Domain-Independent Data Extraction: Person Names

Carl Christensen and Deryle Lonsdale
Brigham Young University
cvchristensen@gmail
lonz@byu.edu
Challenge

• Extraction software and techniques yield good results with domain specific data extraction
• Person names and information rarely domain specific
• Identification and extraction difficult because of noisy data, lack of formatting
WePS task

- Web People Search
- 18 attribute values on person names
  - Date of birth, Birth place, Other name, Occupation, Affiliation, Work, Award, School, Major, Degree, Mentor, Location, Nationality, Relatives, Phone, FAX, Email, Web site
- Training corpus – 17 names, approx. 100 web pages per
- Script given to evaluate performance
- Test corpus of comparable size
- New ground in information extraction
Ontos

- Software developed by BYU data extraction group
- Ontology based method leveraged to organize data
- Off the shelf performance
- Similar uses for obituaries and car ads information extraction
  - Good performance on these tasks
Beloved son, brother and uncle, Robert Gene Larkins

Ontos obituary results

Robert Gene Larkins, 46, passed away Wednesday, October 6, 2004, at his home in Goshen, Utah. Bobby was born December 24, 1957, in Riverside, California. He was the fifth of six children; third son of four, born to Robert A. and Je Ramona High School in Arlington, California. Bobby's passion was anything that would go FAST. When he was 14 years old, the local police pulled him over and issued him a citation for going above the speed limit on people, especially relatives, respectfully referred to Bobby as, "McGuiver." He could fix EVERYTHING with almost ANYTHING, his specialty being cars and motorcycles. Many times he would receive telephone calls very adept at diagnosing the problem and would tell the caller just what they needed to do to fix it. Bobby was an incredible human being. He was the type of person who would give you the shirt off his back if he thought "the going rate" for all the vehicle repairs he made for people. But instead, he chose to serve those who couldn't afford to pay. He will be greatly missed by so many. His absence has already been felt, and will continue to be by many loved ones who have gone on before him; including his older brother Timothy Fred, who passed away two years ago. Also, his beloved dog, Patches. He is survived by his parents, and siblings, Judy Tatum, of Alpine, Utah; John Larkins, of Albany, Oregon; also, many nephews, nieces and friends, who will miss his humor, advice and presence. We appreciate the Warenstl Funeral Home for their assistance. May HEAL. Dear Bobby, we celebrate your life and the great love you shared with us. We know that we shall see you again in time. Funeral services for Robert will be held Monday, October 11, 2004 at 11:00 a.m. at 17 blocks north of the Mt. Tampanogos Temple). Family and friends are invited to call, from 10:00 to 11:00 a.m. Interment for Robert will take place at the Alpine City Cemetery, located at 350 North Grove Street, Alpine.

William Keith Romney, 75, died Wednesday, October 6, 2004 after a lingering illness. He was born on January 17, 1929 in Salt Lake City, Utah, to Lois and Carl Romney. Bill will be remembered for his dedication to his children, his love for skiing, golf, and his appreciation of music and art. He is survived by his soul mate, Joan L. Romney, two sons: William Bruce Bach, Cour d'Alene, ID, Michael (Keri) Keith Romney, Ogden; and Jordan, Heather (Peter) Romney Redgrave, Berlin, Germany; and Laurel Romney, Alta; 12 grandchildren, and three great-grandchildren. Bill is also survived by his brother, Carl (Barbara) Romney, Alexandria, VA, sister and nephews. He is preceded in death by his parents, sister Marie, brother Bob, daughter Valerie, and beloved companion Putter. The family would like to express their appreciation to the V.A. hospital, and his health care providers. Bill's life at his home on Sunday, October 9, 2004 from 1:00-3:00 p.m. In lieu of flowers, please send a donation to the Red Cross in Bill's name. Funeral directors, Myers Mortuary of Ogden. Send condolences to the family.

