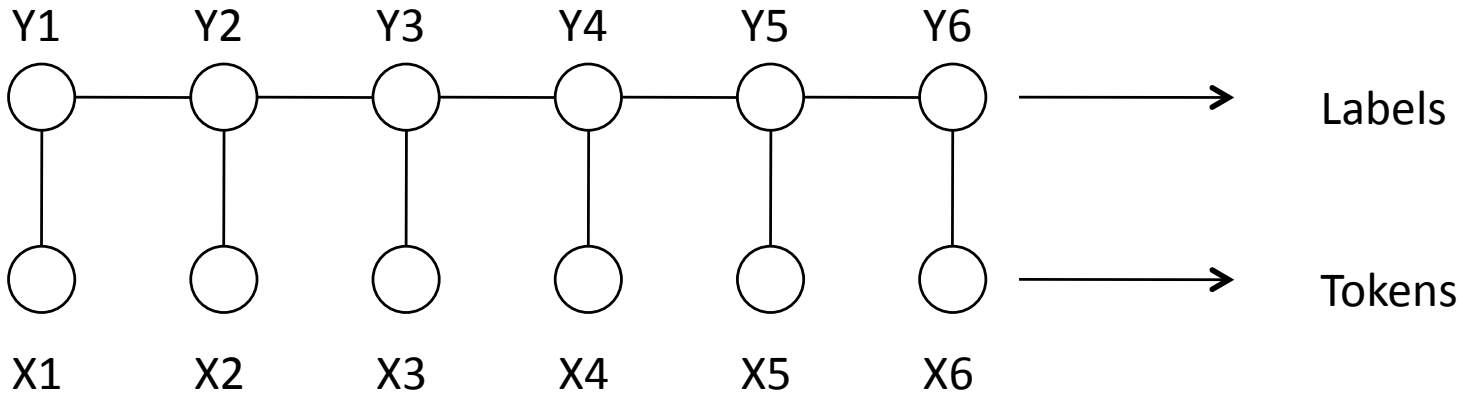
noOfWords>20
noOfWords>50
noOfWords>100
propOfTitleCase<0.2
propOfTitleCase>0.8
overlapWithPageTitle
prefixOverlapWithPageTitle

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]*, Phone)
  - f(New York, Address)
- Conditional random field is a log-linear function over these features

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

Approaches to Extraction: Summary

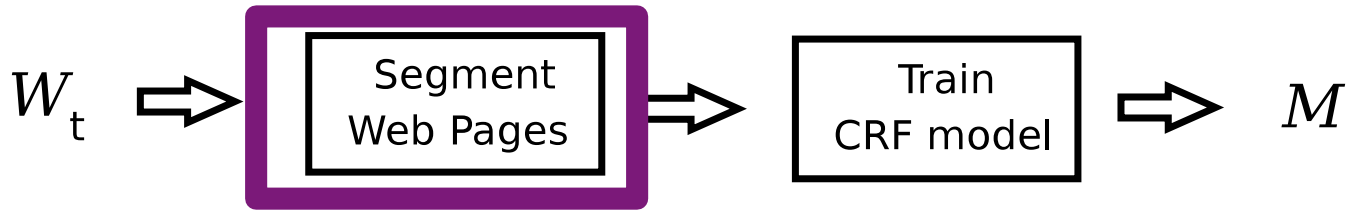
- Wrapper
  - High Precision (> 99%)
  - Large editorial requirement

