Natural Language Processing

NLP

Artificial Intelligence → NLP

Artificial Intelligence & Natural Language Processing

The use of computers for modeling and performing certain problem-solving tasks that were, prior to the invention of the computer, thought to be uniquely human.

The "processing" of our language is one of the most essential of such uniquely human task.

What is

Natural Language Processing

?

The processing of natural language?

Natural Language Processing

Language processing that is natural?

What is a natural language?

- A human language spoken or written in a community
  - English, Swahili, Chinese, …
- Not an artificial language created to serve a specific purpose
  - Matlab, Java, C

Why study NLP?

- Science
  - Computational linguistics
- Engineering
  A. Question answering
  B. Summarization
  C. Machine translation
     - Started with the 1950s effort in the US to translate texts, using computers, from foreign languages to English

A. Question Answering

- "Who was the voice of Miss Piggy?"
- How high is mount Everest?
- Who were the three Stooges?

- Different from "keyword search"
- Three-step approach:
  1. Find relevant documents
  2. Figure out what the question is asking for
  3. Match the question to the document

B. Summarization

- Given a set of documents on the same topic, produce a summary

Tougher problem…

- Automatically "crawl" the Internet looking for documents on the same topic, then produce a summary

Summarization—2002 example

The U.S. Supreme Court agreed Tuesday to hear a case that could determine when hundreds of thousands of books, songs and movies will become freely available over the Internet or in digital libraries. A nonprofit Internet publisher and other plaintiffs argue that Congress sided too heavily with writers and other creators when it passed a law in 1998 that retroactively extended copyright protection by 20 years. On Tuesday, the U.S. Supreme Court announced it would hear a challenge to the 1998 Copyright Term Extension Act, in which Congress extended the term of existing and future copyrights by 20 years. Billions of dollars and the future earning power of some of the nation's most cherished cultural icons are at stake as the U.S. Supreme Court considers a constitutional challenge to a 1998 copyright extension law, legal experts said Wednesday.

NewsBlaster (Columbia U)

Summarization—2011 example

Norway attacks: Police search farm for clues after shooting, explosion leave 94 dead A minute's silence was held in Norway on Monday to remember the people killed in the bomb blast and shooting last week. OSLO, Norway - The self-described perpetrator of Norway's deadly bombing and shooting rampage was ordered held in solitary confinement after calmly telling a court that two other cells of collaborators stood ready to join his murderous campaign. Norwegian police said Monday that the double-counting of bodies in the chaotic aftermath of a shooting spree may have contributed to a dramatic overestimate of the number of people slain, but they offered few other details about the error.

NewsBlaster (Columbia U)

Summarization—another 2011 example

Hugo Chavez cancer: Venezuela government says President Hugo Chavez will remain in charge
Venezuelan President Hugo Chavez's closest associates insisted Friday that their leader would continue running the country even as he continues to undergo medical treatment in Cuba after surgery for the removal of a cancerous tumor. Firefighter-paramedic Dave Chavez is in charge of the Venezuelan government and never was not in charge the vice president , Elias Jaua, said on state television. The image many Venezuelans have is not of a leader on the mend but of a thin, pale president who looked vulnerable and emotional as he delivered a 13-minute address Thursday night in which he revealed his condition.

NewsBlaster (Columbia U)

Subtasks in Summarization

- Web crawling to identify articles on same topic
- Select sentences covering different ideas to be included
- Select sentences covering the same idea and reformulate sentences to express the idea
- Arrange the sentences to form a coherent summary
- A final rewrite for greater fluency

- NewsBlaster currently takes 4-12 hours to produce one summary

C. Machine Translation

- Want to
  - Read books (or webpages) in any language
  - Communicate with anyone in the world
- → Automatically translate from one language to another

How does NLP work?

- Solve many sub-problems
  - Part-of-speech tagging
  - Parsing
  - Word sense disambiguation
  - …

How does NLP work?

- Solve many sub-problems
  - Part-of-speech tagging
  - Parsing—draw tree

Symbolic methods (rule-based)

The cat is on the mat

How does NLP work?

- Solve many sub-problems
  - Part-of-speech tagging
  - Parsing—draw tree
  - Word sense disambiguation

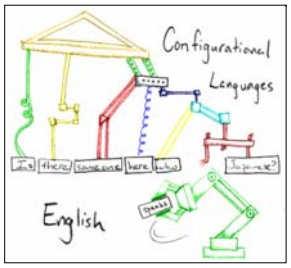
Need context. Use statistical methods?

Those folks on the beach had a ball

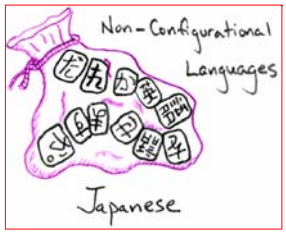
Why is NLP hard?

