Natural Language

Universidad Carlos III de Madrid

AI

Communication

- Definition: [Russell and Norvig, 1995]

  Intentional exchange of information that is carried out through the emision and reception of signs belonging to a conventional system.

Communication

- Definition: [Russell and Norvig, 1995]

  Intentional exchange of information that is carried out through the emision and reception of signs belonging to a conventional system.

- Automatic communication management requires the integration of several disciplines
  - Engineering: voice recognition
  - Computer science: computational models of analysis
  - Linguistics: formal models of language
  - Psychology: intentions, semantics

Motivation

- A complete dialog system
  - input: oral natural language
  - output: oral natural language
- Advantages of oral natural language
  - oral natural language communication is more natural and easy
  - it does not force us to adapt to the computer, as other interfaces do (mouse, keyboard)
  - in some environments, they are safer (vehicles)
  - facilitates the access to elders or disabled people

Motivation

- A complete dialog system
  - input: oral natural language
  - output: oral natural language
- Advantages of oral natural language
  - oral natural language communication is more natural and easy
  - it does not force us to adapt to the computer, as other interfaces do (mouse, keyboard)
  - in some environments, they are safer (vehicles)
  - facilitates the access to elders or disabled people
- HAL-2001:
  `http://www.youtube.com/watch?v=HwBmPiOmEGQ`

Languages

- Types:
  - formal: C, C++, Java, Prolog, Lisp, . . .
  - natural: English, Spanish, . . .

Languages

- Types:
  - formal: C, C++, Java, Prolog, Lisp, . . .
  - natural: English, Spanish, . . .
- They share phrase structures
  - words are made up by joining symbols
  - sentences are made up by joining words
  - there are terminal symbols, non-terminal symbols, and rewriting rules

$$
\begin{array}{ll}
\text{If} & ::= \text{if Cond then Body} \mid \text{if Cond then Body else Body} \\
\text{Cond} & ::= \text{Cond and Cond} \mid \text{Cond or Cond} \mid \text{not Cond} \\
 & ::= \text{Exp} = \text{Exp} \mid \text{Exp} \mathrel{!}= \text{Exp} \mid \text{Exp} > \text{Exp} \mid \ldots \\
\text{Exp} & ::= \text{Exp} + \text{Exp} \mid \text{Exp} - \text{Exp} \mid \text{Exp} * \text{Exp} \mid \text{Exp} / \text{Exp} \mid \ldots
\end{array}
$$

Languages

- Types:
  - formal: C, C++, Java, Prolog, Lisp, . . .
  - natural: English, Spanish, . . .
- They share phrase structures
  - words are made up by joining symbols
  - sentences are made up by joining words
  - there are terminal symbols, non-terminal symbols, and rewriting rules

$$
\begin{array}{lll}
\text{If} & ::= & \text{if Cond then Body} \mid \text{if Cond then Body else Body} \\
\text{Cond} & ::= & \text{Cond and Cond} \mid \text{Cond or Cond} \mid \text{not Cond} \\
 & ::= & \text{Exp} = \text{Exp} \mid \text{Exp} \mathrel{!}= \text{Exp} \mid \text{Exp} > \text{Exp} \mid \ldots \\
\text{Exp} & ::= & \text{Exp} + \text{Exp} \mid \text{Exp} - \text{Exp} \mid \text{Exp} * \text{Exp} \mid \text{Exp} / \text{Exp} \mid \ldots
\end{array}
$$

$$
\begin{array}{lll}
\text{S} & ::= & \text{NP VP} \\
\text{NP} & ::= & \text{det A noun} \mid \text{det noun} \\
\text{VP} & ::= & \text{verb NP} \mid \text{verb adv NP} \\
\text{A} & ::= & \text{adj A} \mid \text{adj}
\end{array}
$$

Complexity: Ambiguity

- Phonology: *flew, flu*

Complexity: Ambiguity

- Phonology: *flew, flu*
- Lexical: *flies: noun, verb, . . .*

Complexity: Ambiguity

- Phonology: *flew, flu*
- Lexical: *flies: noun, verb, . . .*
- Syntactic
  - *I saw the Duero flying to Madrid*
  - *Time flies like an arrow*

Complexity: Ambiguity

- Phonology: *flew, flu*
- Lexical: *flies: noun, verb, . . .*
- Syntactic
  - *I saw the Duero flying to Madrid*
  - *Time flies like an arrow*
- Semantics: *hard (opposite of easy and soft)*

