Community Effort for Semantic Knowledge Discovery, Organization and Use

NSF Sponsored Symposium
Semantic Knowledge Discovery, Organization and Use

Satoshi Sekine
New York University

Outline

• On-Demand Information Extraction
• Semantic Knowledge is needed
• Community Effort
  1. S.K. Discovery
     • Tool
     • Discovery
  2. Organization
  3. Use
Information Extraction

- 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

- Preparation of knowledge for given scenario
  - Pattern: *Company* announced *Person’s* promotion to *Position*
  - Paraphrase: promotion to *Position* = succeed *Position*
  - NE: *Company, Person, Position, (Award, Disease…)*
    - Take time and labor, if we create the knowledge by hand or learn from training data

- Restriction
  - Preparation time (one month in MUC)
  - Work only on specific scenario
On-Demand Information Extraction

- Make **one month** to **one minute** (=automatic)
- **IE on the fly**
- **How?**
  - Use unsupervised knowledge discoveries
    - Pattern discovery, paraphrase discovery
  - Prepare as much knowledge/tools as we can for as many scenarios as possible
    - Extended NE, relation discovery

- (Sekine Coling-ACL06), demo:(Sekine ACL07)

ODIE demo

http://blueberry.cs.nyu.edu:8080/odie/extract-information2.htm
How it works

Description of task

“Management succession” succeed, promote, hire, name

IR system

Pattern discovery

Language Analyzer

Pattern

Paraphrase discovery

Table construction

Table

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
    (Sekine: Person, Organization)

- **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>`

- **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 (TE)**
  - 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
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 them
  - 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** to do Semantic Knowledge …

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

2) **Organization** (Open Archive)

3) **Use** (Evaluation Event)

---

1-a) **Discovery tool**

- **Challenges**
  - Most discoveries involve searching patterns in a large corpus
    - Search from corpus: takes long time
    - Use internet search: inefficient, limitation
  - Most search needs only local contexts
    - Ngram search, instead of full text search
      - Thanks for Google Ngram data
  - Not many people have rich resource (e.g. cluster machine)
    - Ngram Search Engine for the poor
  - We want to see wider contexts of ngrams

- **Future Challenges**
  - More sophisticated search (includes POS, NE, classes)
1-a) Discovery Tool

Ngram Search Engine (current system)

- Search ngram with arbitrary wildcards
- 1B 7grams from 3G word 86 yr. amount of newspaper (no freq. cutoff)
- It also outputs KWIC, original sentences
- Search in 0.02 sec on a single CPU PC-linux with 4G memory
- 2.7TB index
- Demo
  
  http://linserv1.cims.nyu.edu:23232/ngram/

- Copyright issue
- CCB corpus (200MW)

Query Examples:

can not * * because of from * to * via *
it is a * phone
Mr. * said used * * to discover *

1-a) Discovery Tool

Advanced Pattern Search Engine

- Include literals, POS, NE and semantic categories

<POS=VBD> and <POS=VBD>
It is a <POS=ADJ> <syn&hypo=phone>
give|gives|gave PRP$ life|lives to * *
<POS=PRP> attended * <syn&hypo=event> in <NE=loc>

- More generalization, better coverage and precision
1-b) Knowledge Discovery

- Discover SK using LSP, Distributional similarity, alignment
  - From a large corpus (Newspaper, Web, Wikipedia …)

- Knowledge to be discovered includes
  - Synonym (Church 88), Hypermym hyponym (Hearst 92) (Pantel 02) (Snow 06)
  - NE instances (Collins 99) (Yangarber 02) (Shinyama 04) (Sekine 07), NE alias, Attribute of NE (Torisawa 04)
  - Relation between things (Brin 98) (Charniak 99) (Ravichandran 04) (Girju 06) (Hasegawa 04)
  - Paraphrase (Dekang 02) (Idan 08) (Shinyama 02) (Barzilely 01), Textual entailment (Zanzotto 06) Inference rule (Torisawa 06)
  - Relation between events (Chklovski 04) (Girju 02)
  - …more and more…

