CS395-T Topics in Natural Language Processing

Sanda Harabagiu

Department of Computer Sciences

University of Texas at Austin

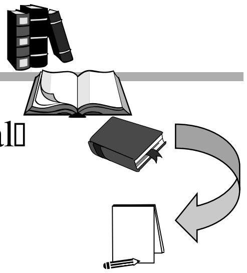
Tu, Th 11AM-12:20 PM

sanda@cs.utexas.edu

http://www.cs.utexas.edu/users/sanda/cs395T.html

http://www.engr.smu.edu/~sanda/cs395T.html

an exciting challenge...

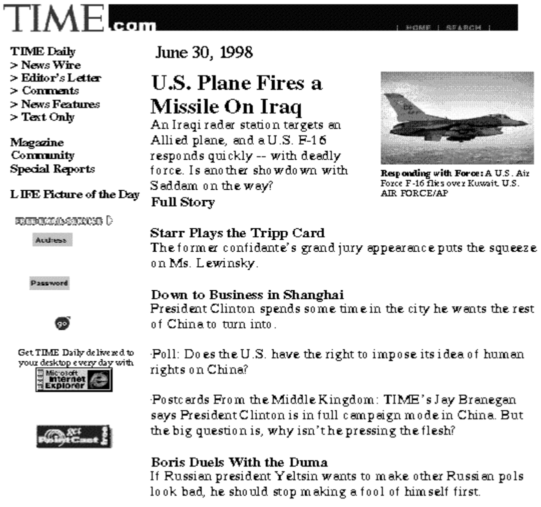


...put a book on the scanner, turn the dial to '2 pages', and read the result...

...download 1000 documents from the web, send them to the summarizer, and select the best ones by reading the summaries of the clusters...

...forward the Japanese email to the summarizer, select '1 par', and skim the translated summary.

Headline news — informing



TIME Daily
> News Wire
> Editor's Letter
> Comments
> News Features
> Text Only

Magazine
Community
Special Reports

LIFE Picture of the Day

Get TIME Daily delivered to your desktop every day with

June 30, 1998

U.S. Plane Fires a Missile On Iraq
An Iraqi radar station targets an Allied plane, and a U.S. F-16 responds quickly -- with deadly force. Is another showdown with Saddam on the way?
Full Story

Responding with Force: A U.S. Air Force F-16 flies over Kuwait. U.S. AIR FORCE/AP

Starr Plays the Tripp Card
The former confidante's grand jury appearance puts the squeeze on Ms. Lewinsky.

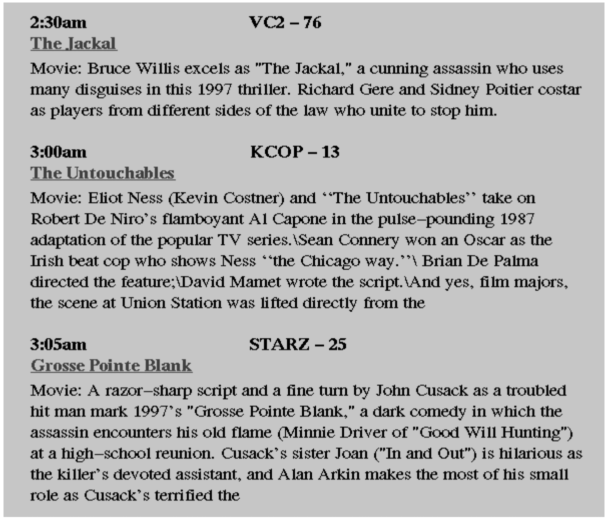
Down to Business in Shanghai
President Clinton spends some time in the city he wants the rest of China to turn into.

·Poll: Does the U.S. have the right to impose its idea of human rights on China?

·Postcards From the Middle Kingdom: TIME's Jay Branegan says President Clinton is in full campaign mode in China. But the big question is, why isn't he pressing the flesh?

Boris Duels With the Duma
If Russian president Yeltsin wants to make other Russian pols look bad, he should stop making a fool of himself first.

