A Reference-Set Approach to Information Extraction from Unstructured, Ungrammatical Data Sources

Matthew Michelson
Ph.D. Defense
Nov. 3\textsuperscript{rd}, 2008
Motivation: Data Integration

Query: Average price for a 3-star crash-rated Honda, and reviews.
Motivation: Data Integration

**Query:** Average price for a 3-star crash-rated Honda, and reviews.

- Structured Sources
  - NHTSA Ratings
- Semi-Structured Sources
  - Car Review
- Unstructured, Ungrammatical Sources
  - Classified ads, Auction listings, Etc.

Diagram with flow of data from User Query through Mediator to integrate with different types of sources.
Unstructured, Ungrammatical Data: “Posts”
Unstructured, Ungrammatical Data: “Posts”

Fri Mar 14

91 Civic SI RHD SHELL - $2900 - (West Covina) pic

2001 Automatic Mazda Millenia Clear Title - $3800 - pic

1984 Ford Tow Truck - $10000 - (Bell)

2004 Audi A4 1.8T - $6800 - pic

1998 International 4700 Tow Truck - $12000 - (Bell)

1994 LEXUS ES 300 >> LEATHER INTERIOR <<< - $3000 - (RESEDA) pic

1987 Chevrolet Tahoe 4x4 just smogged - $1400 - (Palmdale) pic
Query? ...
Information Extraction/Annotation!

MAKE: HONDA (implied!)
MODEL: CIVIC
TRIM: 2 Door SI
YEAR: 1991

[ ALERT - offers to ship cars/trucks are
[ avoid recalled items ] [ success s

Fri Mar 14

91 Civic SI RHD SHELL - $2900 - (West Covina) pic

2001 Automatic Mazda Millenia Clear Title - $3800 - pic

1984 Ford Tow Truck - $10000 - (Bell)
Difficulties

- Unstructured
  - No assumptions on structure
  - “Rule/Pattern” based techniques unsuited
- Ungrammatical
  - Does not conform to English grammar
  - Natural-Language Processing techniques unsuited
Reference-Set Based Extraction/Annotation

91 Civic SI RHD SHELL - $2900 -

Reference Set (s) ➔ Record Linkage ➔ Information Extraction

Annotation

| HONDA | CIVIC | 2 Door SI | 1991 |

Extracted Attributes

| Civic | SI | 91 | $2900 |

Query ➔ Integrate

Introduction • Unsupervised IE • Building Reference Sets • Supervised IE • Conclusion
Reference Sets

- Collections of entities and their attributes
  - List cars -> <make, model, trim, ...>

Scrape make, model, trim, year for all cars from 1990-2005...
Contributions

- Automatic matching and extraction algorithm that exploits a given reference set
  - Automatically select the appropriate reference sets from a repository of reference sets
- Automatic method for building reference sets from the posts themselves
  - Suggest the number of posts required to sufficiently build reference set
  - Algorithm to determine whether automatic method will work, or user should create reference set
- Supervised machine learning for high-accuracy
  - High accuracy, even in the face of ambiguity
## Contributions

3 reference-set based extraction methods

<table>
<thead>
<tr>
<th>Method 1 (ARX) [IJDAR 07]</th>
<th><strong>Summary</strong></th>
<th><strong>Advantages</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Automatically select reference set from repository</td>
<td>● State-of-the-art extraction</td>
</tr>
<tr>
<td></td>
<td>2. Automatic extraction</td>
<td>● Automatic, given reference set</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method 2 (ILA) [JAIR, review]</th>
<th><strong>Summary</strong></th>
<th><strong>Advantages</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Automatically build reference set</td>
<td>● Cannot build reference set (difficult attributes)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Fully automatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Competitive state-of-the-art</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method 3 (Phoebus) [JAIR, 08]</th>
<th><strong>Summary</strong></th>
<th><strong>Advantages</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Supervised approach to extraction</td>
<td>● Highest-accuracy extraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Deals with ambiguity</td>
</tr>
</tbody>
</table>
Automatic method: Three steps