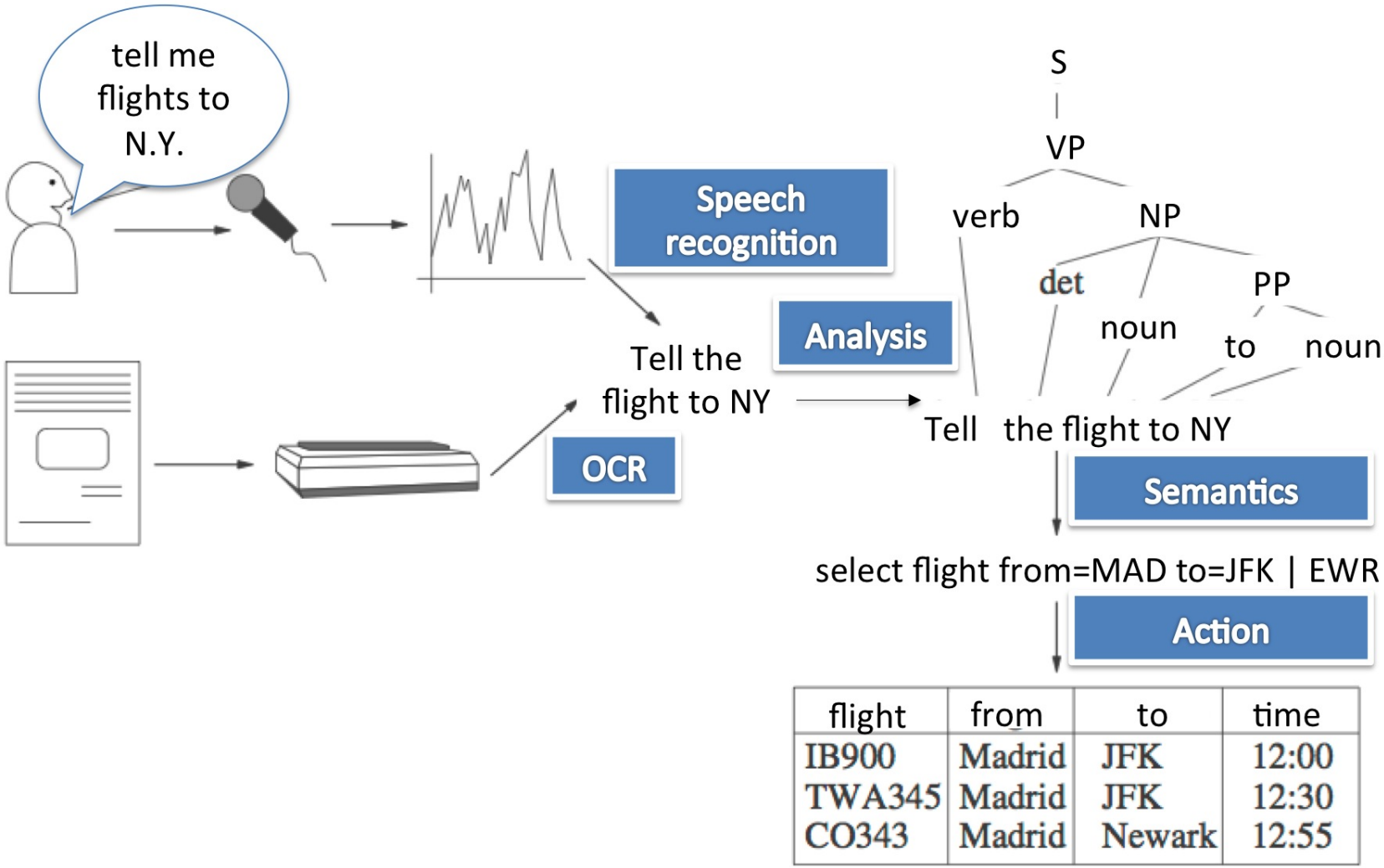
Complexity: Ambiguity

- Phonology: *flew, flu*
- Lexical: *flies: noun, verb, . . .*
- Syntactic
  - *I saw the Duero flying to Madrid*
  - *Time flies like an arrow*
- Semantics: *hard (opposite of easy and soft)*
- Discourse (referential)
  - *I took the cake from the table and cleaned it*
  - *I took the cake from the table and ate it*

