Information Extraction

CIS, LMU München
Winter Semester 2021-2022

Prof. Dr. Alexander Fraser, CIS
Information Extraction – Administravia - I

• Vorlesung
  • Learn the basics of Information Extraction (IE)
  • Weekly lecture from me here in BU 101
  • Exactly the same scientific content as last year
    • "Administravia" is slightly different, see these slides (also uploaded after class)
    • 2020-2021 videos are available, use these to review

• Seminar
  • There are two seminars! You come to just one of the two sessions, either Wednesdays 10:00 (Group 01) or Thursdays 10:00 (Group 02)
    • You should have received an LSF-email stating which day you are registered for
    • Wed is in this building, Thurs is in main building
  • You will present an oral report and then submit a written version later
  • Oral reports will be done in groups of 3 students working on a single topic
Information Extraction – Administravia - II

• Registration:
  • If you are a CIS Student: check whether you are registered for *both* the Vorlesung and the Seminar (these are two things in LSF!)
    • Please ignore the Modulteilprüfung entries
    • Make sure you are registered for the Seminar and the Vorlesung
  • There are a good number of people only in the Vorlesung
  • There are just a couple of people only in the Seminar
Vorlesung and Seminar are two separate courses (in same module for CIS people)
  • However, there will be some shifting around of slots

Vorlesung (Grade):
  • Klausur in February entirely determines the Vorlesung grade

Seminar (Grade):
  • Referat
  • Hausarbeit (write-up of the Referat) (6 pages, independent for each student, due \textbf{3 weeks} after you hold your Referat)

Grades of Vorlesung and Seminar are independent

CIS-ler: No Notenverbesserung
Information Extraction – Administravia - IV

• Syllabus: see WS last year
  • Brief idea at end of this slide deck
• List of Referatsthemen for the Seminar
  • This will be presented *here* in the Vorlesung, in two weeks on Nov 3rd
• Literature:
    • Please read the introduction for next week (it is available on the web page!)
• Questions?
Information Extraction

• An introduction to the course
  • The topic "Information Extraction" means different things to different people
  • In this course we will look at several different perspectives
  • There is unfortunately no comprehensive textbook that includes all of these perspectives
My Biases

• As you may have noticed by now: I am from the US (PhD in Computer Science from USC/ISI, Artificial Intelligence division)
• I am a professor here at CIS
• I do research in the broad area of statistical NLP
  • I mostly work on machine translation, and related structured prediction problems (e.g., treebank-based syntactic parsing, generation using sequence (tagging) models)
  • I also work on other multilingual problems such as cross-language information retrieval
• With respect to rule-based NLP (with manually written rules), I'll try to be as fair as humanly possible
  • I do use these techniques sometimes too
Outline for today

• Motivation
  • Problems requiring information extraction
  • Basic idea of the output
• Abstract idea of the core of an information extraction pipeline
• Course topics
A problem

Mt. Baker, the school district

Baker Hostetler, the company

Baker, a job opening
Bakery Jobs on CareerBuilder.com
www.careerbuilder.com/jobs/keyword/bakery

Baker Jobs, Employment | Indeed.com
www.indeed.com/q-Baker-jobs.html
Jobs 1 - 10 of 16047 – 16047 Baker Jobs available on Indeed.com. one search, all jobs.

Job Openings - Baker University
www.bakeru.edu/jobs
If you are seeking employment in any of these areas, contact Baker University.

Baker, LA Jobs on CareerBuilder.com
www.careerbuilder.com/Jobs/Baker/
Jobs 1 - 25 of 948 – Looking for Baker, LA Jobs? See currently available job openings on CareerBuilder.com. Browse the current listings and fill out job ...

Down Under Bakery Pies: Job Openings at DUB Pies
www.dubpies.com/jobs.php
Listing of job openings at DUB Pies. Down Under Bakery (DUB) Pies is looking for more staff - check out our list of vacancies.