TV-GUIDES — decision making

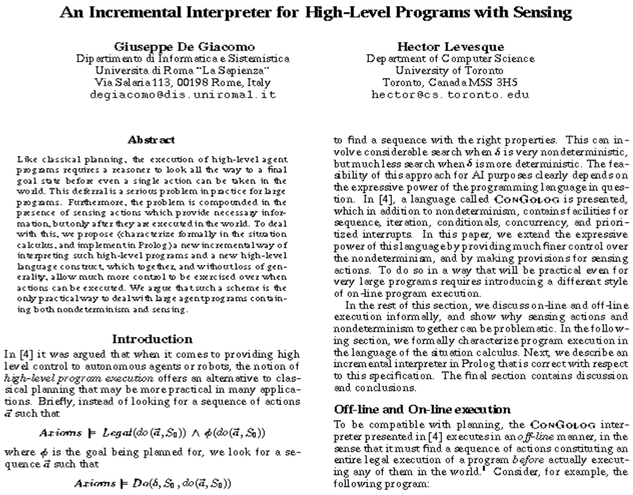


2:30am **VC2 – 76**
The Jackal
Movie: Bruce Willis excels as "The Jackal," a cunning assassin who uses many disguises in this 1997 thriller. Richard Gere and Sidney Poitier costar as players from different sides of the law who unite to stop him.

3:00am **KCOP – 13**
The Untouchables
Movie: Eliot Ness (Kevin Costner) and ''The Untouchables'' take on Robert De Niro's flamboyant Al Capone in the pulse–pounding 1987 adaptation of the popular TV series.\Sean Connery won an Oscar as the Irish beat cop who shows Ness ''the Chicago way.''\ Brian De Palma directed the feature;\David Mamet wrote the script.\And yes, film majors, the scene at Union Station was lifted directly from the

3:05am **STARZ – 25**
Grosse Pointe Blank
Movie: A razor–sharp script and a fine turn by John Cusack as a troubled hit man mark 1997's "Grosse Pointe Blank," a dark comedy in which the assassin encounters his old flame (Minnie Driver of "Good Will Hunting") at a high–school reunion. Cusack's sister Joan ("In and Out") is hilarious as the killer's devoted assistant, and Alan Arkin makes the most of his small role as Cusack's terrified the

Abstracts of papers — time saving



An Incremental Interpreter for High-Level Programs with Sensing

Giuseppe De Giacomo
Dipartimento di Informatica e Sistemistica
Universita di Roma "La Sapienza"
Via Salaria 113, 00198 Rome, Italy
degiacomo@dis.uniroma1.it

Hector Levesque
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3H5
hector@cs.toronto.edu

Abstract

Like classical planning, the execution of high-level agent programs requires a reasoner to look all the way to a final goal state before even a single action can be taken in the world. This deferral is a serious problem in practice for large programs. Furthermore, the problem is compounded in the presence of sensing actions which provide necessary information, but only after they are executed in the world. To deal with this, we propose (characterize formally in the situation calculus, and implement in Prolog) a new incremental way of interpreting such high-level programs and a new high-level language construct, which together, and without loss of generality, allow much more control to be exercised over when actions can be executed. We argue that such a scheme is the only practical way to deal with large agent programs containing both nondeterminism and sensing.

Introduction

In [4] it was argued that when it comes to providing high level control to autonomous agents or robots, the notion of *high-level program execution* offers an alternative to classical planning that may be more practical in many applications. Briefly, instead of looking for a sequence of actions $\vec{a}$ such that

$$Axioms \models Legal(do(\vec{a}, S_0)) \wedge \phi(do(\vec{a}, S_0))$$

where ϕ is the goal being planned for, we look for a sequence $\vec{a}$ such that

