Embracing the New Era of Natural Language Processing

Hang Li
Huawei Noah’s Ark Lab
Outline

• Why Is Natural Language Understanding Extremely Difficult?

• Winning Strategies
  – Task-driven
  – Hybrid

• Machine Translation as Example
Why Is Natural Language Understanding Extremely Difficult?

A Review from Linguistics, Cognitive Science, and Neuroscience
# Neural Computing v.s. Digital Computing

<table>
<thead>
<tr>
<th></th>
<th>Neural Computing</th>
<th>Digital Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scale</strong></td>
<td>10^11 neurons, 10^15 connections</td>
<td>10^10 transistors, sparse connections</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>100us</td>
<td>100ps (10GHz)</td>
</tr>
<tr>
<td><strong>Computing paradigm</strong></td>
<td>Parallel processing</td>
<td>Sequential processing</td>
</tr>
<tr>
<td><strong>Capability</strong></td>
<td>Mathematically Ill-posed problems</td>
<td>Mathematically well-formed problems</td>
</tr>
</tbody>
</table>

Modified from Rajesh Rao & Adrienne Fairhall
Natural Language Understanding by Computer

• Two definitions:
  – Representation-based: if system creates proper internal representation, then we say it “understands” language
  – Behavior-based: if system follows instruction in natural language, then we say it “understands” language, e.g., “bring me a cup of tea”

• We take the latter definition
Five Characteristics of Human Language

• Both Regular and Idiosyncratic
• Recursive
• Metaphorical
• Associated with World Knowledge
• Interactive
Language is a grassroots phenomenon. It is the original wiki, which aggregates the contributions of hundreds of thousands of people who invent jargon, slang and new constructions, some of them get accumulated into the language as people seek out new ways of expressing their thoughts. And that’s how we get a language in the first place.
Language Is Both Regular and Idiosyncratic

• Language has been continuously invented by hundreds of thousands of people over tens of thousands of years

• New lexicons are continuously added and discarded, e.g., “不明觉厉”

• Grammar is more stable, but new constructions are still occasionally invented
  – Simplification, e.g., “I shall” → “I will”
  – Influence from other languages, e.g., in African American English “I working” is legitimate expression

• As result, there always exist rules and exceptions in language, and all grammars leak.
If a grammar has no recursive steps it will be prohibitively complex. If it does have recursive devices, it will produce infinitely many sentences.
Language Is Recursive

• Lexicon may have history of over 1 million years
• Grammar may have history of 50 thousand years
• Champagne may speak some “words”, but they do not speak language, i.e., they do not have ability of recursively constructing language expressions
• Homo Sapiens started to speak language about 50 thousand years ago
• Recursion is key property of language
Most of our ordinary conceptual system is metaphorical in nature. The way we think, what we experience, and what we do every day is very much a matter of metaphor.
Language Is Metaphorical

• Metaphor = connecting two concepts based on their *hidden* similarity, e.g., “在微信里潜水”

• A 4-year-old English-speaking boy was observed to creatively say “open the light”, instead of “turn on the light”

• Language is conventional, once a metaphoric expression is widely accepted, it will be regularly used, e.g., “下厨房、上厕所”

• Metaphor depends on culture, environment, etc, e.g., “warm affection” exists in many languages, but not in those languages in tropical areas
When a person sees or hears a sentence, he makes full use of his knowledge and intelligence to understand it. This includes not only grammar, but also his knowledge about words, the context of the sentence, and most important, his understanding of the subject matter.
Language Is Associated with World Knowledge

• The same area in pre-motor cortex of human gets activated when he does something or imagines doing it
• It is believed that the same part of brain is activated, when you experience something or you imagine doing it
• E.g., seeing “a person is drinking water”, reading from book “a person is drinking water”
• Understanding language = making simulation in brain
• Language understanding involves access to world knowledge stored in whole brain
Language appears to be a central means by which cognitive processes are extended into the world. It may be that language evolved, in part, to enable such extensions of our cognitive resources within actively coupled systems.
Language Is Interactive

• Action-perception loop is key feature of human intelligence

• Language is essentially tool of human human interaction

• Human’s language ability can only be activated when it is used in interaction with environment, cf., wolf child