Field Engineers | Geoscience | Jobs and Careers at Baker Hughes
jobs.bakerhughes.com/
... Oil and Natural Gas? Baker Hughes has career information for you on these, more. ... Search Jobs. Baker Hughes Jobs ... Recent Job Openings. Completion ...

Corner Bakery Job Openings | Glassdoor
www.glassdoor.com/Job/Corner-Bakery-Job-Openings-E297310_P2...
46 Corner Bakery job openings. Search job openings, see if they fit - company salaries, reviews, and more posted by Corner Bakery employees.

Jobs - Baker University
www.bakeru.edu/jobs
See links at left for a complete list of Baker University job openings. It is the policy of Baker University to afford equal opportunity for all persons without distinction ...
A solution
<table>
<thead>
<tr>
<th>Job Title</th>
<th>Company</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food Pantry Workers</td>
<td>Lutheran Social Services</td>
<td>Archbold, OH</td>
</tr>
<tr>
<td>Cooks</td>
<td>Lutheran Social Services</td>
<td>Attleboro, MA</td>
</tr>
<tr>
<td>Bakers Assistants</td>
<td>Fine Catering by Russell Morin</td>
<td>United States</td>
</tr>
<tr>
<td>Baker's Helper</td>
<td>Bird-in-Hand</td>
<td>Maryland Heights, MO</td>
</tr>
<tr>
<td>Assistant Baker</td>
<td>Gourmet To Go</td>
<td>Beaverton, OR</td>
</tr>
<tr>
<td>The School District</td>
<td></td>
<td>Huntsville, UT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Garden Grove, CA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Houma, LA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nisswa, MN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Big Sky, MT</td>
</tr>
<tr>
<td>Cake Decorator/Baker</td>
<td>Mandalay Bay Hotel and Casino</td>
<td>Las Vegas, NV</td>
</tr>
<tr>
<td>Shift Supervisors</td>
<td>Brueggers Bagels</td>
<td>Minneapolis, MN</td>
</tr>
</tbody>
</table>
Title: Ice Cream Guru
Description: If you dream of cold creamy…
Contact: susan@foodscience.com
Category: Travel/Hospitality
Function: Food Services
Another Problem
Often structured information in text

0.44 CT ROUND CUT DIAMOND PENDANT 14 K WHITE GOLD

Classic style and beauty, this comfortable 14 K White gold pendant contains:
An Ideal cut Round 0.44 CT Diamond, in a magnificent high polish bezel.
- Color: F
- Clarity: SI-1
- Setting: 14 K White Gold
- Chain: 16 inches 14 K White Gold
- Weight: 3.4 g
- Measurements: 10 mm x 10 mm

Retail Price: $2,319.00
Close Out Price: $889.00

Slide from Cohen/McCallum

Peter Norvig Robert Wilesky University of California, Berkeley Computer... Thirteenth International Conference on Computational Linguistics, Volume 3

Abstract: This paper critically evaluates three recent abductive interpretation models, those of Charniak and Goodman (1989), Hobbs, Stichel, Martin and Edwards (1988), and Ng and Moore (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions.

Context of citations in this paper: More

... (break slight modification of the one given in [Ng and Moorey, 1990] The new definition remedies the anomaly reported in [Norvig and Wileisky, 1990] of occasionally preferring spurious interpretations of greater depth... Table 1: Empirical Results Computing Coherence and...

... costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wileisky (1990). The use of abduction in disambiguation is discussed in Kay et al. (1990). We will assume the following: 13) a. Only literals...

Cited by: More
Translation Mismatch in a Hybrid MT System - Gavron (1992) (Correct)
Abduction and Mismatch in Machine Translation - Gavron (1999) (Correct)
Interpretation as Abduction - Hobbs, Stichel, Aspelt, Martin (1990) (Correct)

Active bibliography (related documents): More All
0.1: Decision Analytic Networks in Artificial Intelligence - Metzkevish, Abramson (1995) (Correct)
0.1: A Probabilistic Network of Predicators - Tachau, Lin (1993) (Correct)
Information Extraction (IE) is the process of extracting structured information (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).