$$Axioms \models Do(\delta, S_0, do(\vec{a}, S_0))$$

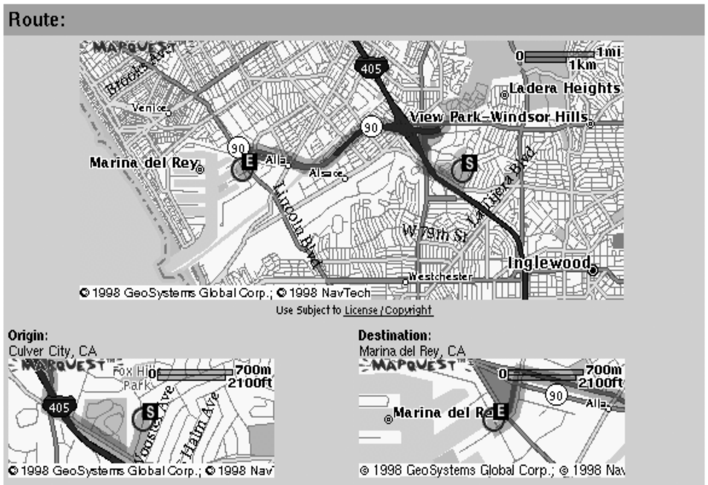
to find a sequence with the right properties. This can involve considerable search when δ is very nondeterministic, but much less search when δ is more deterministic. The feasibility of this approach for AI purposes clearly depends on the expressive power of the programming language in question. In [4], a language called CONGOLOG is presented, which in addition to nondeterminism, contains facilities for sequence, iteration, conditionals, concurrency, and prioritized interrupts. In this paper, we extend the expressive power of this language by providing much finer control over the nondeterminism, and by making provisions for sensing actions. To do so in a way that will be practical *even* for very large programs requires introducing a different style of on-line program execution.

In the rest of this section, we discuss on-line and off-line execution informally, and show why sensing actions and nondeterminism together can be problematic. In the following section, we formally characterize program execution in the language of the situation calculus. Next, we describe an incremental interpreter in Prolog that is correct with respect to this specification. The final section contains discussion and conclusions.

Off-line and On-line execution

To be compatible with planning, the CONGOLOG interpreter presented in [4] executes in an *off-line* manner, in the sense that it must find a sequence of actions constituting an entire legal execution of a program *before* actually executing any of them in the world.[1] Consider, for example, the following program:

Graphical maps — orienting



Route:

© 1998 GeoSystems Global Corp.; © 1998 NavTech
Use Subject to License / Copyright

