What is NLP?

A possible definition:

> *The attempt to get computers to process* human *languages in* textual *form in a way that utilises* knowledge of language *in order to perform some useful task.*

Note:

- NLP deals with human languages (English, Chinese, Swahili)
  - ◇ *not* artificial languages (Modula-2, C, Predicate Calculus)
- NLP deals with language in its textual or graphical form
  - ◇ the related field of Speech Processing deals with spoken language
- NLP addresses tasks requiring the use of *knowledge of language*
  - ◇ tasks such as string matching or generation of pre-stored text are not usually considered NLP

Disciplines Studying Human Natural Language

- Human language also studied by various other disciplines, including:
  - ◇ *Linguistics*:
    - studies language as a formal system, i.e. the structure of languages
    - uses intuitions of well-formedness and meaning
  - ◇ *Psycholinguistics*:
    - studies processes of human language production and comprehension
    - uses experimental investigation of performance
  - ◇ *Philosophy*:
    - studies nature of meaning, how words carry it, how words refer to objects in the world; what it means to have beliefs and intentions
    - uses philosophical debate rooted in intuitions of meaning
  - ◇ *Computational Linguistics*:
    - develops computational theories/models of aspects of language and language processing
    - draws on above other disciplines + computer science and maths

Why Study NLP ?

Linguistic or cognitive science motivation

- to understand how humans communicate using language
- hard to underestimate the significance of language:
  - ◇ Language is a distinguishing feature of human beings
    - Other animals can communicate
    - But no other animal can, e.g build arbitrarily complex sentence structures, refer to non-present times/places or to non-existent situations
  - ◇ Language is the key medium of social interaction
    - Hard to think of a social structure that is NOT predicated on language: business, education, courts, government, news, . . .
    - Most scholars agree writing systems arose as a means of recording business transactions
    - Dunbar (1996) argues gossip replaces grooming for humans

      Up to 10-20% of apes time spent grooming (necessary not for hygiene but to maintain social cohesion, defuse conflict)

      Speech took over this role in hairless apes with abnormal vocal fluency

Why Study NLP ?

Linguistic or cognitive science motivation (cont)

- hard to underestimate the significance of language (cont):
  - ◇ Language is the basis of collective memory/culture
    - Written language literally makes history possible
    - Facilitates building on ideas/discoveries of previous generations without which we would not have witnessed the spectacular expansion of the human species over the past 5,000 years
  - ◇ Language is intimately bound up with thought
    - Philosophers have long argued whether non-linguistic thought is possible i.e. whether all thought takes place in language
      Even those who reject this strong claim (e.g. Dennett) still argue that language constrains the brain/thinking
    - Psychologists, such as Vygotsky, interested in the cognitive development of the child, argue that learning word meaning and concept acquisition in the child go hand in hand

Why Study NLP ?

Technological or engineering motivation

- to build computer systems that can perform tasks that require understanding of textual language — application areas include:
    - ◇ Machine Translation
    - ◇ NL Database Interfaces
    - ◇ Information Retrieval
    - ◇ Information Extraction
    - ◇ Automatic Summarisation
    - ◇ Question Answering
    - ◇ Text Categorisation/Clustering
    - ◇ Plagiarism/Authorship Detection
    - ◇ NL Generation
    - ◇ Dialogue /Integrated Speech and Language Systems

Why Language Processing is so hard for a Computer?

Language is

- dynamic - new words ("ecotourist") and word senses ("literally","gay")
- highly ambiguous
  - ◇ multiple possible meanings for words ("crane", "bank")
  - ◇ multiple possible syntactic structures for word sequences
    "One morning I shot an elephant in my pajamas.
    How he got into my pajamas I dont know." (Groucho Marx)
- complex
  - ◇ no as yet discovered comprehensive set of grammar rules
  - ◇ 10s of thousands of lexical items
- partial - requires the reader/listener to supplement the text/utterance with world/conceptual knowledge in order to recover the message
  - ◇ "He put the computer on the chair. It was (wobbly | still hot)".

