Information Extraction: I
Predetermined Relations

David Israel
SRI (Emeritus)
Sapienza (Visiting)
The BIG Vision of Machine Reading
DARPA’s Vision

• “Building the Universal Text-to-Knowledge Engine”
• “A universal engine that captures knowledge from naturally occurring text and transforms it into of the formal representations used by AI reasoning systems.”
• “Machine Reading is the Revolution that will bridge the gap between textual and formal knowledge.”
• That is how the Program Manager for DARPA’s Machine Reading Program described the goal of the program – both to us researchers and to his superiors at DARPA.
Information Extraction:
The Input

• A Much **MUCH** More Limited Goal
  – At least at the beginning of work in IE – late ‘80s, early 90’s

• Fix a single kind of state or event, or a very small set of these, of the type described in new stories
  – Terrorist Attacks
  – Joint Ventures
  – Management changes at companies
IE: The Output

• Design a DB schema or a DB-like Attribute-Value template
  – Some fields (attributes) come with a fixed set of alternative value-options
    • simplest case: **Yes, No**, or with
    • constraints on values (fillers)
      – E..g, filler must be a date in : 2-digit day month (abbrev) 4-digit year format: 1 Dec 2015

• Automatically populate that DB or those templates
IE vs. TU

• In the original conception of IE: predetermined set of relations/events of interest (“closed domain”)
• Only a fraction of a text was likely to be relevant and some texts were irrelevant
• Output format was highly constrained
• Only interested in a subset of facts relevant to the focal relations/events: those representable within the fixed output-schema
  – SO, if there is no column or slot for people’s ages, then ignore their ages
San Salvador, 19 Apr 89 (ACAN-EFE) -- [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

... 
Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

... 
Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

... 
According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.
### Example: Filled A-V Template

<table>
<thead>
<tr>
<th>Incident: Date</th>
<th>19 Apr 89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incident: Location</td>
<td>El Salvador: San Salvador (CITY)</td>
</tr>
<tr>
<td>Incident: Type</td>
<td>Bombing</td>
</tr>
<tr>
<td>Perpetrator: Individual ID</td>
<td>urban guerrillas</td>
</tr>
<tr>
<td>Perpetrator: Organization ID</td>
<td>FMLN</td>
</tr>
<tr>
<td>Perpetrator: Confidence</td>
<td>Suspected or Accused by Authorities: FMLN</td>
</tr>
<tr>
<td>Physical Target: Description</td>
<td>vehicle</td>
</tr>
<tr>
<td>Physical Target: Effect</td>
<td>Some Damage: vehicle</td>
</tr>
<tr>
<td>Human Target: Name</td>
<td>Roberto Garcia Alvarado</td>
</tr>
<tr>
<td>Human Target: Description</td>
<td>attorney general: Roberto Garcia Alvarado driver bodyguards</td>
</tr>
<tr>
<td>Human Target: Effect</td>
<td>Death: Roberto Garcia Alvarado No Injury: driver Injury: bodyguards</td>
</tr>
</tbody>
</table>
Another Example: The Text

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.
The Output

<table>
<thead>
<tr>
<th>TIE-UP-1:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relationship:</strong></td>
<td>TIE-UP</td>
</tr>
<tr>
<td><strong>Entities:</strong></td>
<td>Bridgestone Sports Co.</td>
</tr>
<tr>
<td></td>
<td>a local concern</td>
</tr>
<tr>
<td></td>
<td>a Japanese trading house</td>
</tr>
<tr>
<td><strong>Joint Venture Company:</strong></td>
<td>Bridgestone Sports Taiwan Co.</td>
</tr>
<tr>
<td><strong>Activity:</strong></td>
<td>ACTIVITY-1</td>
</tr>
<tr>
<td><strong>Amount:</strong></td>
<td>NT$20000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVITY-1:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity:</strong></td>
<td>PRODUCTION</td>
</tr>
<tr>
<td><strong>Company:</strong></td>
<td>Bridgestone Sports Taiwan Co.</td>
</tr>
<tr>
<td><strong>Product</strong></td>
<td>iron and `metal wood' clubs</td>
</tr>
<tr>
<td><strong>Start Date:</strong></td>
<td>DURING: January 1990</td>
</tr>
</tbody>
</table>
Applications of IE

• Automatic population of KB’s, DB’s with ground atomic facts
• First and foremost: Assistance to analysts
  – Intelligence analysis
  – Competitive and market analysis
  – Directories of people and organizations
  – Various tasks in bioinformatics
  – Shopping db’s
  – Stock analysis
  – Etc., etc.
**Who** did **What** to **Whom/What** **Where** and **When**?

- Note the centrality of proper names, in an extended sense, in news stories

- Names of
  - Dates
  - Places
  - People
    - Including Titles
  - Organizations
Named Entity Tagging

- [LOC San Salvador], [TIM 19 Apr 89] (ACAN-EFE) -- [TEXT] [PER [LOC Salvadoran] [TITLE President-elect] Alfredo Cristiani] condemned the terrorist killing of [PER [TITLE Attorney General] Roberto Garcia Alvarado] and accused the [ORG [PER Farabundo Martí] National Liberation Front (FMLN)] of the crime.

- ...

- [PER Garcia Alvarado], [AGE 56], was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown [LOC San Salvador].

- ...

- [PER [TITLE Vice President-elect] Francisco Merino] said that when the attorney general's car stopped at a light on a street in downtown [LOC San Salvador], an individual placed a bomb on the roof of the armored vehicle.

- ...

