Outline

- Task Definition
- Supervised Models
  - Basic Features
  - World Knowledge
  - Learning Models
- Joint Inference
- Semi-supervised Learning
- Domain-independent Relation Extraction
Relation Extraction: Task

relation: a semantic relationship between two entities

<table>
<thead>
<tr>
<th>ACE relation type</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent-Artifact</td>
<td>Rubin Military Design, the <em>makers</em> of the <em>Kursk</em></td>
</tr>
<tr>
<td>Discourse</td>
<td><em>each</em> of <em>whom</em></td>
</tr>
<tr>
<td>Employment/ Membership</td>
<td>Mr. Smith, a senior <em>programmer</em> at <em>Microsoft</em></td>
</tr>
<tr>
<td>Place-Affiliation</td>
<td><em>Salzburg</em> Red Cross <em>officials</em></td>
</tr>
<tr>
<td>Person-Social</td>
<td><em>relatives</em> of the <em>dead</em></td>
</tr>
<tr>
<td>Physical</td>
<td>a <em>town</em> some 50 miles south of <em>Salzburg</em></td>
</tr>
<tr>
<td>Other-Affiliation</td>
<td><em>Republican senators</em></td>
</tr>
</tbody>
</table>
A Simple Baseline with K-Nearest-Neighbor (KNN)

Train Sample

Test Sample

Train Sample

Train Sample

$K=3$
1. If the heads of the mentions don’t match: +8
2. If the entity types of the heads of the mentions don’t match: +20
3. If the intervening words don’t match: +10
Typical Relation Extraction Features

- **Lexical**
  - Heads of the mentions and their context words, POS tags

- **Entity**
  - Entity and mention type of the heads of the mentions
  - Entity Positional Structure
  - Entity Context

- **Syntactic**
  - Chunking
  - Premodifier, Possessive, Preposition, Formulaic
  - The sequence of the heads of the constituents, chunks between the two mentions
  - The syntactic relation path between the two mentions
  - Dependent words of the mentions

- **Semantic Gazetteers**
  - Synonyms in WordNet
  - Name Gazetteers
  - Personal Relative Trigger Word List

- **Wikipedia**
  - If the head extent of a mention is found (via simple string matching) in the predicted Wikipedia article of another mention

References: Kambhatla, 2004; Zhou et al., 2005; Jiang and Zhai, 2007; Chan and Roth, 2010, 2011
Using Background Knowledge (Chan and Roth, 2010)

• Features employed are usually restricted to being defined on the various representations of the target sentences
• Humans rely on background knowledge to recognize relations
• Overall aim of this work
  • Propose methods of using knowledge or resources that exists beyond the sentence
    • Wikipedia, word clusters, hierarchy of relations, entity type constraints, coreference
    • As additional features, or under the Constraint Conditional Model (CCM) framework with Integer Linear Programming (ILP)
David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
David Brian Cone (born January 2, 1963) is a former Major League Baseball pitcher. He compiled an 8–3 postseason record over 21 postseason starts and was a part of five World Series championship teams (1992 with the Toronto Blue Jays and 1996, 1998, 1999 & 2000 with the New York Yankees). He had a career postseason ERA of 3.80. He is the subject of the book A Pitcher’s Story: Innings With David Cone by Roger Angell. Fans of David are known as "Cone-Heads."

Cone lives in Stamford, Connecticut, and is formerly a color commentator for the Yankees on the YES Network.[1]

Contents
[hide]
1 Early years
2 Kansas City Royals
3 New York Mets

Partly because of the resulting lack of leadership, after the 1994 season the Royals decided to reduce payroll by trading pitcher David Cone and outfielder Brian McRae, then continued their salary dump in the 1995 season. In fact, the team payroll, which was always among the league’s highest, was sliced in half from $40.5 million in 1994 (fourth-highest in the major leagues) to $18.5 million in 1996 (second-lowest in the major leagues)
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.

<table>
<thead>
<tr>
<th>细粒度</th>
<th>粗粒度</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment: Staff</td>
<td>0.20</td>
</tr>
<tr>
<td>Employment: Executive</td>
<td>0.15</td>
</tr>
<tr>
<td>Personal: Family</td>
<td>0.10</td>
</tr>
<tr>
<td>Personal: Business</td>
<td>0.10</td>
</tr>
<tr>
<td>Affiliation: Citizen</td>
<td>0.20</td>
</tr>
<tr>
<td>Affiliation: Based-in</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Using Background Knowledge

David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Knowledge\textsubscript{1} : Wikipedia\textsubscript{1}
(as additional feature)

- We use a Wikifier system (Ratinov et al., 2010) which performs context-sensitive mapping of mentions to Wikipedia pages
- Introduce a new feature based on:

\[
\begin{align*}
  w_i(m_i, m_j) &= \begin{cases} 
    1, & \text{if } A_{m_i}(m_j) \text{ or } A_{m_j}(m_i) \\
    0, & \text{otherwise}
  \end{cases} 
\end{align*}
\]
- Introduce a new feature by combining the above with the coarse-grained entity types of $m_i, m_j$
Knowledge\textsubscript{1}: Wikipedia\textsubscript{2} (as additional feature)

- Given $m_i, m_j$, we use a Parent-Child system (Do and Roth, 2010) to predict whether they have a parent-child relation.
- Introduce a new feature based on:
  \[ w_2(m_i, m_j) = \begin{cases} 
  1, & \text{if parent-child}(m_i, m_j) \\
  0, & \text{otherwise}
\end{cases} \]
- Combine the above with the coarse-grained entity types of $m_i, m_j$. 
Knowledge\textsubscript{2}: Word Class Information
(as additional feature)

