Ontology-based Web Information Extraction in Practice

eRecruitment – eTourism - eProcurement

Japan-Austria Joint Workshop on “ICT”

Tokyo, October 18-19, 2010

Institute for Application Oriented Knowledge Processing

a.Univ.-Prof. Dr. DI Birgit Pröll
bproell@faw.jku.at
Contents

- **Motivation**
- Web Information Extraction (WebIE) by Examples
  - General Architecture
  - Web Crawler
  - Ontology Aware WebIE
  - Structure Analysis: Page Segementation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work
Web Information Extraction (WebIE)

...extracting structured data from Web pages

- accommodation's name: Alpenrose
- phone: ++43 (0)524352930
- address: A-6212 Maurach
- pool facility: -

templates
WebIE Projects in cooperation with Austrian Industry

**TourIE**
Tiscover AG

**Application area**
- eTourism

**JobOlize**
JoinVision E-Services GmbH
FFG (grant 813202)

**Application area**
eRecruitment

**Marlies**
Tech2select GmbH
FFG (grant 817789)

**Application area**
eManufacturing, supply-chain-management
Projects’ Requirements and Approach Taken

Some WebIE peculiarities in the given projects
- Heterogeneously designed Web pages
- Mixture of (semi-)structured data and full text
- Significant structural aspects, e.g.,
  - location of information on Web page
  - information “hidden” in Web tables
- Information scattered over several Web pages
- Web site evolution

WebIE Approaches
- Screen scraping approaches (wrapper generation)
- Automatically trainable systems (machine learning)
- Knowledge-engineering approach
  + Web crawler + structural analysis + …

[Appelt et al., 1999]
Contents

- Motivation
- Web Information Extraction (WebIE) by Examples
  - General Architecture
  - Web Crawler
  - Ontology Aware WebIE
  - Structure Analysis: Page Segementation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work
Overall Architecture

Pre-Processing

Crawler

Web sites

Information Extraction

IE-Pipeline (GATE *)

Tokenizer
Sentence-Splitter (e.g.)
Gazetteer-Lists
Ontology-Plugin
Transducers

Gazetteer lists
Domain Ontology

Knowledge Base

Rules

Post-Processing

Output

Annotated Web pages
XML

*) [Cunningham et al, 2006]
Web Crawler

- Collects relevant Web pages
- Classifies Web pages
  - Home page, price pages, location pages, etc.
  - Based on Support Vector Machine
- Recognises language
  - Using meta-tags and an n-gram based algorithm
Overall Architecture

Types of annotations
- syntactical, morphological
- ontological
- structural
- relevance judging

Crawler
- Pre-Processing
- Web sites

IE-Pipeline (GATE)
- Tokenizer
- Sentence-Splitter (e.g.)
- Gazetteer-Lists
- Ontology-Plugin
- Transducers

Gazetteer lists
- Domain Ontology
- Knowledge Base

Rules

Output
- Annotated Web pages
- XML
- Manual Evaluation
- Manual Correction

Post-Processing
Regular Expressions & Gazetteeer Lookup

**Rule: Phone1**

```plaintext
( {Token.string=="+") {Token.kind==number}
{SpaceToken.kind==space}*)
{Token.string=="("}
{Token.kind==number}
{Token.string==")")
{SpaceToken.kind==space}*)
{Token.kind==number})
):phone
 -->
 :phone.MyPhone={}
```

**Gazetteer list `phone keywords`**

- Phone
- Telephone
- Tel.
- Tel.:
- Telefon

Phone: +43 (0)5243 52930
Fax: +43 (0)5243 5466
info@alpenrose.at

Telefon +43(0)7213/6365
Fax +43 (0)7213/6365-8
info@sternsteinhof.at

Tel: 0043/6432/84 75
Fax 0043/6432/84 75 70
E-mail: info@alpina-hotel.com
Internet: www.alpina-hotel.com
Ontology-Aware Entity Recognition (1/2)

Unser Hallenbad – klein aber fein – ist eine Oase des Friedens und der Ruhe. Im 33°C warmen Granderwasser mit Gegenströmanlage können Sie Ihren Kreislauf und die Muskulatur stärken.