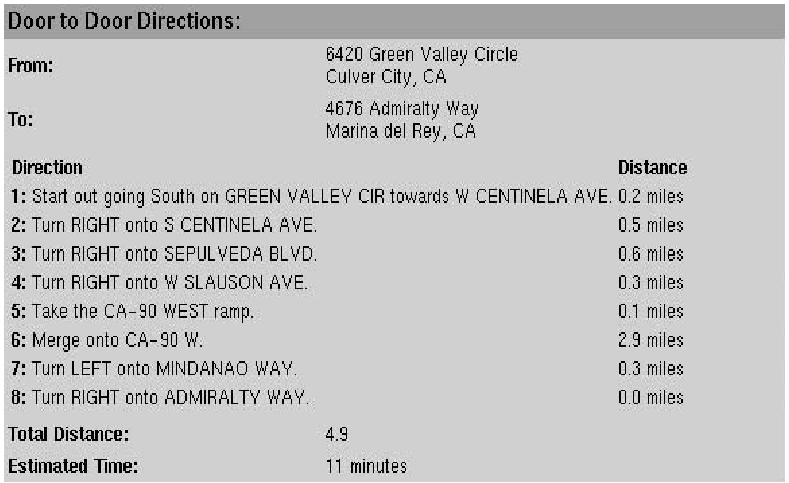
Origin:
Culver City, CA

© 1998 GeoSystems Global Corp.; © 1998 Nav

Destination:
Marina del Rey, CA

© 1998 GeoSystems Global Corp.; © 1998 Nav

Textual Directions — planning



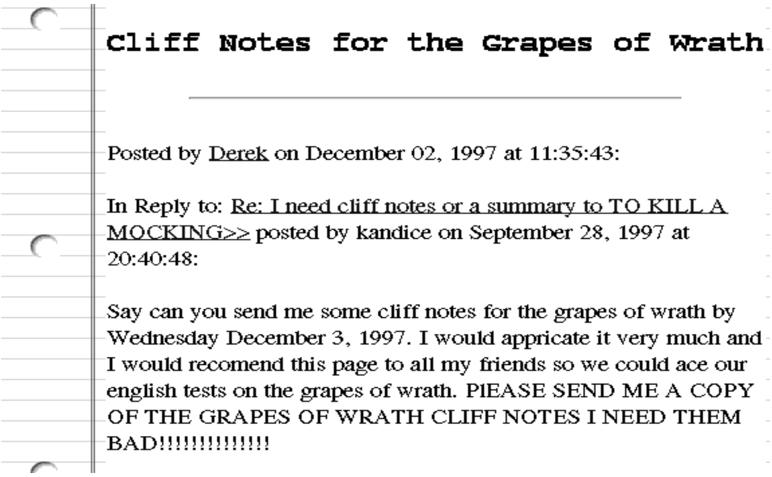
Door to Door Directions:

From:	6420 Green Valley Circle Culver City, CA
To:	4676 Admiralty Way Marina del Rey, CA

Direction	Distance
1: Start out going South on GREEN VALLEY CIR towards W CENTINELA AVE.	0.2 miles
2: Turn RIGHT onto S CENTINELA AVE.	0.5 miles
3: Turn RIGHT onto SEPULVEDA BLVD.	0.6 miles
4: Turn RIGHT onto W SLAUSON AVE.	0.3 miles
5: Take the CA-90 WEST ramp.	0.1 miles
6: Merge onto CA-90 W.	2.9 miles
7: Turn LEFT onto MINDANAO WAY.	0.3 miles
8: Turn RIGHT onto ADMIRALTY WAY.	0.0 miles

Total Distance:	4.9
Estimated Time:	11 minutes

Cliff notes — Laziness support



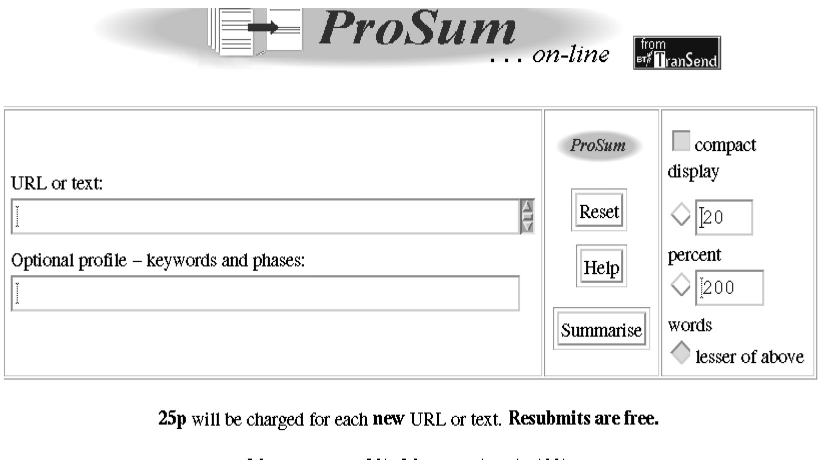
Cliff Notes for the Grapes of Wrath

Posted by Derek on December 02, 1997 at 11:35:43:

In Reply to: Re: I need cliff notes or a summary to TO KILL A MOCKING>> posted by kandice on September 28, 1997 at 20:40:48:

Say can you send me some cliff notes for the grapes of wrath by Wednesday December 3, 1997. I would appricate it very much and I would recomend this page to all my friends so we could ace our english tests on the grapes of wrath. PIEASE SEND ME A COPY OF THE GRAPES OF WRATH CLIFF NOTES I NEED THEM BAD!!!!!!!!!!!!!!

Real systems — Money making



URL or text:

Optional profile – keywords and phases:

□ compact display

◇ 20 percent

◇ 200 words

◆ lesser of above

25p will be charged for each **new** URL or text. **Resubmits are free.**

Questions

- **What kind** of natural language processing do people need?
  - What are *summarization, machine translation, question answering*...?
- **How sophisticated** must NLP systems be?
  - Are statistical techniques sufficient?
  - Or do we need symbolic techniques and deep understanding as well?
- **What milestones** would mark quantum leaps in Natural Language Processing theory and practice?
  - What do we need to build an NLP system?
  - What NLP systems are "hot" today?
  - Is it any NLP at google.com or askJeeves.com?

