Natural Language Processing (NLP)

10/30/02
NLP Definition

- a range of computational techniques
NLP Definition (cont’d)

• a range of computational techniques

• for analyzing and representing naturally occurring texts
NLP Definition (cont’d)

• a range of computational techniques
• for analyzing and representing naturally occurring texts

• at one or more levels of linguistic analysis
Levels of Language Understanding

- Pragmatic
- Discourse
- Semantic
- Syntactic
- Lexical
- Morphological
NLP Definition (cont’d)

• a range of computational techniques
• for analyzing and representing naturally occurring texts
• at one or more levels of linguistic analysis

• for the purpose of achieving human-like language processing
NLP Definition (cont’d)

• a range of computational techniques
• for analyzing and representing naturally occurring texts
• at one or more levels of linguistic analysis
• for the purpose of achieving human-like language processing

• for knowledge intensive applications
Goals of Information Extraction

- A robust information extraction system
Goals of Information Extraction

• A robust information extraction system
  – Recognize concepts and the implicit relations amongst them
Goals of Information Extraction

• A robust information extraction system
  – Recognize concepts and the implicit relations amongst them
  – Convert vast amounts of textual data into a semantic representation
Goals of Information Extraction

- A robust information extraction system
  - Recognize concepts and the implicit relations amongst them
  - Convert vast amounts of textual data into a semantic representation
  - Provide knowledge discovery tools for multiple analyst activities
Goals of Information Extraction

• **A robust information extraction system**
  – Recognize concepts and the implicit relations amongst them
  – Convert **vast** amounts of textual data into a semantic representation
  – Provide knowledge discovery tools for multiple analyst activities
    • **visual exploration**
    • **data-mining via NLP queries**
    • **link analysis**
High Level Task Description

- Evaluate the application of automatic knowledge extraction to link analysis
High Level Task Description

- Evaluate the application of automatic knowledge extraction to link analysis
  - Specialization of generic relations
  - Prototype IE to Link Analysis tool
High Level Task Description

• Evaluate the application of automatic knowledge extraction to link analysis
  – Specialization of generic relations
  – Prototype IE to Link Analysis tool

• Identify current technological barriers
High Level Task Description

- Evaluate the application of automatic knowledge extraction to link analysis
  - Specialization of generic relations
  - Prototype IE to Link Analysis tool

- Identify current technological barriers

- Establish high-payoff research directions
High Level Task Description

• Evaluate the application of automatic knowledge extraction to link analysis
  – Specialization of generic relations
  – Prototype IE to Link Analysis tool

• Identify current technological barriers

• Establish high-payoff research directions

• Produce substantive report on current state-of-the-art
KNOW-IT Overview

- Automatically identifies and extracts concepts and relations involving people, events, places, and organizations, etc from massive volumes of digital textual data.
- For purpose of building / adding to Knowledge Bases for use by human & automated reasoners.
- General technology capability
  - currently used for various text types & domains
  - can be specialized for specific applications.
KNOW-IT’s Building Blocks:

Natural Language Processing

+  

Knowledge Extraction

+  

Graphical Visualization
KNOW-IT components

• Concepts
  – 60 + Proper Noun Categories
Proper Noun Categorization Scheme

Geographic Entity
- City
- Port
- Airport
- Island
- County
- Province
- Country
- Continent
- Region
- Water
- Geo. Misc.

Affiliation
- Religion
- Nationality

Organization
- Company
- Company Type
- Government
- U.S. Government
- Organization

Human Document
- Person
- Title

Equipment
- Document
- Software
- Hardware
- Machines

Scientific
- Drugs
- Chemicals

Temporal
- Date
- Time

Misc.

KNOW-IT components

• Concepts
  – 60 + Proper Noun Categories
  – WordNet Synsets
Current Synset: (3571) attack, onslaught, onset, onrush
Definition: the beginning of an offensive; “the attack began at dawn”
KNOW-IT components

• Concepts
  – 60 + Proper Noun Categories
  – WordNet Synsets

• Relations
  – 40 + generic semantic relations
## Semantic Relations

<table>
<thead>
<tr>
<th>Relations</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGNT (act, animate)</td>
<td>animate is performer (agent) of action</td>
</tr>
<tr>
<td>PART (entity-x, entity-y)</td>
<td>entity-x has part entity-y</td>
</tr>
<tr>
<td>PTIM (T, time)</td>
<td>$T$ occurred at specific time</td>
</tr>
<tr>
<td>CAUS (state-x, state-y)</td>
<td>$x$ has a cause $y$</td>
</tr>
<tr>
<td>PURP (act-x, act-y)</td>
<td>act-x has purpose act-y</td>
</tr>
<tr>
<td>(state/entity, act-y)</td>
<td>state has purpose act-y</td>
</tr>
</tbody>
</table>
KNOW-IT components