- Supervised systems face an issue of data sparseness (of lexical features)
- Use class information of words to support generalization better: instantiated as word clusters in our work
  - Automatically generated from unlabeled texts using algorithm of (Brown et al., 1992)
Knowledge$_2$: Word Class Information

- Supervised systems face an issue of data sparseness (of lexical features)
- Use class information of words to support generalization better: instantiated as word clusters in our work
  - Automatically generated from unlabeled texts using algorithm of (Brown et al., 1992)
Knowledge\textsubscript{2}: Word Class Information

- Supervised systems face an issue of data sparseness (of lexical features)
- Use class information of words to support generalization better: instantiated as word clusters in our work
  - Automatically generated from unlabeled texts using algorithm of (Brown et al., 1992)
Knowledge$_2$: Word Class Information

- All lexical features consisting of single words will be duplicated with its corresponding bit-string representation.
Constraint Conditional Models (CCMs)
(Roth and Yih, 2007; Chang et al., 2008)

\[ \arg\max_{y} \lambda \cdot F'(x, y) \]

weight vector for “local” models

collection of classifiers
Constraint Conditional Models (CCMs)
(Roth and Yih, 2007; Chang et al., 2008)

\[
\arg\max_y \lambda \cdot F(x, y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)})
\]

- **Weight vector for “local” models**
- **Collection of classifiers**
- **Penalty for violating the constraint**
- **How far y is from a “legal” assignment**
Constraint Conditional Models (CCMs)

(Roth and Yih, 2007; Chang et al., 2008)

\[ \argmax_y \lambda \cdot F(x, y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)}) \]

- Wikipedia
- word clusters
- hierarchy of relations
- entity type constraints
- coreference
David Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.
Constraint Conditional Models (CCMs)
(Roth and Yih, 2007; Chang et al., 2008)

- Key steps
  - Write down a linear objective function
  - Write down constraints as linear inequalities
  - Solve using integer linear programming (ILP) packages
Knowledge$_3$: Relations between our target relations

- personal
  - family
  - biz
  - ...

- employment
  - staff
  - executive
  - ...
  - ...

...
Knowledge$_3$: Hierarchy of Relations

```
personal  employment  
family    biz        staff    executive  ...
```

coarse-grained classifier

fine-grained classifier
Knowledge_3: Hierarchy of Relations
Knowledge$_3$: Hierarchy of Relations

- **personal**
  - family
  - biz
- **employment**
  - staff
  - executive
Knowledge_3: Hierarchy of Relations

- personal
  - family
  - biz
  - ...

- employment
  - staff
  - executive
  - ...

- ...
  - ...
  - ...
  - ...
Knowledge\textsubscript{3}: Hierarchy of Relations

- personal
- employment
  - family
  - biz
  - staff
  - executive
  - ...
  - ...
  - ...
  - ...
  - ...
Knowledge$_3$: Hierarchy of Relations

- personal
- employment
- family
- biz
- staff
- executive
- ...
Knowledge$_3$: Hierarchy of Relations

```
  personal
    ├── family
    │    └── biz
    │            ...  
  employment
    ├── staff
    │    └── executive
    │            ...  
    └── ...  
```
Knowledge_3: Hierarchy of Relations

Write down a linear objective function

$$\max \sum_{R \in R} \sum_{rc \in L_{Rc}} p_R(rc) \cdot x_{R,rc} + \sum_{R \in R} \sum_{rf \in L_{Rf}} p_R(rf) \cdot y_{R,rf}$$

coarse-grained prediction probabilities
fine-grained prediction probabilities
Knowledge\textsubscript{3}: Hierarchy of Relations

Write down a linear objective function

\[
\max \sum_{R \in R} \sum_{rc \in L_{Rc}} p_R(rc) \cdot x_{\langle R,rc \rangle} + \sum_{R \in R} \sum_{rf \in L_{Rf}} p_R(rf) \cdot y_{\langle R,rf \rangle}
\]

coarse-grained prediction probabilities

coarse-grained indicator variable

fine-grained prediction probabilities

fine-grained indicator variable

indicator variable \(==\) relation assignment
Knowledge 3: Hierarchy of Relations

Write down constraints

- If a relation $R$ is assigned a coarse-grained label $rc$, then we must also assign to $R$ a fine-grained relation $rf$ which is a child of $rc$.

\[ x_{\langle R,rc \rangle} \Rightarrow y_{\langle R,rf_1 \rangle} \lor y_{\langle R,rf_2 \rangle} \cdots \lor y_{\langle R,rf_n \rangle} \]

- (Capturing the inverse relationship) If we assign $rf$ to $R$, then we must also assign to $R$ the parent of $rf$, which is a corresponding coarse-grained label

\[ y_{\langle R,rf \rangle} \Rightarrow x_{\langle R,parent(rf) \rangle} \]
Knowledge\textsubscript{4}: Entity Type Constraints

(\textit{Roth and Yih, 2004, 2007})

- Entity types are useful for constraining the possible labels that a relation $R$ can assume.

\begin{itemize}
  \item Employment:Staff
  \item Employment:Executive
  \item Personal:Family
  \item Personal:Business
  \item Affiliation:Citizen
  \item Affiliation:Based-in
\end{itemize}
Knowledge$_4$: Entity Type Constraints

(Roth and Yih, 2004, 2007)

- Entity types are useful for constraining the possible labels that a relation $R$ can assume.
**Knowledge$^4$: Entity Type Constraints**

(Roth and Yih, 2004, 2007)