Complexity: Ambiguity

- Phonology: *flew, flu*
- Lexical: *flies: noun, verb, . . .*
- Syntactic
  - *I saw the Duero flying to Madrid*
  - *Time flies like an arrow*
- Semantics: *hard (opposite of easy and soft)*
- Discourse (referential)
  - *I took the cake from the table and cleaned it*
  - *I took the cake from the table and ate it*
- Pragmatics: *Can you open the door?*

Handling natural language



Speech recognition

- Given some signal that corresponds to a spoken sentence, return the sentence
- Techniques
  - neural networks
  - HMMs: $P(W_{t+1} \mid W_t), P(A \mid W)$
    - A: acoustic model (signal)
    - W: language model (phonemes)
- Tools
  - Sphinx (`http://cmusphinx.sourceforge.net/`)
  - HTK (`http://htk.eng.cam.ac.uk/`)
  - Loquendo (`http://loquendo-speech-suite.software.informer.com/`)
  - Jasper (`http://jasperproject.github.io/`)

Analysis steps

- Lexical: how words are formed

  gives: (verb give, present, third person, singular)

Analysis steps

- Lexical: how words are formed

  gives: (verb give, present, third person, singular)

- Syntax: how words join in more complex structures (*grammars*)

  The (article) big (adjective) man (noun) runs (verb)

Analysis steps

- Lexical: how words are formed

  gives: (verb give, present, third person, singular)

- Syntax: how words join in more complex structures (*grammars*)

  The (article) big (adjective) man (noun) runs (verb)

- Semantics: meaning of sentences

  Colorless green ideas sleep furiously (Chomsky)

Analysis steps

- Lexical: how words are formed

  gives: (verb give, present, third person, singular)

- Syntax: how words join in more complex structures (*grammars*)

  The (article) big (adjective) man (noun) runs (verb)

- Semantics: meaning of sentences

  Colorless green ideas sleep furiously (Chomsky)

- Discourse: how are sentences combined

  He gave him the book. I told him

Analysis steps

- Lexical: how words are formed

  gives: (verb give, present, third person, singular)

- Syntax: how words join in more complex structures (*grammars*)

  The (article) big (adjective) man (noun) runs (verb)

- Semantics: meaning of sentences

  Colorless green ideas sleep furiously (Chomsky)

- Discourse: how are sentences combined

  He gave him the book. I told him

- Pragmatics: what you really wanted to say

  Do you know the time?

Lexical analysis

- Use of a dictionary
  - an already built one (WordNet, `http://wordnet.princeton.edu/`)
  - automatically built from corpora

Lexical analysis

- Use of a dictionary
  - an already built one (WordNet, `http://wordnet.princeton.edu/`)
  - automatically built from corpora
- Part-of-speech (POS) tagging: determine words and their lexical labels (*move* $\rightarrow$ VERB, NOUN)

Lexical analysis

- Use of a dictionary
  - an already built one (WordNet, `http://wordnet.princeton.edu/`)
  - automatically built from corpora
- Part-of-speech (POS) tagging: determine words and their lexical labels (*move* $\rightarrow$ VERB, NOUN)
- Stemming: remove parts of words that do not provide meaning (*move, moves, moved* $\rightarrow$ *mov*)

Lexical analysis

- Use of a dictionary
  - an already built one (WordNet, `http://wordnet.princeton.edu/`)
  - automatically built from corpora
- Part-of-speech (POS) tagging: determine words and their lexical labels (*move* $\rightarrow$ VERB, NOUN)
- Stemming: remove parts of words that do not provide meaning (*move, moves, moved* $\rightarrow$ *mov*)
- Optional: remove Stop words (*a, the,* . . .)

Syntactic and semantics analysis. First steps (1950-75)

- Pattern recognition
  - ELIZA (Weizenbaum)
  - PARRY (Colby)
- Generative linguistics (Chomsky)
- Augmented Transition Networks (Woods)

Dialog with ELIZA (Emacs-doctor)

http://www-ai.ijs.si/eliza/eliza.html

I have problems with my father.

Tell me something about your family.

My mother sleeps well.

You seem to dwell on your family.

Yes, it is a problem.

Why do you say yes it is a problem?

Because I do not like my family.

family what ?

dog

Is it because you do not like your family that you came to me?

Yes.

