On-Demand Information Extraction and Linguistic Knowledge Acquisition

Satoshi Sekine
New York University
Introduction
(http://nlp.cs.nyu.edu/sekine)
Research topics

On-demand IE

IE pattern Discovery
Paraphrase Discovery
Preemptive IE

Relation Discovery
ENE Tagger
ENE Design

ENE Attributes

Query log & class

Knowledge Discovery

Ngram Search Engine
ENE Design

English Analyzer (OAK)
ENE

Web People Search
What is Information Extraction

- To automatically extract information on specific scenario from unstructured text and put it into table format
- Ex. Management succession

<table>
<thead>
<tr>
<th>Date</th>
<th>person</th>
<th>Company</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/1/2003</td>
<td>John Smith</td>
<td>Smith Trade corporation</td>
<td>COE</td>
</tr>
<tr>
<td>7/2/2003</td>
<td>Bill Brown</td>
<td>Bank of Manhattan</td>
<td>President</td>
</tr>
</tbody>
</table>
Problem in conventional IE

<table>
<thead>
<tr>
<th>Preparation of knowledge for given scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Pattern, lexicon and semantic knowledge</td>
</tr>
<tr>
<td>- Company announced Person’s promotion to Position</td>
</tr>
<tr>
<td>- Writing patterns by hand or creating training data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Preparation time (one month in MUC)</td>
</tr>
<tr>
<td>- Specific/Limited scenario/domain</td>
</tr>
</tbody>
</table>
Our goal

• Make **one month** to **one minute** (=automatic)
• IE on the fly
• How?
  – Use unsupervised learning methods
    • Pattern discovery, paraphrase discovery
  – Prepare as much knowledge/tools as we can for as many scenarios as possible
    • Extended NE, semantic knowledge, relation
ODIE demo
http://blueberry.cs.nyu.edu:8080/odie/extract-information2.htm
Description of task

“Executive succession” succeed, promote, hire, name

Pattern discovery

IR system

Pattern sets

Paraphrase discovery

Co-reference

Table construction

Table

<table>
<thead>
<tr>
<th>doc</th>
<th>person</th>
<th>Company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>0123</td>
<td>Mike Smith</td>
<td>XYZ</td>
<td>CEO</td>
</tr>
<tr>
<td>0567</td>
<td>Westfield</td>
<td>ABC</td>
<td>VP</td>
</tr>
</tbody>
</table>

DocID:200712100123
Citibank named the next Chairman …

P1: Company announced Person’s promotion to Position
P2: Company announced the next Position to Person
P3: Person stepped down from Company’s Position
Automatic Discovery of IE patterns
(Sudo et.al HLT01) (Sudo et.al ACL03) (Sudo thesis 04)

• Finding important IE pattern for the task
  Company announced Person’s promotion to Position

• Key Idea (TFIDF for patterns)
  – Retrieve documents on the given topic
  – Collect relatively frequent patterns in the retrieved documents
Pattern format and Results

- Three pattern formats (dependency tree)
  - Subtree: Computation time (11 years of newspaper)
    - Tree mining algorithms: 14 minutes
      (Abe et. al KDD 2002)
    - Count-all-in-advance strategy: 30 seconds
  - Near human performance was achieved
IE Pattern Discovery - Result

Succession

Train: 1 yr. newspaper
Test: 148 (87 relevant, 61 irrelevant)

Slots:
  Person: 135
  Org: 172
  Pot: 215
Automatic acquisition of paraphrase
(Sekine Gengo01) (Shinyama et al HLT02) (Shinyama et al IWP03)

• Purpose: Relationship between IE patterns
• Key idea
  – Events are usually reported in different newspapers on the same day
  – However, NE and numeric, date expressions are identical
  – Use them as anchor to acquire paraphrase
• We observed encouraging results, but we found that it is not so easy a task
Procedure

Newspaper A

Find Comparable Articles

Article A

NE tagging Coref.

