Semantic NLU
The Missing LNKR™
Brian Garr

- CEO and founder LinguaSys, Inc.
- Managed IBM’s Speech Recognition and MT Group
- Created WTS in 2000
- CTO of Globalink (an MT company) from 1995–1998
- 1998 Smithsonian Institution’s “Heroes in Technology” award for his work in Machine Translation
- Chapter contributor; *Where Humans Meet Machines: Innovative Solutions for Knotty Natural-Language Problems*, edited by Amy Neustein and Judith A. Markowitz (Jun 30, 2013)
Characteristics of Words

- They have meaning(s)
- They may have gender
- They may be countable
- They may change form when pluralized
- They generally have a morphology
- They may change meaning when paired with other words
- They may describe animate or inanimate things
- If a Verb, is it transitive?
- ........
More about words

- **Words have;**
  - **Synonyms** = Car, Automobile
  - **Homonyms** = Tank (armed vehicle), Tank (blow a test), Tank (a container of gasoline)
  - **Hyponyms** = Tank < Armored Vehicles, Military Vehicles
  - **Hypernyms** = Tank > Panzer, Sherman
  - **Domains** = Toppings for Pizzas > Anchovies, Pepperoni, Meatballs....
Words in multiple languages

- Physical boundaries? Not always
- Connection to the same concept in other languages?
  - One to One
  - One to Many
  - Many to Many
  - Many to One
  - Many to None
  - One to None
My definition of Semantics*
for purposes of this discussion

- The understanding of the concepts and ideas being presented in an utterance (textual or vocal) of words
  - What does the source intend it to mean?
    - Evaluate all of the words in the utterance
    - Look at the relationship to the other words in the utterance
    - Look for word phases
    - Disambiguate highly ambiguous words?
I need to fill the tank of the tank.
The Missing LNKR

- A=1 T=301 \%A=136357 T=112358 Y13=M
- ED \%A=452 T=323 \%A=128399 T=108811
- Y13=HIGH \%A=117 T=309 \%A=3509686 T=28805
- Y1 = JAR Y13 = LOW \%A=2761459 T=324
- A=37343 T=34500 Y13=MED

<table>
<thead>
<tr>
<th>Language</th>
<th>Family ID</th>
<th>Lemma / Description</th>
<th>Note</th>
<th>Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>26925</td>
<td>tank</td>
<td></td>
<td>3006</td>
</tr>
<tr>
<td>English</td>
<td>34499</td>
<td>tank</td>
<td></td>
<td>3734</td>
</tr>
<tr>
<td>English</td>
<td>34500</td>
<td>tank</td>
<td>a military tank</td>
<td>3734</td>
</tr>
<tr>
<td>English</td>
<td>34503</td>
<td>tank</td>
<td></td>
<td>3734</td>
</tr>
<tr>
<td>English</td>
<td>83371</td>
<td>tank</td>
<td></td>
<td>8880</td>
</tr>
<tr>
<td>English</td>
<td>107475</td>
<td>tank</td>
<td>&quot;tank animal refuse&quot;</td>
<td>12555</td>
</tr>
<tr>
<td>English</td>
<td>117887</td>
<td>tank</td>
<td></td>
<td>14856</td>
</tr>
<tr>
<td>English</td>
<td>28805</td>
<td>tank</td>
<td>gas tank</td>
<td>350968</td>
</tr>
</tbody>
</table>
The Carabao Data Manager
Carabao Runtime Architecture

Carabao–based application

Carabao Engine

Unstructured Content

Analysis

Synthesis

Semantic Objects or Transformed Content

Carabao Linguistic Database

English
Spanish
Brazilian Portuguese
German,
Chinese (Simplified & Traditional)
Japanese
Thai
Vietnamese
Malay
Pashto
Russian
Persian
Urdu
Hebrew
Arabic
French

Applications
English
Spanish
Brazilian Portuguese
German,
Chinese (Simplified & Traditional)
Japanese
Thai
Vietnamese
Malay
Pashto
Russian
Persian
Urdu
Hebrew
Arabic
French
## Carabao Engine