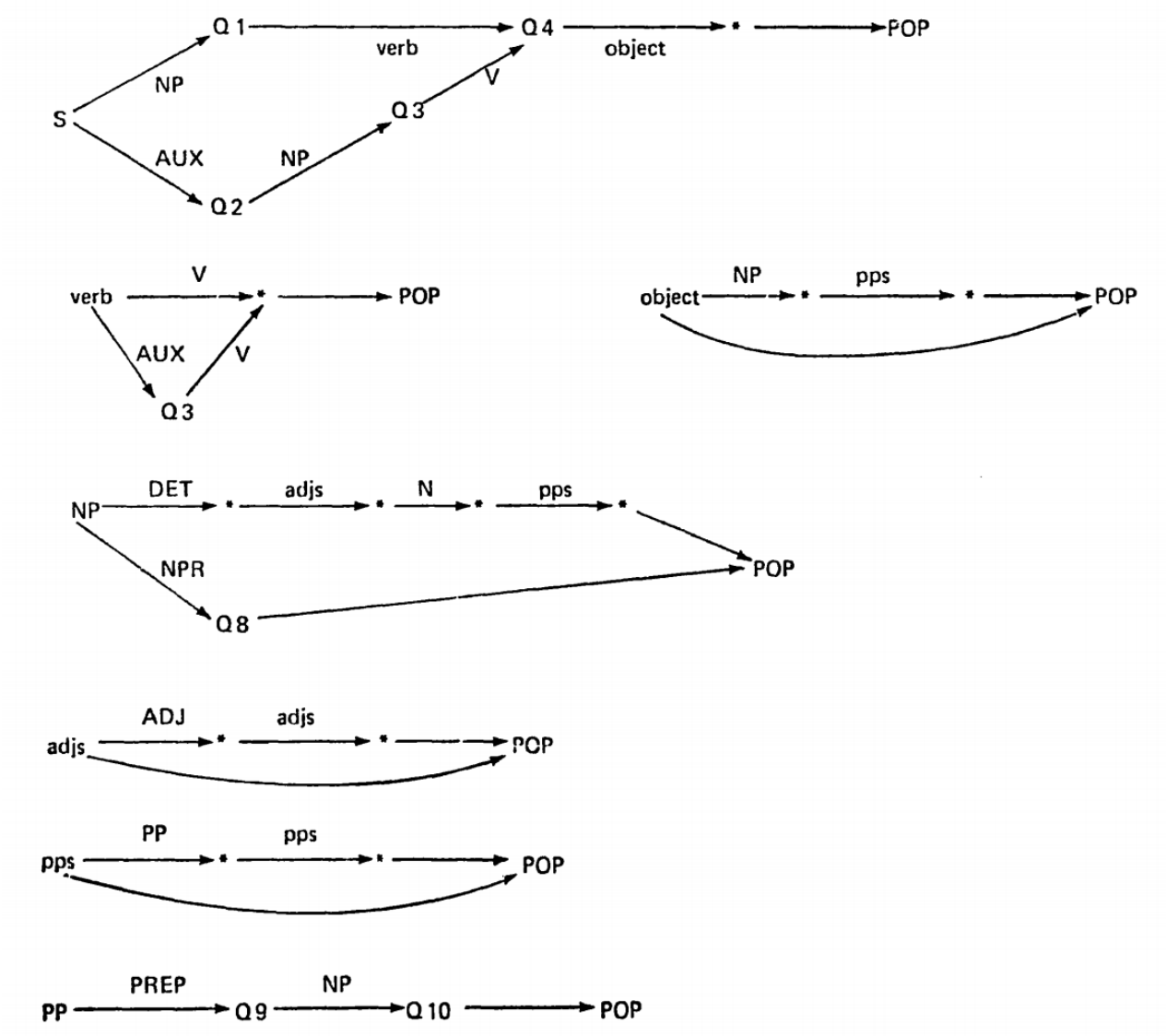
I see... Well, what makes you believe this is so?

...

Generative linguistics: Chomsky

S ::= NP VP
NP ::= det A noun | det noun
VP ::= verb NP | verb adv NP
A ::= adj A | adj

Augmented Transition Networks (ATN) [Woods, 1970]



Knowledge-based (1975-85)

- Conceptual dependency (Shank) *Luis moved the table*

Knowledge-based (1975-85)

- Conceptual dependency (Shank) *Luis moved the table*

- Case grammar/frames (Fillmore)

Frame moved	
agent	Luis
beneficiary	?
object	table
time	past

Frame table	
category	noun
number	singular
gender	female

Knowledge-based (1975-85)

- Conceptual dependency (Shank) *Luis moved the table*

- Case grammar/frames (Fillmore)

