Information Extraction

• Information extraction (IE) systems
  • Find and understand limited relevant parts of texts
  • Gather information from many pieces of text
  • Produce a structured representation of relevant information:
    • relations (in the database sense), a.k.a.,
    • a knowledge base
• Goals:
  1. Organize information so that it is useful to people
  2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

• IE systems extract clear, factual information
  • Roughly: Who did what to whom when?
  • E.g.,
    • Gathering earnings, profits, board members, headquarters, etc. from company reports
    • The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    • headquarters("BHP Billiton Limited", "Melbourne, Australia")
  • Learn drug-gene product interactions from medical research literature

Low-level information extraction

• Is now available – and I think popular – in applications like Apple or Google mail, and web indexing

The Los Alamos Robotics Board of Directors is having a potluck dinner Friday, January 6, 2012, and PBC (MYTH) 2012. You are back and it was a...
Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Evaluation of Named Entity Recognition

- The uses:
  - Named entities can be indexed, linked off, etc.
  - Sentiment can be attributed to companies or products
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.

- Concretely:
  - Many web pages tag various entities, with links to bio or topic pages, etc.
  - Apple/Google/Microsoft/… smart recognizers for document content
  - Dialogue systems, like Alexa, Google Home, etc

The Named Entity Recognition Task

Task: Predict entities in a text

<table>
<thead>
<tr>
<th>Person</th>
<th>Date</th>
<th>Location</th>
<th>Organization</th>
</tr>
</thead>
</table>

- Standard evaluation is per entity, not per token

- Foreign ORG
- Ministry ORG
- spokesman O
- Shen PER
- Guofang PER
- told O
- Reuters ORG
Precision/Recall/F1 for IE/NER

• Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
• The measure behaves a bit funny for IE/NER when there are boundary errors (which are common):
  • First Bank of Chicago announced earnings ...

The ML sequence model approach to NER

Training
1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing
1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Sequence Models for Named Entity Recognition

Encoding classes for sequence labeling

<table>
<thead>
<tr>
<th>ID encoding</th>
<th>IOB encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred</td>
<td>PER</td>
</tr>
<tr>
<td>showed</td>
<td>O</td>
</tr>
<tr>
<td>Sue</td>
<td>PER</td>
</tr>
<tr>
<td>Mengjia</td>
<td>PER</td>
</tr>
<tr>
<td>Huang</td>
<td>PER</td>
</tr>
<tr>
<td>'s</td>
<td>O</td>
</tr>
<tr>
<td>new</td>
<td>O</td>
</tr>
<tr>
<td>painting</td>
<td>O</td>
</tr>
</tbody>
</table>

Features for sequence labeling

• Words
  • Current word (essentially like a learned dictionary)
  • Other kinds of inferred linguistic classification
  • Part-of-speech tags
• Label context
  • Previous (and perhaps next) label
Features: Word substrings
- Drug
- company
- movie
- place
- person

Cotrimoxazole
Wethersfield
Alien Fury: Countdown to Invasion

Features: Word shapes
- Word Shapes
- Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster
Xx-xxx
mRNA
XXX
CPA1
XXd

Maximum Entropy Sequence Models
- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences...
- We can think of our task as one of labeling each item

Sequence problems
- Maximum Entropy Sequence Models
- MEMM inference in systems
- For a Conditional Markov Model (CMM) a.k.a. Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions

<table>
<thead>
<tr>
<th>Local Context</th>
<th>Decision Point</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Dow</td>
<td>22.6</td>
<td></td>
</tr>
<tr>
<td>News</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drug</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)
Example: POS Tagging

- Scoring individual labeling decisions is no more complex than standard classification decisions
- We have some assumed labels to use for prior positions
- We use features of those and the observed data (which can include current, previous, and next words) to predict the current label

<table>
<thead>
<tr>
<th>Local Context</th>
<th>Decision Point</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 1 0</td>
<td>DT NNP VBD JJJ</td>
<td>??</td>
</tr>
<tr>
<td>The Dow fell 22.6%</td>
<td></td>
<td>??</td>
</tr>
</tbody>
</table>

Example: POS Tagging

- POS tagging features can include:
  - Current, previous, next words in isolation or together.
  - Previous one, two, three tags.
  - Word-internal features: word types, suffixes, dashes, etc.

<table>
<thead>
<tr>
<th>Local Context</th>
<th>Decision Point</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1 0</td>
<td>DT NNP VBD JJJ</td>
<td>??</td>
</tr>
<tr>
<td>The Dow fell 22.6%</td>
<td></td>
<td>??</td>
</tr>
</tbody>
</table>

Inference in Systems

- Sequence Data → Sequence Level → Sequence Model → Inference
- Local Level

Greedy Inference

- Greedy inference:
  - We just start at the left, and use our classifier at each position to assign a label.
  - The classifier can depend on previous labeling decisions as well as observed data.
- Advantages:
  - Fast, no extra memory requirements
  - Very easy to implement
  - With rich features including observations to the right, it may perform quite well.
- Disadvantage:
  - Greedy, we make commit errors we cannot recover from.
Beam Inference

- Beam inference:
  - At each position keep the top $k$ complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the $k$ slots at the next position.