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

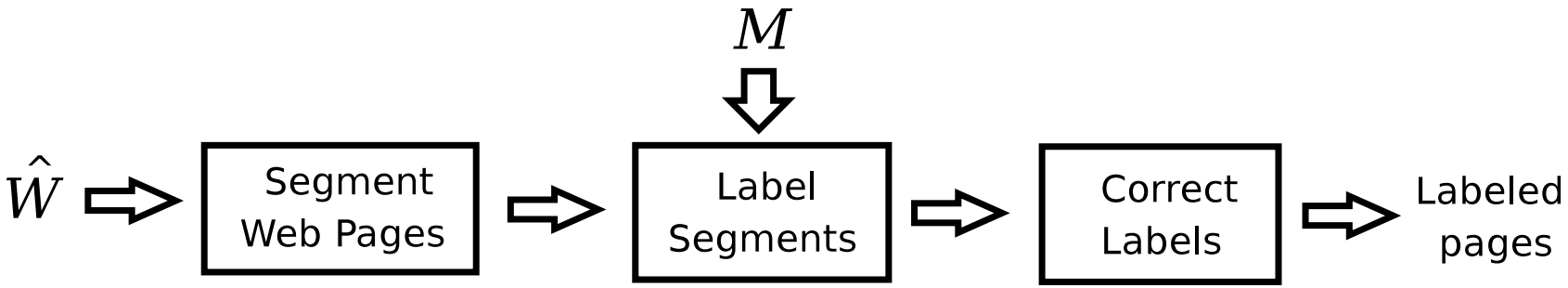
Problem Definition

- 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

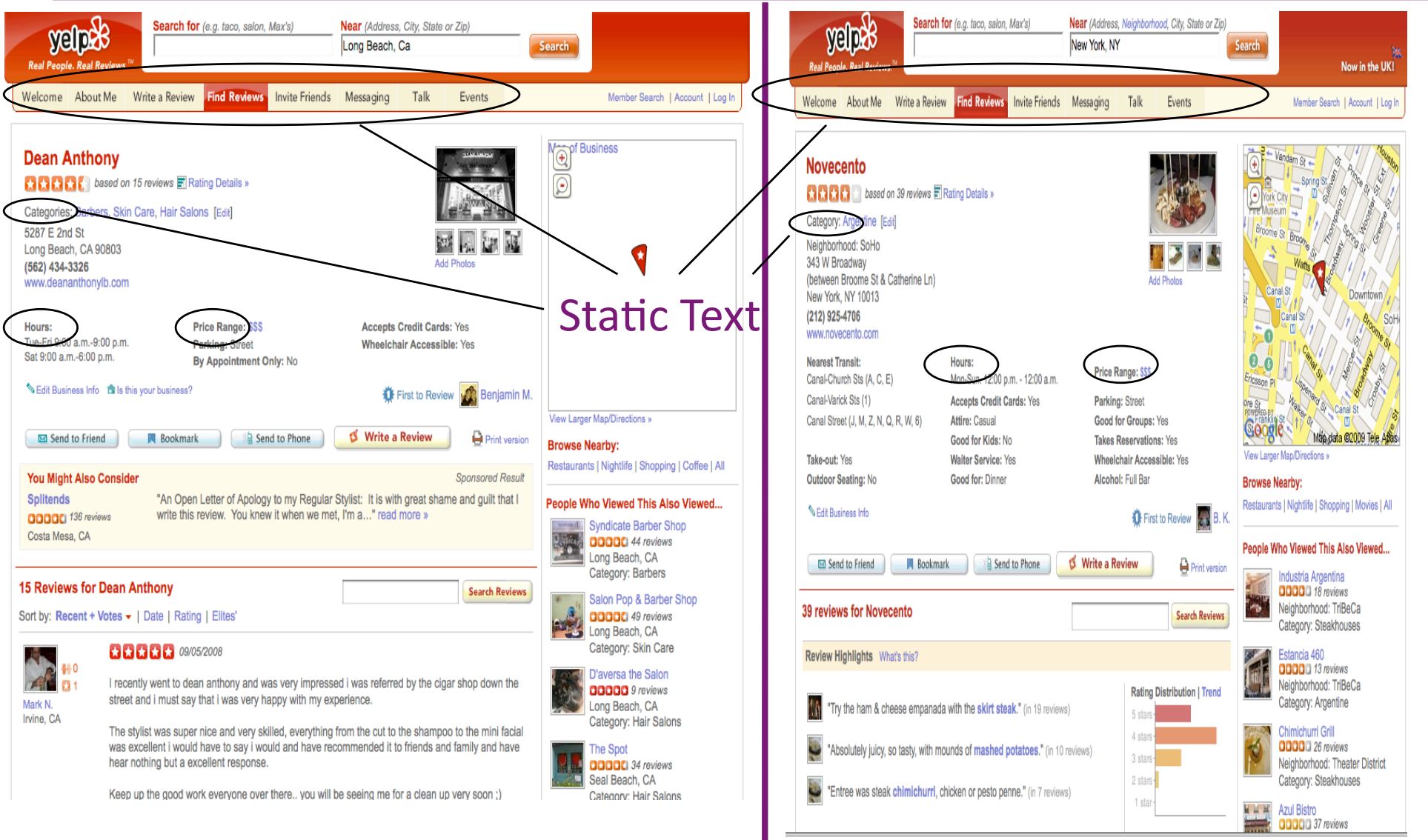


(a) Training Phase

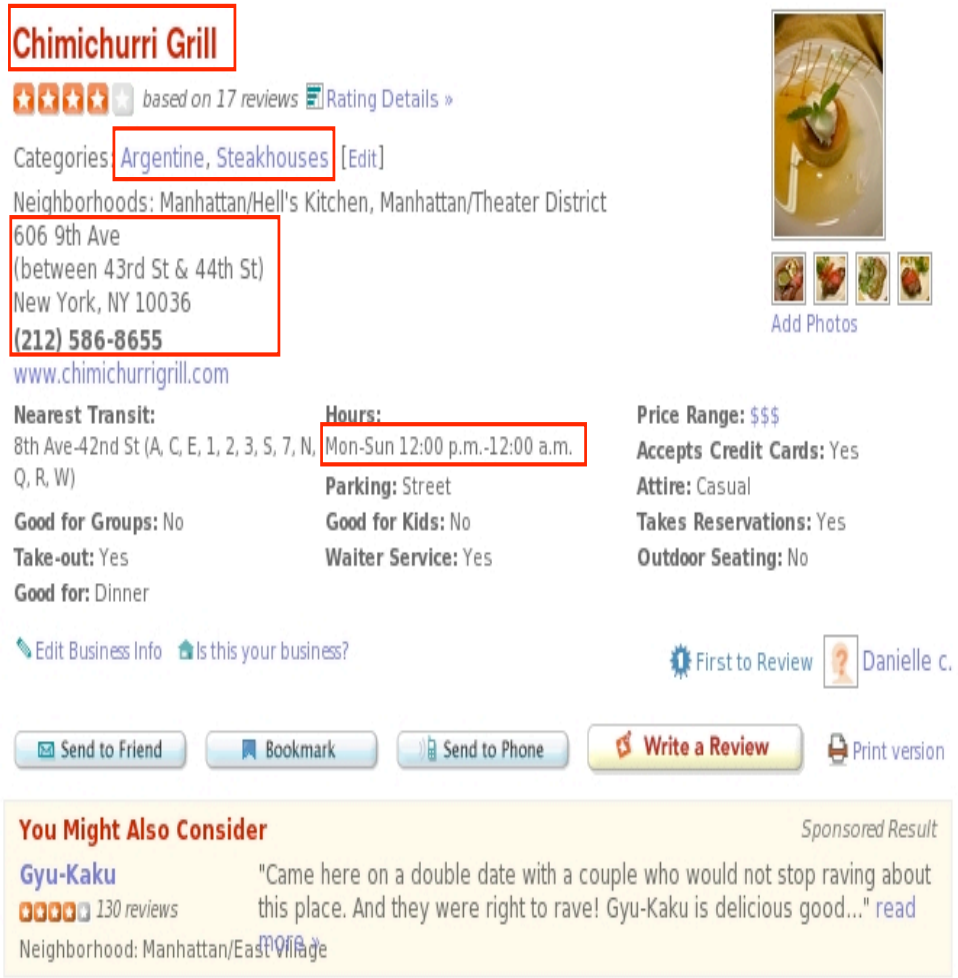


(b) Labeling Phase

Static Text in Scripted Pages



Segmentation