<table>
<thead>
<tr>
<th>Language Neutral Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique to LinguaSys. In the NLP space, not aware of any commercially available product that draws 100% of language processing information from the language models</td>
</tr>
<tr>
<td><strong>Benefit:</strong></td>
</tr>
<tr>
<td>Cheaper and faster to add new languages and modifications</td>
</tr>
<tr>
<td>No programming or application QA is required</td>
</tr>
<tr>
<td>Work can be done by trained computational linguists, not software engineers</td>
</tr>
<tr>
<td><strong>For Example:</strong></td>
</tr>
<tr>
<td>There is no incremental work to do to add new language support to an existing application, once the language model is finished</td>
</tr>
<tr>
<td>Language</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td>Spanish</td>
</tr>
<tr>
<td>Russian</td>
</tr>
<tr>
<td>S. Chinese</td>
</tr>
</tbody>
</table>

All applications support all languages immediately.
FORGET YOU SAW THAT CHART
Natural Language Understanding

- Can be hosted or embedded
- Understands 15 languages
  - 4 months to add new language
- No statistical models or embedded grammars
- Concept based
- Create new NLU apps in days, not months
NLU 2.0

- TEXT or SPEECH
- HOSTED or EMBEDDED
- 17 LANGUAGES
  - More each Q
- CONCEPT BASED NLU
- 10X DECREASE IN COST/SPEED TO BUILD
SEMANTIC UNDERSTANDING

• RESPONSE
  Convertible, 9/26/2013, 10/3/2013

• REQUEST
  Cabriolet, tomorrow, next Thu
Hotel Booking

Today’s date Tuesday, 10/08/13

- Accepts input in all supported languages
- Dialog management
- No SLMs
- No Embedded Grammars
  - But Out of Grammar messages
- Simple XML scripts
- Hotel Booking demo is 576 lines!
Analyze/Disambiguate every word/phrase

<table>
<thead>
<tr>
<th>Language</th>
<th>Family ID</th>
<th>Lemma / Description</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>91982</td>
<td>next</td>
<td>“next the doctor examined his back”</td>
</tr>
<tr>
<td>English</td>
<td>95926</td>
<td>next</td>
<td>“the following day”; “next in line”; “the next day”</td>
</tr>
<tr>
<td>English</td>
<td>97332</td>
<td>next</td>
<td>“had adjacent rooms”; “in the next room”</td>
</tr>
<tr>
<td>English</td>
<td>100859</td>
<td>next</td>
<td>“our next president”</td>
</tr>
</tbody>
</table>

 próximo [detected as FamilyID#95926, RecordID#2088372] at 27
 miércoles [detected as FamilyID#90929, RecordID#2103089] at 35
 por [detected as FamilyID#440, RecordID#2082661] at 45
 4 [detected as FamilyID#1999999999, RecordID#285] at 49
 4 noches [FamilyID#1101, RecordID#2179274] at 51
 4 noches [FamilyID#1101, RecordID#2176516] at 58
 4 noches [FamilyID#10946, RecordID#2103102] at 61
 en [detected as FamilyID#431, RecordID#2082469] at 64
 el [detected as FamilyID#309, RecordID#2023781] at 69
 Mamott [detected as FamilyID#138148, RecordID#2933774] at 73
 Orlando [detected as FamilyID#58564, RecordID#12938514] at 73

Changing an existing dictionary entry

<table>
<thead>
<tr>
<th>Family ID:</th>
<th>Language:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1101</td>
<td>ENG</td>
</tr>
</tbody>
</table>

number, identity 1
  category: large integer, identity 4
  number, identity 3
  (possible gap)
  allowed: regular noun
  allowed: adjective
  (possible gap)
plural, countable, regular noun, priority 1 (head), identity 2
not allowed: regular noun
English, Arabic, Traditional Chinese, Simplified Chinese, German, French, Hebrew, Japanese, Malay, Spanish, Pashto, Persian, Portuguese, Russian, Thai, Vietnamese, Urdu
Faster, Cheaper, and Better
Text and Speech Input
Multiple languages