- Advantages:
  - Fast: beam sizes of 3–5 are almost as good as exact inference in many cases.
  - Easy to implement (no dynamic programming required).

- Disadvantage:
  - Inexact: the globally best sequence can fall off the beam.

Advantages:
- Fast; beam sizes of 3–5 are almost as good as exact inference in many cases.
- Easy to implement (no dynamic programming required).

Disadvantage:
- Inexact: the globally best sequence can fall off the beam.

Viterbi Inference

- Viterbi inference:
  - Dynamic programming or memoization.
  - Requires small window of state influence (e.g., past two states are relevant).

- Advantage:
  - Exact: the global best sequence is returned.

- Disadvantage:
  - Harder to implement long-distance state-state interactions (but beam inference tends not to allow long-distance resurrection of sequences anyway).

Relation Extraction

- Extracting relations from text
  - Company report: "International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)."
  - Extracted Complex Relation:
    Company-Founding
    Company: IBM
    Location: New York
    Date: June 16, 1911
    Original-Name: Computing-Tabulating-Recording Co.
  - But we will focus on the simpler task of extracting relation triples
    Founding-year(IBM,1911)
    Founding-location(IBM,New York)

Why Relation Extraction?

- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support question answering
  - The granddaughter of which actor starred in the movie "E.T."
    (acted-in ?x "E.T."):(is-a ?x actor):(granddaughter-of ?x ?y)
- But which relations should we extract?

Automated Content Extraction (ACE)

17 relations from 2008 "Relation Extraction Task"
Automated Content Extraction (ACE)

• Physical-Located PER-GPE
  He was in Tennessee
• Part-Whole-Subsidiary ORG-ORG
  XYZ, the parent company of ABC
• Person-Social-Family PER-PER
  John’s wife Yoko
• Org-AFF-Founder PER-ORG
  Steve Jobs, co-founder of Apple...

UMLS: Unified Medical Language System

• 134 entity types, 54 relations

<table>
<thead>
<tr>
<th>Injury</th>
<th>disrupts</th>
<th>Physiological Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily Location</td>
<td>location-of</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>Anatomical Structure</td>
<td>part-of</td>
<td>Organism</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>causes</td>
<td>Pathological Function</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>treats</td>
<td>Pathologic Function</td>
</tr>
</tbody>
</table>

Extracting UMLS relations from a sentence

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

Echocardiography, Doppler DIAGNOSES Acquired stenosis

Databases of Wikipedia Relations

Relation databases that draw from Wikipedia

• Resource Description Framework (RDF) triples
  subject predicate object
  Golden Gate Park location San Francisco
dbpedia:Golden_Gate_Park dbpedia-owl:location dbpedia:San_Francisco
• DBPedia: 1 billion RDF triples, 385 from English Wikipedia
• Frequent Freebase relations:
  people/person/nationality, people/person/profession,
  biology/organism_higher_classification film/film/genre

Ontological relations

Examples from the WordNet Thesaurus

• IS-A (hypernym): subsumption between classes
  • Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal...
• Instance-of: relation between individual and class
  • San Francisco instance-of city
How to build relation extractors

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
   - Bootstrapping (using seeds)
   - Distant supervision
   - Unsupervised learning from the web

Rules for extracting IS-A relation

Early intuition from Hearst (1992)
- "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use"
  - What does Gelidium mean?
  - How do you know?

Hearst’s Patterns for extracting IS-A relations

(Hearst, 1992): Automatic Acquisition of Hyponyms

- "Y such as X ((, X)* (, and|or) X)"
- "such Y as X"
- "X or other Y"
- "X and other Y"
- "Y including X"
- "Y, especially X"

Hand-written Patterns

Rules for extracting IS-A relation

Early intuition from Hearst (1992)
- "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use"
  - What does Gelidium mean?
  - How do you know?

Hearst’s Patterns for extracting IS-A relations

(Hearst, 1992): Automatic Acquisition of Hyponyms

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td>such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
Extracting Richer Relations Using Rules

• Intuition: relations often hold between specific entities
  • located-in (ORGANIZATION, LOCATION)
  • founded (PERSON, ORGANIZATION)
  • cures (DRUG, DISEASE)
  • Start with Named Entity tags to help extract relation!

Named Entities aren’t quite enough. Which relations hold between 2 entities?

Drug
Cure?
Prevent?
Cause?
Disease

What relations hold between 2 entities?

Founder?
Investor?
Member?
Employee?
President?

Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what organization?

PERSON, POSITION of ORG
  • George Marshall, Secretary of State of the United States
PERSON(named|appointed|chose|etc.) PERSON Prep? POSITION
  • Truman appointed Marshall Secretary of State
PERSON(be)\?(named|appointed|etc.) Prep? ORG POSITION
  • George Marshall was named US Secretary of State

Hand-built patterns for relations

• Plus:
  • Human patterns tend to be high-precision
  • Can be tailored to specific domains
• Minus
  • Human patterns are often low-recall
  • A lot of work to think of all possible patterns!
  • Don’t want to have to do this for every relation!
  • We’d like better accuracy

Supervised machine learning for relations

• Choose a set of relations we’d like to extract
• Choose a set of relevant named entities
• Find and label data
  • Choose a representative corpus
  • Label the named entities in the corpus
  • Hand-label the relations between these entities
  • Break into training, development, and test
• Train a classifier on the training set
How to do classification in supervised relation extraction

1. Find all pairs of named entities (usually in same sentence)
2. Decide if 2 entities are related
3. If yes, classify the relation
   • Why the extra step?
     • Faster classification training by eliminating most pairs
     • Can use distinct feature-sets appropriate for each task.