Anchors identified

Find Comparable Sentences

Extract Paraphrase

Anchors identified

phrase A

phrase B
## Evaluation

- **Two techniques** (co-reference, Argument Structure Database)
- **One year of two Japanese newspaper**
- **195 article pairs on arrest of murder suspects**

<table>
<thead>
<tr>
<th></th>
<th>Coref</th>
<th>ASD</th>
<th>obtained</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(93)</td>
<td>w/o</td>
<td></td>
<td>55</td>
<td>75%(41)</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>with</td>
<td></td>
<td>75</td>
<td>69%(52)</td>
<td>56%</td>
</tr>
<tr>
<td><strong>Phrases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&gt;100)</td>
<td>w/o</td>
<td>w/o</td>
<td>106</td>
<td>24%(25)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>w/o</td>
<td>with</td>
<td>32</td>
<td>56%(18)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>with</td>
<td>with</td>
<td>37</td>
<td>62%(23)</td>
<td></td>
</tr>
</tbody>
</table>
Relation discovery
(Hasegawa et al ACL04)

• Motivation
  – Discovering particular relations between NE’s
  – Country & President, Merged Companies

• Unsupervised method
  – Not known the relationships in advance
  – As many relationships as possible

• Idea
  Context based clustering
  – Pairs of entities with similar context would be in the same relation
  – Similar contexts could also characterize the relations
Procedure

1. Tag ENE on a large text, select two ENE types
2. Collecting expressions within 5 words, and make the context between them as context vector
3. Cluster ENE pairs based on context vectors
4. Label each cluster based on frequent context words

**Company A**

Accumulated contexts

is offering to buy
’s interest in
is negotiating to acquire
’s planned purchase of
…

**Company B**
<table>
<thead>
<tr>
<th>Major relations</th>
<th>Ratio</th>
<th>Common words (Relative frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>17/23</td>
<td>President (1.0), president (0.415), …</td>
</tr>
<tr>
<td>Senator</td>
<td>19/21</td>
<td>Sen. (1.0), Republican (0.214), …</td>
</tr>
<tr>
<td>Prime Minister</td>
<td>15/16</td>
<td>Minister (1.0), minister (0.875), Prime (0.875), …</td>
</tr>
<tr>
<td>Governor</td>
<td>15/16</td>
<td>Gov. (1.0), governor (0.458), Governor (0.3), …</td>
</tr>
<tr>
<td>Secretary</td>
<td>6/7</td>
<td>Secretary (1.0), secretary (0.143), …</td>
</tr>
<tr>
<td>Republican</td>
<td>5/6</td>
<td>Rep. (1.0), Republican (0.667), …</td>
</tr>
<tr>
<td>Coach</td>
<td>5/5</td>
<td>coach (1.0), …</td>
</tr>
<tr>
<td>M&amp;A</td>
<td>10/11</td>
<td>buy (1.0), bid (0.382), …</td>
</tr>
<tr>
<td>M&amp;A</td>
<td>9/9</td>
<td>acquire (1.0), acquisition (0.583), buy (0.583), …</td>
</tr>
<tr>
<td>Parent</td>
<td>7/7</td>
<td>parent (1.0), unit (0.476), own (0.143), …</td>
</tr>
</tbody>
</table>
Evaluation and Error analysis

• Evaluation result
  Labeling accuracy: $10/10 = 100\%$ (Cluster level)
  $108/121 = 89\%$ (NE pair level)
  (cf. recall: $140/(177+65) = 58\%$ (All clusters))

• False Alarm (Mis-clustered NE pairs):
  – Common words in another cluster which existed in
    5 words contexts by accident (noise words)
    ... Chechnya war may exhaust President Boris
    Yeltsin ...

• Miss (Undetected NE pairs):
  – Absence of common words in 5 words contexts
  – Outer context might be helpful
    ... Boris Yeltsin on end the fighting in Chechnya ...
Another Paraphrase Discovery via Relation Discovery
(Hasegawa et.al Gnengo-05)

- Expressions found in the same relation should contain paraphrase
- Two filters
  - Used in more than one pair of instance
  - Expressions contain frequent term

Ex) A bought B / A has agreed to buy B / A, which is buying B / A’s proposed acquisition of B / A’s acquisition of B / A’s agreement to buy B / A’s purchase of B / A bid for B / A’s takeover of B / A merger with B / A succeeded in buying B / B, which was acquired by A / B would become a subsidiary of A / B agreed to be bought by A
Yet Another Paraphrase Discovery using context and keywords
(Sekine IPW-05)

• To eliminate the frequency threshold for the vector model in Hasegawa’s method
• Algorithm
  – Find keywords in context of each pair of NEs using TFIDF
  – Cluster phrases based on NE instances (buy – acquire / buy – agree / buy – purchase)
ODIE – Flow Chart