• We gather information on entity type constraints from ACE-2004 documentation and impose them on the coarse-grained relations
  • By improving the coarse-grained predictions and combining with the hierarchical constraints defined earlier, the improvements would propagate to the fine-grained predications
Knowledge$_5$: Coreference

$m_i$

- Employment: Staff
- Employment: Executive
- Personal: Family
- Personal: Business
- Affiliation: Citizen
- Affiliation: Based-in

$m_j$
Knowledge$_5$: Coreference

• In this work, we assume that we are given the coreference information, which is available from the ACE annotation.
Experiment Results

<table>
<thead>
<tr>
<th></th>
<th>All nwire</th>
<th>10% of nwire</th>
</tr>
</thead>
<tbody>
<tr>
<td>BasicRE</td>
<td>50.5%</td>
<td>31.0%</td>
</tr>
</tbody>
</table>

F1% improvement from using each knowledge source
Most Successful Learning Methods: Kernel-based

- Consider different levels of syntactic information
  - Deep processing of text produces structural but less reliable results
  - Simple surface information is less structural, but more reliable

- Generalization of feature-based solutions
  - A kernel (kernel function) defines a similarity metric $\Psi(x, y)$ on objects
  - No need for enumeration of features

- Efficient extension of normal features into high-order spaces
  - Possible to solve linearly non-separable problem in a higher order space

- Nice combination properties
  - Closed under linear combination
  - Closed under polynomial extension
  - Closed under direct sum/product on different domains

- References: Zelenko et al., 2002, 2003; Aron Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhao and Grishman, 2005; Che et al., 2005; Zhang et al., 2006; Qian et al., 2007; Zhou et al., 2007; Khayyamian et al., 2009; Reichartz et al., 2009
Kernel Examples for Relation Extraction

1) Argument

\[ \psi_1(R_1, R_2) = \sum_{i=1,2} K_E(R_1 \text{.arg}_i, R_2 \text{.arg}_i), \]

where

\[ K_E(E_1, E_2) = K_T(E_1 \text{.tk}, E_2 \text{.tk}) + I(E_1 \text{.type}, E_2 \text{.type}) + I(E_1 \text{.subtype}, E_2 \text{.subtype}) + I(E_1 \text{.role}, E_2 \text{.role}) \]

\( K_T \) is a token kernel defined as:

\[ K_T(T_1, T_2) = I(T_1 \text{.word}, T_2 \text{.word}) + I(T_1 \text{.pos}, T_2 \text{.pos}) + I(T_1 \text{.base}, T_2 \text{.base}) \]

2) Local dependency

\[ \psi_2(R_1, R_2) = \sum_{i=1,2} K_D(R_1 \text{.arg}_i \text{.dseq}, R_2 \text{.arg}_i \text{.dseq}), \]

where

\[ K_D(dseq, dseq') = \sum_{0 \leq i < dseq \text{.len}} \sum_{0 \leq j < dseq' \text{.len}} (I(\text{arc}_i \text{.label}, \text{arc'}_j \text{.label}) + K_T(\text{arc}_i \text{.dw}, \text{arc'}_j \text{.dw})) \]

3) Path

\[ \psi_3(R_1, R_2) = K_{path}(R_1 \text{.path}, R_2 \text{.path}), \]

where

\[ K_{path}(path, path') = \sum_{0 \leq i < path \text{.len}} \sum_{0 \leq j < path' \text{.len}} (I(\text{arc}_i \text{.label}, \text{arc'}_j \text{.label}) + K_T(\text{arc}_i \text{.dw}, \text{arc'}_j \text{.dw})) \]

Composite Kernels:

\[ \Phi_1(R_1, R_2) = (\psi_1 + \psi_2) + (\psi_1 + \psi_2)^2 / 4 \]

(Zhao and Grishman, 2005)
Bootstrapping for Relation Extraction

Occurrences of seed tuples:

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MICROSOFT</td>
<td>REDMOND</td>
</tr>
<tr>
<td>IBM</td>
<td>ARMONK</td>
</tr>
<tr>
<td>BOEING</td>
<td>SEATTLE</td>
</tr>
<tr>
<td>INTEL</td>
<td>SANTA CLARA</td>
</tr>
</tbody>
</table>

Computer servers at Microsoft’s headquarters in Redmond...
In mid-afternoon trading, share of Redmond-based Microsoft fell...
The Armonk-based IBM introduced a new line...
The combined company will operate from Boeing’s headquarters in Seattle.
Intel, Santa Clara, cut prices of its Pentium processor.

Initial Seed Tuples

Generate New Seed Tuples

Augment Table

Generate Extraction Patterns
Learned Patterns:

- `<STRING1>`’s headquarters in `<STRING2>`
- `<STRING2>`-based `<STRING1>`
- `<STRING1>`, `<STRING2>`
Bootstrapping for Relation Extraction (Cont’)

<table>
<thead>
<tr>
<th>ORGANIZATION</th>
<th>LOCATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG EDWARDS</td>
<td>ST LUIS</td>
</tr>
<tr>
<td>157TH STREET</td>
<td>MANHATTAN</td>
</tr>
<tr>
<td>7TH LEVEL</td>
<td>RICHARDSON</td>
</tr>
<tr>
<td>3COM CORP</td>
<td>SANTA CLARA</td>
</tr>
<tr>
<td>3DO</td>
<td>REDWOOD CITY</td>
</tr>
<tr>
<td>JELLIES</td>
<td>APPLE</td>
</tr>
<tr>
<td>MACWEEK</td>
<td>SAN FRANCISCO</td>
</tr>
</tbody>
</table>

Generate new seed tuples; start new iteration
State-of-the-art and Remaining Challenges