Today

- Introductions
- Course Overview
- Administrivia
- What is NLP
- An Annotated Corpus Example
- Why TreeBanks are important?
- Fun Examples

Syllabus

First 7 lectures: SYNTAX
Next 6 lectures: SEMANTICS
Next 7 lectures: Information Extraction, Question Answering,
Summarization
Last 3 lectures: Discourse and Dialogue Processing

No Mid-Term –No Final!!!
3 Homework assignments + Project

GRADE ???
Class participation = 15%
Homeworks = 60% (each 20%)
Project = 25%

Goals of the Course

- Learn about:
  - Design, development and use of some fun NLP systems e.g. "Virtual Friends", "Talking Appliances"
  - Practical and theoretical foundations of basis NLP infrastructure
  - What makes NLP difficult? Practical solutions
  - Hands-on experience with NLP systems

13

Readings

- Textbook
  - *Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Daniel Juraksy* and James Martin, Prentice Hall, 2000.
- Code
  - *Brill's Tagger*
  - *WordNet*
- Extra
  - *Natural Language Processing,* James Allan, Benjamin/Cummings Publishing Co, 1994
  - Papers from recent NLP conferences

14

Why is NLP important ?

- plays an increasing role in curbing the information explosion on Internet.

- used for building NL interfaces to databases,machine translation,, and others.

NLP is a difficult, and largely unsolved problem. One reason for this is its multidisciplinary nature:

- **Linguistics :** How words, phrases, and sentences are formed.

- **Psycholinguistics :** How people understand and communicate using human language.

Multidisciplinary ???

- **Computational linguistics :** Deals with models and computational aspects of N.L. (e.g algorithms)

- **Philosophy:** Relates to the semantics of language ; notation of meaning, how words identify objects, NLP require considerable knowledge about the world.

- Computer Science model formation and implementation using modern methods.

- Artificial intelligence ; issues related to knowledge representation and reasoning.

- NL Engineering : implementation of large, realistic systems.

Applications of NLP

Text- based applications:

- finding documents on certain topics (document classification)
- information retrieval; search for keywords or concepts.
- information extraction; relevant to a topic.
- text comprehension
- translation from a language to another
- summarization
- knowledge management

- Dialogue-based applications:

- human-machine communication
- question-answering
- tutoring systems
- problem solving

- Speech - processing(not considered in this class)

Basic levels of language processing

- **Phonetic** - how words are related to the sounds that realize them. Essential for speech processing.
- **Morphological Knowledge** - how words are constructed : e.g friend, friendly, unfriendly, friendliness.
- **Syntactic Knowledge** - how words can be put together to form correct sentences, and the role of each play's in the sentence. e.g John ate the cake.
- **Semantic Knowledge** - Words and sentence meaning:

  They saw a log

  They saw a log yesterday.

  He saws a log.

Things get more complicated!!!

- **Pragmatic Knowledge** - how sentences are used in different situations (contexts)

  Mary grabbed her umbrella.

  A) It is a cloudy day.

  B) She was afraid of dogs.

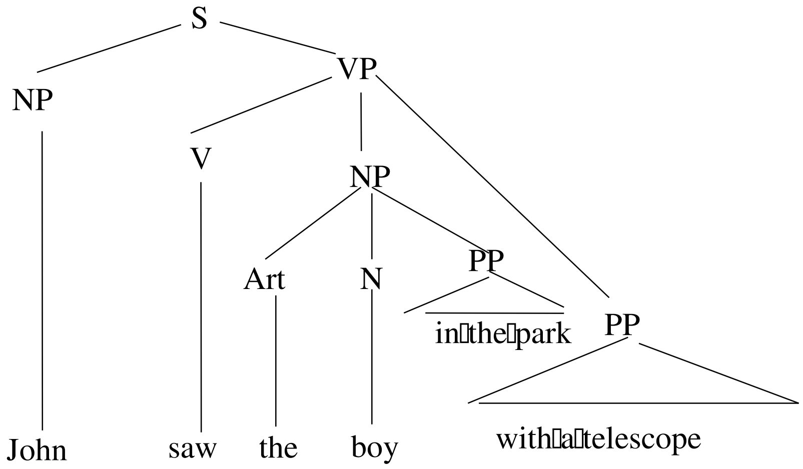
- **Discourse Knowledge** - how the meaning of words and sentences is effected by the previous sentences, e.g. pronoun resolution.