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

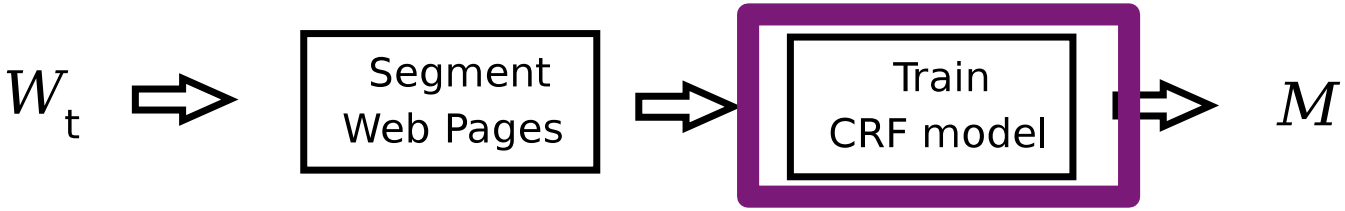
Segmented Sequence

- [Chimichuri Gril]
- [based on 17 reviews]
- { Rating details }
- { Categories }
- [Steakhouses, Argentine]
- { Neighbourhoods }
- [Theatre district, Kitchen]
- [603 9th Ave]
- [.....]
- [(212) 586-8655]
- [ww.chimichurigril.com]
- { Nearest Transit: }
- [8th Ave .....]
- [.....]