Description of task

“Executive succession” succeed, promote, hire, name

IR system

Pattern discovery

Paraphrase discovery

Table construction

Table

<table>
<thead>
<tr>
<th>doc</th>
<th>person</th>
<th>Company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>0123</td>
<td>Mike Smith</td>
<td>XYZ</td>
<td>CEO</td>
</tr>
<tr>
<td>0567</td>
<td>Westfield</td>
<td>ABC</td>
<td>VP</td>
</tr>
</tbody>
</table>
ODIE - Evaluation Result

- Evaluate the system as a support tool for IR.
- Out of 20 topics, tables created for 2 topics are judged very useful (This can replace the IR), 12 are useful (Supportive for IR), 6 are not useful.
- Correctness of the table fillers

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Number of rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>84</td>
</tr>
<tr>
<td>Partially correct</td>
<td>4</td>
</tr>
<tr>
<td>Incorrect</td>
<td>12</td>
</tr>
</tbody>
</table>
ODIE demo

http://blueberry.cs.nyu.edu:8080/odie/extract-information2.htm
Preemptive Information Extraction
(Shinyama, Sekine NAACL06) (Shinyama thesis 07)

- We can even delete queries...
- Find a topics using document clustering method, then create table preemptive manner
  - Use NE as a clue to the clustering
- Finding the table of interest can be done by a search
- Interesting system was created
Extended Named Entity
(Sekine et.al LREC02) (Sekine et.al LREC04)

- 7~8 categories can’t cover the world
- We need a reasonable set of NE categories
- Named Entity with 200 categories
- Definition (150 page) was released
- Automatic tagger
  - 240K entries (dictionary)
  - 2000 rules (e.g. Mr. <Capital> → person name)
  - 70% accuracy
  - Improving...
- [http://nlp.cs.nyu.edu/ene](http://nlp.cs.nyu.edu/ene)
Attributes of ENE

- Attributes are essential and interesting information for names.
- Also, in order to define ENE categories, it is very useful.
  - Water body (ver. 6) ↔ River, Lake (ver. 7)
- Also, it is useful for disambiguation (people name disambiguation = WePS).
- Attributes are important for opinion mining.
- We define attributes for 110 categories using encyclopedia examples.
## Attributes for ENE (ver. 7.0.1)

**example for PERSON**

<table>
<thead>
<tr>
<th>Attribute(20)</th>
<th>Example of value</th>
<th>Freq. (%)</th>
<th>ENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocation</td>
<td>professional baseball player, economist, poet</td>
<td>46(100)</td>
<td>Vocation</td>
</tr>
<tr>
<td>Nationality</td>
<td>American, Chinese, Japanese</td>
<td>29(63)</td>
<td>Country</td>
</tr>
<tr>
<td>Career</td>
<td>A professor at Yale University, The Princess of Wales</td>
<td>26(57)</td>
<td>Vocation</td>
</tr>
<tr>
<td>Masterpiece</td>
<td>Guernica, Mona Lisa</td>
<td>25(54)</td>
<td>Product, Facility</td>
</tr>
<tr>
<td>Graduate</td>
<td>M.A. in German at Cambridge, MK High School</td>
<td>20(44)</td>
<td>School</td>
</tr>
<tr>
<td>Hometown</td>
<td>Paris, Manchester, Shanghai</td>
<td>19(41)</td>
<td>City</td>
</tr>
<tr>
<td>Native Province</td>
<td>State of Illinois, Sichuan</td>
<td>18(39)</td>
<td>Province</td>
</tr>
<tr>
<td>Previous stay</td>
<td>England, New York</td>
<td>12(26)</td>
<td>Location</td>
</tr>
<tr>
<td>Mentor</td>
<td>Andrea del Verrocchio, Michelangelo di Lodovic Buonarroti Simoni</td>
<td>10(22)</td>
<td>Person</td>
</tr>
<tr>
<td>Death date</td>
<td>04/23/1704, 04/23/1704, unknown</td>
<td>10(22)</td>
<td>date</td>
</tr>
<tr>
<td>Era</td>
<td>Edo period, the 11th century</td>
<td>8(17)</td>
<td>Era</td>
</tr>
<tr>
<td>Award</td>
<td>Academy Award, MVP, Nobel Prize</td>
<td>8(17)</td>
<td>Award</td>
</tr>
<tr>
<td>Real Name</td>
<td>Saint Nicholas</td>
<td>8(17)</td>
<td>Person</td>
</tr>
<tr>
<td>Another name</td>
<td>Santa, father Christmas</td>
<td>8(17)</td>
<td>Person</td>
</tr>
<tr>
<td>Title</td>
<td>Knight, an honorary degree at Yale</td>
<td>6(13)</td>
<td>Title</td>
</tr>
<tr>
<td>Competition</td>
<td>World Series, 1955 piano competition in Paris</td>
<td>6(13)</td>
<td>Game</td>
</tr>
<tr>
<td>Place of birth</td>
<td>New York, Birmingham</td>
<td>5(11)</td>
<td>Location</td>
</tr>
<tr>
<td>Father</td>
<td>John B. Kelly, Sr.</td>
<td>5(11)</td>
<td>Person</td>
</tr>
<tr>
<td>Cause of death</td>
<td>Car accident, Guillotine</td>
<td>5(11)</td>
<td></td>
</tr>
</tbody>
</table>
Research topics