1) Select reference set(s)
2) Find best matches (unsupervised)
3) Extraction using matches (unsupervised)

ARX: Automatic Reference-set based eXtraction
Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts…

FORD Thunderbird - $4700

2001 White Toyota Corrolla CE Excellent Condition - $8200

SIM:0.7

Cars  Hotels  Restaurants
Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts…

FORD Thunderbird - $4700
2001 White Toyota Corrolla CE Excellent Condition - $8200

SIM:0.7 SIM:0.4

Cars Hotels Restaurants
Selecting the Reference Set(s)

Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts...

FORD Thunderbird - $4700

2001 White Toyota Corolla CE Excellent Condition - $8200

SIM:0.7  SIM:0.4  SIM:0.3

Cars  Hotels  Restaurants
Selecting the Reference Set(s)
Vector space model: set of posts are 1 doc, reference sets are 1 doc

Select reference set most similar to the set of posts...

FORD Thunderbird - $4700
2001 White Toyota Corrolla CE Excellent Condition - $8200

Cars 0.7  \( PD(C,H) = 0.75 > T \)
Hotels 0.4  \( PD(H,R) = 0.33 < T \)
Restaurants 0.3
Avg. 0.47
Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!
Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}
{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}

{NISSAN, ALTIMA, 2007}
Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!
Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}
{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}

Vector-based matching

{NISSAN, ALTIMA, 2007}
Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!
Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}
{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}
{BMW, M3, 2 Dr STD Convertible, 2002}

{LINCOLN, TOWN CAR, 4 Dr, 2001}

{RENAULT, LE CAR, 2 Dr, 1987}
Unsupervised matching between the posts and reference set

new 2007 altima

02 M3 Convertible .. Absolute beauty!!!
Awesome car for sale! Cheap too!

{NISSAN, ALTIMA, 4 Dr 3.5 SE Sedan, 2007}
{NISSAN, ALTIMA, 4 Dr 2.5 S Sedan, 2007}
{NISSAN, 2007}
{BMW, M3, 2 Dr STD Convertible, 2002}
{LINCOLN, TOWN CAR, 4 Dr, 2001}
{RENAULT, LE CAR, 2 Dr, 1987}

Vector-based matching

Prune false positives!
Unsupervised Extraction

91 Civic SI RHD SHELL - $2900 -

make model trim year
Civic SI 91

Clean Whole Attribute
Results: Information Extraction

- State-of-the-art comparison
  1. Conditional Random Field (structure)
     1. CRF-Orth
        - Orthographic features: cap, start-num, etc.
     2. CRF-Win
        - CRF-Orth + 2-word sliding window
          - more structure!
  2. Amilcare
     - NLP
     - “Gazetteers” (list of hotels, etc.)
- ARX = automatic, others = supervised
- Field-level extractions
  - All tokens required, no extras (strict!)
## Results: Information Extraction

### Craigs Cars Posts (Craigslist)

<table>
<thead>
<tr>
<th></th>
<th>ARX</th>
<th>CRF-Orth</th>
<th>CRF-Win</th>
<th>Amilcare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make</td>
<td>97.95</td>
<td>83.66</td>
<td>78.67</td>
<td>94.57</td>
</tr>
<tr>
<td>Model</td>
<td>88.61</td>
<td>74.25</td>
<td>68.72</td>
<td>81.24</td>
</tr>
<tr>
<td>Trim</td>
<td>49.70</td>
<td>47.88</td>
<td>38.75</td>
<td>35.94</td>
</tr>
<tr>
<td>Year</td>
<td>86.47</td>
<td>88.04</td>
<td>84.52</td>
<td>88.97</td>
</tr>
</tbody>
</table>