- Needs clean-up by human
  - Accuracy is never be perfect. Some(most?) needs clean-up
  - Is it plausible?
    - Priority for needed and/or easier resource first

2) Organization

- Create Open Archive
  - Objective: share (contribute and re-use) SK
  - Should not be just a “list of links”
  - Mechanism of “rewards”
    - Easily record and display user’s experience, comments and appreciations (a.k.a. Web 2.0, e.g SourceForge.net)

- It has to be carefully maintained
  - Find, Encourage and Recruit SK
  - Categorize, Format
  - Easy to search, browse
  - Relation to LDC, ELRA
2) Organization

• Basically “Open” for type, theory and format
• Type:
  – Tool (Semantic engine, inference tool …)
  – Knowledge (NE list, paraphrase list …)
  – Annotations (Text with annotated knowledge)
• Theory:
  – As neutral as possible
• Format
  – Simple plain text (easy to read)
  – XML, XCES
  – UIMA, GATE
• Challenges
  – Link different SK
  – Differences between the same kind of knowledge
  – Use the knowledge pool for further discovery

3) Use

• Conduct Evaluation Events (such as ACE, TREC, TAC etc)
  – The task would be something which use SK
    • IE, Textual Entailment, Q&A, Dialogue??
  – Encourage people to use SK
  – Promote contributions of SK

• Collaborative Evaluation
  – Current (competitive) evaluations have limitation
    • Because of poor SK resources:
      – Quick and dirty systems
      – ML based systems with bag-of-word features
  – 2 Ideas
    • Participants start from a certain level using the SK archive
    • Common training/test data annotated by many tools
3) Use
Common training/test (just an idea)

- NANTC (general) – 500M words (LDC2008T16, $300 for non-member)
- BLLIP parsed, add more and more information (NE, coref., relation…)
- Use a part of this for training, development and test for the evaluation

<table>
<thead>
<tr>
<th>Canadian</th>
<th>NNP</th>
<th>B-NP</th>
<th>B-GPE</th>
<th>B-ORG</th>
<th>B-NATIONALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>River</td>
<td>NNP</td>
<td>I-NP</td>
<td>O</td>
<td>I-ORG</td>
<td>O</td>
</tr>
<tr>
<td>Expeditions</td>
<td>NNP</td>
<td>I-NP</td>
<td>O</td>
<td>I-ORG</td>
<td>O</td>
</tr>
<tr>
<td>offers</td>
<td>VBZ</td>
<td>B-VP</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>B-NP</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>change</td>
<td>NN</td>
<td>I-NP</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>From</td>
<td>IN</td>
<td>B-PP</td>
<td>O</td>
<td>O</td>
<td>B-DATE_RANGE</td>
</tr>
<tr>
<td>July</td>
<td>NNP</td>
<td>B-NP</td>
<td>B-DATE</td>
<td>B-MON</td>
<td>I-DATE_RANGE</td>
</tr>
<tr>
<td>1</td>
<td>CD</td>
<td>I-NP</td>
<td>I-DAY</td>
<td>I-DAY</td>
<td>I-DATE_RANGE</td>
</tr>
<tr>
<td>to</td>
<td>TO</td>
<td>B-PP</td>
<td>O</td>
<td>O</td>
<td>I-DATE_RANGE</td>
</tr>
<tr>
<td>August</td>
<td>NNP</td>
<td>B-NP</td>
<td>B-DATE</td>
<td>B-MON</td>
<td>I-DATE_RANGE</td>
</tr>
</tbody>
</table>

Summary

- This effort relates many NLP activities
  - Community efforts is needed!

- This is a Win-Win game for creators and users
  - Let’s Discover, Organize and Use Semantic Knowledge to raise the level of NLP!

THANK YOU!!!