A Very Sketchy History of NLP

- 1940's
  - ◇ First electronic digital computers
  - ◇ Computers used in code breaking (A. Turing)
- 1950's
  - ◇ First transistorised computers
  - ◇ Lisp invented
  - ◇ Turing proposes language generation/understanding as key test of intelligence (*Turing Test*)
  - ◇ Automata theory; Chomsky's work on formal language theory and *Syntactic Structures*; Harris's Transformations and Discourse Analysis
  - ◇ Machine Translation (MT) efforts begin
- 1960's
  - ◇ Minicomputers introduced
  - ◇ Transformational Grammar; Brown Corpus
  - ◇ Weizenbaum's *Eliza*
  - ◇ *Systrans* first commercial MT system

A Very Sketchy History of NLP (ctd)

- 1970's
  - ◇ Microcomputers introduced
  - ◇ Prolog invented
  - ◇ Procedural grammars; Schank's model of conceptual knowledge; Montague semantics; statistical models for speech recognition
  - ◇ Winograd's SHRDLU; Wood's LUNAR
- 1980's
  - ◇ Rapid expansion of *Internet*
  - ◇ Move to declarative representations in grammar
  - ◇ Growing interest in corpus-based approaches
  - ◇ Machine readable dictionaries and online mono/multi-lingual corpora
- 1990's
  - ◇ Explosion of *Internet* and availability of on-line text
  - ◇ Hybrid formal/statistical approaches in NLP
  - ◇ Rise of quantitative evaluation regimes for NLP systems
  - ◇ Appearance of speech recognition products

A Very Sketchy History of NLP (ctd)

- 2000's
  - ◇ Mass storage costs plummet allowing researchers access to massive, local digital text collections
  - ◇ Explosion of work on machine learning, corpus-based NLP approaches
  - ◇ NLP technologies in:
    - search engines (spelling correction, ranking, summarisation, MT, named entity recognition)
    - mobile phones (word expansion)
    - word processors (grammar and spelling checkers, summarisation)
  - ◇ Advent of the "Semantic Web"
    - possibilities for NLP as enabling technology

Early Systems: Eliza – A Non-directive Therapist

```
(U = User; E = Eliza)
U:   Men are all alike.
E:   IN WHAT WAY?
U:   They’re always bugging us about something.
E:   CAN YOU THINK OF A SPECIFIC EXAMPLE?
U:   Well my boyfriend made me come here
E:   YOUR BOYFRIEND MADE YOU COME HERE?
U:   He says I’m depressed much of the time.
E:   I’M SORRY TO HEAR YOU ARE DEPRESSED.
U:   It’s true.  I’m unhappy.
E:   DO YOU THINK THAT COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?
U:   I need some help, that much seems certain.
E:   WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP?
U:   Perhaps I could learn to get along with my mother.
E:   TELL ME MORE ABOUT YOUR FAMILY.
```

Eliza – A Non-directive Therapist (ctd)

U: My mother takes care of me.

E: WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?

U: My father.

E: YOUR FATHER?

U: You are like my father in some ways.

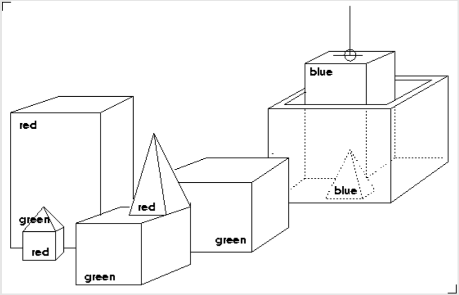
E: WHAT RESEMBLANCE DO YOU SEE?

U: You are not very aggressive but I think you don't want me to notice that.

E: WHAT MAKES YOU THINK I AM NOT VERY AGRESSIVE?

....

Early Systems: SHRDLU – A Blocks World System



U: Pick up a big red block.

S: OK.

U: Grasp the pyramid.

S: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

U: Find a block which is taller than the one you are holding and put it in the box.

S: BY 'IT', I ASSUME YOU MEAN "THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING".

U: What does the box contain?

S: THE BLUE PYRAMID AND THE BLUE BLOCK.

U: What is the blue pyramid supported by?

S: THE BOX.