~27,000 cars: Edmunds/ Super Lamb Auto

### BFT Posts (biddingfortravel.com)

<table>
<thead>
<tr>
<th></th>
<th>ARX</th>
<th>CRF-Orth</th>
<th>CRF-Win</th>
<th>Amilcare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Rating</td>
<td>91.03</td>
<td>94.77</td>
<td>94.21</td>
<td>96.46</td>
</tr>
<tr>
<td>Hotel Name</td>
<td>73.46</td>
<td>67.47</td>
<td>41.33</td>
<td>62.91</td>
</tr>
<tr>
<td>Local Area</td>
<td>71.98</td>
<td>70.19</td>
<td>33.07</td>
<td>68.01</td>
</tr>
</tbody>
</table>

~130 hotels: BiddingForTravel.com

**Automatic, state-of-the-art extraction on posts**

- **ARX**
  - Automatic & better than supervised on 5/7 attributes
  - Cases where ARX underperforms
    - w/in 5%
    - Strong numeric component
  - Recall issue

- **CRF-Win**
  - Worst on 6/7
  - Can’t rely on structure!
Automatic construction of reference sets

- What if there isn’t already a reference set?

- What about coverage?

<table>
<thead>
<tr>
<th>Ford</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dodge</td>
<td>Caravan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HP Pavillion DV2000 laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gateway ML6230, Intel Cel</td>
</tr>
</tbody>
</table>

| ACURA TL 3.2 VTEC - 1999 |
Automatic construction of reference sets

- What if there isn’t already a reference set?
  - HP Pavilion DV2000 laptop
  - Gateway ML6230, Intel Cel ...

- What about coverage?

<table>
<thead>
<tr>
<th>Ford</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dodge</td>
<td>Caravan</td>
</tr>
</tbody>
</table>

1) Select reference set(s)
2) Automatic matching
3) Automatic extraction using matches
Automatic construction of reference sets

- What if there isn’t already a reference set?
  - HP Pavillion DV2000 laptop
  - Gateway ML6230, Intel Cel …

- What about coverage?
  - Ford Focus
  - Dodge Caravan
  - ?
  - ACURA TL 3.2 VTEC - 1999

1) Automatically build reference set
2) Automatic matching
3) Automatic extraction using matches
Build reference sets from posts

JAIR, review

Step 1
Construct Bi-Grams

Step 2
Create hierarchies

Form reference set

91 Civic SI RHD ...
{91 Civic}
{Civic SI}
{SI RHD}
...
Constructing entity hierarchies

- Sanderson & Croft heuristic
  - \( x \text{ SUBSUMES } y \text{ IF } P(x|y) \geq 0.75 \) \& \( P(y|x) \leq P(x|y) \)

- Merge heuristic
  - \( \text{MERGE}(x,y) \text{ IF } x \text{ SUBSUMES } y \) \& \( P(y|x) \geq 0.75 \)
Constructing entity hierarchies

- Sanderson & Croft heuristic
  - $x \text{ SUBSUMES } y$ \text{ IF } $P(x|y) \geq 0.75$ \& $P(y|x) \leq P(x|y)$

- Merge heuristic
  - $\text{MERGE}(x,y)$ \text{ IF } $x \text{ SUBSUMES } y$ \& $P(y|x) \geq 0.75$

Honda civic is cool
Honda civic is nice
Honda accord rules
Honda accord 4 u!

$P(\text{Honda}|\text{civic}) = 2/2 = 1$
$P(\text{civic}|\text{Honda}) = 2/4 = 0.5 \rightarrow \text{SUBSUME, not MERGE}$
Constructing entity hierarchies

- Sanderson & Croft heuristic
  - \( x \) **SUBSUMES** \( y \) **IF** \( P(x|y) \geq 0.75 \) & \( P(y|x) \leq P(x|y) \)

- Merge heuristic
  - **MERGE**(\( x \), \( y \)) **IF** \( x \) **SUBSUMES** \( y \) & \( P(y|x) \geq 0.75 \)

Honda civic is cool
Honda civic is nice
Honda accord rules
Honda accord 4 u!