John gave his bike to Bill. He didn't care much of it anyway.

Classical Example

`John saw the boy in the park with a telescope.`

Syntactic Parse



Language Ambiguities

Lexical ambiguity- when a word has more than one part of speech.:

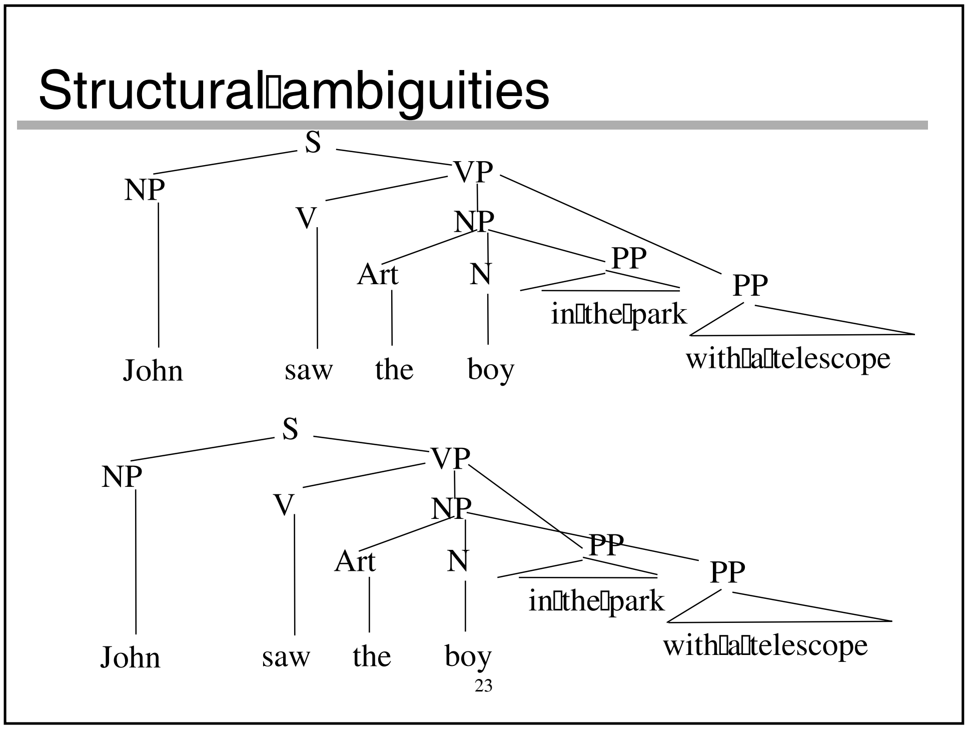
Rice flies like sand.

Note that these synthetic ambiguities lead to different parse structures.