On-demand IE

- IE pattern Discovery
- Paraphrase Discovery
- Preemptive IE

Knowledge Discovery

- Relation Discovery
- Ngram Search Engine
- ENE Tagger
- Query log & class
- ENE Attributes
- ENE Design

Web People Search

ENE

English Analyzer (OAK)
Why Semantic Knowledge?

- Knowledge is essential for semantic analysis.
  - 鳥取: 0xc4 0xbb 0xbc 0xe8
- We have to have knowledge behind sentences or words
- We have to create it in advance
- Semantic knowledge is huge
- We have to (semi-)automatically discover it
  (otherwise we have to create it by hand)
In the past decade, we have had successful experiences using Machine Learning!

- FOR: POS tagger, NE tagger, Parser, WSD...
- BY: HMM, DT, DL, MaxEnt, SVM, CRFs...
- These tasks can be translated into labeling problem with a handful classes
Supervised ML does not work on Semantic Knowledge!!

- Limitations of ML in semantic problems
- because of
  - Enormous number of classes (if we can enumerate)
  - Limitation of training data
  - It is a lexically dependent problem (sparseness)

  Hence, it is reasonable to believe that it is inherent rather than accidental that those semantic tasks are irreconcilable with supervised ML methods
Challenges

• **Named Entity Recognition**
  – 200 category named entity (Sekine LREC04)
  – Product name (“I can not believe it’s not a butter”)
  – Event name (“the Cardiff Singer of the World competition”)
  – Ambiguity between different classes
    (Toyota: Person, Company, City)

• **Coreference / Name alias**
  – Noun coref.: Prof. Sekine ⇔ NLP researcher
  – Name alias: Japan ⇔ Tokyo, USA ⇔ Obama (dynamic)

• **Attributes of name**
  – Sports team: Players, League,,,,, Team Color, Mascot,
Challenges

• **Coverage of patterns**
  – `<PERSON>` was ambushed (for attack event)

• **Coverage of paraphrase**
  – `<PERSON>` gave his life to `<WAR>`
  – `<PERSON>` was killed at `<WAR>`
Challenges

- **WSD**
  - <PERSON-1> attacked <PERSON-2>
    - PERSON1: a politician => Verbal attack
    - PERSON1: a robber => Physical attack

- **Domain**
  - He hits a victim
  - He hits a ball

- **Script (Textual Entailment)**
  - A Moscow politician had left his house in his van at 7 am. A few minutes later, three heavily armed men forced him to get out of his car and get into a Renault.
  - Kidnap
    - Abduct the victim
      - Move the victim from his place to kidnapper’s place
        » If his is in a car, it involves getting him out of his car
<table>
<thead>
<tr>
<th></th>
<th>Same meaning</th>
<th>Typology</th>
<th>Meronymy</th>
<th>Description</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thing</td>
<td>Synonym</td>
<td>Hierarchy, hypernym-hyponym</td>
<td>Part-of</td>
<td>Adjective</td>
<td>(Many…)</td>
</tr>
<tr>
<td>(noun)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Name alias</td>
<td>Membership</td>
<td>Loc-of, Family, …</td>
<td>Attribute</td>
<td>Attribute</td>
</tr>
<tr>
<td>(proper noun)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event</td>
<td>Paraphrase</td>
<td>tropology</td>
<td>Sub-event, Script</td>
<td>Adverb</td>
<td>Causal, Temporal, …</td>
</tr>
<tr>
<td>(verb)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-target knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>• Thing-Event</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Information Extraction pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Telic role, agentive role (generative lexicon)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>• Name-Event</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>– Information Extraction result</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problem egg