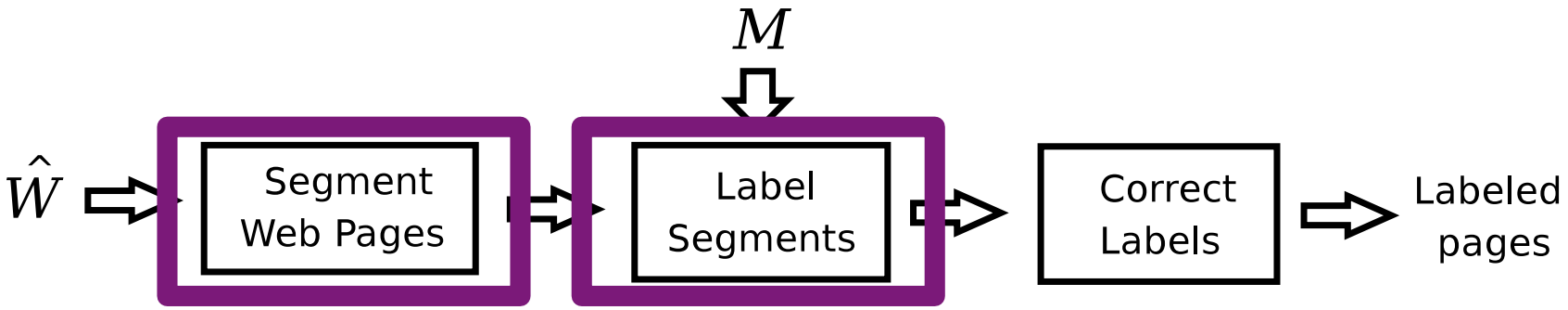
Benefits of Static Text and Segmentation