Vitaltempel - Sauna-experience-pool area

- 45°C rock steam sauna
- NEW - 47°C steam bath and 48°C vapour room
- 50°C caldarium
- 60°C laconicum
- 80°C Tyrolean pine sauna
- Whirl pools
- Salt grotto set in Tyrolean rock
- "Stille Alm", an oasis of relaxation with water beds kept at optimum temperature and a magnificent view of the stunning mountain panorama
- Our new "Stille Hochalm" - the crème de la crème of quiet rooms with water beds
- Labyrinth of the Senses
- "Spring garden with moated castle", a 180 square metre adventure temperature of 28°C and an outdoor pool temperature of 31°C
- Cosy block sauna with quiet room in the 'Residenz-Garten'
- Swimming pond with glorious sunbathing area covering 800 square meter
Ontology-Aware Entity Recognition (2/2)

We offer a wonderful 2500m² wellness area, lead by a trained wellness team. Indoor swimming pools, new heated natural outdoor pool with sandy beach, open air whirlpool with a wonderful view of lake Caldaro, large sauna world, and our private beach directly at lake Caldaro, full fill all wishes!

Rule: PoolsWithAttributes
{
  {{PoolAttribute}}
  +:attributes
  {{PoolFacility}}:facilities
}

RHS (List attributes, Annotation facilities){
  create new FeatureMap features;
  for(Annotation attribute : attributes){
    add attribute to features;
  }
  create new Annotation MyPool;
  add features to MyPool;
  return MyPool;
### Structure Analysis: Web Page Segmentation

**Top part**
- **Job Title**
  - Senior Java Developer

**Content part**
- **IT Skills + Level**
  - JAVA + perfect
  - MySQL + basic
- **Operation Area**
  - SW programming, testing
- **Language Skills**
  - English fluently
- **Contact**
  - -

**Bottom part**
- **[Debnath et al., 2005]**
- **[Chakrabarti et al., 2007]**
### Senior JAVA Developer

**Job Name:**

**Senior JAVA Developer Needed to Make an Impact in a Small Growing Organization!**

Join a dynamic team as a Senior JAVA developer for a leading-edge integrated advertising platform in a Linux/MySQL environment. To be considered, you must have solid Object Oriented development experience. Must have experience taking ownership of development initiatives for entire project lifecycle; the ability to work independently and with a team and the ability to conceptualize, design and implement business requirements. **PERMANENT Position! Apply today!**

**Responsibilities**

- JAVA: Perform detailed software design, documentation, development and testing for Java applications.
- Analyze customer needs and develop overall concept and design objectives.

**Required Skills**

- Extensive knowledge of Java
- MySQL
- Preferred skills include: LAMP, Ruby on Rails, Linux

**Benefits/Perks**

- Flexible hours; work from home 1-2 days a week
- Very independent organization

**Application Requirements**

Client will only consider local candidates. Client requires Green Card or U.S. Citizenship for this position.
Structure Analysis: Table Data Extraction in Marlies

<table>
<thead>
<tr>
<th>machine type</th>
<th>description</th>
<th>Ø (mm)</th>
<th>to weight</th>
<th>module</th>
<th>max.</th>
<th>quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>tothing machines:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pfauter P 400</td>
<td>hobbing machine</td>
<td>15 - 400</td>
<td>6</td>
<td></td>
<td></td>
<td>8 - 9</td>
</tr>
<tr>
<td>Pfauter P 900</td>
<td>hobbing machine</td>
<td>30 - 900</td>
<td>10</td>
<td></td>
<td></td>
<td>8 - 9</td>
</tr>
<tr>
<td>Pfauter P 900</td>
<td>hobbing machine</td>
<td>30 - 900</td>
<td>10</td>
<td></td>
<td></td>
<td>8 - 9</td>
</tr>
<tr>
<td>Pfauter PE 1200</td>
<td>hobbing machine</td>
<td>30 - 1200</td>
<td>20</td>
<td></td>
<td></td>
<td>6 - 7</td>
</tr>
<tr>
<td>Pfauter P 1250</td>
<td>hobbing machine</td>
<td>50 - 1250</td>
<td>16</td>
<td></td>
<td></td>
<td>8 - 9</td>
</tr>
<tr>
<td>Lorenz SN 8</td>
<td>shaping machine</td>
<td>30 - 750</td>
<td>10</td>
<td></td>
<td></td>
<td>8 - 9</td>
</tr>
</tbody>
</table>

Result

| Hobbing machine – Pfauter P 400 | Diameter (min): 15 mm Diameter (max): 400 mm |

[Yang et al., 2002] [Gatterbauer et al., 2007]
Structure Analysis: Table Data Extraction in TourIE