Relation Extraction
Classify the relation between two entities in a sentence

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Word Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Parse Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Named Entity Type and Mention Level
Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Parse Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Word Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Named Entity Type and Mention Level
Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Parse Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
Gazeteer and trigger word features for relation extraction

- Trigger list for family: kinship terms
  - parent, wife, husband, grandparent, etc. (from WordNet)
- Gazeteer:
  - Lists of useful geo or geopolitical words
    - Country name list
    - Other sub-entities

Classifiers for supervised methods

- Now you can use any classifier you like
  - MaxEnt
  - Naïve Bayes
  - SVM
  - ...
- Train it on the training set, tune on the dev set, test on the test set

Summary: Supervised Relation Extraction

- Can get high accuracies with enough hand-labeled training data, if test similar enough to training
- Labeling a large training set is expensive
- Supervised models are brittle, don’t generalize well to different genres
Semi-supervised and Unsupervised Relation Extraction

Seed-based or bootstrapping approaches to relation extraction

• No training set? Maybe you have:
  • A few seed tuples or
  • A few high-precision patterns

• Can you use those seeds to do something useful?
  • Bootstrapping: use the seeds to directly learn to populate a relation

Relation Bootstrapping (Hearst 1992)

• Gather a set of seed pairs that have relation \( R \)
• Iterate:
  1. Find sentences with these pairs
  2. Look at the context between or around the pair and generalize the context to create patterns
  3. Use the patterns for grep for more pairs

Bootstrapping

• \(<\text{Mark Twain}, \text{Elmira}>\) Seed tuple
  • Grep (google) for the environments of the seed tuple
    “Mark Twain is buried in Elmira, NY.”
    X is buried in Y
    “The grave of Mark Twain is in Elmira.”
    The grave of X is in Y
  • “Elmira is Mark Twain’s final resting place”
    Y is X’s final resting place.

• Use those patterns to grep for new tuples
• Iterate

Dipre: Extract \(<\text{author}, \text{book}>\) pairs

• Start with 5 seeds:

<table>
<thead>
<tr>
<th>Author</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isaac Asimov</td>
<td>The Robots of Dawn</td>
</tr>
<tr>
<td>David Brin</td>
<td>Startide Rising</td>
</tr>
<tr>
<td>James Gleick</td>
<td>Chaos: Making a New Science</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td>Great Expectations</td>
</tr>
<tr>
<td>William Shakespeare</td>
<td>The Comedy of Errors</td>
</tr>
</tbody>
</table>

• Find Instances:
  The Comedy of Errors, by William Shakespeare, was
  The Comedy of Errors, one of William Shakespeare’s earliest attempts
  The Comedy of Errors, one of William Shakespeare’s most

• Extract patterns (group by middle, take longest common prefix/suffix)
  \( x \text{ by } y \), \( x \text{, one of } y \text{’s} \)

• Now iterate, finding new seeds that match the pattern

Snowball: Extracting Relations from Large Plain-Text Collections.

• Similar iterative algorithm

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location of Headquarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>Redmond</td>
</tr>
<tr>
<td>Exxon</td>
<td>Irving</td>
</tr>
<tr>
<td>IBM</td>
<td>Armonk</td>
</tr>
</tbody>
</table>

• Group instances w/similar prefix, middle, suffix, extract patterns
  • But require that X and Y be named entities
  • And compute a confidence for each pattern

.69 \( \text{[I’s, in, headquarters]} \) LOCATION
.75 LOCATION \( \text{[is, based]} \) ORGANIZATION
### Distant Supervision

-Distinctly labeled learning of relation extraction patterns

1. For each relation: **Born-in**
   - \(<\text{Edwin Hubble}, \text{Marshfield}>)
   - \(<\text{Albert Einstein}, \text{Ulm}>\)

2. For each tuple in big database
   - Hubble was born in Marshfield
   - Einstein, born (1879), Ulm
   - Hubble’s birthplace in Marshfield

3. Find sentences in large corpus with both entities
   - `PER was born in LOC`
   - `PER, born (XXXX), LOC`
   - `PER’s birthplace in LOC`

4. Extract frequent features (parse, words, etc)
   - `P(born-in 1, f_1, f_2, ... f_70000)`

5. Train supervised classifier using thousands of patterns

---

### Distant supervision paradigm

- Like supervised classification:
  - Uses a classifier with lots of features
  - Supervised by detailed hand-created knowledge
  - Doesn’t require iteratively expanding patterns

- Like unsupervised classification:
  - Uses very large amounts of unlabeled data
  - Not sensitive to genre issues in training corpus