Application Areas

- Machine Translation
- NL Database Interfaces
- Information Retrieval
- Information Extraction
- Automatic Summarisation
- Question Answering
- Text Categorisation/Clustering
- Plagiarism/Authorship Detection
- NL Generation
- Dialogue /Integrated Speech and Language Systems

Machine Translation

- Try "round trip" translation using Google Translate.
  - ◇ E.g. try translating:

    We have not the faintest idea whether the first words spoken were uttered 20,000 years ago or 200,000 years ago. What is certain is that mankind did little except procreate and survive for 100,000 generations. (For purposes of comparison, only about eighty generations separate us from Christ.) Then suddenly, about 30,000 years ago, there burst forth an enormous creative and cooperative effort which led to the cave paintings at Lascaux, the development of improved, lightweight tools, the control of fire, and many other cooperative arrangements. It is unlikely that any of this could have been achieved without a fairly sophisticated system of language.

    from English to French and back again.
  - ◇ Now try it from English to Chinese ...

Information Extraction/Text Mining

- The recognition of **entities** and **relations** in text
  - ◇ *Tata's victory over Brazil's CSN last week in a dramatic auction for Corus created the world's fifth-biggest steel maker.*
- Typically recognised information marked up using XML

```
<entity id=1 type=company>Tata</entity>'s victory
over
<entity id=2 type=country>Brazil</entity>'s
<entity id=3 type=company>CSN</entity>
last week in a dramatic auction for
<entity id=4 type=company>Corus</entity>
created the world's fifth-biggest steel maker.
<relation id=1 type=takeover arg1=1 arg2=4>
```

  or mapped into a structured representation called a **template**.
- Extracted information can be used:
  - ◇ to generate summaries
  - ◇ for document search, navigation and browsing ("semantic web")
  - ◇ for mining to discover trends or patterns

Question Answering

- Supply precise answers to questions rather than retrieving documents
  How tall is the Eiffel Tower?
- See, e.g. Wolfram Alpha's on-line system:
  `http://www.wolframalpha.com`
- Watch IBM's Watson win at Jeopardy in February 2011:
  `https://www.youtube.com/watch?v=WFR3lOm_xhE`
- Try Apple's Siri: `https://www.apple.com/uk/ios/siri/`

Natural Language Generation – Adaptive Hypermedia

- ILEX: the **I**ntelligent **L**abelling **EX**plorer:
  - generates dynamic hypertext describing objects in a ‘virtual gallery’
    - the jewellery collection of the National Museums of Scotland
  - uses a single, abstract (i.e. non-natural language) description of the objects, facts about them, and relations between them
  - from this, generates different textual descriptions to take account of:
    - differing interests of different visitors, e.g. style vs. artist:

> *This jewel is a necklace and is in the Organic style.*
> *It was made in 1976. It is made from opals, diamonds and pearls.*
>
> *This jewel is a necklace and was made by Gerda Flockinger, who was a designer and was English. The jewel, which is in the Organic style, was made in 1976.*

    - differing amounts of time available for visit
    - navigation history: compare to objects already viewed / avoid repitition
- Abstract underlying representation: allows generation to >1 language
- Other NL generation applications:
  - technical manuals (aircraft, software)
  - intelligent tutoring systems (tropical medicine, mathematics)

Spoken Dialogue Systems

- A Spoken Dialogue System requires a speech recognition component, but this is not enough: once the words are recognized, must
  - ◇ understand the content of the user utterances
  - ◇ *manage* the dialogue, so as to provide information efficiently/effectively
- Sample successful dialogue from the Dutch public transport information service's Voice Input Output System (VIOS):

Computer	From which station to which station do you want to travel?
Caller	I want to go from Amsterdam to Maastricht
Computer	On which day do you want to travel from Amsterdam CS to Maastricht?
Caller	Tomorrow
Computer	At what time do you want to travel tomorrow?
Caller	I want to leave at about 11 o'clock
Computer	So you want to leave between 10:30 and 11:30?
Caller	Yes

Spoken Dialogue Systems (ctd)

⋮

Computer Departure from Amsterdam at 10:33 arrival in Maastricht at 13:14 . . . Do want another travel advice?

Caller No

Computer OVR wishes you a pleasant trip, until next time.

- Other Spoken Dialogue System applications:
  - ◇ voice access to email

    e.g. AT&T's ELVIS: EmaiL Voice Interactive System
  - ◇ call centre automation
  - ◇ toys and games

Reading

Major sources:

- C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 2003
- D. Jurafsky and J. Martin, Speech and Language Processing, Prentice-Hall, 2007 (2nd edn)
- S. Bird, E. Klein, and E. Loper. Natural Language Processing Analyzing Text with Python and the Natural Language Toolkit, http://www.nltk.org/book. Also published by OReilly.