- 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

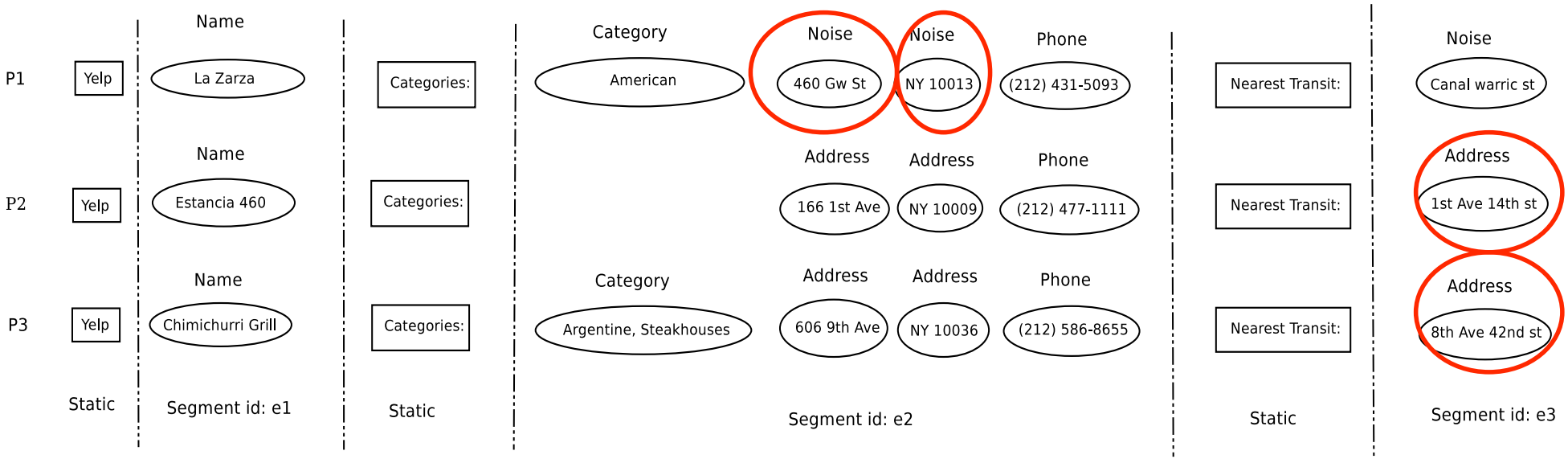


(a) Training Phase



(b) Labeling Phase

CRF Labeling



Identify attribute labels at segment level

seg("address") = *e2*

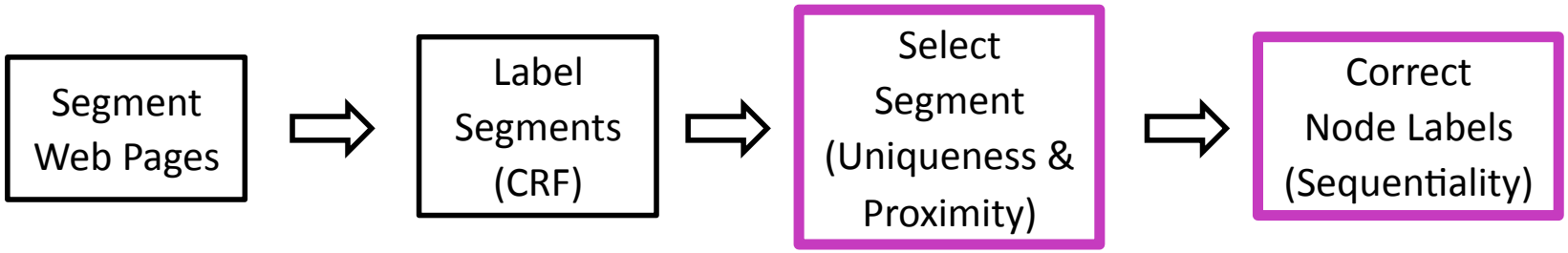
Use *Attribute Uniqueness & Proximity*