\[
P(\text{Honda}|\text{civic}) = \frac{2}{2} = 1
\]
\[
P(\text{civic}|\text{Honda}) = \frac{2}{4} = 0.5 \rightarrow \text{SUBSUME, not MERGE}
\]

- Construct hierarchies, then flatten

<table>
<thead>
<tr>
<th>Honda</th>
<th>CIVIC</th>
<th>ACCORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HONDA</td>
<td>CIVIC</td>
<td>ACCORD</td>
</tr>
</tbody>
</table>

Introduction ● Unsupervised IE ● Building Reference Sets ● Supervised IE ● Conclusion
Construction issues

- \{a, y\}, \{b, y\}, \{c, y\} \rightarrow y \text{ is “general token”}
  - Instead use \(P(\{a \cup b \cup c\} | y)\)
  - e.g. car trims: Pathfinder LE, Corolla LE, …

- How many posts are enough?
- Lock attributes (tree levels)
  - Lock out noise
  - Need only enough posts until lock all levels

Key: redundancy. At some point you’ve gotten all you can from the posts
## Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set

(ARX for extraction)

<table>
<thead>
<tr>
<th>Craig’s Cars: 4,400 posts</th>
<th>Make</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>78.19</td>
<td>84.52</td>
<td>81.23</td>
<td></td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>92.51</td>
<td>99.52</td>
<td>95.68</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Craig’s Cars: 4,400 posts</th>
<th>Model</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>64.25</td>
<td>82.79</td>
<td>72.35</td>
<td></td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>79.50</td>
<td>91.86</td>
<td>85.23</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Craig’s Cars: 4,400 posts</th>
<th>Trim</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>23.45</td>
<td>52.17</td>
<td>32.35</td>
<td></td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>38.01</td>
<td>63.69</td>
<td>47.61</td>
<td></td>
</tr>
</tbody>
</table>
## Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set

(ARX for extraction)

<table>
<thead>
<tr>
<th></th>
<th>Craig’s Cars: 4,400 posts</th>
<th>Make</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ILA (580)</strong></td>
<td></td>
<td>78.19</td>
<td>84.52</td>
<td></td>
<td>81.23</td>
</tr>
<tr>
<td><strong>Edmunds (27,006)</strong></td>
<td></td>
<td>92.51</td>
<td>99.52</td>
<td></td>
<td>95.68</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ILA (580)</strong></td>
<td></td>
<td>64.25</td>
<td>82.79</td>
<td></td>
<td>72.35</td>
</tr>
<tr>
<td><strong>Edmunds (27,006)</strong></td>
<td></td>
<td>79.50</td>
<td>91.86</td>
<td></td>
<td>85.23</td>
</tr>
<tr>
<td><strong>Trim</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ILA (580)</strong></td>
<td></td>
<td>23.45</td>
<td>52.17</td>
<td></td>
<td>32.35</td>
</tr>
<tr>
<td><strong>Edmunds (27,006)</strong></td>
<td></td>
<td>38.01</td>
<td>63.69</td>
<td></td>
<td>47.61</td>
</tr>
</tbody>
</table>

Number of reference set tuples discovered

27,000 → wasted effort!
## Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set (ARX for extraction)

Determined by locking

<table>
<thead>
<tr>
<th>Craig’s Cars: 4,400 posts</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Make</strong></td>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>ILA (580)</td>
<td>78.19</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>92.51</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>ILA (580)</td>
<td>64.25</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>79.50</td>
</tr>
<tr>
<td><strong>Trim</strong></td>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>ILA (580)</td>
<td>23.45</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>38.01</td>
</tr>
</tbody>
</table>
## Results: Information Extraction

Iterative Locking Algorithm (ILA) vs. manual reference set (ARX for extraction)