- State-of-the-art: About 71% F-score on perfect mentions, and 50% F-score on system mentions

- Single human annotator: 84% F-score on perfect mentions

- Remaining Challenges
  - Context generalization to reduce data sparsity
    - Test: “ABC’s Sam Donaldson has recently been to Mexico to see him”
    - Training: PHY relation ( “arrived in”, “was traveling to”, …)
  - Long context
    - Davies is leaving to become chairman of the London School of Economics, one of the best-known parts of the University of London
  - Disambiguate fine-grained types
    - “U.S. citizens” and “U.S. businessman” indicate “GPE-AFF” relation while “U.S. president” indicates “EMP-ORG” relation
  - Parsing errors
Event Extraction

- Task Definition
- Basic Event Extraction Approach
- Advanced Event Extraction Approaches
  - Information Redundancy for Inference
  - Co-training
- Event Attribute Labeling
- Event Coreference Resolution
Event Mention Extraction: Task

- An event is a specific occurrence that implies a change of states
- **event trigger**: the main word which most clearly expresses an event occurrence
- **event arguments**: the mentions that are involved in an event (participants)
- **event mention**: a phrase or sentence within which an event is described, including trigger and arguments
- Automatic Content Extraction defined 8 types of events, with 33 subtypes

<table>
<thead>
<tr>
<th>ACE event type/subtype</th>
<th>Event Mention Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life/Die</td>
<td>Kurt Schork died in Sierra Leone yesterday</td>
</tr>
<tr>
<td>Transaction/Transfer</td>
<td>GM sold the company in Nov 1998 to LLC</td>
</tr>
<tr>
<td>Movement/Transport</td>
<td>Homeless people have been moved to schools</td>
</tr>
<tr>
<td>Business/Start-Org</td>
<td>Schweitzer founded a hospital in 1913</td>
</tr>
<tr>
<td>Conflict/Attack</td>
<td>the attack on Gaza killed 13</td>
</tr>
<tr>
<td>Contact/Meet</td>
<td>Arafat’s cabinet met for 4 hours</td>
</tr>
<tr>
<td>Personnel/Start-Position</td>
<td>She later recruited the nursing student</td>
</tr>
<tr>
<td>Justice/Arrest</td>
<td>Faison was wrongly arrested on suspicion of murder</td>
</tr>
</tbody>
</table>
Supervised Event Mention Extraction: Methods

- **Staged classifiers**
  - **Trigger Classifier**
    - to distinguish event instances from non-events, to classify event instances by type
  - **Argument Classifier**
    - to distinguish arguments from non-arguments
  - **Role Classifier**
    - to classify arguments by argument role
  - **Reportable-Event Classifier**
    - to determine whether there is a reportable event instance
- Can choose any supervised learning methods such as MaxEnt and SVMs

*(Ji and Grishman, 2008)*
Typical Event Mention Extraction Features

- **Trigger Labeling**
  - **Lexical**
    - Tokens and POS tags of candidate trigger and context words
  - **Dictionaries**
    - Trigger list, synonym gazetteers
  - **Syntactic**
    - the depth of the trigger in the parse tree
    - the path from the node of the trigger to the root in the parse tree
    - the phrase structure expanded by the parent node of the trigger
    - the phrase type of the trigger
  - **Entity**
    - the entity type of the syntactically nearest entity to the trigger in the parse tree
    - the entity type of the physically nearest entity to the trigger in the sentence

- **Argument Labeling**
  - **Event type and trigger**
    - Trigger tokens
    - Event type and subtype
  - **Entity**
    - Entity type and subtype
    - Head word of the entity mention
  - **Context**
    - Context words of the argument candidate
  - **Syntactic**
    - the phrase structure expanding the parent of the trigger
    - the relative position of the entity regarding to the trigger (before or after)
    - the minimal path from the entity to the trigger
    - the shortest length from the entity to the trigger in the parse tree

*(Chen and Ji, 2009)*
Why Trigger Labeling is so Hard?

- DT this “this is the largest pro-troops demonstration that has ever been in San Francisco”
- RP forward “We've had an absolutely terrific story, pushing forward north toward Baghdad”
- WP what “what happened in”
- RB back “his men back to their compound”
- IN over “his tenure at the United Nations is over”
- IN out “the state department is ordering all non-essential diplomats”
- CD nine eleven “nine eleven”
- RB formerly “McCarthy was formerly a top civil servant at”
Why Trigger Labeling is so Hard?

- A suicide bomber detonated explosives at the entrance to a crowded medical teams carting away dozens of wounded victims.
- dozens of Israeli tanks advanced into the northern Gaza Strip.
- Many nouns such as “death”, “deaths”, “blast”, “injuries” are missing.
Why Argument Labeling is so Hard?

- Two 13-year-old children were among those killed in the Haifa bus bombing, Israeli public radio said, adding that most of the victims were youngsters.
- Israeli forces staged a bloody raid into a refugee camp in central Gaza targeting a founding member of Hamas.
- Israel's night-time raid in Gaza involving around 40 tanks and armoured vehicles.
- Eight people, including a pregnant woman and a 13-year-old child were killed in Monday's Gaza raid.
- At least 19 people were killed and 114 people were wounded in Tuesday's southern Philippines airport.
- The waiting shed literally exploded.

Wikipedia “A shed is typically a simple, single-storey structure in a back garden or on an allotment that is used for storage, hobbies, or as a workshop.”
Why Argument Labeling is so Hard?