Elvis Presley was a famous rock singer.

Mary once remarked that the only attractive thing about the painter Elvis Hunter was his first name.

"Seeing the Web as a table"
Defining an IE problem

• In what I will refer to as "classic" IE, we are converting documents to one or more table entries
  • There are other kinds of IE, we will talk about those later
• The **design** of these tables is usually determined by some business need
• Let's look at the table entries for a similar set of examples to the ones we just saw
## Motivating Examples

<table>
<thead>
<tr>
<th>Title</th>
<th>Type</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business strategy Associate</td>
<td>Part time</td>
<td>Palo Alto, CA</td>
</tr>
<tr>
<td>Registered Nurse</td>
<td>Full time</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Motivating Examples

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthplace</th>
<th>Birthdate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elvis Presley</td>
<td>Tupelo, MI</td>
<td>1935-01-08</td>
</tr>
</tbody>
</table>

Elvis Presley, in the humblest of circumstances, was born to Vernon and Gladys Presley in a two-room house in Tupelo, Mississippi on January 8, 1935. His twin brother, Jessie Garon, was stillborn, leaving Elvis to grow up as an only child. He and his parents moved to Memphis, Tennessee in 1948, and Elvis graduated from Humes High School there in 1953.
Motivating Examples

Information Extraction: Techniques and Challenges

Ralph Grishman

Information Integration Papers


<table>
<thead>
<tr>
<th>Author</th>
<th>Publication</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grishman</td>
<td>Information Extraction...</td>
<td>2006</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Motivating Examples

<table>
<thead>
<tr>
<th>Product</th>
<th>Type</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynex 32”</td>
<td>LCD TV</td>
<td>$1000</td>
</tr>
</tbody>
</table>
Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents.
## Information Extraction

**Traditional definition:** Recovering structured data from text

What are some of the sub-problems/challenges?

### Management Team

- **Board of Directors**
  - Our Firm & WOMMA
  - FAQs
  - Contact Us
  - Careers

### Board Members

- **Itzhak Fisher**
  - Chairman of Nielsen BuzzMetrics
- **Thom Mastrelli**
  - Executive Vice President/Corporate Development, VNU
- **Jonathan Carson**
  - CEO of Nielsen BuzzMetrics
- **Mahendra Vora**
  - CEO and Owner, Vora Technology Park
- **Ori Levy**
  - President of Nielsen BuzzMetrics
  - Israel
- **Ron Schoeler**
  - Senior Vice President and General Manager, Nielsen Ventures
- **James O’Hara**
  - Senior Vice President and Chief Financial Officer, VNU’s Media Measurement and Information Group

*Slide from Nigam/Cohen/McCallum*
Information Extraction?

• Recovering structured data from text
  • Identifying fields (e.g. named entity recognition)
Information Extraction?

- Recovering structured data from text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)
Information Extraction?

• Recovering structured data from text
  • Identifying fields (e.g. named entity recognition)
  • Understanding relations between fields (e.g. record association)
  • Normalization and deduplication
Information extraction

• Input: Text Document
  • Various sources: web, e-mail, journals, ...
• Output: Relevant fragments of text and relations possibly to be processed later in some automated way
Not all documents are created equal...

- Varying regularity in document collections
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated

Slide from McCallum
BACKGROUND: The most challenging aspect of revision hip surgery is the management of bone loss. A reliable and valid measure of bone loss is important since it will aid in future studies of hip revisions and in preoperative planning. We developed a measure of femoral and acetabular bone loss associated with failed total hip arthroplasty. The purpose of the present study was to measure the reliability and the intraoperative validity of this measure and to determine how it may be useful in preoperative planning. METHODS: From July 1997 to December 1998, forty-five consecutive patients with a failed hip prosthesis in need of revision surgery were prospectively followed. Three general orthopaedic surgeons were taught the radiographic classification system, and two of them classified standardized preoperative anteroposterior and lateral hip radiographs with use of the system. Interobserver testing was carried out in a blinded fashion. These results were then compared with the intraoperative findings of the third surgeon, who was blinded to the preoperative ratings. Kappa statistics (unweighted and weighted) were used to assess correlation. Interobserver reliability was assessed by examining the agreement between the two preoperative raters. Prognostic validity was assessed by examining the agreement between the assessment by either Rater 1 or Rater 2 and the intraoperative assessment (reference standard). RESULTS: With regard to the assessments of both the femur and the acetabulum, there was significant agreement (p < 0.0001) between the preoperative raters (reliability), with weighted kappa values of >0.75. There was also significant agreement (p < 0.0001) between each rater's assessment and the
Partially Structured:
Seminar Announcements

Extract time, location, speaker, etc.

We will finish the CSE AI research seminar this Monday, November 26th, with speaker Dave Kauchak from the UCSD AI lab. We meet in AP&M 4882 at 12:10PM. Free pizza!

Title:
Boosting for information extraction

Abstract:

In this talk I will examine Boosted Wrapper Induction (BWI, Freitag & Kushmerick) as an exemplar of recent rule-based information extraction (IE) techniques. Results will be shown for BWI on a wider variety of tasks than has previously been studied, including several natural text document collections. I will examine these results and show how the tests performed allow for a systematic analysis of how each of BWI's algorithmic components, particularly boosting, contributes to its performance over comparable methods. I will also present a new metric, the SWLE Ratio, which is a quantitative measure of the regularity of an extraction task, and

First up is Gary Cottrell.

A Neural Network that Perceives and Categorizes Facial Expressions

Abstract:

How do we perceive emotions in facial expressions? On the one hand, findings show that we map facial expressions into discrete categories, as in color and phoneme perception, with sharp boundaries between emotions and better discrimination between pairs of stimuli that straddle a category boundary. On the other hand, there is good evidence
Highly Structured: Zagat’s Reviews

Extract restaurant, location, cost, etc.
Landscape of IE Tasks: Document Formatting

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University.

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC’s Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Non-grammatical snippets, rich formatting & links

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Email</th>
<th>Department</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barto, Andrew G.</td>
<td>(413) 545-2109</td>
<td><a href="mailto:barto@cs.umass.edu">barto@cs.umass.edu</a></td>
<td>CS276</td>
<td></td>
</tr>
<tr>
<td>Berger, Emer D.</td>
<td>(413) 577-4211</td>
<td><a href="mailto:emery@cs.umass.edu">emery@cs.umass.edu</a></td>
<td>CS344</td>
<td></td>
</tr>
<tr>
<td>Brock, Oliver</td>
<td>(413) 577-0334</td>
<td><a href="mailto:oli@cs.umass.edu">oli@cs.umass.edu</a></td>
<td>CS246</td>
<td></td>
</tr>
<tr>
<td>Clarke, Lori A.</td>
<td>(413) 545-1328</td>
<td><a href="mailto:clarke@cs.umass.edu">clarke@cs.umass.edu</a></td>
<td>CS304</td>
<td></td>
</tr>
</tbody>
</table>

Table

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 - 9:30 AM</td>
<td>Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty</td>
</tr>
<tr>
<td>9:30 - 10:00 AM</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:00 - 11:30 AM</td>
<td>Technical Paper Sessions</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Slide from McCallum
Landscape of IE Tasks
Intended Breadth of Coverage

Web site specific
Formatting
Amazon.com Book Pages

Genre specific
Layout
Resumes

Wide, non-specific
Language
University Names
Landscape of IE Tasks: Complexity of entities/relations

**Closed set**
- U.S. states
  - He was born in Alabama…
  - The big Wyoming sky…
- U.S. postal addresses
  - University of Arkansas
    - P.O. Box 140
      - Hope, AR
  - Headquarters:
    - 1128 Main Street, 4th Floor
      - Cincinnati, Ohio 45210

**Regular set**
- U.S. phone numbers
  - Phone: (413) 545-1323
  - The CALD main office is 412-268-1299

**Complex pattern**
- U.S. postal addresses
  - University of Arkansas
    - P.O. Box 140
      - Hope, AR
  - Headquarters:
    - 1128 Main Street, 4th Floor
      - Cincinnati, Ohio 45210

**Ambiguous patterns, needing context and many sources of evidence**
- Person names
  - …was among the six houses sold by Hope Feldman that year.
  - Pawel Opalinski, Software Engineer at WhizBang Labs.

