Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Khoa CNTT – Trường ĐHBKHN
Email: huonglt-fit@mail.hut.edu.vn

General Information

Lecturer: Lê Thanh Hương

- Department of Information Systems, Faculty of Information Technology
- Office: R325 C1
- Email: huonglt-fit@mail.hut.edu.vn, huongthanh@gmail.com
- Mobile: 0904674102

References

1. Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.
2. Dan Jurafsky and James Martin. 2000. *Speech and Language Processing.* PrenticeHall.
3. James Allen. 1994. *Natural Language Understanding*. The Benajmins/Cummings Publishing Company Inc.

General Information

- **Assessment**
  - Attendance: 10%
  - Assignment: 90%
- Assignment**:**
  - Writing a literature review report or implementating an application
  - Defending results on the last two weeks of the semester
- **Website:** http://is.hut.vn/~huonglt/NLP

What is NLP?

What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
Blah blah GINGER blah blah blah blah blah blah GINGER blah blah blah blah blah

- Goal: understand multi-languages
- It's not simple as string processing or keyword mapping

Applications:

- Difficult: speech processing, machine translation, information extraction, dialog interface, question answering, intelligent Web searching, text summarization, text generation, …
- Easy: spell checking, document classification, …

ELIZA

- http://en.wikipedia.org/wiki/ELIZA
- http://www.cs.bham.ac.uk/research/projects/cogaff/eliza/

Talk with Eliza

- E: how are you today?
- H: not so good
- E: why do you say not so good?
- H: i suffer from allergies?
- E: that's interesting
- H: i come out in spots
- E: i'm not sure i know what you're talking about
- H: spots on my face
- E: your face
- H: that's right
- E: is it really that right?

Talk with Eliza

- H: you’re a program, aren’t you?
- E: does it please you to believe i am a program, am not i?
- H: it sure does
- E: i see
- H: yup
- E: a while back you talk about your face
- H: what about your face?
- E: what is it you really want to know?
- H: are you my kind of girl?
- E: perhaps in your fantasies i am your kind of girl
- H: in my fantasies you are blonde

Talk with Eliza

- E: what makes you think i am blonde?
- H: i don’t know
- E: don’t you really know?
- H: no
- E: are you saying no just to be negative?
- H: no
- E: you are being a bit negative
- H: no i’m not
- E: how long have you been not?

Course Goals

- Learn the basic principles and theoretical approaches underlying NLP
- Learn techniques and tools which can be used to develop practical, robust systems that can (partly) understand text or communicate with users in one or more languages
- Gain insight into many of the open research problems in natural language

Topics in NLP

- **Levels of Analysis:** syntax, semantics, discourse, pragmatics, world knowledge...
- **Subproblems:** part-of-speech tagging, syntactic parsing, word sense disambiguation, discourse processing...
- **Algorithms and Methodologies:** corpus-based methods, knowledge-based techniques,...
- **Applications:** information extraction, information retrieval, machine translation, question answering, natural language understanding....

Levels of Analysis and Knowledge Used in NLP

- **Morphology (hình thái học):** how words are constructed; prefixes & suffixes
- **Syntax (cú pháp):** structural relationships between words
- **Semantics (ngữ nghĩa):** meanings of words, phrases, and expressions
- **Discourse (diễn ngôn):** relationships across different sentences or thoughts
- **Pragmatic (thực chứng):** the purpose of a statement; how we use language to communicate
- **World Knowledge (tri thức thế giới):** facts about the world at large; common sense

Morphology

English: metamorphosis (biến hình), multisyllable

- kick, kicks, kicked, kicking
- sit, sits, sat, sitting
- murder, murders

v: nhồi nhét; n: những cái đã ăn, hẻm núi

But it’s not just as sir dding and deleting endings...

rực rỡ

- gorge, gorgeous
- arm, army

Cánh tay

Quân đội

Vietnamese: non-metamorphosis, monosyllable → word segmentation

(Read Chapter 3 - Speech and Language Processing)

Word segmentation

- A phrase may have n word compositions, but only one of them are correct.
- Simple solution: determines the longest syllable sequence which starts at the current position and is listed in the lexicon.
- Problems: overlapping candidate words
  - Học sinh | học sinh | học.
  - Học sinh | học | sinh học.