Fix node labels

"Noise" -> "Address" in Segment *e2*

Use *Sequentiality*

Label Correction

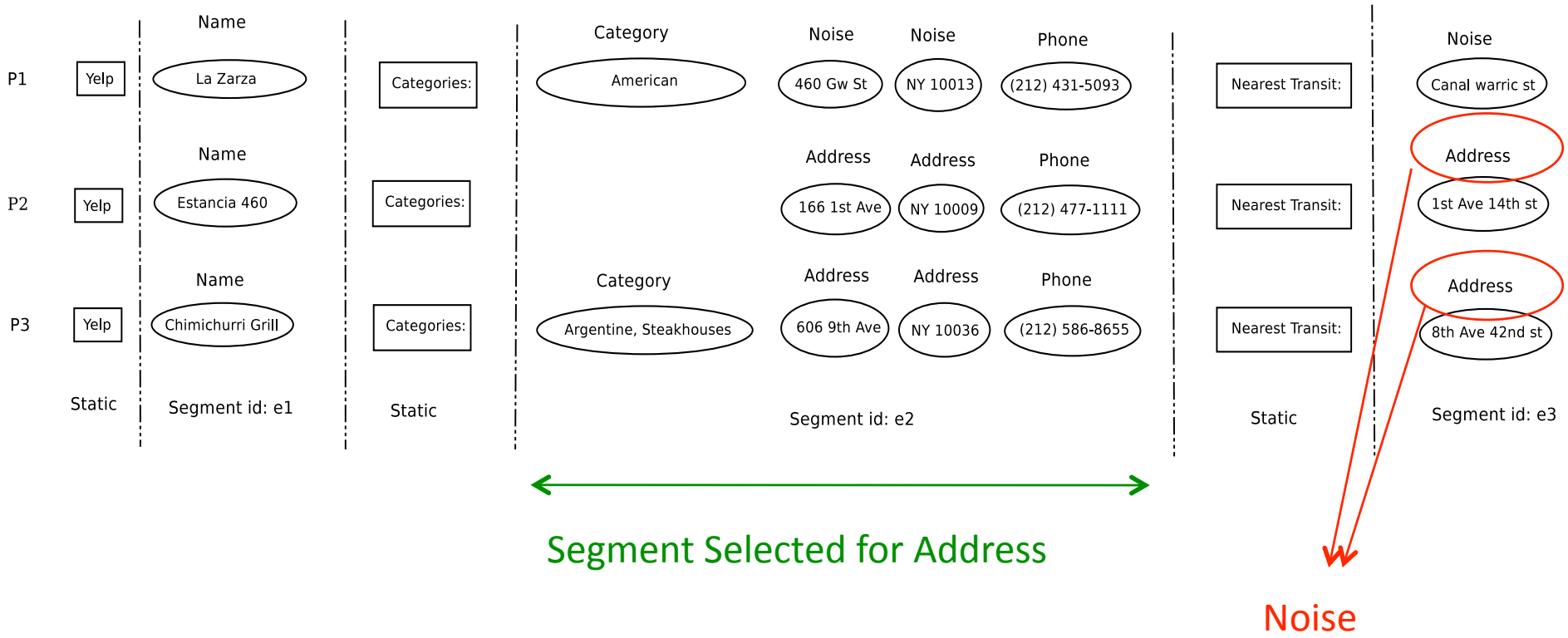


Segment Selection

- 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

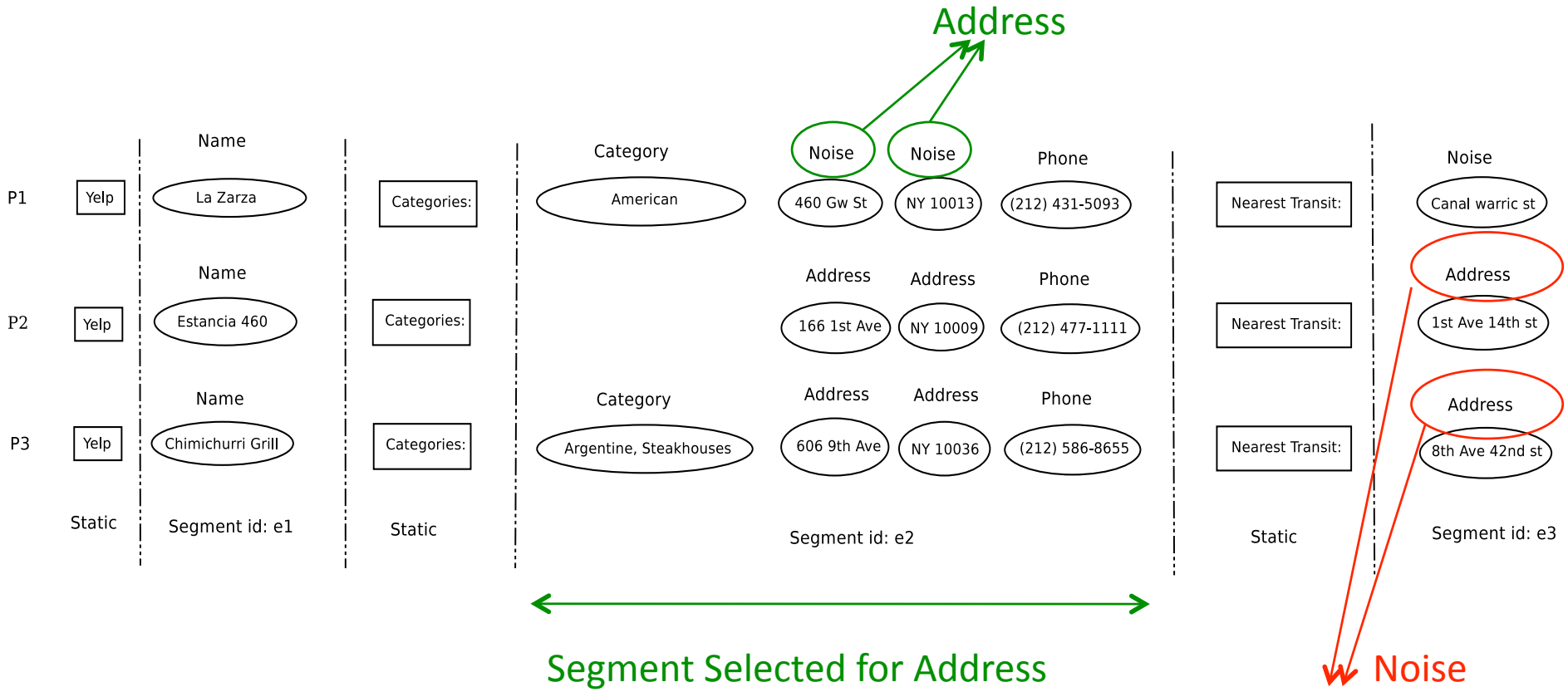
$$w_e = \sum_f dist(e, f) \cdot |attr(f)|;$$