Slide from McCallum
Landscape of IE Tasks: Arity of relation

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: Jack Welch
Person: Jeffrey Immelt
Location: Connecticut

Binary relationship

Relation: Person-Title
Person: Jack Welch
Title: CEO
Relation: Company-Location
Company: General Electric
Location: Connecticut

N-ary record

Relation: Succession
Company: General Electric
Title: CEO
Out: Jack Welsh
In: Jeffrey Immelt

"Named entity" extraction

Slide from McCallum
Association task = Relation Extraction

• Checking if groupings of entities are instances of a relation

1. Manually engineered rules
   • Rules defined over words/entities: “<company> located in <location>”
   • Rules defined over parsed text:
     • “((Subj<company>) (Verb located) (*) (Obj <location>))”

2. Machine Learning-based
   • Supervised: Learn relation classifier from examples
   • Partially-supervised: bootstrap rules/patterns from “seed” examples

Slide modified from Manning
May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire, is finding itself hard pressed to cope with the crisis...

<table>
<thead>
<tr>
<th>Date</th>
<th>Disease Name</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 1995</td>
<td>Malaria</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>July 1995</td>
<td>Mad Cow Disease</td>
<td>U.K.</td>
</tr>
<tr>
<td>Feb. 1995</td>
<td>Pneumonia</td>
<td>U.S.</td>
</tr>
</tbody>
</table>
We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

Slide from Manning
John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tuesday, May 29, 1917, at 3:00 pm, the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; Rose, in turn, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political figure who was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline for ten years and attended Edward Devotion School, Noble and Greenough Lower School, and the Dexter School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, New York, where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become President). Kennedy spent summers with his family at their home in Hyannisport, Massachusetts, and Christmas and Easter holidays with his family at their winter home in Palm Beach, Florida. For the 5th through 7th grade, Kennedy attended Riverdale Country School, a private school for boys. For 8th grade in September 1930, the 13-year old Kennedy attended Canterbury School in New Milford, Connecticut.
### Rough Accuracy of Information Extraction

<table>
<thead>
<tr>
<th>Information type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>90-98%</td>
</tr>
<tr>
<td>Attributes</td>
<td>80%</td>
</tr>
<tr>
<td>Relations</td>
<td>60-70%</td>
</tr>
<tr>
<td>Events</td>
<td>50-60%</td>
</tr>
</tbody>
</table>

- Errors cascade (error in entity tag $\rightarrow$ error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks
What we will cover in this class (briefly)

• PART I: basic information extraction (through Named Entity Recognition)
  • History of IE, Related Fields
  • Source Selection
  • Tokenization and Normalization
  • Named Entity Recognition (NER)
What we will cover in this class (briefly)

• PART II: machine learning in depth (mostly tagging models used for named entities)
  • Decision Trees and Overfitting
  • Linear Models
  • Feature Engineering
  • Word Embeddings
  • Deep Learning (Non-Linear Models)
What we will cover in this class (briefly)

- PART III: advanced information extraction
  - Instance Extraction
  - Fact/Event Extraction
  - Ontological IE/Open IE
Last words

• The seminar tomorrow is cancelled
• The topics will be presented on November 3rd
• Also, don't forget the reading for next week!
• Sarawagi: Information Extraction (available from web page)
  Read the introduction!
• These slides will be uploaded as well
  • The video of this lecture from WS 2020-2021 is available, note that the "Administravia" is different
• Thank you for your attention!