<table>
<thead>
<tr>
<th>Make</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>78.19</td>
<td>84.52</td>
<td>81.23</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>92.51</td>
<td>99.52</td>
<td>95.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>64.25</td>
<td>82.79</td>
<td>72.35</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>79.50</td>
<td>91.86</td>
<td>85.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trim</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (580)</td>
<td>23.45</td>
<td>52.17</td>
<td>32.35</td>
</tr>
<tr>
<td>Edmunds (27,006)</td>
<td>38.01</td>
<td>63.69</td>
<td>47.61</td>
</tr>
</tbody>
</table>

Craig’s Cars: 4,400 posts

**Competitive: fully automatic…**
# Results: Information Extraction

## Laptops (Craigslist): 2,400 posts

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (295)</td>
<td>60.42</td>
<td>74.35</td>
<td>66.67</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>84.41</td>
<td>95.59</td>
<td>89.65</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Recall</td>
<td>Prec.</td>
<td>F-Mes.</td>
</tr>
<tr>
<td>ILA (295)</td>
<td>61.91</td>
<td>76.18</td>
<td>68.31</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>43.19</td>
<td>80.88</td>
<td>56.31</td>
</tr>
<tr>
<td><strong>Model Num.</strong></td>
<td>Recall</td>
<td>Prec.</td>
<td>F-Mes.</td>
</tr>
<tr>
<td>ILA (295)</td>
<td>27.91</td>
<td>81.08</td>
<td>41.52</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>6.05</td>
<td>78.79</td>
<td>11.23</td>
</tr>
</tbody>
</table>

## Skis (eBay): 4,600 posts

<table>
<thead>
<tr>
<th>Brand</th>
<th>Recall</th>
<th>Prec.</th>
<th>F-Mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILA (1,392)</td>
<td>60.84</td>
<td>55.26</td>
<td>57.91</td>
</tr>
<tr>
<td>Skis.com (213)</td>
<td>83.62</td>
<td>87.05</td>
<td>85.30</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Recall</td>
<td>Prec.</td>
<td>F-Mes.</td>
</tr>
<tr>
<td>ILA (1,392)</td>
<td>51.33</td>
<td>48.93</td>
<td>50.10</td>
</tr>
<tr>
<td>Skis.com (213)</td>
<td>28.12</td>
<td>67.95</td>
<td>39.77</td>
</tr>
<tr>
<td><strong>Model Spec.</strong></td>
<td>Recall</td>
<td>Prec.</td>
<td>F-Mes.</td>
</tr>
<tr>
<td>ILA (1,392)</td>
<td>39.14</td>
<td>56.35</td>
<td>46.29</td>
</tr>
<tr>
<td>Skis.com (213)</td>
<td>18.28</td>
<td>59.44</td>
<td>27.96</td>
</tr>
</tbody>
</table>
Results: Information Extraction

<table>
<thead>
<tr>
<th>Laptops (Craigslist): 2,400 posts</th>
<th>Skis (eBay): 4,600 posts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manufacturer</strong></td>
<td><strong>Brand</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>60.42 74.35 66.67</td>
<td>60.84 55.26 57.91</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>84.41 95.59 89.65</td>
<td>83.62 87.05 85.30</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td><strong>Model</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>61.91 76.18 68.31</td>
<td>51.33 48.93 50.10</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>43.19 80.88 56.31</td>
<td>28.12 67.95 39.77</td>
</tr>
<tr>
<td><strong>Model Num.</strong></td>
<td><strong>Model Spec.</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>27.91 81.08 41.52</td>
<td>39.14 56.35 46.29</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>6.05 78.79 11.23</td>
<td>18.28 59.44 27.96</td>
</tr>
</tbody>
</table>

Overstock: new laptops do not cover used ones for sale

Ski Brands: Many models found as brands. Again, specific attributes
Results: Information Extraction