Ernesto J. Sandoval, 80, passed away October 7, 2004 at his home surrounded by his family. Born November 9, 1923 to Adelaida and Onofre Aguilar Sandoval in La Puente, New Mexico. Survived by his daughters:

Annotation Results

<table>
<thead>
<tr>
<th>DeceasedName</th>
<th>Age</th>
<th>DeathDate</th>
<th>BirthDate</th>
<th>Relationship</th>
<th>IntermentDate</th>
<th>IntermentAddress</th>
<th>FuneralAddress</th>
<th>FuneralTime</th>
<th>FuneralDate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robert Gene Larkins</td>
<td>46</td>
<td>October 6, 2004</td>
<td>December 24, 1957</td>
<td>Sibling</td>
<td>Show</td>
<td>350 North Grove Street, Alpine, Utah</td>
<td>1776 North 900 East, American Fork, Utah</td>
<td>11:00</td>
<td>October 11, 2004</td>
</tr>
<tr>
<td>Relationship</td>
<td>RelativeName</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sibling</td>
<td>Gloria Larkins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>Jean Larkins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>niece</td>
<td>CANNOT HEAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViewingDate</td>
<td>ViewingAddress</td>
<td>BeginningTime</td>
<td>EndingTime</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goshen, Utah</td>
<td></td>
<td>10:00</td>
<td></td>
<td>11:00 a.m.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>William Keith Romney</td>
<td>75</td>
<td>October 6, 2004</td>
<td>January 17, 1929</td>
<td>Show</td>
<td></td>
<td></td>
<td></td>
<td>1:00</td>
<td>October 9, 2004</td>
</tr>
<tr>
<td>Ernesto J. Sandoval</td>
<td>80</td>
<td>October 7, 2004</td>
<td>November 9, 1923</td>
<td>Show</td>
<td></td>
<td></td>
<td></td>
<td>174 East 900 South</td>
<td>10:00</td>
</tr>
</tbody>
</table>
WePS extraction process

WePS ontology

Knowledge sources
Ontos
Annotated results
Person webpage
Text file output
Evaluation script
Results report
Data frames

XML description of extraction ontology components
Knowledge files

- Names, cities, countries, hypocoristics, occupations, etc.

- Knowledge gathered from extracting and formatting online databases
  - Live Journal
  - Wikipedia
  - Bureau of Labor Statistics
  - etc.

- Approximately 80,000 school names and 30,000 occupations
  - 66% of total schools in U.S. in 2003

- All possible options for some files, small subset for others
  - e.g. Occupations, hypocoristics
Constraints

• Required context expressions
  
  `<RequiredContextExpression>
  <ExpressionText>`BirthTime`</ExpressionText></RequiredContextExpression>`

• Cardinality
  
  Search Person [0:*] has Occupation [1:*];
  Search Person [0:*] has Affiliation [1:*];
  Search Person [0:*] has Email [1];

• Regular expressions
  
  Email : `[w]+[d]*@[w]+[.]{1}[w]+`
King David Lodge No. 93

Work:
Canadian

Installation of officers:
December

District:
District No. 17

Lodgehall:
1763 Bellevue Ave., West Vancouver

Meeting day:
1st Thursday

Website:
www.kingdavidlodge93.org
Family Photo Album Dec02/Jan03

CLICK ON THE PICTURES BELOW TO SEE THE LARGER PHOTO

At Home.............hanging around.........

On Holidays, December/January 2003........horning around............


**PEOPLE**

**Rt Hon Michael Howard QC MP**

Member of Parliament for Folkestone & Hythe

**ABOUT MICHAEL**

Michael was born in 1941 and educated at Llanelli Grammar School and Peterhouse, Cambridge. In 1962 he was elected President of the Cambridge Union. He was called to the Bar in 1964 and was appointed a QC in 1982.

He was elected Member of Parliament for Folkestone and Hythe in 1983. In 1984 he was appointed Parliamentary Private Secretary to the Solicitor General. The following year he entered the Government as Parliamentary Under Secretary of State at the Department of Trade and Industry with responsibility for corporate and consumer affairs.

In 1987 he moved to the Department of the Environment, first as Minister of State for Local Government and then as Minister of State for Water and Planning. In 1990 he entered the Cabinet as Secretary of State for Employment, abolishing the closed shop and playing a crucial role in negotiating the UK's opt-out from the Social Chapter at Maastricht.

Following the 1992 election, Michael was appointed Secretary of State for the Environment. In this position he played a major role in securing the participation of the USA at the Earth Summit in Rio which he attended on behalf of the Government.

