Language Processing with Perl and Prolog
Chapter 1: An Overview of Language Processing

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/
Applications of Language Processing

- Spelling and grammatical checkers: *MS Word*, e-mail programs, etc.
- Text indexing and information retrieval on the Internet: *Google*, *Microsoft Bing*, *Yahoo*, or software like *Apache Lucene*
- Translation: *Google Translate*, *SYSTRAN*
- Spoken interaction: Apple Siri, Google Now, *Tellme.com*, or *SJ* (trains in Sweden)
- Speech dictation of letters or reports: *IBM ViaVoice*, *Windows Vista*
Applications of Language Processing (ctn’d)

- Direct translation from spoken English to spoken Swedish in a restricted domain: SRI and SICS
- Voice control of domestic devices such as tape recorders: Philips or disc changers: MS Persona
- Conversational agents able to dialogue and to plan: TRAINS
- Spoken navigation in virtual worlds: Ulysse, Higgins
- Generation of 3D scenes from text: Carsim
- Question answering: IBM Watson and Jeopardy!
Linguistics Layers

- Sounds
- Phonemes
- Words and morphology
- Syntax and functions
- Semantics
- Dialogue
Sounds and Phonemes

Serious

C’est par là ‘It is that way’
Lexicon and Parts of Speech

The big cat ate the gray mouse

The/article big/adjective cat/noun ate/verb the/article gray/adjective mouse/noun
Le/article gros/adjectif chat/nom mange/verbe la/article souris/nom grise/adjectif
Die/Artikel große/Adjektiv Katze/Substantiv ißt/Verb die/Artikel graue/Adjektiv Maus/Substantiv
### Morphology

<table>
<thead>
<tr>
<th>Word</th>
<th>Root form</th>
</tr>
</thead>
<tbody>
<tr>
<td>worked</td>
<td><em>to work</em> + verb + preterit</td>
</tr>
<tr>
<td>travaillé</td>
<td><em>travailler</em> + verb + past participle</td>
</tr>
<tr>
<td>gearbeitet</td>
<td><em>arbeiten</em> + verb + past participle</td>
</tr>
</tbody>
</table>
The noun phrase

article

The

noun

boy

verb phrase

verb

hit

noun phrase

article

the

noun

ball
Syntax: A Classical View

A graph of dependencies and functions

Subject → Verb → Object

The boy hit the ball
Semantics

As opposed to syntax:
1. Colorless green ideas sleep furiously.
2. *Furiously sleep ideas green colorless.

Determining the logical form:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Logical representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank is writing notes</td>
<td>writing(Frank, notes).</td>
</tr>
<tr>
<td>François écrit des notes</td>
<td>écrit(François, notes).</td>
</tr>
<tr>
<td>Franz schreibt Notizen</td>
<td>schreibt(Franz, Notizen).</td>
</tr>
</tbody>
</table>
Word senses:

1. **note** (*noun*) short piece of writing;
2. **note** (*noun*) a single sound at a particular level;
3. **note** (*noun*) a piece of paper money;
4. **note** (*verb*) to take notice of;
5. **note** (*noun*) of note: of importance.
1. Sentence
   Pierre wrote notes

2. Logical representation
   wrote(pierre, notes)

3. Real world
   refers to

   Louis
   Pierre
   Charlotte

   refers to

   operating systems
   language processing
   Prolog programming
Ambiguity

Many analyses are ambiguous. It makes language processing difficult. Ambiguity occurs in any layer: speech recognition, part-of-speech tagging, parsing, etc.

Example of an ambiguous phonetic transcription:

*The boys eat the sandwiches*

That may correspond to:

*The boy seat the sandwiches; the boy seat this and which is; the buoys eat the sand which is*
Linguistics has produced an impressive set of theories and models
Language processing requires significant resources
Models and tools have matured. Resources are available.
Tools involve notably finite-state automata, regular expressions, logic, statistics, and machine learning.
## The Carsim System: A Text-to-Scene Converter

<table>
<thead>
<tr>
<th>Texts</th>
<th>XML Templates</th>
<th>3D Animation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Véhicule B venant de ma gauche, je me trouve dans le carrefour, à faible vitesse environ 40 km/h, quand le véhicule B, percute mon véhicule, et me refuse la priorité à droite. Le premier choc atteint mon aile arrière gauche,</td>
<td>// Static Objects Static [ ROAD TREE ] // Dynamic Objects Dynamic [ Vehicle [ ID = vehicule_b; INITDIRECTION = east;</td>
<td>![3D Animation Image]</td>
</tr>
</tbody>
</table>

NLP engine  Java 3D animation program
A conversation with Peedy

<table>
<thead>
<tr>
<th>Turn</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>User:</td>
<td>Good morning, Peedy.</td>
</tr>
<tr>
<td>Peedy:</td>
<td>Good morning.</td>
</tr>
<tr>
<td>User:</td>
<td>Let’s do a demo.</td>
</tr>
<tr>
<td>Peedy:</td>
<td>Your wish is my command, what would you like to hear?</td>
</tr>
<tr>
<td>User:</td>
<td>What have you got by Bonnie Raitt?</td>
</tr>
<tr>
<td>Peedy:</td>
<td>I have “The Bonnie Raitt Collection” from 1990.</td>
</tr>
<tr>
<td>User:</td>
<td>Pick something from that</td>
</tr>
<tr>
<td>Peedy:</td>
<td>How about “Angel from Montgomery”?</td>
</tr>
</tbody>
</table>
User: Sounds good.
[Peedy drops note on pile]

Peedy: OK.

User: Play some rock after that.
[Peedy scans the notes again, selects one]

Peedy: How about “Fools in love”?

User: Who wrote that?
[Peedy cups one wing to his ‘ear’]

Peedy: Huh?

User: Who wrote that?
[Peedy looks up, scrunches his brow]

Peedy: Joe Jackson

User: Fine.
[ Drops note on pile]
Persona System Architecture

Figure 1: System diagram of the Persona conversational assistant

IBM Watson: A system that can answer questions better than any human.

Video:
https://www.youtube.com/watch?v=WFR31Om_xhE


Bottom line: Text is the repository of human knowledge.
IBM Watson: Simplified Architecture

Question processing

Passage retrieval

Answer extraction

Question parsing and classification:
*Syntactic parsing, entity recognition, answer classification*

Document retrieval.
Extraction and ranking of passages:
*Indexing, vector space model.*

Extraction and ranking of answers:
*Answer parsing, entity recognition*