- Two 13-year-old children were among those killed in the Haifa bus bombing, Israeli public radio said, adding that most of the victims were youngsters.
- Fifteen people were killed and more than 30 wounded Wednesday as a suicide bomber blew himself up on a student bus in the northern town of Haifa.
- Two 13-year-old children were among those killed in the Haifa bus bombing.
State-of-the-art and Remaining Challenges

State-of-the-art Performance (F-score)

- English: Trigger 70%, Argument 45%
- Chinese: Trigger 68%, Argument 52%
- Single human annotator: Trigger 72%, Argument 62%

Remaining Challenges

- Trigger Identification
  - Generic verbs
  - Support verbs such as “take” and “get” which can only represent an event mention together with other verbs or nouns
  - Nouns and adjectives based triggers

- Trigger Classification
  - “named” represents a “Personnel_Nominate” or “Personnel_Start-Position”?
  - “hacked to death” represents a “Life_Die” or “Conflict_Attack”?

- Argument Identification
  - Capture long contexts

- Argument Classification
  - Capture long contexts
  - Temporal roles

(Ji, 2009; Li et al., 2011)
IE in Rich Contexts

Information Networks

Human Collaborative Learning
Capture Information Redundancy

• When the data grows beyond some certain size, IE task is naturally embedded in rich contexts; the extracted facts become inter-dependent

• Leverage Information Redundancy from:
  - Large Scale Data (Chen and Ji, 2011)
  - Background Knowledge (Chan and Roth, 2010; Rahman and Ng, 2011)
  - Inter-connected facts (Li and Ji, 2011; Li et al., 2011; e.g. Roth and Yih, 2004; Gupta and Ji, 2009; Liao and Grishman, 2010; Hong et al., 2011)
  - Diverse Documents (Downey et al., 2005; Yangarber, 2006; Patwardhan and Riloff, 2009; Mann, 2007; Ji and Grishman, 2008)
  - Diverse Systems (Tamang and Ji, 2011)
  - Diverse Languages (Snover et al., 2011)
  - Diverse Data Modalities (text, image, speech, video…)

• But how? Such knowledge might be overwhelming…
Cross-Sent/Cross-Doc Event Inference Architecture
Baseline Within-Sentence Event Extraction

1. Pattern matching
   - Build a pattern from each ACE training example of an event
   - British and US forces reported gains in the advance on Baghdad
     → PER report gain in advance on LOC

2. MaxEnt models
   ① Trigger Classifier
      - to distinguish event instances from non-events, to classify event instances by type
   ② Argument Classifier
      - to distinguish arguments from non-arguments
   ③ Role Classifier
      - to classify arguments by argument role
   ④ Reportable-Event Classifier
      - to determine whether there is a reportable event instance
Global Confidence Estimation

- Within-Sentence IE system produces local confidence
- IR engine returns a cluster of related docs for each test doc

- Document-wide and Cluster-wide Confidence
  - Frequency weighted by local confidence
  - $XDoc$-$Trigger$-$Freq(trigger, etype)$: The weighted frequency of string $trigger$ appearing as the trigger of an event of type $etype$ across all related documents
  - $XDoc$-$Arg$-$Freq(arg, etype)$: The weighted frequency of $arg$ appearing as an argument of an event of type $etype$ across all related documents
  - $XDoc$-$Role$-$Freq(arg, etype, role)$: The weighted frequency of $arg$ appearing as an argument of an event of type $etype$ with role $role$ across all related documents
  - $Margin$ between the most frequent value and the second most frequent value, applied to resolve classification ambiguities
  - ……
Cross-Sent/Cross-Doc Event Inference Procedure

- Remove triggers and argument annotations with local or cross-doc confidence lower than thresholds
  - *Local-Remove*: Remove annotations with low local confidence
  - *XDoc-Remove*: Remove annotations with low cross-doc confidence

- Adjust trigger and argument identification and classification to achieve document-wide and cluster-wide consistency
  - *XSent-Iden/XDoc-Iden*: If the highest frequency is larger than a threshold, propagate the most frequent type to all unlabeled candidates with the same strings
  - *XSent-Class/XDoc-Class*: If the margin value is higher than a threshold, propagate the most frequent type and role to replace low-confidence annotations
Experiments: Data and Setting

- Within-Sentence baseline IE trained from 500 English ACE05 texts (from March – May of 2003)
- Use 10 ACE05 newswire texts as development set to optimize the global confidence thresholds and apply them for blind test
- Blind test on 40 ACE05 texts, for each test text, retrieved 25 related texts from TDT5 corpus (278,108 texts, from April-Sept. of 2003)
Selecting Trigger Confidence Thresholds
to optimize Event Identification F-measure on Dev Set
Selecting Argument Confidence Thresholds to optimize Argument Labeling F-measure on Dev Set

![Graph showing recall and precision for different methods, with points marked for Local-Remove-Arg, Local-Remove-Role, XSent-Iden, XDoc-Remove-Arg, and XDoc-Remove-Role. Points are labeled with F-measure values.]
# Experiments: Trigger Labeling

<table>
<thead>
<tr>
<th>System/Human</th>
<th>Performance</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Sent IE (Baseline)</td>
<td></td>
<td>67.6</td>
<td>53.5</td>
<td>59.7</td>
</tr>
<tr>
<td>After Cross-Sent Inference</td>
<td></td>
<td>64.3</td>
<td>59.4</td>
<td>61.8</td>
</tr>
<tr>
<td>After Cross-Doc Inference</td>
<td></td>
<td>60.2</td>
<td>76.4</td>
<td>67.3</td>
</tr>
<tr>
<td>Human Annotator 1</td>
<td></td>
<td>59.2</td>
<td>59.4</td>
<td>59.3</td>
</tr>
<tr>
<td>Human Annotator 2</td>
<td></td>
<td>69.2</td>
<td>75.0</td>
<td>72.0</td>
</tr>
<tr>
<td>Inter-Adjudicator Agreement</td>
<td></td>
<td>83.2</td>
<td>74.8</td>
<td>78.8</td>
</tr>
</tbody>
</table>
### Experiments: Argument Labeling