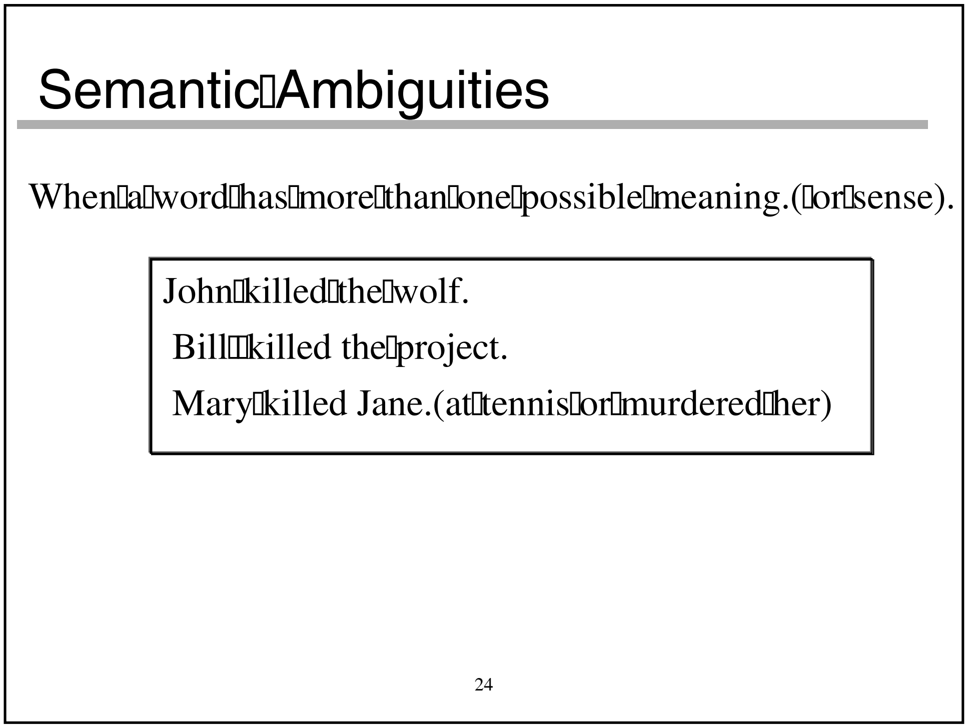
Sometimes it is possible to use grammar rules (like subject verb arguments) to disambiguate:

Flying planes are dangerous.

Flying planes is dangerous.



Structural ambiguities



Semantic Ambiguities

When a word has more than one possible meaning.(or sense).

John killed the wolf.

Bill killed the project.

Mary killed Jane.(at tennis or murdered her)

A (poor) organization of NLP System

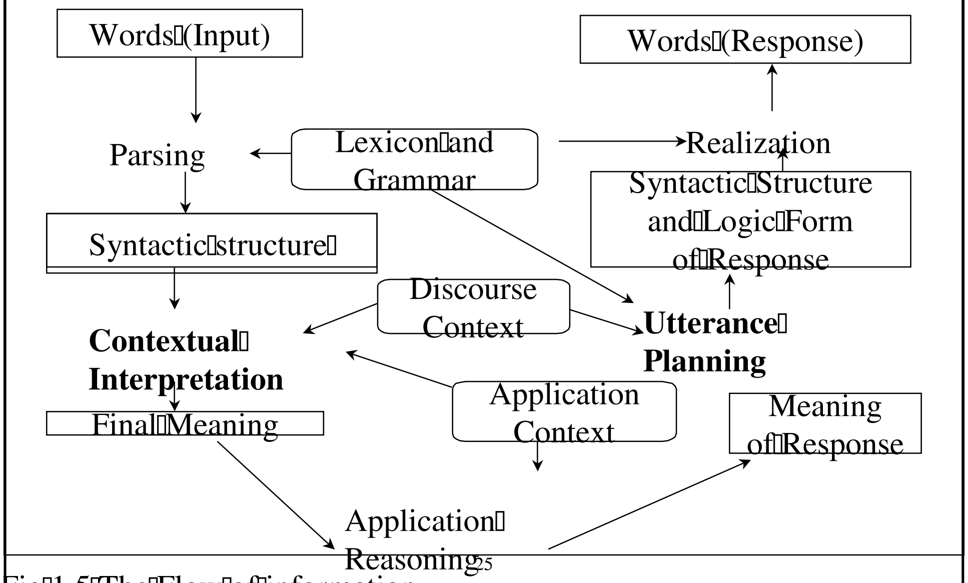


Fig 1.5 The Flow of information

State of the art in NLP Research

- Association of Computational Linguistics (ACL)

-AAAI -every year /IJCAI -every second year

-Natural Language Engineering (journal).

-Information retrieval/ Extraction

MUC - Message Understanding Conf.

DUC – Document Understanding Conf.