- **Concepts**
  - 60 + Proper Noun Categories
  - WordNet Synsets

- **Relations**
  - 40 + generic semantic relations
  - Concept-Relation-Concept
Concept-Relation Extraction

HEADLINE: Albanian suspected to have links to bin Laden arrested
SOURCE: Agence France Presse, 01/10/99

Maksim Ciciku was arrested by the Albanian police in Tirana. … Ciciku met Osama bin Laden in April 1994.
Maksim Ciciku was arrested by the Albanian police in Tirana. … Ciciku met Osama bin Laden in April 1994.

CG_1  OBJ ( arrest, Maksim Ciciku|person )
AGNT ( arrest, Albanian police )
CHRC ( police, Albanian|nationality )
LOC ( arrest, Tirana|city )

CG_2  AGNT ( meet, Maksim Ciciku|person )
OBJ ( meet, Osama bin Laden|person )
PTIM ( meet, April 1994 )
Adapting KNOW-IT for Link Analysis

- Extraction in KNOW-IT is **broad and shallow**
  - based on linguistic regularities
  - not domain-dependent rules

- But the technology can be extended to **narrow and deep** applications for Link Analysis
  - ‘terrorism’ domain for HPKB
Specialization Methodology

- Map 2 or more general C-R-C extraction rules into a more specific link rule, e.g. for SUPPORT:
Specialization Methodology

• Map 2 or more *general* C-R-C extraction rules into a more specific link rule, e.g. for **SUPPORT**:

\[
\begin{align*}
C_1 - R - C_2 \\
+ \\
C_2 - R - C_3
\end{align*}
\]
Specialization Methodology:

- Map 2 or more general C-R-C extraction rules into a more specific link rule, e.g. for SUPPORT:

  C1 - AGNT - C2
  +
  C2 - OBJ - C3
Specialization Methodology

• Map 2 or more general C-R-C extraction rules into a more specific link rule, e.g. for SUPPORT:

```
C1 - AGNT - C2
  +
C2 - OBJ - C3
```

```<international agent*> AGNT <support verb*>  
  +
<support verb*> OBJ <X|54>```
Where,

International agent = any Proper Noun whose category is an element of the set \{7, 40, 411, 412, 413, 414, 415, 416, 417, 50, 501, 51, 52, 53, 54\}

AND

Support verb = any element of the synsets containing verbs such as: \{fund, back, support, aid, help, assist, sponsor, subsidize, patronize, cosponsor, bankroll, champion, defend\}
Then,

…. extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...
Then,

…. extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

AGENT (back, Osama bin Laden|person)

OBJECT (back, Hartakul Jihad|terrorist_group group)
Then,

…. extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe … the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden …
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe … the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden …

… the\DT extremist\JJ Harkatul_Jihad\NP\1 group\NN ,|, reportedly\RB backed\VBD by\IN Saudi\NP\2 dissident\IN Osama_bin_Lade\NP\3 …
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe … the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden …

… the|DT extremist|JJ Harkatul_Jihad|NP|1 group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP|2 dissident|IN Osama_bin_LadeNP|3 …

<PN>
1 54 Harkatul Jihad
2 17 Saudi
3 30 Osama bin Laden
</PN>
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe
… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

… the|DT extremist|JJ Harkatul_Jihad|NP|1 group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP|2 dissident|IN OsamabinLadeNP|3 …

<PN>
1  54  Harkatul Jihad
2  17  Saudi
3  30  Osama bin Laden
</PN>

CG0:  AGNT  (back , Osama bin Laden|person)
OBJECT  (back , Harkatul Jihad|terrorist_group group)
CHRC  (Harkatul Jihad|terrorist_group group , extremist)
MANR  (back , reportedly)
ISA  (Osama bin Laden|person , Saudi|nationality dissident)
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe
… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

… the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

<PN>
1  54  Harkatul Jihad
2  17  Saudi
3  30  Osama bin Laden
</PN>

CG0: AGNT  (back , Osama bin Laden|person|
OBJECT (back , Harkatul Jihad|terrorist_group group)
CHRC  (Harkatul Jihad|terrorist_group group , extremist)
MANR  (back , reportedly)
ISA  (Osama bin Laden|person , Saudi|nationality dissident)

CG0: AGNT  (back , Osama bin Laden|person|
OBJECT (back , Harkatul Jihad|terrorist_group group)
... SUPPORT(Osama bin Laden|person , Harkatul Jihad|terrorist_group group)
03/14/1999 (AFP) Bangladesh bomb blast toll 10, opposition wants judicial probe … the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

support (Osama bin Laden|person, Harkatul Jihad|terrorist_group group)
03/12/1999 (AFP) Bangladesh arrest Afghan war veteran over bomb attack
… from Monirul Hassan, a member of the Harkatul Jihad group who was reportedly trained by the Taliban militia in Afghanistan ...
03/12/1999 (AFP) Bangladesh arrest Afghan war veteran over bomb attack
… from Monirul Hassan, a member of the Harkatul Jihad group who was reportedly
trained by the Taliban militia in Afghanistan ...

CG0

ISA (Monirul Hassan|person, member)
AFFL (member, Harkatul Jihad|terrorist_group group)
AGNT (train, Taliban militia|organization)
OBJ (train, Monirul Hassan|person)
LOC (train, Afghanistan|country)
03/12/1999 (AFP) Bangladesh arrest Afghan war veteran over bomb attack
… from Monirul Hassan, a member of the Harkatul Jihad group who was reportedly
trained by the Taliban militia in Afghanistan ...

CG0

ISA (Monirul Hassan|person , member)
AFFL (member , Harkatul Jihad|terrorist_group group)

affiliate (Monirul Hassan|person , Harkatul Jihad|terrorist_group group)

AGNT (train , Taliban militia|organization)
OBJ (train , Monirul Hassan|person)

prep (Taliban militia|organization , Monirul Hassan|person)
03/12/1999 (AFP) Bangladesh arrest Afghan war veteran over bomb attack
… from Monirul Hassan, a member of the Harkatul Jihad group who was reportedly
trained by the Taliban militia in Afghanistan ...

affiliate  (Monirul Hassan|person, Harkatul Jihad|terrorist_group group)
prep     (Taliban militia|organization, Monirul Hassan|person)
03/12/1999 (AFP) Bangladesh arrest Afghan war veteran over bomb attack
… from Monirul Hassan, a member of the Harkatul Jihad group who was reportedly trained by the Taliban militia in Afghanistan ...

affiliate  (Monirul Hassan|person, Harkatul Jihad|terrorist_group group)
prep    (Taliban militia|organization, Monirul Hassan|person)
03/08/1999 (AFP) 16 soldiers killed, 21 wounded in Algerian ambush

The Salafist Group for Preaching and Combat (GSPC), led by Hassan Hattab, recently distributed …

Created at the instigation of bin Laden, the group is especially active ...
03/08/1999 (AFP) 16 soldiers killed, 21 wounded in Algerian ambush
The Salafist Group for Preaching and Combat (GSPC), led by Hassan Hattab, recently distributed …

Created at the instigation of bin Laden, the group is especially active ...

head   (Hassan Hattab|person , GSPC|terrorist_group)
support (Osama bin Laden|person , GSPC|terrorist_group)
03/08/1999 (AFP) 16 soldiers killed, 21 wounded in Algerian ambush
The Salafist Group for Preaching and Combat (GSPC), led by Hassan Hattab, recently distributed …

Created at the instigation of bin Laden, the group is especially active ...

head (Hassan Hattab|person, GSPC|terrorist_group)
support (Osama bin Laden|person, GSPC|terrorist_group)
02/15/1999 (AFP) Bin Laden held to be behind an armed Algerian Islamic movement

Mohamed Berrachad had worked for Hattab, who is himself a dissident from the Armed Islamic Group (GIA) …

In his testimony, Berrachad said Bin Laden and Hattab communicated by satellite telephone and that he had heard their conversations, said to hinge on the discrediting of Antar Zouabri's GIA by its savage massacres of civilians …

Diagram:

- Antar Zouabri lead GIA, disagree Mohamed Berrachad
- Taliban Militia prep Monirul Hassan
- Osama bin Laden support Harkatul Jihad group
- Hassan Hattab support head GSPC
- Mohamed Berrachad affiliate Harkatul Jihad group
- Antar Zouabri disagree Sandy Zouabri

CS470/670 – NLP (10/30/02)
As a link analyzer, KNOW-IT

- Assists analysts in appraising a potential crisis situation by determining the key players and the nature of their relations to one another
- Automatically filters, extracts, organizes, and analyzes textual intelligence data
- Generates and visualizes networks from relevant, unstructured text
- Allows analysts to specialize the links by easy-to-write specification & generalization rules
- Provides rich output to visualization tools