The problem

- Synonym
- Word hierarchy
- Names
- Paraphrase
- WSD
- Script
- Dialogue ('60s~)
- IR ('70s~)
- IE ('80s~)
- Summarization ('80s~)
- Q&A ('00~)
- TE ('02~)
- ...
Knowledge Discovery Technologies
(JNLP tutorial 2009: see my homepage)

Unsupervised (Fully automatic)
- Distributional similarity
- Lexico-syntactic pattern
- Rewrite and verify
- Clustering
- Bootstrapping
- Alignment
- Co-occurrence, collocation
- General knowledge collection from text analysis
- Knowledge acquisition from semi-structured texts

Semi-supervised
- Active learning
- Cloud/crowds
- Hand-made data and annotation
1. Distributional Similarity

• Harris’s distributional Hypothesis

*Words that occurred in the same contexts tend to be similar*

• Find similar words (synonym, NE instance)
  – London, Tokyo, New York, …
• Find similar phrases (paraphrase)
  – X found a solution to Y
  – X solved Y
1. Distributional Similarity
(Dekang and Pantel KDD2001)

- **DIRT - Discovery of Inference Rules from Text**
- Analyze corpus (SUSANNE corpus/1G news) by dependency analyzer (Minipar)
- Extract certain form of paths (links) and two words
  - N:subj:V<-find->V:obj:N->solution->N:to:N
  - X finds solution to Y
- Find similarities of the links
  - Count frequencies of link-SlotX, link-SlotY
  - Use mutual information to compute similarity
- Evaluation on TREC Questions
  - Top 40 candidates: 35~93% accuracy
- Example
  - X is author of Y  ~  X wrote Y
  - X solved Y        ~  X found a solution to Y
2. Lexico-syntactic Pattern

- Use typical expression for given relation
  - Hypernym-hyponym (Hearst 92)
    - X, Y and other C, C such as X, Y and Z
    - XやYなどのC, XといったC
  - Happens before (VerbOcean)
    - to X and then Y
    - Xed and eventually Yed
  - Causal relation
    - X caused Y, Y, because of X
    - XによりY, Xを反映し、Y

- Ideas for improvements
  - Use syntactic structures
  - Use multiple evidences
2. Lexico-Syntactic Pattern
(Sakaji et al. PAKM2008)
3. Rewrite and Verify

- Check if alternative exists
  - NP structure analysis (Nakov & Hearst 05)
    - “(brain stem) cell” OR “brain (stem cell)”
    - Stem cells in the brain -> right
    - Stem cells from the brain -> right
    - Cells from the brain stem -> left*
  - Non-referential pronoun recognition (Dekang 08)
    - make it in advance -> make them in advance
    - make it in Hollywood -> make them in Hollywood*
4. Clustering

- Clustering words, expressions based on
  - not only the distribution of the context
  - but also other features, such as
    - Bag-of-words
    - Topic

  to find class of words, similar relation etc
4. Clustering
(Hasegawa et al. ACL04)

• **Context based clustering**
  – Tag NE on a large corpus
  – Pairs of entities with similar context would be in the same relation
  – Similar contexts could also characterize the relations

```
<table>
<thead>
<tr>
<th>Named entity</th>
<th>5 words or less</th>
<th>Named entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company A</td>
<td>is offering to buy</td>
<td>Company B</td>
</tr>
<tr>
<td>Accumulated contexts</td>
<td>’s interest in</td>
<td>’s planned purchase of</td>
</tr>
<tr>
<td></td>
<td>is negotiating to acquire</td>
<td></td>
</tr>
</tbody>
</table>
```
5. Bootstrapping

• Use two independent clues
  – (Pairs of relation, Expression of the relation)
    • (people – birthday, “X was born in Y”)
    • (book title – author, “X which is written by Y”)
  – (Spelling clue of NE, Context of NE)
    • (“Mr. X”, “X, a professor at”)

• Combination of Distributional similarity and LSP
  – User gives several seeds as initial sample
  – The acquired data will be used to acquire new data
5. Bootstrapping
(Ravichandran and Hovy ACL2002)

1. Select pair {Mozart, 1756}, and search the web
2. Find frequent patterns which contains the pair
   – “Mozart (1756), “Mozart was born on 1756”, …
3. Generalize the pattern
   – “<NAME> (<BY>”, “<NAME> was born on <BY>”
4. Find more examples, go back to 2.