**Annotation Results**

<table>
<thead>
<tr>
<th>BirthDate</th>
<th>BirthPlace</th>
<th>OtherName</th>
<th>Occupation</th>
<th>Affiliation</th>
<th>Award</th>
<th>School</th>
<th>Major</th>
<th>Degree</th>
<th>Mentor</th>
<th>Nationality</th>
<th>Relatives</th>
<th>Phone</th>
<th>Fax</th>
<th>Email</th>
<th>WebSite</th>
<th>Foo</th>
<th>Atte</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phone
- 01302 253524
- 01302 251061

Email
- Howardm@parliament.uk

WebSite
- http://michaelhowardmp.com
Precision/Recall

Recall = \frac{|\{\text{relevant information}\} \cap \{\text{retrieved information}\}|}{|\{\text{relevant information}\}|}

Precision = \frac{|\{\text{relevant information}\} \cap \{\text{retrieved information}\}|}{|\{\text{retrieved information}\}|}

F_{\alpha} = \frac{(1 + \alpha) \times \text{precision} \times \text{recall}}{((\alpha \times \text{precision}) + \text{recall})}

\alpha = 0.5 \text{ for WePS}
### WePS2-AE matching Result for

<table>
<thead>
<tr>
<th>Attribute</th>
<th>MATCH</th>
<th>MISS1</th>
<th>MISS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>affiliation</td>
<td>Royal Society of Literature&lt;br&gt;University of Birmingham</td>
<td>Birmingham&lt;br&gt;History&lt;br&gt;Home</td>
<td>Silver Nymph at the International Television Festival&lt;br&gt;Sunday Express Book of the Year</td>
</tr>
<tr>
<td>award</td>
<td>[CBE]//CBE&lt;br&gt;[Hawthornden Prize]//Hawthornden Prize&lt;br&gt;[Royal Television Society's Award for best Drama serial in the year 1989]//Royal Television Society's Award for best Drama serial in the year 1989&lt;br&gt;[Yorkshire Post Fiction Prize]//Yorkshire Post Fiction Prize</td>
<td>Booker Prize&lt;br&gt;Commonwealth Writers Prize&lt;br&gt;Criticism&lt;br&gt;Nice Work&lt;br&gt;Romance&lt;br&gt;Silver Nymph</td>
<td>Silver Nymph at the International Television Festival&lt;br&gt;Sunday Express Book of the Year</td>
</tr>
<tr>
<td>birthplace</td>
<td></td>
<td>MA</td>
<td>London, England</td>
</tr>
<tr>
<td>dateofbirth</td>
<td></td>
<td></td>
<td>January 28, 1935</td>
</tr>
<tr>
<td>degree</td>
<td>BA&lt;br&gt;PhD</td>
<td></td>
<td>MA</td>
</tr>
<tr>
<td>major</td>
<td></td>
<td>Modern&lt;br&gt;Humanities</td>
<td></td>
</tr>
<tr>
<td>nationality</td>
<td>British</td>
<td></td>
<td></td>
</tr>
<tr>
<td>occupation</td>
<td>Fellow&lt;br&gt;Honorary Professor&lt;br&gt;author&lt;br&gt;full-time writer</td>
<td>Author&lt;br&gt;Booker&lt;br&gt;Novelist&lt;br&gt;Reader&lt;br&gt;Writers&lt;br&gt;literary critics&lt;br&gt;page&lt;br&gt;satirists</td>
<td></td>
</tr>
<tr>
<td>school</td>
<td>[University of Birmingham]//University of Birmingham</td>
<td></td>
<td>University College London</td>
</tr>
<tr>
<td>website</td>
<td></td>
<td><a href="http://en.wikipedia.org/wiki/David_Lodge_%28author%29">http://en.wikipedia.org/wiki/David_Lodge_%28author%29</a></td>
<td></td>
</tr>
</tbody>
</table>
Challenges

- Smaller/larger match preference
  - Preference for “Arizona” as place over “University of Arizona” as school
- DOM parser
  - Unofficial HTML tags cause system to fail
- Text formatting
  - Record detection for individuals intractable
- System functionality
  - Cardinality bounds, system output file
Performance

- Very low initial precision/recall: < 1%
- Increased drastically with knowledge engineering and system constraints
  - 27% recall with some person results approaching 40% recall
  - Approaching 10% precision
- Nothing to measure against
  - Official WePS results will be released in April
Future work

- Ontos robustness
- Machine learning for constraints/knowledge files
- Person name disambiguation
- Keyword probability values