☞ List all possible segmentations and design a strategy to select the most probable correct one.

Syntax: part-of-speech tagging (gán nhãn từ loại)

The boy threw a ball to the brown dog.

- The/DT boy/NN threw/VBD a/DT ball/NN to/IN the/DT brown/JJ dog/NN./.

DT – determiner
VBD – verb, past tense
JJ – adjective

NN – noun, single or mass
IN – preposition, sub-conj
. – sentence final punc

Syntax: structural ambiguity (part of speech)

Time flies like an arrow.

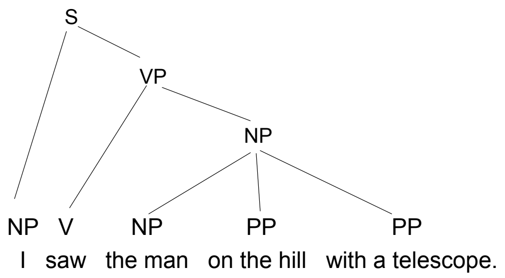
Time // flies like an arrow.

VBZ comparative proposition (IN)

Time flies // like an arrow.

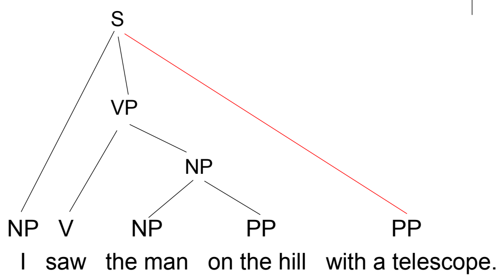
NNS VBP

Syntax: structural ambiguity (attachment)



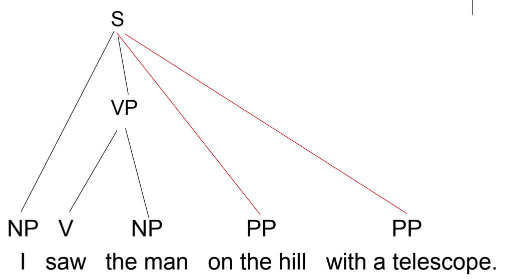
S
VP
NP
NP V NP PP PP
I saw the man on the hill with a telescope.

Syntax: structural ambiguity (attachment)



S
VP
NP
NP V NP PP PP
I saw the man on the hill with a telescope.

Syntax: structural ambiguity (attachment)



S
VP
NP V NP PP PP
I saw the man on the hill with a telescope.

But syntax doesn’t tell us much about meaning

- Colorless green ideas sleep furiously. [Chomsky]
- fire match arson hotel
- plastic cat food can cover

Semantics: lexical ambiguity

- I walked to the bank ...
  - of the river.
  - to get money.
- The bug in the room ...
  - was planted by spies.
  - flew out the window.
- I work for John Hancock ...
  - and he is a good boss.
  - which is a good company.

Discourse: coreference

President John F. Kennedy was assassinated.

The president was shot yesterday.

Relatives said that John was a good father.

JFK was the youngest president in history.

His family will bury him tomorrow.

Friends of the Massachusetts native will hold a candlelight service in Mr. Kennedy’s home town.

Pragmatics

What should you conclude from the fact that I said something? How should you react?

Rules of Conversation

- Can you tell me what time it is?
- Could I please have the salt?

Speech Acts

- I bet you $50 that the Jazz will win.

World Knowledge

John went to the diner. He ordered a steak. He left a tip and went home.

- What did John eat for dinner?
- Who brought John his food?
- Who cooked the steak?
- Did John pay his bill?

Knowledge of language: What do we know about this sequence?

- Words must appear in a certain order:
  *Dogs icecream ate
- Parts and divisions:
  dogs = Subject; ate icecream = Predicate
- Who did what to whom:
  agent(dogs), action(ate), object(ice-cream)

Anything else?

- The two sentences "John claimed the dogs ate icecream" and "John denied the dogs ate ice-cream" are logically incompatible
- Sentence & the world: know whether the sentence is true or not - perhaps whether in some particular situation (possible world) the dogs did indeed eat icecream
- "I had espresso this morning, but John is intelligent" looks odd.