- Use the pattern to find BY of given person (Q&A)
- Q&A performance (MRR):
  - Birthday: 0.69, Inventor 0.58, Discoverer 0.88
6. Alignment

- **Target**
  - Acquire paraphrase, translation knowledge
  - Named entity knowledge
- **Resource**
  - Newspaper from the same day about the same event
  - Multiple translations of the same book
- **Key**
  - Alignment clue: NE, known word pairs
  - The same named entities appear periodically similarly across different newspapers
6. Alignment
(Shinyama, Sekine IWP2004)
7. Co-occurrence, collocation

- Find attribute/value of specific NE or NE class
- Find typical phrase
- Find typical argument frame
- Modifiers of NE or NE class (sentiment analysis)
- Speech Recognition
- Spelling correction, OCR
- Word sense disambiguation
- Word selection in generation
- Word prediction
8. General knowledge collection from text analysis

- MindNet (MSR)
  - Automatically-constructed knowledge base from free text
    • Detailed dependency analysis for each sentence, aggregated into arbitrarily large graph
    • Named Entities, morphology, temporal expressions, etc.
    • Frequency-based weights on subgraphs
    • Path exploration algorithms, learned lexical similarity function
    • Built from arbitrary corpora: Encarta, web chunks, dictionaries, etc.

- Knext (Lenhart Schubert)
  - Accumulation of general knowledge
    • “Room may have windows”, “People may want to be rid of dictator”
  - Technique & Results
    • 80 rules for mapping phrases to logic form
    • Obtained 117,000 distinct factoids from Brown corpus, 6M factoid from BNC
    • 60% judged reasonable general claims about the world by human
9. Knowledge Discovery from Semi-structured data

- Wikipedia+WordNet
  - Yago (Yet Another Greater Ontology)
    - Source: Wikipedia Category + Info box
      - (Elvis Presley, is-a, singer)
      - ((Elvis Presley, hasWonPrize, Grammy Award), in, 1974)

- Tables on Web (Yoshida et al. 02)
  - Extract attribute-values from tables
  - Create ontology from extracted results
  - Ex) human class
    - Name – John Smith, Ichiro Suzuki, Mary
    - Gender – Male, Female
    - Blood type – A, AB…
## Knowledge Discovery Technologies
(Semi-Automatic Acquisition)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Active Learning</td>
</tr>
<tr>
<td>2.</td>
<td>Crowd/Cloud</td>
</tr>
<tr>
<td>3.</td>
<td>Role of hand-made Knowledge and annotation</td>
</tr>
</tbody>
</table>
1. Active Learning

- Supervised Machine Learning, but requires much smaller training data
- Human intervenes during the process
- Select most useful data for learning

- This paradigm is important because most of the Minimally supervised methods can’t produce perfect result
- We need efficient and minimal human intervention
2. Crowds
(Open Mind Common Sense)

- Starts 2000, now has 250K assertions
2. Crowds
(AMAZON Mechanical Turk)

• https://www.mturk.com/mturk/welcome
• A significantly cheaper and faster method for collecting annotations from broad base of paid non-expert contributors over the Web
  • Affect recognition, word similarity, recognizing TE, event temporal ordering and WSD
  • High agreement
e.g. TE: 8000 labels by $8.00
  WSD: 1770 labels by $1.76
• 4 (2-9) non-expert to emulate expert-level label quality
3. Role of hand-made knowledge and annotation

• We have
  – a lot of hand-made knowledge
    • WordNet, FrameNet, Comlex, Nomlex
    • Cyc, EDR
    • Wikipedia, Wiktionary, Wikimedia Commons…
    • Commercial dictionary etc. . . . .
  – and annotated corpora
    • Penn Treebank, Propbank, Nombank, Timebank, Penn Discourse Treebank, “Pie in the sky”
    • BNC, ANC, BCCWJ, EDR-corpus
    • MUC, DUC, AQUAINT, Conell, SemEval, WePS, IREX…

• How can we utilize it? (can’t afford to ignore it)
Research topics

On-demand IE

Knowledge Discovery

IE pattern Discovery
Paraphrase Discovery
Preemptive IE
ENE Tagger
Query log & class
ENE Design
ENE Attributes
ENE

Ngram Search Engine
Relation Discovery
ENE Design
English Analyzer (OAK)
Web People Search
Discovery tool