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

Correct Node Labels

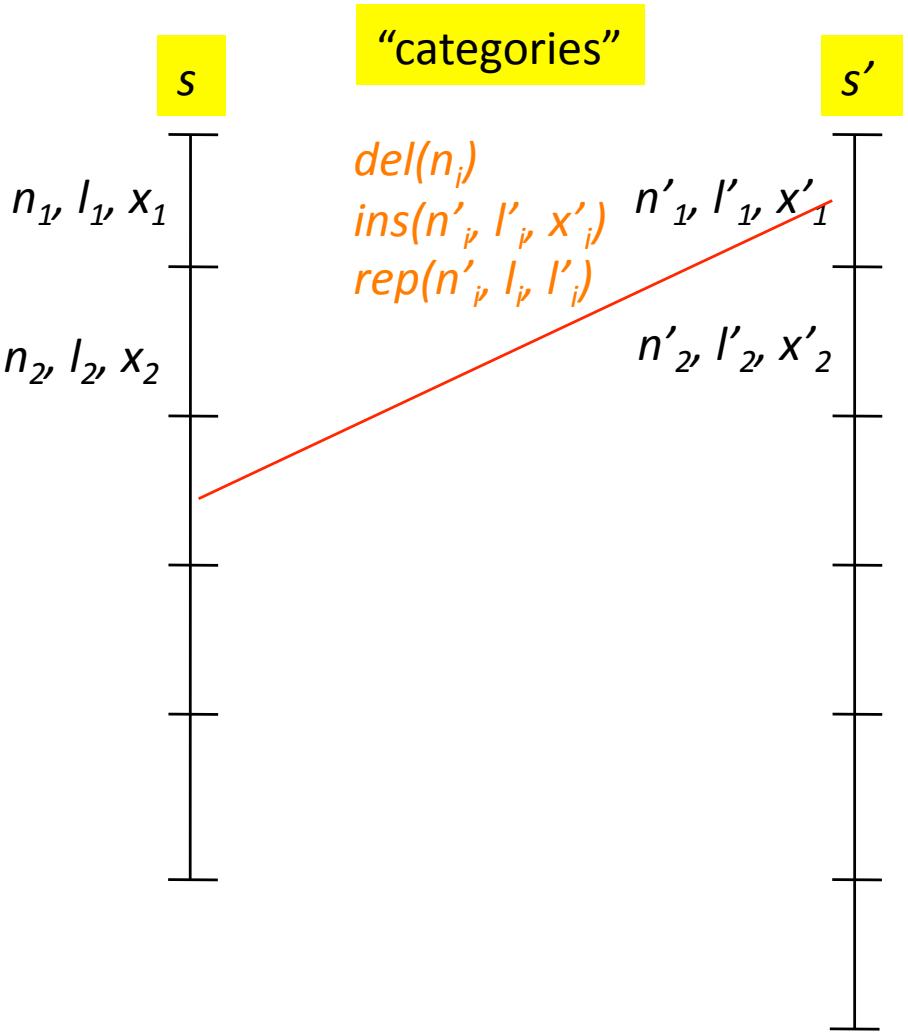
- Inter-page Constraint (Site Knowledge)
  - Same Template: Since pages are script generated, they follow same template
- Label Variations across same segment will be minor and primarily due to
  - Missing or additional nodes in certain segments
  - Incorrectly labeled nodes in some segments
- Intuition: 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