What is the character of this knowledge?

- Some of it must be memorized:
  - Singing → Sing+ing; Bringing → bring+ing
- ***Duckling → ?? Duckl +ing***
- So, must know *duckl* is not a word
- But it can't all be memorized because there is too much to know

Besides memory, what else do we need?

English plural:

- Toy+s -> toyz ; add z
- Book+s -> books ; add s
- Church+s -> churchiz ; add iz
- Box+s-> boxiz ; add iz

➢ *must be a rule system to generate/process infinite #examples*

"Parsing" = mapping from surface to underlying representation

- What makes NLP hard: there is not a 1-1 mapping between any of these representations!
- We have to know the data structures and the algorithms to make this efficient, despite exponential complexity at every point

LSAT / (former) GRE Analytic Section Questions

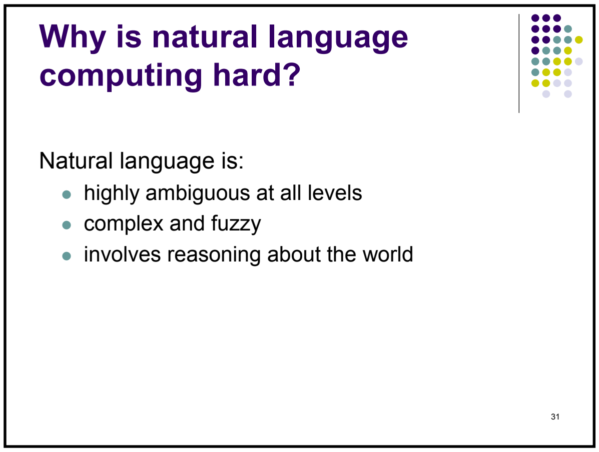
- Six sculptures – C, D, E, F, G, H – are to be exhibited in rooms 1, 2, and 3 of an art gallery.
  - Sculptures C and E may not be exhibited in the same room.
  - Sculptures D and G must be exhibited in the same room.
  - If sculptures E and F are exhibited in the same room, no other sculpture may be exhibited in that room.
  - At least one sculpture must be exhibited in each room, and no more than three sculptures may be exhibited in any room.
- If sculpture D is exhibited in room 3 and sculptures E and F are exhibited in room 1, which of the following may be true?

A. Sculpture C is exhibited in room 1
B. Sculpture H is exhibited in room 1
C. Sculpture G is exhibited in room 2
D. Sculptures C and H are exhibited in the same room
E. Sculptures G and F are exhibited in the same room

Reference Resolution

U: Where is A Bug's Life playing in Mountain View?
S: A Bug's Life is playing at the Summit theater.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show.
How much would that cost?

- Knowledge sources:
  - Domain knowledge
  - Discourse knowledge
  - World knowledge

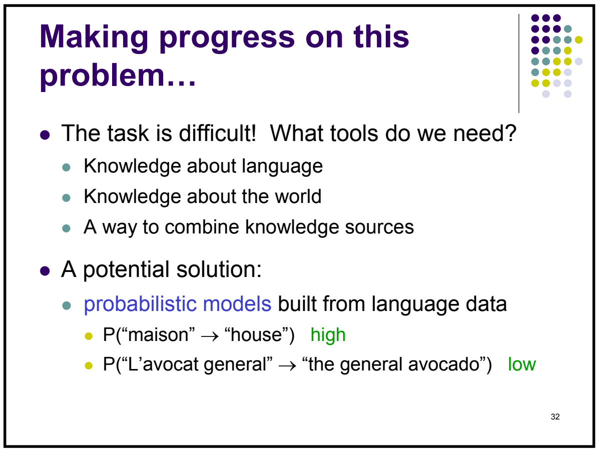


Why is natural language computing hard?

Natural language is:

- highly ambiguous at all levels
- complex and fuzzy
- involves reasoning about the world

31



Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- A potential solution:
  - probabilistic models built from language data
    - P("maison" → "house")   high
    - P("L'avocat general" → "the general avocado")   low

32