<table>
<thead>
<tr>
<th>Laptops (Craigslist): 2,400 posts</th>
<th>Skis (eBay): 4,600 posts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manufacturer</strong></td>
<td><strong>Brand</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>Recall: 60.42, Prec.: 74.35, F-Mes.: 66.67</td>
<td>Recall: 60.84, Prec.: 55.26, F-Mes.: 57.91</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>Recall: 84.41, Prec.: 95.59, F-Mes.: 89.65</td>
<td>Recall: 83.62, Prec.: 87.05, F-Mes.: 85.30</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td><strong>Model</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>Recall: 61.91, Prec.: 76.18, F-Mes.: 68.31</td>
<td>Recall: 51.33, Prec.: 48.93, F-Mes.: 50.10</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>Recall: 43.19, Prec.: 80.88, F-Mes.: 56.31</td>
<td>Recall: 28.12, Prec.: 67.95, F-Mes.: 39.77</td>
</tr>
<tr>
<td><strong>Model Num.</strong></td>
<td><strong>Model Spec.</strong></td>
</tr>
<tr>
<td>ILA (295)</td>
<td>ILA (1,392)</td>
</tr>
<tr>
<td>Recall: 27.91, Prec.: 81.08, F-Mes.: 41.52</td>
<td>Recall: 39.14, Prec.: 56.35, F-Mes.: 46.29</td>
</tr>
<tr>
<td>Overstock (279)</td>
<td>Skis.com (213)</td>
</tr>
<tr>
<td>Recall: 6.05, Prec.: 78.79, F-Mes.: 11.23</td>
<td>Recall: 18.28, Prec.: 59.44, F-Mes.: 27.96</td>
</tr>
</tbody>
</table>

Overstock: new laptops do not cover used ones for sale

Ski Brands: Many models found as brands. Again, specific attributes

Fully automatic method that is competitive with supervised methods

<table>
<thead>
<tr>
<th>ILA vs. CRF-Win</th>
<th>ILA vs. CRF-Ortho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outperforms</td>
<td>Outperforms</td>
</tr>
<tr>
<td>4/9</td>
<td>1/9</td>
</tr>
<tr>
<td>Within 10%</td>
<td>Within 10%</td>
</tr>
<tr>
<td>7/9</td>
<td>4/9</td>
</tr>
</tbody>
</table>
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue

- 5 bigram-types:
  - … brand new Land Rover Discovery for…
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue

- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”,

Introduction ● Unsupervised IE ● Building Reference Sets ● Supervised IE ● Conclusion
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue
- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”, “SAME ATTR”,
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue

- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”,

“DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”,

“DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”,
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue

- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”, “JUNK ATTR”,
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott …
    - “Boundary” issue

- 5 bigram-types:
  - … brand new Land Rover Discovery for…

“DIFF ATTR”, “SAME ATTR”, “ATTR JUNK”, “JUNK ATTR”, “JUNK JUNK”
ILA’s Applicability

- Difficulty: multi-token, multi-attribute domains
  - BFT: 2.5* Courtyard Rancho Cordova Marriott ...
  - “Boundary” issue

- 5 bigram-types:
“Bootstrap-Compare”

- Easily decide to use ILA

---

Bootstrap labels

- Distribution of 5 bigram types

- KL-Divergence (Cars/Laptops/Skis)

Manually Build Reference set

---------

Posts

2002 Honda Accord EX …
2002 Accord for sale …

< T

Can run ILA

---

Honda Accord 2002 …
Label 1 post
“Bootstrap-Compare”

- Easily decide to use ILA

---

**Introduction**

**Unsupervised IE**

**Building Reference Sets**

**Supervised IE**

**Conclusion**

---

- **Easily decide to use ILA**

- **Bootstrap labels**
  - Distribution of 5 bigram types

- **KL-Divergence (Cars/Laptops/Skis)**
  - < T

- **Can run ILA**

---

**Manually Build Reference set**

**Posts**

**Distribution of 5 bigram types**

**Bootstrap labels**

**2002 Honda Accord EX**

**2002 Accord for sale**

---

**Honda Accord 2002**

**Label 1 post**
“Bootstrap-Compare”

- Easily decide to use ILA

Source | Can build? | Classification
--- | --- | ---
Digicams (eBay) | Yes, good extraction | ILA: 18/20
Cora (references) | No, poor extraction | Manual: 20/20
Supervised Machine Learning for Extraction from Posts

JAIR, 2008

- Require **highest-accuracy** extraction
  - Ambiguity: 626, Mazda or car price?