- After all, kids can do it!
- NLP is hard because of
  - Ambiguity in natural languages
  - The need for world knowledge or context
  - Data sparsity
  - Complexity of specific languages

A cartoon comparison of configurational and non-configurational languages



A standard "Subject Verb Object" (SVO) language



By Phineas Q. Phlogiston

Phonological ambiguity

It's very hard to recognize speech
It's very hard to wreck a nice beach

Syntactic ambiguity

- I saw a man with a telescope
- Visiting relatives can be annoying
- I made her duck

Real headlines!

Doctor Helps Dog Bite Victim
British Left Waffles on Falkland Islands
Kids Make Nutritious Snacks
Local High School Dropouts Cut in Half

More ambiguity

- No left turn Wednesdays 4-6PM except transit vehicles
- Thank you for not smoking, drinking, eating, or playing radios without headphones while on the bus
  - *Thank you for not eating without headphones?*
- My neighbor's hat was taken by the wind. He tried to catch it.

Jan Hajic

North American Computational Linguistics Olympiad (NACLO)

- North American branch of the international competition

Maasai

by Doris L. Payne

Maasai is a language spoken by about 800,000 people in East Africa, mostly in Kenya and Tanzania.

As with many languages in East Africa, "tone" is very important in Maasai. The different tones are written as marks above some letters. For example, the letters á, í and ó are all pronounced with high tone. The letters à, ì and ò are all pronounced with low tone. If there is no mark over a letter, it is pronounced with "mid tone," half way in between high and low.

There are also some letters in the Maasai alphabet that are not used in English. For example, "ɔ" is a sound like the English word "awe." "ɛ" is similar to the vowel sound in "let," "ʊ" is like the vowel sound in "hood" and "ɪ" is like the vowel sound in "lit." You don't need to be able to pronounce these words in order to solve the problem, however, you should pay very close attention to the letters and the tone marks.

The following are some sentences in Maasai, and the English translations in random order. Indicate which translation goes with each Maasai sentence by placing the letter of the correct translation in the space provided:

1. éósh ɔlmʊraní ɔlásʊráì ____
2. áadɔ́l ɔlásʊráí ____
3. áaósh ɔlmʊraní ____
4. ídɔ́l ɔlmʊránì ____
5. íóshokí ɔlmʊránì ɔlásʊráì ____
6. ádúŋokí ɔlmʊránì ɔlcɛtá ____
7. ádúŋ ɔlcɛtá ____
8. áaduŋokí ɔlmʊraní ɔlcɛtá ____
9. áadúŋ ɔlmʊraní ____
10. édúŋ ɔlmʊraní ____

English translations in *random order*

A. 'The warrior cuts me.'
B. 'The warrior cuts the tree for me.'
C. 'The warrior cuts it.'
D. 'I cut the tree for the warrior.'
E. 'The warrior hits me.'
F. 'You see the warrior.'
G. 'The warrior hits the snake.'
H. 'The snake sees me.'
I. You hit the snake for the warrior.'
J. 'I cut the tree.'

Machine Translation

AI → NLP → Machine Translation

Is machine translation possible?

- Why not?
  - Poetry, literature (Shakespeare, Goethe, …)
  - "AI-Complete" – a computational problem as difficult as solving the central AI problem of making computers intelligent like human
- So what counts as success?
  - Is the first pass "good enough"?
  - Add a post-editing process to make the result useful?

Machine translation approaches

"I went to the red river's bank"

- Word-for-word translation (dictionary lookup)
- Rule-based
- Statistical
- Interlingua vs. language-to-language

Applications of machine translation

- One to many
  - Translate instruction manuals into many languages
- Many to one
  - Read newspapers in any language
- Many to many
  - European parliamentary proceedings
- Cross-lingual information retrieval
- Speech-to-speech

Where is research focused?

- Initial research was on Russian-English
- Now focused on most widely-spoken languages (Mandarin, English, Hindi/Urdu, Spanish, etc.)

Evaluation

- What is a "good" translation?

- Grammatical?
  - Easy: translate every sentence as "The cat is on the mat"
- Faithful?
  - Word-for-word translation (but what about idioms, words with multiple meanings, etc.)
- BLEU score
  - Compares a translation with reference translations statistically. Correlates well to human judgment (for an overall translation, not individual sentences)

Challenges

- Word order
  - Reporters said IBM bought Lotus
  - Reporters IBM Lotus bought said
- Word sense
  - River bank vs. money bank
  - House plant vs. power plant
- Idioms
  - Kick the bucket
  - Have a ball
- Pronouns
  - Esta aqui
    - It is here; she is here; he is here; you are here

More challenges in spoken language

The emphasis gives additional information…

- "**I** never said she stole my money"
- "I **never** said she stole my money"
- "I never **said** she stole my money"
- "I never said **she** stole my money"
- "I never said she **stole** my money"
- "I never said she stole **my** money"
- "I never said she stole my **money**"

Statistical machine translation

- Currently the dominant strategy
- Need lots and lots of data
- More data → better translations

- More later when we discuss *machine learning*