<table>
<thead>
<tr>
<th>System/Human</th>
<th>Argument Identification</th>
<th>Argument Classification Accuracy</th>
<th>Argument Identification + Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Sent IE</td>
<td>P 47.8 R 38.3 F 42.5</td>
<td>P 86.0</td>
<td>P 41.2 R 32.9 F 36.3</td>
</tr>
<tr>
<td>After Cross-Sent Inference</td>
<td>P 54.6 R 38.5 F 45.1</td>
<td>P 90.2</td>
<td>P 49.2 R 34.7 F 40.7</td>
</tr>
<tr>
<td>After Cross-Doc Inference</td>
<td>P 55.7 R 39.5 F 46.2</td>
<td>P 92.1</td>
<td>P 51.3 R 36.4 F 42.6</td>
</tr>
<tr>
<td>Human Annotator 1</td>
<td>P 60.0 R 69.4 F 64.4</td>
<td>P 85.8</td>
<td>P 51.6 R 59.5 F 55.3</td>
</tr>
<tr>
<td>Human Annotator 2</td>
<td>P 62.7 R 85.4 F 72.3</td>
<td>P 86.3</td>
<td>P 54.1 R 73.7 F 62.4</td>
</tr>
<tr>
<td>Inter-Adjudicator Agreement</td>
<td>P 72.2 R 71.4 F 71.8</td>
<td>P 91.8</td>
<td>P 66.3 R 65.6 F 65.9</td>
</tr>
</tbody>
</table>
Global Knowledge Based Inference for Event Extraction

- Cross-document inference (Ji and Grishman, 2008)
- Cross-event inference (Liao and Grishman, 2010)
- Cross-entity inference (Hong et al., 2011)
- All-together (Li et al., 2011)
Within a cluster of topically-related documents, the distribution is much more convergent; closer to its distribution in the collection of topically related documents than the uniform training corpora. E.g. In the overall information networks only 7% of “fire” indicate “End-Position” events; while all of “fire” in a topic cluster are “End-Position” events. E.g. “Putin” appeared as different roles, including “meeting/entity”, “movement/person”, “transaction/recipient” and “election/person”, but only played as an “election/person” in one topic cluster.

Topic Modeling can enhance information network construction by grouping similar objects, event types and roles together.
Bootstrapping Event Extraction

• Both systems rely on expensive human labeled data, thus suffers from **data scarcity**
  (much more expensive than other NLP tasks due to the extra tagging tasks of entities and temporal expressions)

**Questions:**

• Can the monolingual system benefit from bootstrapping techniques with a **relative small set of** training data?
• Can a monolingual system (in our case, the Chinese event extraction system) benefit from the **other resource-rich** monolingual system (English system)?
Intuition:

- The same event has different “views” described in different languages, because the lexical unit, the grammar and sentence construction differ from one language to the other.
- Satisfy the **sufficiency** assumption
Cross-lingual Co-Training for Event Extraction

(Chen and Ji, 2009)

- Bootstrapping: n=1: trust yourself and teach yourself
- Co-training: n=2 (Blum and Mitchell, 1998)
  - the two views are individually **sufficient** for classification
  - the two views are conditionally **independent** given the class
Cross-lingual Projection

- A key **operation** in the cross-lingual co-training algorithm
- In our case, project the **triggers** and the **arguments** from one language into the other language according to the alignment information provided by bitexts.
Experiments (Chen and Ji, 2009)

Data

- ACE 2005 corpus
  - 560 English documents
  - 633 Chinese documents

- LDC Chinese Treebank English Parallel corpus
  - 159 bitexts with manual alignment
Experiment results

- **Self-training, and Co-training** (English-labeled & Combined-labeled) for Trigger Labeling
  - F-score:
    - Self-training: 30.844
    - English-labeled: 30.862
    - Combined-labeled: 30.704

- **Self-training, and Co-training** (English-labeled & Combined-labeled) for Argument Labeling
  - F-score:
    - Self-training: 31.324
    - English-labeled: 32.439
    - Combined-labeled: 33.974

- For Argument Labeling:
  - F-score:
    - Self-training: 22.035
    - English-labeled: 22.006
    - Combined-labeled: 22.206

- For Argument Labeling:
  - F-score:
    - Self-training: 21.878
    - English-labeled: 22.741
    - Combined-labeled: 24.306
Analysis

- **Self-training**: a little gain of 0.4% above the baseline for trigger labeling and a loss of 0.1% below the baseline for argument labeling. The deterioration tendency of the self-training curve indicates that entity extraction errors do have counteractive impacts on argument labeling.

- **Trust-English method**: a gain of 1.7% for trigger labeling and 0.7% for argument labeling.

- **Combination method**: a gain of 3.1% for trigger labeling and 2.1% for argument labeling. The third method outperforms the second method.
1. An explosion in a cafe at one of the capital's busiest intersections killed one woman and injured another Tuesday.

2. Police were investigating the cause of the explosion in the restroom of the multistory Crocodile Cafe in the commercial district of Kizilay during the morning rush hour.