SIGIR – Special Interest Group in IR

Resources

Machine Readable Dictionaries (MRD)

WordNet (www.cogsci.princeton.edu/~wn)

Large corpora:

Penn Treebank –

www.cis.upenn.edu/~treebank

All Treebank data is released through The Language Data Consortium (LDC)

Building the Penn Treebank

- Corpus of 4.5 million words of American English
- Part_of_Speech tagged
- Syntactic Bracketing

PoS Tagging

1.	CC	Coordinating conjunction	13.	NNS	Noun, plural
2.	CD	Cardinal number	14.	NNP	Proper noun, singular
3.	DT	Determiner	15.	NNPS	Proper noun, plural
4.	EX	Existential *there*	16.	PDT	Predeterminer
5.	FW	Foreign word	17.	POS	Possesive ending
6.	IN	Preposition/subord. conjunction	18.	PRP	Personal pronoun
7.	JJ	Adjective	19.	PP$	Possesive pronoun
8.	JJR	Adjective, comparative	20.	RB	Adverb
9.	JJS	Adjective, superlative	21.	RBR	Adverb, comparative
10.	LS	List item marker	22.	RBS	Adverb, superlative
11.	MD	Modal	23.	RP	Particle
12.	NN	Noun, singular or mass	24.	SYM	Symbol (mathematical or scientific)

PoS Tagging

25.	TO	*to*	37.	#	Pound sign
26.	UH	Interjection	38.	$	Dollar sign
27.	VB	Verb, base form	39.	.	Sentence-final punctuation
28.	VBD	Verb, past tense	40.	,	Comma
29.	VBG	Verb,gerund/present participle	41.	:	Colon, semi-colon
30.	VBN	Verb, past participle	42.	(	Left bracket character
31.	VBP	Verb, non-3rd ps.sing. present	43.	)	Right bracket character
32.	VBZ	Verb, 3rd ps.sing. present	44.	"	Straight double quote
33.	WDT	*wh*-determiner	45.	‘	Left open single quote
34.	WP	*wh*-pronoun	46.	“	Left open double quote
35.	WP$	Possesive *wh*-pronoun	47.	’	Right close single quote
36.	WRB	*wh*-adverb	48.	”	Left close double quote

The Syntactic Tag Set

	Tags	
1.	ADJP	Adjective phrase
2.	ADVP	Adverb phrase
3.	NP	Noun phrase
4.	PP	Prepositional phrase
5.	S	Simple declarative clause
6.	SBAR	Clause introduced by subordinating conjunction or 0 (see below)
7.	SBARQ	Direct question introduced by *wh*-word or *wh*-phrase
8.	SINV	Declarative sentence with subject-aux inversion
9.	SQ	Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase
10.	VP	Verb phrase
11.	WHADVP	*wh*-adverb phrase
12.	WHNP	*wh* -noun phrase
13.	WHPP	*wh* -prepositional phrase
14.	X	Constituent of unknown or uncertain category
	Null elements	
1.	*	"Understood" subject of infinitive or imperative
2.	0	Zero variant of *that* in subordinate clauses
3.	T	Trace-marks position where moved *wh*-constituent is interpreted
4.	NIL	Marks position where preposition is interpreted in pied-piping contexts

Sample Bracketed Text

```
((S
  (NP Battle-tested industrial managers
      here)
   always
  (VP buck
      up
      (NP nervous newcomers
      (PP with
          (NP the tale
              (PP of
                  (NP (NP the
                          (ADJP first
                                (PP of
                                    (NP their countrymen)))
                          (S (NP *)
                             to
                             (VP visit
                                 (NP Mexico)))
                      ,
                      (NP (NP a boatload
                              (PP of
                                  (NP (NP warriors)
                                      (VP-1 blown
                                          ashore
                                          (ADVP (NP 375 years)
                                                 ago)))))
                          (VP-1 *pseudo-attach*))))))))
.)
```

The Penn Treebank

Description	Tagged for Part-of-Speech	Skeletal Parsing
	(Tokens)	(Tokens)
Dept. of Energy abstracts	231,404	231,404
Dow Jones Newswire stories	3,065,776	1,061,166
Dept. of Agriculture bulletins	78,555	78,555
Library of America texts	105,652	105,652
MUC-3 messages	111,828	111,828
IBM Manual sentences	89,121	89,121
WBUR radio transcripts	11,589	11,589
ATIS sentences	19,832	19,832
Brown Corpus, retagged	1,172,041	1,172,041
Total:	**4,885,798**	**2,881,188**