Natural Language Processing

Berlin Chen 2004
Textbooks & References

• Textbooks
  – C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999

• References
  – X. Huang, A. Acero, H. Hon, Spoken Language Processing, Prentice Hall, 2001
Motivation for NLP

• **Academic**: Explore the natural of linguistic communication
  – Obtain a better understanding of how language work

• **Practical**: Enable effective human-machine communication
  – Conversational agents are becoming an important form of human-computer communication
  – Revolutionize the way computers are used
    • More flexible and intelligent
Motivation for NLP

• Different Academic Disciplines: Problems and Methods
  – Electrical Engineering, Statistics
  – Computer Science
  – Linguistics
  – Psychology

• Many of the techniques presented were first developed for speech and then spread over into NLP
  – E.g. Language models in speech recognition
Turing Test

- Alan Turing, 1950

  Predicted at the end of 20 century a machine with 10 gigabytes of memory would have 30% chance of fooling a human interrogator after 5 minutes of questions

  Does it come true?
Hollywood Cinema

- Computers/robots can listen, speak, and answer our questions
  - E.g.: HAL 9000 computer in “2001: A Space Odyssey”
    (2001太空漫遊)
State of the Art

- Canadian computer program accepted daily weather data and generated weather reports (1976)
- MIT Spoken dialogue systems for information of restaurant, air travel, etc. (1991~)
- AT&T, How May I Help You?
  - Read student essays and grade them
  - Automated reading tutor
  - ......
State of the Art

- CMU Universal Speech Interface
Statistical and Non-Statistical NLP

• The dividing line between the two has become much more fuzzy recently
  – An increasing number of non-statistical researches use corpus evidence and incorporate quantitative methods
    • Corpus: “a body of texts” (大量的文稿)

  – Statistical NLP needs to start with all the scientific knowledge available about a phenomenon when building a probabilistic model, rather than closing one’s eye and taking a clean-slate approach
Models and Algorithms for NLP

• Models
  - Finite-state automata
  - Finite-state transducers
  - Markov models
  - Hidden Markov models

• Algoirithms
  - Search:
    - Dynamic programming, depth-first search, best-first search, A* search
  - Learning/Training Methods

Logic

First order logic (predicate calculus)
Semantic networks
Conceptual dependency

Regular grammars
Regular relations
Context-free grammars
Feature-augmented grammars

Finite-state automata
Finite-state transducers
Markov models
Hidden Markov models

State Machines

NLP

Formal Rule Systems

Speech Guys

Linguistics Guys

At Guys

Knowledge

Pragmatics

discourse

Syntax

morphology

Phonetics/Phonology

Speech Guys Linguistics Guys
Major Topics for NLP

• Probability Theory/Statistics
  – Supervised/Unsupervised Machine Learning Techniques

• Words
  – Morphology
  – Regular expressions
  – Automata, Finite-State Transducers

• Syntax
  – Part-of-Speech Tagging
  – (Probabilistic) Context-Free Grammar
  – Parsing
Major Topics for NLP

• Semantics/meaning
  – Representation of Meaning
  – Semantic Analysis
  – Word Sense Disambiguation

• Pragmatics
  – Natural Language Generation
  – Discourse, Dialogue and Conversational Agents
  – Machine Translation
# Topic List and Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:20</td>
<td><strong>Course Overview &amp; Introduction</strong>&lt;br&gt;<strong>Linguistic Essentials</strong></td>
</tr>
<tr>
<td>2:21</td>
<td>Linguistic Essentials&lt;br&gt;Regular Expressions and Automata</td>
</tr>
<tr>
<td>3:05</td>
<td>Mathematical Foundations 朱思錦</td>
</tr>
<tr>
<td>3:12</td>
<td>Part-of-Speech Tagging</td>
</tr>
<tr>
<td>3:19</td>
<td><strong>Break</strong> (ICDAT 2004)</td>
</tr>
<tr>
<td>3:26</td>
<td>Collections 張志豪&lt;br&gt;Parsing with Context-Free Grammars</td>
</tr>
<tr>
<td>4:00</td>
<td><strong>Break</strong></td>
</tr>
<tr>
<td>4:16</td>
<td>Word Sense Disambiguation 劉理論</td>
</tr>
<tr>
<td>4:23</td>
<td>Midterm</td>
</tr>
<tr>
<td>4:30</td>
<td>Text Categorization 鄭永幸</td>
</tr>
<tr>
<td>5:07</td>
<td>Probabilistic Context-Free Grammars</td>
</tr>
<tr>
<td>5:14</td>
<td>Paper Survey</td>
</tr>
<tr>
<td>5:24</td>
<td><strong>Break</strong> (ICASSP 2004)</td>
</tr>
<tr>
<td>5:28</td>
<td>Paper Survey</td>
</tr>
<tr>
<td>6:11</td>
<td>Semantics and Logical Form</td>
</tr>
<tr>
<td>6:18</td>
<td><strong>Final</strong></td>
</tr>
</tbody>
</table>

---

13
Applications of NLP

• Speech Recognition
• Information Retrieval and Extraction
• Summarization
• Question Answering
• Conversational Agents
• Machine (Speech/Language) Translation
• Spelling Check
• Segmentation and Alignment
• Bioinformatics
• ....
Resources

• Corpora (Speech/Language resources)
  – Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them
    • LDC - Linguistic Data Consortium
    • The Association for Computational Linguistics and Chinese Language Processing
Resources

• Institutes/People
  – Foreign
    • MIT
    • CU
    • CMU
    • JHU
    • UMass
    • Cambridge
    • Microsoft
    • IBM
    • MITRE
    • HP
    • ……………
Resources

• Conferences and Journals
  – **ACL**: Association for Computational Linguistics
  – **COLING**: International Conference on Computational Linguistics
  – **Computational Linguistics**
  – **Natural Language Engineering**
  – **ICSLP**: International Conference on Spoken Language Processing
  – **EUROSPEECH**: European Conference on Speech Communication and Technology
  – **ICASSP**: IEEE International Conference on Acoustics, Speech, Signal processing
  – **Speech Communication**
  – **Computer Speech and Language**
  – **IEEE Transactions on Speech and Audio Processing**