3. The blast shattered walls and windows in the building.

4. Ankara police chief Ercument Yilmaz visited the site of the morning blast.

5. The explosion comes a month after.

6. A bomb exploded at a McDonald's restaurant in Istanbul, causing damage but no injuries.

7. Radical leftist, Kurdish and Islamic groups are active in the country and have carried out the bombing in the past.
# Typical Event Mention Pair Classification Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event type</td>
<td>type_subtype</td>
<td>pair of event type and subtype</td>
</tr>
<tr>
<td>Trigger</td>
<td>trigger_pair</td>
<td>trigger pairs</td>
</tr>
<tr>
<td></td>
<td>pos_pair</td>
<td>part-of-speech pair of triggers</td>
</tr>
<tr>
<td></td>
<td>nominal</td>
<td>if the trigger of EM2 is nominal</td>
</tr>
<tr>
<td></td>
<td>exact_match</td>
<td>if the triggers exactly match</td>
</tr>
<tr>
<td></td>
<td>stem_match</td>
<td>if the stems of triggers match</td>
</tr>
<tr>
<td></td>
<td>trigger_sim</td>
<td>trigger similarity based on WordNet</td>
</tr>
<tr>
<td>Distance</td>
<td>token_dist</td>
<td>the number of tokens between triggers</td>
</tr>
<tr>
<td></td>
<td>sentence_dist</td>
<td>the number of sentences between event mentions</td>
</tr>
<tr>
<td></td>
<td>event_dist</td>
<td>the number of event mentions between EM₁ and EM₂</td>
</tr>
<tr>
<td>Argument</td>
<td>overlap_arg</td>
<td>the number of arguments with entity and role match</td>
</tr>
<tr>
<td></td>
<td>unique_arg</td>
<td>the number of arguments only in one event mention</td>
</tr>
<tr>
<td></td>
<td>diffrole_arg</td>
<td>The number of coreferential arguments but role mismatch</td>
</tr>
</tbody>
</table>
# Incorporating Event Attribute as Features

<table>
<thead>
<tr>
<th>Event Attributes</th>
<th>Event Mentions</th>
<th>Attribute Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modality</td>
<td>Toyota Motor Corp. said Tuesday it will promote Akio Toyoda, a grandson of the company's founder who is widely viewed as a candidate to <em>some day</em> head Japan's largest automaker.</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Managing director Toyoda, 46, grandson of Kiichiro Toyoda and the eldest son of Toyota honorary chairman Shoichiro Toyoda, <em>became</em> one of 14 senior managing directors under a streamlined management system set to be…</td>
<td>Asserted</td>
</tr>
<tr>
<td>Polarity</td>
<td>At least 19 people were <em>killed</em> in the first blast</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>There were <em>no</em> reports of <em>deaths</em> in the blast</td>
<td>Negative</td>
</tr>
<tr>
<td>Genericity</td>
<td>An <em>explosion</em> in a cafe at one of the capital's busiest intersections killed one woman and injured another Tuesday</td>
<td>Specific</td>
</tr>
<tr>
<td></td>
<td>Roh has said <em>any</em> pre-emptive <em>strike</em> against the North's nuclear facilities could prove disastrous</td>
<td>Generic</td>
</tr>
<tr>
<td>Tense</td>
<td><em>Israel holds the Palestinian leader responsible for the latest violence, even though the recent attacks were</em> carried out by Islamic militants</td>
<td>Past</td>
</tr>
<tr>
<td></td>
<td><em>We are warning Israel not to exploit this war against Iraq to carry out more attacks against the Palestinian people in the Gaza Strip and destroy the Palestinian Authority and the peace process.</em></td>
<td>Future</td>
</tr>
</tbody>
</table>

- Attribute values as features: Whether the attributes of an event mention and its candidate antecedent event conflict or not; 6% absolute gain (Chen et al., 2009)
Clustering Method 1: Agglomerative Clustering

Basic idea:

- Start with singleton event mentions, sort them according to the occurrence in the document

- Traverse through each event mention (from left to right), iteratively merge the active event mention into a prior event (largest probability higher than some threshold) or start the event mention as a new event
Clustering Method 2: Spectral Graph Clustering

- (Chen and Ji, 2009)
Spectral Graph Clustering

cut(A,B) = 0.1 + 0.2 + 0.2 + 0.3 = 0.8
Spectral Graph Clustering (Cont’)

- Start with full connected graph, each edge is weighted by the coreference value

- Optimize the normalized-cut criterion (Shi and Malik, 2000)

\[
\min NCut(A, B) = \frac{cut(A, B)}{vol(A)} + \frac{cut(A, B)}{vol(B)}
\]

- \( vol(A) \): The total weight of the edges from group \( A \)
  - Maximize weight of within-group coreference links
  - Minimize weight of between-group coreference links
State-of-the-art Performance

- MUC metric does not prefer clustering results with many singleton event mentions (Chen and Ji, 2009)
The performance bottleneck of event coreference resolution comes from the poor performance of event mention labeling.
Beyond ACE Event Coreference

Annotate events beyond ACE coreference definition

- ACE does not identify Events as coreferents when one mention refers only to a part of the other
- In ACE, the plural event mention is not coreferent with mentions of the component individual events.
- ACE does not annotate:
  “Three people have been convicted…Smith and Jones were found guilty of selling guns…”
  “The gunman shot Smith and his son. ..The attack against Smith.”
CMU Event Coref Corpus

- Annotate related events at the document level, including subevents.