- **Challenges**
  - Most discoveries involve searching patterns in a large corpus
    - Search from corpus: takes long time
    - Use internet search: inefficient, limitation
- **Possible solutions**
  - Most search needs only local contexts
    - Ngram search, instead of full text search (e.g. Google Ngram data)
  - Open Search Engine for NLP
    - TSUBAKI (Shinzato et al. 08)
  - Resource poor search engine
    - Ngram Search Engine (Sekine 08)
Discovery Tool

Ngram Search Engine
- Search ngram with arbitrary wildcards with POS, chunk, NE restrictions
- 1B 7grams from 1.7G word Wikipedia (no freq. cutoff)
- It also outputs KWIC, original sentences
- Search in 0.10 sec on a single CPU PC-linux with 4G memory
- 1.3TB index
- Demo
  http://nlp.cs.nyu.edu/nsearch

Query Examples:
- can not #VB * because of
- from #LOC to #LOC via #LOC
- it is a #ADJ phone
- Mr. * said
- used * * to discover *
Discovery Tool (Japanese)
http://languagecraft.jp

日本語N-GRAM検索エンジン

大量な文章データから、任意の単語列を高速に検索します。ウィルドカードを含む単語列を指定でき、言語の知識発見、日本語教育、人工知能などへの応用が可能です。

![Image of Discovery Tool](http://languagecraft.jp)

<table>
<thead>
<tr>
<th>データベース</th>
<th>サポート</th>
<th>など</th>
<th>の</th>
<th>基盤</th>
</tr>
</thead>
<tbody>
<tr>
<td>プレミアム</td>
<td>サポート</td>
<td>など</td>
<td>の</td>
<td>基盤</td>
</tr>
</tbody>
</table>

Content-type: text/html; charset=euc-ja

ifreq = 83
tfreq = 92

<table>
<thead>
<tr>
<th>項目</th>
<th>書類</th>
<th>など</th>
<th>の</th>
<th>基盤</th>
</tr>
</thead>
<tbody>
<tr>
<td>データベース</td>
<td>サポート</td>
<td>など</td>
<td>の</td>
<td>基盤</td>
</tr>
<tr>
<td>プレミアム</td>
<td>サポート</td>
<td>など</td>
<td>の</td>
<td>基盤</td>
</tr>
</tbody>
</table>

This is a screenshot of the Discovery Tool's interface.
Large Corpus

- **English**
  - Newspaper from LDC (86 years: 2G words)
  - Web data (???: NYU=2G words)
  - Wikipedia (1.7GW) [http://nlp.cs.nyu.edu/wikipedia-data](http://nlp.cs.nyu.edu/wikipedia-data)
  - ICWSM blog data (2 month amount; 400MB XML-file)
  - CCB corpus (English-French 200M words)

- **Japanese**
  - Mainichi (2B char; 1991-2007; 120K yen each; Nichigai) [http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html](http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html)
  - Web data
    - LC2003; 12 Billion char.
    - NTCIR-5 (2004); 1TB
    - Kawahara Corpus; 0.5B sentences
    - I-explosion; 0.9T sentences
  - Wikipedia
NSF Sponsored Symposium

Semantic Knowledge Discovery, Organization and Use

– November 14-15, 2008 at New York University

– http://nlp.cs.nyu.edu/sk-symposium

Invited Speakers

Ido Dagan (Bar Ilan U.)
Bill Dolan (MSR)
Oren Etzioni (U. Washington)
Christiane Marti (UC Berkeley)
Kentaro Inui (NAIST)
Dekang Lin (Google)
Bernardo Magnini (ITC-irst)
Dan Moldovan (U. Texas)
Patrick Pantel (Yahoo! Labs)
Marius Pasca (Google)
Peter Turney (NRC)
Satoshi Sekine (NYU)
Pioneer of LSP technique (Hearst Coling92)

Three Tricks

1. Lots o’ Text
   - Bigger is better than smarter! (Banko and Brill ACL01)

2. Unambiguous Cues
   - Such NP as {NP ,}* {(or|and} NP

3. Rewrite&Verify
   - “brain stem cell” -> (brain stem) cell / brain (stem cell)
     - stem cells in the brain
     - stem cells from the brain
     - cells from the brain stem