- According to the police and [PER Garcia Alvarado]'s driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.
Our 2\textsuperscript{nd} Example

• [\textbf{ORG} Bridgestone Sports Co.] said [\textbf{TIM} Friday] it has set up a joint venture in [\textbf{LOC} Taiwan] with a local concern and a [\textbf{LOC} Japanese] trading house to produce golf clubs to be shipped to [\textbf{LOC} Japan].

• The joint venture, [\textbf{ORG} Bridgestone Sports [\textbf{LOC} Taiwan] Co.], capitalized at [\textbf{NUM} 20 million new [\textbf{LOC} Taiwan] dollars], will start production in [\textbf{TIM} January 1990] with production of [\textbf{NUM} 20,000 iron and "metal wood" clubs] a month.
What’s in a Name?

• Think about … dates
  – Simple rules for describing/generating
  – 7 days in a week, all named
  – 12 months, all named
  – Years, all numbered
  – Same goes for clock times
    • Two options: 12-hr clock / 24-hr. clock
• Now think about People’s Names and Organization Names
• Much more complex than dates & times, but
• Everything is pretty local!
• No long-distance dependencies! No memory!
• Finite-state methods seem to work well enough
One Slide on Co-Reference ?!? 

- We have been talking about “mentions” of named-entities: particular occurrences of names
- But, of course, some of these are co-referential
- Multiple approaches: sub-string identity, sameness of category-tag, etc.
- My current favorite?
- Stanford’s Multi-Pass Sieve Approach
- Check it out at: http://nlp.stanford.edu/software/dcoref.shtml
Inspiration for FASTUS

• The success of finite-state methods at Name Recognition
• Given the centrality of NER in the IE task
• Why not go all the way and use only finite-state methods for the whole IE task?
• Design of FASTUS: a “cascade” – sequence – of non-deterministic finite-state transducers
  – (Mealy machines) FSMs that produce output based on input and state
;;; For <company> appoints <person> <position>

(defpattern appoint
  "np-sem(C-company)? rn? sa? vg(C-appoint) np-sem(C-person) \',\'? to-be? np(C-position) to-succeed?::
  company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes
  position-at=8.attributes |
...

(defun when-appoint (phrase-type)
  (let ((person-at (binding 'person-at))
    (company-entity (entity-bound 'company-at))
    (person-entity (essential-entity-bound 'person-at 'C-person))
    (position-entity (entity-bound 'position-at))
    (predecessor-entity (entity-bound 'predecessor-at))
    (new-event)
    (not-an-antecedent position-entity)
    ;; if no company is specified for position, use agent
  ...
  )
Moving on ...

- IE went from one or two event-types, to multiple relation types:
- From yet another government-funded and run-competition (ACE: Automatic Content Extraction)
  - **Role**: relates a person to an organization or a geopolitical entity. Subtypes: member, owner, affiliate, client, citizen
  - **PART**: generalized containment. Subtypes: subsidiary, physical part-of, set membership
  - **AT**: permanent and transient locations. Subtypes: located, based-in, residence
  - **SOCIAL**: social relations among persons. Subtypes: parent, sibling, spouse, grandparent, associate
And then along came Bio-informatics, etc., etc.

• Five basic Methods for dealing with more challenging coverage requirements
  – 1. Hand-built patterns meant for “full coverage”
    • Time-consuming for each relation
    • And now there are lots and lots of relations!
  – 2. Bootstrapping Methods
  – 3. Supervised Methods
  – 4. Distant Supervision
  – 5. Unsupervised Methods
Bootstrapping Methods

• Appropriate when (i) not enough annotated data, but lots of unannotated data and (ii) you have a small seed set of either rules/patterns or even just relation instances (usually binary)
• Target Relation: burial place
• Seed tuple [Mark Twain, Elmira]
• Grep/Google for “Mark Twain” and for “Elmira”, and return:
  • “Mark Twain is buried in Elmira, NY”
    – X is buried in Y
  • “The grave of Mark Twain is in Elmira.”
    – The grave of X / X’s grave is in Y
  • “Elmira is Mark Twain’s final resting place.”
    – Y is X’s (final) resting place
• Use those patterns to search for new tuples
Pattern-Based Relation Extraction

A Picture
Problems with Bootstrapping

• Requires seeds for each relation – and there may be lots!
  – Also, results are sensitive to quality of original seeds
  – And to tweaking & tuning of lots of parameters
• Problem of drift at each iteration – hard to anticipate
• Precision tends to be low
• No probabilistic interpretation or account
  – So hard to know how confident to be in results of each iteration or… their aggregate
Supervised Methods

• Define the task, via specification of output labels:
  – Binary for relation detection
  – Multiple for classification: located-in, employee-of, etc.
• Collect a corpus and annotate a sample
• Specify a feature representation
  – Words, entity types, etc., etc.
• Choose a classifier: Naïve Bayes, MaxEnt, SVM, ...
• Execute, Evaluate, ... Repeat!
Features

• Lightweight features – require little pre-processing
  – Bags-of-words; bigrams between, before and after entity-mentions
  – Stemmed version of the above
  – Entity-types
  – Distance (in words) between entity mentions

• Middleweight features – require base-phrase chunking (FSA)
  – Base-phrase paths
  – Bags of base-phrase heads

• Heavyweight features – require full parsing
  – Dependency-tree paths
  – Constituent-tree paths
  – Presence of particular constructions in constituent structure

• STRIKING FINDING: The heavyweight features had very little impact!
Summary on Supervised Methods for Relation Extraction

• Can achieve high *accuracy* – sometimes
  – If we have a lot of hand-labeled training data
• BUT Significant Limitations
  – Very little generalizability beyond a typically small pre-determined relation set
  – Doesn’t really scale, say to 100’s, let alone 1,000’s of relations
• So researchers moved toward
  – Distant (indirect/weak) supervision
  – Unsupervised relation extraction