Examples:
- “drug war” (contains subevents: attacks, crackdowns, bullying…)
- “attacks” (contains subevents: deaths, kidnappings, assassination, bombed…)

Applications

- Complex Question Answering
  - Event questions: Describe the drug war events in Latin America.
  - List questions: List the events related to attacks in the drug war.
  - Relationship questions: Who is attacking who?
Drug War events

We don't know who is winning the drug war in Latin America, but we know who's losing it -- the press. Over the past six months, six journalists have been killed and 10 kidnapped by drug traffickers or leftist guerrillas -- who often are one and the same -- in Colombia. Over the past 12 years, at least 40 journalists have died there. The attacks have intensified since the Colombian government began cracking down on the traffickers in August, trying to prevent their takeover of the country.

- drug war (contains subevents: attacks, crackdowns, bullying …)
  - lexical anchor: drug war
- crackdown
  - lexical anchor: cracking down
  - arguments: Colombian government, traffickers, August
- attacks (contains subevents: deaths, kidnappings, assassination, bombed..)
  - attacks (set of attacks)
  - lexical anchor: attacks
  - arguments: (inferred) traffickers, journalists
Events to annotate

• Events that happened
  “Britain **bombed** Iraq last night.”

• Events which did not happen
  “Hall did not speak about the **bombings**.”

• Planned events
  planned, expected to happen, agree to do…
  “Hall planned to **meet** with Saddam.”
Other cases

- Event that is pre-supposed to have happened
  *Stealing event*
  “It may well be that theft will become a bigger problem.”

- Habitual in present tense
  “It opens at 8am.”
Annotating related entities

• In addition to event coreference, we also annotate entity relations between events.
  
  e.g. Agents of bombing events may be related via an “ally” relation.
  
  e.g. “the four countries cited, Colombia, Cuba, Panama and Nicaragua, are not only where the press is under greatest attack ...”
  
  Four locations of attack are annotated and the political relation (CCPN) is linked.
Other Features

Arguments of events

• Annotated events may have arguments.
• Arguments (agent, patient, location, etc.) are also annotated.

Each instance of the same event type is assigned a unique id.
  e.g. attacking-1, attacking-2
Emergent Events in Social Media (Li and Ji, 2014)

Event 1

**Attitude**
- **Emotion**
  - Sad, Anger, Fear, Happiness
- **Sentiment**
  - Objective, Subjective; Positive, Negative
- **Aspect**
  - Active, Passive
- **Confidence**
  - Percentage of Intent / Uncertainty

**Arguments**

**Other Participants**
- Agent, Victim, etc.

**Environment**
- Place and Time

**Narrator**
- **Profile**
  - Gender, Age, Occupation, Location, etc.
- **Property**
  - Individual, Group (news)
- **Role**
  - Reporter, Propagator, Witness, Audience, Participant, Commenter

**Relations**
- Temporal; Precursor
- Hierarchy; Elaboration
- Presupposition; Causal
- Contradiction; Co-reference

Social Norm

Perceived Control

(1) Obtain Warning
(2) Understand Warning
(3) Trust Warning
(4) Personalize Warning
(5) Seek/obtain Confirmation
(6) Take Action
Annotating multiple intersecting meaning layers

Three types of annotations have been added with the GATE tool

• What events are related? Which events are subevents of what events? (Event Coreference)
• What type of relationships between entities? (Entity Relations)
• How certain are these events to have occurred? (Committed Belief)
Domain-independent IE

- Traditional IE assumes the scenario and event types are known in advance so that the corresponding training data and seeds can be prepared.

- Open IE (Banko et al., 2007)
  - Learn a general model of how relations are expressed (in a particular language), based on unlexicalized features such as part-of-speech tags and domain-independent regular expressions; e.g. “E1 verb E2 (X established Y) “
  - The identities of the relations to be extracted are unknown and the billions of documents found on the Web necessitate highly scalable processing.

- On-demand IE (Sekine, 2006):
- Pre-emptive IE (Shinyama et al., 2006): hierarchical pattern clustering

Advantages
- Can extract unknown relations and events from heterogeneous corpora

Disadvantages
- Low recall, cannot incorporate complicated long distance patterns

- Automatic event type and template discovery for new scenarios
  - Using clustering and semantic role labeling techniques (Li et al., 2010)
  - Template discovery (Chambers and Jurafsky, 2011)
# Summary of IE Methods

<table>
<thead>
<tr>
<th>IE Methods</th>
<th>Supervised Learning</th>
<th>Bootstrapping</th>
<th>Distant Supervision</th>
<th>Open IE</th>
<th>Template Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach Overview</td>
<td>Learn rules or supervised model from labeled data</td>
<td>Send seeds to extract common patterns from unlabeled data</td>
<td>Project large database entries into unlabeled data to obtain annotations</td>
<td>Open-domain IE based on syntactic patterns</td>
<td>Automatically discover scenarios, event types and templates</td>
</tr>
<tr>
<td>Requirement of labeled data</td>
<td>Large unstructured labeled data</td>
<td>Small seeds</td>
<td>Large seeds</td>
<td>Small unstructured labeled data</td>
<td>Little labeled data</td>
</tr>
<tr>
<td>Quality Precision</td>
<td>High</td>
<td>Moderate</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Quality Recall</td>
<td>High</td>
<td>Difficult to measure</td>
<td>Moderate</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Portability</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Scalability</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Examples</td>
<td>(Mccallum, 2003; Ahn, 2006; Hardy et al., 2006; Ji and Grishman, 2008)</td>
<td>(Riloff, 1996; Brin, 1999; Agichtein and Gravano, 2000; Etzioni et al., 2004; Yangarber, 2000)</td>
<td>(Mintz et al., 2009; Wu and Weld, 2010).</td>
<td>(Sekine, 2006; Shinyama et al., 2006 Banko et al., 2007)</td>
<td>(Li et al., 2010; Chambers and Jurafsky, 2011)</td>
</tr>
</tbody>
</table>