Natural Language Processing: An Introduction

NLP: The Ultimate Goal (1990)

The Ultimate Goal – For computers to use NL as effectively as humans do....

"Natural language, whether spoken, written, or typed, is the most natural means of communication between humans, and the mode of expression of choice for most of the documents they produce. As computers play a larger role in the preparation, acquisition, transmission, monitoring, storage, analysis, and transformation of information, endowing them with the ability to understand and generate information expressed in natural languages becomes more and more necessary."

NLP: Grand Challenges (1990)

Reading and writing text
  • Abstracting
  • Monitoring
  • Extraction into Databases

Interactive Dialogue: Natural, effective access to computer systems
  • Informal Speech Input and Output

Translation: Input and Output in Multiple Languages

Significant Advances In NLP I

• Web-scale information extraction & question answering
  • IBM’s Watson

• Interactive Dialogue Systems
  • Apple’s Siri
  • (Microsoft Cortana)
  • (Amazon Echo)
  • (Google Now)

Significant Advances In NLP II

Automatic Machine Translation

Early Successes: Human Machine Interfaces

• SHRDLU (Winograd, 1969)
  • A fragile demonstration of the fundamental vision

• LUNAR (Woods, Webber, Kaplan 1971)
  • Answering geologist’s questions about the Apollo 11 moon rocks
Review: SHRDLU: A demonstration proof

LUNAR – William Woods 1971

- NLP interface to database of analyses of Apollo 11 moon rocks
- Examples
  - What is the average concentration of aluminum in high alkali rocks?
  - How many breccias contain olivine?
  - Give me the modal analyses of those samples for all phases.
- Handled 78% of sentences typed by geologists at 1971 Lunar Rocks conference
  - (90% after "minor fixes")

The Past: Crucial flaws in the paradigm

These and other later systems worked well, BUT
1. Person-years of work to port to new applications
2. Very limited coverage of English

Crucially, they worked well because of a magical fact:

People automatically adapt and limit their language given a small set of exemplars if the underlying linguistic generalizations are HABITABLE

This won’t handle pre-existing text!

The State of NLP

NLP Past before 1995:
- Rich Representations

NLP Present:
- Powerful Statistical Disambiguation

1995: A breakthrough in parsing

10^6 words of Treebank Annotation + Machine Learning = Robust Parsers (Magerman '95)

1990 Best hand-built parsers: ~40-60% accuracy (guess)
• 1995* Statistical parsers: >90% accuracy (both on short sentences)

The Penn Treebank: 1988-94

- Wall Street Journal: 1.3 million words
- Brown Corpus: 1 million words
- Switchboard: 1 million words
- All Tagged with Part-of-Speech & Syntactic Structure
- Developed '88-94 (Marcus, Santorini, Taylor, Bies, …)
- Finished before it had any practical use!
A Few Core Technologies

1. Named Entity Recognition & Information Extraction
2. Machine Translation
3. Text Summarization

Information Extraction

- Information extraction is the identification, in text, of specified classes of Named Entities +
  - Relations
  - Events
- For relations and events, this includes finding the participants and modifiers (date, time, location, etc.).
- Goal: fill out a database with given relation or event types:
  - People's jobs
  - People's whereabouts
  - Merger and acquisition activity
  - Disease outbreaks
  - Genomics relation

Named Entity Recognition

The task: identify atomic elements of information in text
- Flag the who, where, when & how much in text
- Person names
- Company/organization names
- Locations
- Dates & times
- Percentages
- Monetary amounts

Information Extraction & Named Entity Recognition

Extraction Example

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
<th>Location</th>
<th>Person</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>European Information Services, Inc.</td>
<td>London</td>
<td>George Garrick</td>
<td>Out</td>
</tr>
<tr>
<td>CEO</td>
<td>Nielsen Marketing Research</td>
<td>USA</td>
<td>George Garrick</td>
<td>In</td>
</tr>
</tbody>
</table>

Lexicalized parsing results
(Labeled Constituent Precision/Recall F1)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFGs (Charniak 97)</td>
<td>73.0%</td>
</tr>
<tr>
<td>Conditional Models – Decision Trees (Magerman 95)</td>
<td>84.2%</td>
</tr>
<tr>
<td>Lexical Dependencies (Collins 96)</td>
<td>85.5%</td>
</tr>
<tr>
<td>Conditional Models – Logistic (Ratnaparkhi 97)</td>
<td>86.9%</td>
</tr>
<tr>
<td>Generative Lexicalized Model (Charniak 97)</td>
<td>86.7%</td>
</tr>
<tr>
<td>Generative Lexicalized Model (Collins 97)</td>
<td>88.2%</td>
</tr>
<tr>
<td>Logistic-inspired Model (Charniak 99)</td>
<td>89.6%</td>
</tr>
<tr>
<td>Boosting (Collins 2000)</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

(Chris Manning, Stanford)
Won't simple lists solve the problem?

- too numerous to include in dictionaries
- changing constantly
- appear in many variant forms
- subsequent occurrences might be abbreviated

⇒ list search/matching doesn't perform well

Information Extraction from Propositions

Propositions are normalized connections from the parse trees.

Entities and relations are extracted statistically from propositions.

Statistical Machine Translation

(For more on this topic, check out courses taught by Prof. Chris Callison-Burch)

(Next several slides from Language Weaver)
Translating a New Document

Broadcast Monitoring
BBN MAPS & Language Weaver MT

Language Weaver Hybrid Translation Technology

Text Summarization
(For more on this topic, check out courses taught by Prof. Ani Nenkova)