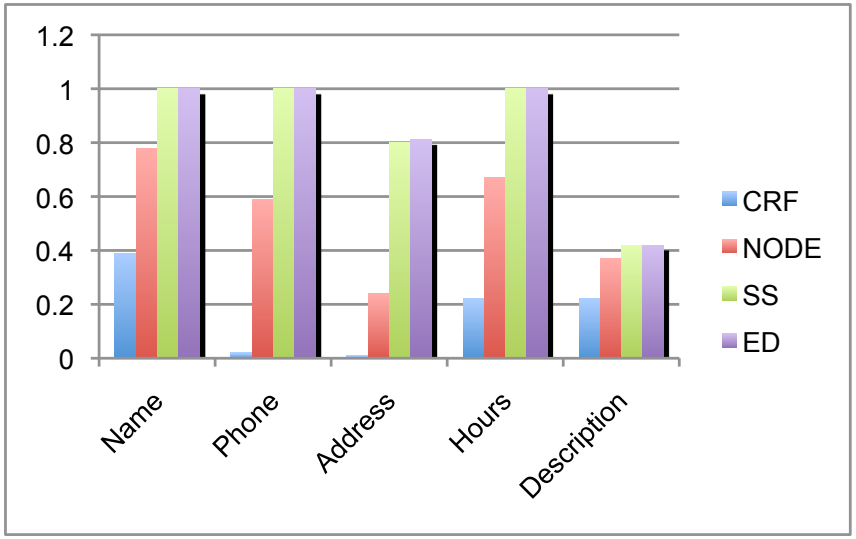
- Attributes Spanning Segments
  - Cluster the segments
  - Select cluster whose average weight is minimum
- Missing Static Nodes
  - Insert Static node at appropriate position using Edit Distance

Experiments

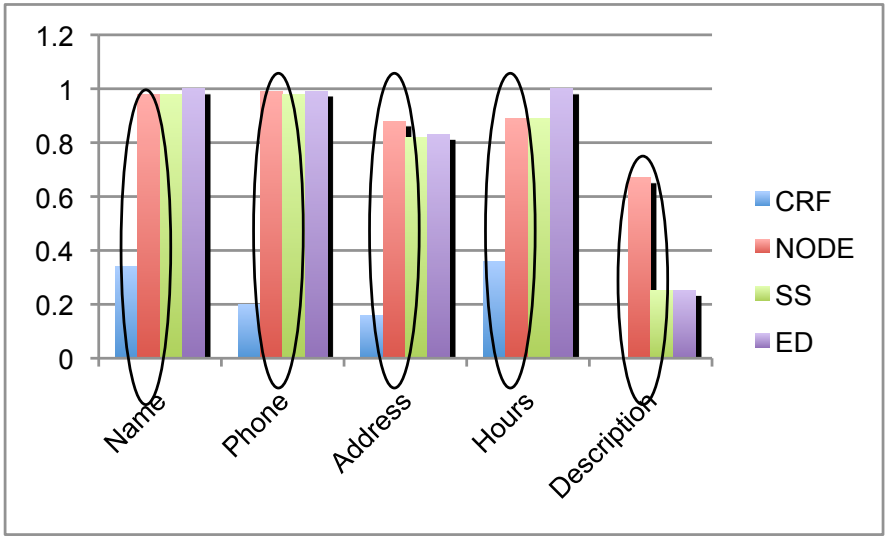
- Dataset
  - 5 restaurant sites, ~ 100 pages from each site
  - Attribute: Name, Address, Phone, Hours, Description
  - Attribute order: NAPHD, NHAPD, NAPDH, NAPH
- Features
  - Lexicon, Regex, Node-level
- Experiment
  - Learn on four sites, Label the fifth
- Baselines
  - Full-page CRF
  - HCRF

Results

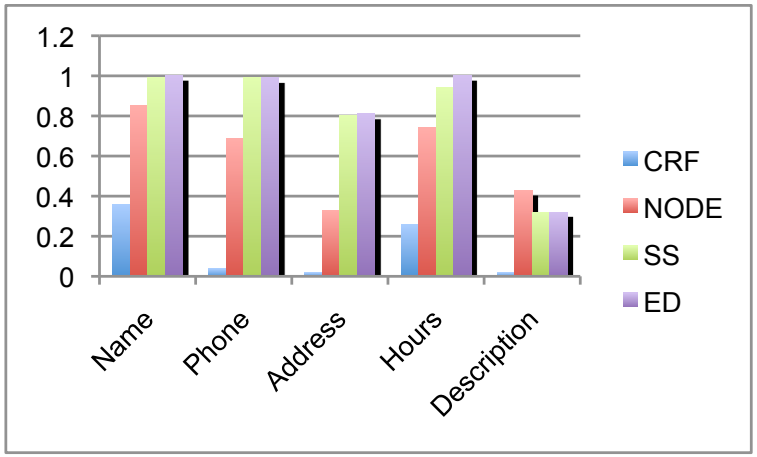
Precision



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