Cognitive Linguistics -> Path Model
• Use ngram to acquire deep knowledge
  – Preposition (benefit * the community: of, to)
  – Context sensitive spell checker (cite, sight, site)
  – Non-referential pronoun recognition
    • make it in advance -> make them in advance
    • Make it in Hollywood -> make them in Hollywood*
  – Gender probability
    • Ababibu explained himself, Takeru bought his car
    • Noun * pronoun/possesive: Pat (58% M, 31% F)
  – Anti-reflexive
    • John needs his friend (his=John’s)
    • John needs his support (his =/= John’s)
    • Count compatibility for “* needs * support”
  – Distributional similarity based on PMI -> clustering
Ido Dagan

*It's time for a semantic engine*

- Propose “Semantic Engine (API)”
- Textual Entailment as a framework
  - Do generic inference
  - Applications: QA, IE, MT eval, Tutoring, Summarization
- Joint community Resource

**NLP Problem**

- Meaning Representation
- Variability
- Ambiguity

**Inference and Entailment**

- Inference
- Interpretation
- Textual Entailment
Bill Dolan

Reasons to avoid reasoning:
Where does NLP stop and AI begin?

  - Automatically constructed knowledge base from encarta, dictionary, web chunks
    - isa, part-of, locn_of, purpose, type_obj,
  - QA with MindNet
    - Worked beautifully,,, just not very often…
      - Linguistic alternations, paraphrase, discourse, coreference
      - Extra linguistic problems: genre conventions, mathematical reasoning, extensive text comprehension

- What is our goal?
  - Learned Pattern matching
  - Follow the MT paradigm
    - Statistically mature technology, Large parallel corpora, Automated metric, rapid training-test cycle
    - Need large training test dataset
  - Entailment (deeper reasoning) vs. Paraphrase (language)
  - We should focus on paraphrase (can be approached by today’s technology)
On-Demand Information Extraction
- IE on the fly (show the salient relation of query in one minute; type and instance)
- Challenge (Need more knowledge)
  - Extended NE, Event pattern, paraphrase, coreference, WSD, domain, script...

We need semantic knowledge
Community effort on
- Semantic knowledge discovery
  - Discovery tool (ngram/pattern search engine)
  - Knowledge discovery
- Organization
  - Open archive for re-use
- Use
  - Collaborative evaluation event
We need Semantic Knowledge!

- **Challenge**
  - Semantic Knowledge (SK) is too diverse and vast to be created by a single academic institution

- **Situation**
  - It will take all my time until retirement to make all of this SK
  - Someone in this room may have created the knowledge I need
  - Someone in this room may have created the knowledge you need
  - Someone in this room may look for the knowledge you created

- **Solution**
  - If all the knowledge created by the people in this room is available, I can retire now
  - OR even better, I can start from the point of retirement (Extending my life span)
I propose a **Community Effort** for Semantic Knowledge ...

1) Discovery
   1-a) Discovery Tool
   1-b) Knowledge Discovery

2) **Organization** (Open Archive)
3) **Use** (Evaluation Event)
Research topics

On-demand IE

- IE pattern Discovery
- Paraphrase Discovery
- Preemptive IE

Knowledge Discovery

- Ngram Search Engine
- Relation Discovery
- ENE Tagger
- Query log & class
- ENE Design
- ENE Attributes
- ENE

English Analyzer (OAK)

Web People Search
WePS-3

- WePS-1 (2006-2007)
  - People disambiguation
- WePS-2 (2008-2009)
  - People disambiguation
  - Attribute extraction
- WePS-3 (2010)
  - People disambiguation & attribute extraction
  - Organization disambiguation on twitter entries

http://nlp.uned.es/weps
Apply the HLT to the real world

- Language Craft Corporation
  - Question Answering System
  - Document clustering
  - Spelling variation detection
  - Proofreader
  - Solve language tests
  - Japanese ngram search engine
  - Sentiment analysis
  - Building language data and dictionaries

- [http://languagecraft.jp](http://languagecraft.jp)
Bibliography

• 関根聡、乾健太郎、鳥澤健太郎「提唱「コーパスベース知識工学」言語処理学会第13回年次大会ワークショップ（2007）
• 鳥澤健太郎、関根聡、乾健太郎「言語処理学事典 知識獲得」
• Dekang Lin, Ken Church, Satoshi Sekine. John Hopkins University, CLSP summer workshop “Unsupervised Acquisition of Lexical Knowledge from N-Grams” http://www.clsp.jhu.edu/workshops/ws09/

• ANC, FLaReNet, MASTAR

• Thanks: 乾健太郎, 鳥澤健太郎, Ellen Riloff, Ralph Grishman, 坂地泰紀, 小林聡雄

THANK YOU!!!