Frame moved	
agent	Luis
beneficiary	?
object	table
time	past

Frame table	
category	noun
number	singular
gender	female

- Definite Clause Grammars [Pereira and Warren, 1980]

```
sentence(sentence(S,V)) - ->
        subject(Num,Per,S),
        verb(Num,Per,V).
```

Definite Clause Grammars (DCG)

sentence(s(NP, VP)) → noun_phrase(NP), verb_phrase(VP).
noun_phrase(np(Det,Noun,Rel))→ determiner(Det), noun(Noun),
rel_clause(Rel).
noun_phrase(np(Name)) → name(Name).
verb_phrase(vp(TV,NP)) → trans_verb(TV), noun_phrase(NP).
verb_phrase(vp(IV)) -, intrans_verb(IV).
rel_clause(rel(that,VP)) → [that], verb_phrase(VP).
rel_clause(rel(nil)) → [].
determiner(det(W)) [W], is_determiner(W).
noun(n(W)) → [W], is_noun(W).
name(name(W)) –. [W], is_name(W).
trans_verb(tv(W)) → [W], is_trans(W).
intrans_verb(iv(W)) → [W], is_intrans(W).
is_determiner(every).
is_noun(man).
is-name(mary).
is_trans(loves).
is_intrans(lives).

Currently (1985-)

- Statistical approach: based on analyzing huge quantities of texts
  - machine learning
  - HMM
    - *n*-grams: sequences of *n* continuous words
    - Bayes theorem
    - probabilistic automata
  - Probabilistic grammars
- Hybrid techniques
- Overview of tools
  `http://www-nlp.stanford.edu/links/statnlp.html`

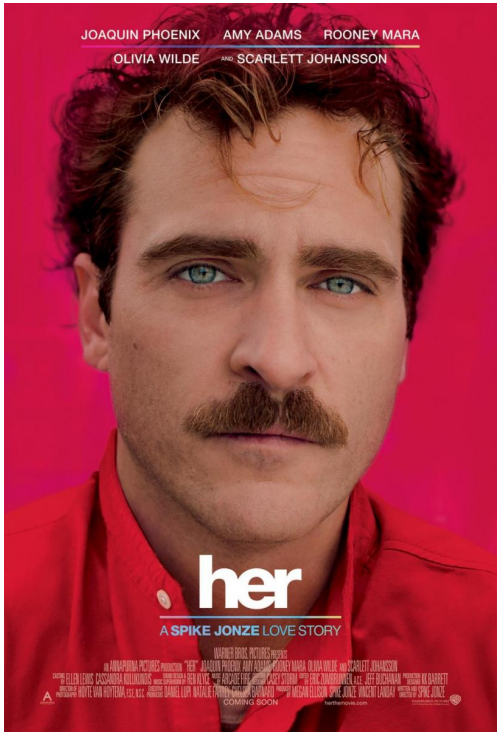
Applications

- Computer control
- Phone applications: Siri
- Dictation systems: ViaVoice (IBM), Voice Xpress (Lernout & Hauspie)
- Vehicles control: AutoPC (Clarion)
- Machine translation: Systrans
- Information extraction: question answering
- Text summarization
- Text mining
- Ontology learning
- Orthographic and grammatical correction
- OCR
- Conversational agents
- DNA analysis

Some examples

- Simple applications
  - word counters (wc in UNIX)
  - orthographic and grammatical correctors
  - text prediction (cell phones)
  - chatbots: A.L.I.C.E. (`http://alicebot.blogspot.com/`)
- Bigger applications
  - Siri
  - Yahoo, Google, Microsoft: information retrieval
  - Monster.com, HotJobs.com (Job finders): matching
  - Systran (Babelfish, `http://www.babelfish.com`): machine translation
  - Ask Jeeves (`http://es.ask.com/`), Quora (`https://www.quora.com/`): question answering
  - Myspace, Facebook, Blogspot: contents mining
  - all big ones have research groups: IBM, Microsoft, AT&T, Xerox, Sun, etc.

Her



http://www.youtube.com/watch?v=WzV6mXIOVl4

References

Fernando C. N. Pereira and David H. D. Warren.

Definite clause grammars for language analysis – A survey of the formalism and a comparison with augmented transition networks.
Artificial Intelligence, 13:231–278, 1980.

Stuart Russell and Peter Norvig.

Artificial Intelligence: A Modern Approach.
Prentice Hall, 1995.

W. A. Woods.

Transition network grammars for natural language analysis.
Communications of the ACM, 13, 1970.