<table>
<thead>
<tr>
<th>Room type</th>
<th>7.5. - 27.6.</th>
<th>27.6. - 6.9.</th>
<th>6.9. - 20.9.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type A</td>
<td>34,00</td>
<td>36,00</td>
<td>33,00</td>
</tr>
<tr>
<td>Double room with TV, shower/WC, partly with a balcony</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type B</td>
<td>37,00</td>
<td>38,00</td>
<td>35,00</td>
</tr>
<tr>
<td>Double room with TV, bath room with shower/WC, hairdryer, partly with a balcony</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type Fritzerkogel</td>
<td>42,00</td>
<td>43,00</td>
<td>40,00</td>
</tr>
<tr>
<td>Bigger double room with 1 double bed and 1 extra bed, bath tube/WC, hairdryer, TV, telephone, radio, CD, balcony</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Children's reduction summer 2009 in the room of the parents:

- 0 - 2 years 100%
- 3 - 5 years 50%
- 6-12 years 30%
- 13-18 years 20%

Prices per person, per night in the double room including half-board - min. stay 3 nights for 1 or 2 nights surcharge of Euro 3,50 per person per night.

Price domain board: time period, room name + room type, price value without currency, detailed description of room and its amenities.
Contents

- Motivation
- Web Information Extraction (WebIE) by Examples
  - General Architecture
  - Web Crawler
  - Ontology Aware WebIE
  - Structure Analysis: Page Segementation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work
Evaluation: TourIE

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>0.6667</td>
<td>0.5833</td>
<td>0.6481</td>
<td>0.4167</td>
</tr>
<tr>
<td>Address</td>
<td>0.7632</td>
<td>0.6170</td>
<td>0.7286</td>
<td>0.3830</td>
</tr>
<tr>
<td>Pool facilities</td>
<td>0.7500</td>
<td>0.4286</td>
<td>0.6522</td>
<td>0.5714</td>
</tr>
<tr>
<td>Aggregated</td>
<td>0.7333</td>
<td>0.5556</td>
<td>0.6892</td>
<td>0.4444</td>
</tr>
</tbody>
</table>

- Evaluation results were satisfactory with respect to the preliminary study.
- Pool facility extraction quality was poor because of incomplete ontology.
Evaluation: JobOlize

Page segmentation & block identification considerably rises precision.
Evaluation: Marlies

Marlies Ontology
Classes: 2313
Instances: 2661
Assignments of object properties to instances: 42791

Preliminary results for recall:

» Work in progress (e.g., table extraction).
Manual Correction via Rich Client GUI

Jobs

We are one of the largest Austrian software development companies.

Junior Software Developer (m/w)

Main Responsibilities:

Design and development of new .NET and Java modules and refactoring of already existing modules. Creation and execution of test cases including defect tracking, furthermore ensure of overall quality of the product.

Requirements:

- C#, ASP.NET and Java
- Web services
- XML, SVG and Javascript
- Basic knowledge of DB systems MySQL and Oracle
- University degree in Computer science
- Good spoken and written command of English language

We offer:

- Use of cutting-edge technologies
- Exciting and challenging environment
- Attractive benefits

Send us your CV if you want to be part of our team!

Contact Information:

E-mail: jobs@ourcompany.at
Address: Our Company, Company Str.100, 1000 Vienna, Austria
powered by TypeIt
Contents

- Motivation
- Web Information Extraction (WebIE) by Examples
  - General Architecture
  - Web Crawler
  - Ontology Aware WebIE
  - Structure Analysis: Page Segementation, Table Extraction
- Evaluation & Manual Correction of Results
- Lessons Learned & Future Work
Lessons Learned

- Today’s Web pages do not adhere to standards or semantic Web proposals.
  - Only a few RDF resources available; proposed microformats rarely used
  - Poor HTML, e.g., tables used for layout purposes
  - Web 2.0 coded Web pages in progress; content-based image retrieval & OCR

- Development & maintenance of knowledge-based WebIE systems is expensive.
  - Domain experts & knowledge engineers are needed.
  - Rule-coding is tedious and error-prone.
  - Evaluation of numerous methods & algorithms; multiplied due to multilinguality
  - Manual evaluation is time consuming.

- WebIE performance considerably depends on quality of domain ontology.

- We have to observe (evolving) legal issues
  - Robots exclusion standard, Sitemap etc.
  - Further processing of extracted data
Future Work: Ontosophia

Ontology-driven IE Supported by (Semi-) Automatic Corrective Feedback

1. Extraction Domain Ontology
2. Ontology Optimization
3. Visual Error Trace Back
4. Evaluation Support
Thank you for your Attention!

Acknowledgements to:

Christina Feilmayr
Stefan Parzer
Christina Buttinger
Michael Guttenbrunner