---

**Set of posts**

**Record Linkage**
1. Blocking (candidate matches)
2. Matching: supervised ML

**Information Extraction**
(supervised ML)
**Supervised Machine Learning for Extraction**

**Record Level Similarity + Field Level Similarities**

1. **Record Linkage**

   \[ V_{RL} = \langle RL\_scores(post, \text{attribute}_1, \text{attribute}_2, \ldots, \text{attribute}_n), \]
   \[ RL\_scores(post, \text{attribute}_1), \]
   \[ \ldots, \]
   \[ RL\_scores(post, \text{attribute}_n) \rangle \]

2. **Supervised Extraction**

   Compare to match’s attributes

   Multiclass-SVM / CRF
## Results: Information Extraction

<table>
<thead>
<tr>
<th>Domain</th>
<th>Num. of Attributes with Max F-Mes.</th>
<th>Total Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phoebus</td>
<td>PhoebusCRF</td>
</tr>
<tr>
<td>BFT</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>eBay Comics</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Craig’s Cars</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>All</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

- **Phoebus/PhoebusCRF**
  - Best 12/16 attributes (> ARX > other methods)
  - Different extraction methods → reference set makes difference
- **CRF-Win max: Comics price attribute**
  - Not statistically significant…
  - CRFs outperformed
    - No structure to rely on!
- **Amilcare/ARX use reference sets**
  - Every max F-mes. used reference set
Related Work

- **Semantic Annotation**
  - Require grammar/structure (Cimiano, Handschuh & Staab, 2004; Dingli, Ciravegna, & Wilks, 2003; Handschuh, Staab & Ciravegna, 2002; Vargas-Vera, et. al., 2002)

- **Record Linkage**
  - Decomposed attributes (Fellegi & Sunter, 1969; Bilenko & Mooney, 2003)
  - WHIRL (Cohen, 2000): simple matching

- **Data Cleaning**
  - Tuple-to-Tuple (Lee, et. al., 1999; Chaudhuri, et. al., 2003)

- **BSL**
  - Other work focuses on methods, not choosing attributes (Baxter, Christen, & Churches, 2003; McCallum, Nigam, & Ungar, 2000; Winkler, 2005)
  - Bilenko, Kamath, & Mooney, 2006: graphical set covering
Related Work (2)

- **Unstructured information extraction**
  - DataMold (Borkar, Deshmukh, & Sarawagi, 2001), CRAM (Agichtein & Ganti, 2004): no junk tokens
  - Semi-CRF methods (Cohen & Sarawagi, 2004): dictionary component, but look-up

- **Ontology based IE**
  - requires ontology management (Embley, et. al., 1999; Ding, Embley & Liddle, 2006; Muller, et. al., 2004)

- **Ontology creation**
  - Use web pages to build single hierarchies (Sanderson & Croft, 1999; Schmitz, 2006; Comiano, Hotho & Staab, 2004; Dupret & Piwowarski, 2006; Makrehchi & Kamel, 2007)
    - I build many and flatten them
Conclusion: Contributions

- Automatic, state-of-the-art extraction on posts given reference set(s)
- Automatically build reference set for cases where difficult to do so manually
- Supervised extraction on posts with highest accuracy
Conclusion: Future Work

● Applications
  ● Information Retrieval
    ● Source classification ➔ page of “cars”
  ● Ontology alignment
    ● Match 2 ontologies to posts, then transitive closure
  ● Semantic Web mark-up

● Research
  ● More robust automatic creation
  ● Weakly (semi?) supervised approach to IE
  ● Information Fusion
    ● Larger documents? NER?
  ● Data mining the results
    ● Create portals
    ● User decision support
Questions?

THANK YOU

Thanks