• Language understanding needs understanding of environment
The Kitten Carousel Experiment
- Held and Hein, 1963

- Twin cats are born; one is taken as “active cat”, the other “passive cat”
- On day time, they are put onto a carousel; active cat is on the ground and can walk, but passive cat is on a basket and cannot walk; carousel can be rotated by active cat
- On evening time, they are put into a dark place, eating and sleeping
- Two months later, both cats are released; active cat can walk normally, but passive cat cannot
Five Characteristics of Human Language

• Both Regular and Idiosyncratic
• Recursive
• Metaphorical
• Associated with World Knowledge
• Interactive
Natural Language Understanding by Computer Is Extremely Difficult

• It is still not clear whether it is possible to realize human language ability on computer

• On modern computer
  – The rule-exception and recursion characteristics imply combinatorial computation
  – The metaphor, knowledge, and interaction characteristics imply exhaustive computation

• Big question: can we invent new computer closer to human brain?
Why Does Natural Language Understanding Appear to be Easy?

• Language has regular and recursive structures, which parts can be relatively easily captured by rules
• People tend to ignore the fact that language understanding is largely conducted in sub-consciousness
• Feeling that language understanding is easy is an illusion, just like rainbow
• Human can learn native language without difficulty before 12 year-old
Outline

• Why Is Natural Language Understanding Extremely Difficult?

• Winning Strategies
  – Task-driven
  – Hybrid

• Machine Translation as Example
Two Strategies: Task-Driven and Hybrid
Task-Driven Strategy

• Simplified tasks
  – Question answering, including search
  – Machine translation

• Restricted tasks
  – Multi-turn dialogue

• Avoid the difficult challenge of language understanding

• Machines can “pretend” to understand language

• AI Loop: natural language processing systems can be continuously improved in the loop of system-user-data-algorithm
Simplified Problem Definition
- Question Answering

Question answering, including search, can be practically performed, because it is simplified.
Simplified Problem Definition - Machine Translation

Machine translation can be practically performed, because it is simplified.
Multi-turn dialogue may be practically performed, if it is restricted to specific domain.
Advancement in AI, including NLP can be made through the closed loop.
Hybrid Strategy

• NLP Technologies
  – First generation: Rule-based NLP
  – Second generation: Statistical NLP
  – Third generation: Neural NLP
• Hybrid = rule-based + statistical + neural
• Will converge to human’s language ability
Like Cubism Arts

Pablo Picasso,
1910, Girl with a Mandolin
Prediction: Five to Ten Years from Now

• Question answering: widely used in work and life, just like search today
• Translation: widely used in travel and several scenarios
• Multi-turn dialogue: widely used in a number of domains such as hotel booking
Outline

• Why Is Natural Language Understanding Extremely Difficult?

• Strategies
  – Task-driven
  – Hybrid

• Machine Translation as Example
Researchers

Zhaopeng Tu
Zhengdong Lu
Xiaohua Liu
Xiao Chen
Noah Neural Machine Translation (NMT) System
Noah NMT System

A cat is sitting on the mat

- Coverage vector can eliminate over-translation and under-translation
- Context gates can dynamically change the attention mechanism
- One more unpublished technique

一只猫坐在那张垫子上
Experimental Result

- Coverage vector, context gates can continuously improve accuracy of translation
- Training Data: 1.25 million LDC data (Chinese-English)
- Test Data: NIST datasets (Chinese-English)
Experimental Result

- Google NMT system works better, apparently due to its larger training data and more powerful computing architecture
- Google NMT system also employs coverage mechanism
Big Topic: Combination of NMT, SMT, and even Rule-based MT
Experimental Result

- NMT is good at fluency and SMT is good at adequacy
- A hybrid approach of NMT and SMT works better
- Training Data: 1.25 million LDC data (Chinese-English)
- Test Data: NIST datasets (Chinese-English)
Summary

• Why Is Natural Language Understanding Extremely Difficult?

• Winning Strategies
  – Task-driven
  – Hybrid

• Machine Translation as Example
References

• Steven Pinker. *Linguistics as a Window to Understanding the Brain, Big Think*, 2013.
• George Lakoff, Mark Johnson, *Metaphors We Live by*, 1980.
Thank you!