Chapter5 contd...

Introduction to NLP

Basanta Joshi, PhD
basanta@ioe.edu.np
Lecture notes can be downloaded from
www.basantajoshi.com.np

Communication ***Typical communication episode***

S (speaker) wants to convey P (proposition) to H (hearer) using W (words in a formal or natural language)

1. Speaker

- **Intention:** S wants H to believe P
- **Generation:** S chooses words W
- **Synthesis:** S utters words W

2. Hearer

- **Perception:** H perceives words W" (ideally W" = W)
- **Analysis:** H infers possible meanings P1,P2,…,Pn for W"
- **Disambiguation:** H infers that S intended to convey Pi (ideally Pi=P)
- **Incorporation:** H decides to believe or disbelieve Pi

Natural Language - General

Natural Language is characterized by

- a common or shared set of signs alphabet; lexicon
- a systematic procedure to produce combinations of signs

syntax

- a shared meaning of signs and combinations of signs

(constructive) semantics

Natural Language Processing (NLP)

1. Natural Language Understanding
   - Taking some spoken/typed sentence and working out what it means
2. Natural Language Generation
   - Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Natural Language Processing Overview

- Speech Recognition
- Natural Language Processing
  - Syntax
  - Semantics
  - Pragmatics
- Spoken Language

Background

- The HAL 9000 computer in Stanley Kubrick's film *2001: A Space Odyssey*
  - HAL is an artificial agent capable of such advanced language processing behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips.
- The language-related parts of HAL
  - Speech recognition
  - Natural language understanding (and, of course, lip-reading),
  - Natural language generation
  - Speech synthesis
  - Information retrieval
  - information extraction and
  - inference

Background

- Solving the language-related problems and others like them, is the main concern of the fields known as Natural Language Processing, Computational Linguistics, and Speech Recognition and Synthesis, which together we call **Speech and Language Processing(SLP)**.
- Applications of language processing
  - **spelling correction**,
  - **grammar checking**,
  - **information retrieval**, and
  - **machine translation**.

Natural Language and Speech

- Speech Recognition
  - acoustic signal as input
  - conversion into phonemes and written words
- Natural Language Processing
  - written text as input; sentences (or 'utterances')
  - syntactic analysis: parsing; grammar
  - semantic analysis: "meaning", semantic representation
  - pragmatics: dialogue; discourse; metaphors
- Spoken Language Processing
  - transcribed utterances
  - Phenomena of spontaneous speech

Natural language understanding

Raw speech signal

↓ • **Speech recognition**

Sequence of words spoken

↓ • **Syntactic analysis** using knowledge of the grammar

Structure of the sentence

↓ • **Semantic analysis** using info. about meaning of words

Partial representation of meaning of sentence

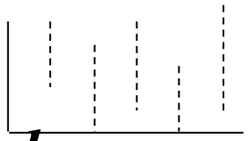
↓ • **Pragmatic analysis** using info. about context

Final representation of meaning of sentence

Natural Language Understanding

- Input/Output data — Processing stage — Other data used

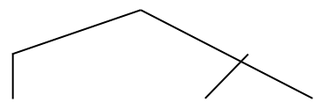
Frequency spectrogram



Word sequence

"He loves Mary"

Sentence structure



Partial Meaning

Ξx loves(x,mary)

Sentence meaning

loves(john,mary)

speech recognition. ← sounds, freq. of diff.

syntactic analysis. ← language, grammar of

semantic analysis ← words, meanings of

pragmatics. ← utterance, context of

Some Brief History

- Speech and language processing encompasses a number of different but overlapping fields in these different departments:
  - **computational linguistics** in linguistics,
  - **natural language processing** in computer science,
  - **speech recognition** in electrical engineering,
  - **computational psycholinguistics** in psychology.

Some Brief History

Foundational Insights: 1940s and 1950s

- Two foundational paradigms:
  - the **automaton** and
  - **probabilistic** or **information-theoretic models**
- Turing's work led first to the **McCulloch-Pitts neuron** (McCulloch and Pitts, 1943),
  - a simplified model of the neuron as a kind of computing element that could be described in terms of propositional logic,
- And then to the work of Kleene (1951) and (1956) on
  - finite automata and regular expressions.
- Shannon (1948) applied probabilistic models of discrete Markov processes to automata for language. (*continued*)

Some Brief History

Foundational Insights: 1940s and 1950s

- Chomsky (1956), drawing the idea of a finite state Markov process from Shannon's work, first considered finite-state machines as a way to characterize a grammar, and defined a finite-state language as a language generated by a finite-state grammar.
- These early models led to the field of **formal language theory**, which used algebra and set theory to define formal languages as sequences of symbols.
  - This includes the context-free grammar, first defined by Chomsky (1956) for natural languages but independently discovered by Backus (1959) and Naur et al. (1960) in their descriptions of the ALGOL programming language.

Some Brief History

Foundational Insights: 1940s and 1950s

- The second foundational insight of this period was the development of probabilistic algorithms for speech and language processing, which dates to Shannon's other contribution:
  - the metaphor of the **noisy channel** and **decoding** for the transmission of language through media like communication channels and speech acoustics.
  - Shannon also borrowed the concept of **entropy** from thermodynamics as a way of measuring the information capacity of a channel, or the information content of a language, and performed the first measure of the entropy of English using probabilistic techniques.
  - It was also during this early period that the sound spectrograph was developed (Koenig et al., 1946), and foundational research was done in instrumental phonetics that laid the groundwork for later work in speech recognition.
    - This led to the first machine speech recognizers in the early 1950s.

Some Brief History

The Two Camps: 1957–1970

- By the end of the 1950s and the early 1960s, SLP had split very cleanly into two paradigms: *symbolic* and *stochastic*.
- The symbolic paradigm took off from two lines of research.
  - The **first** was the work of Chomsky and others on formal language theory and generative syntax throughout the late 1950s and early to mid 1960s, and the work of many linguistics and computer scientists on parsing algorithms, initially top-down and bottom-up and then via dynamic programming.
  - One of the earliest complete parsing systems was Zelig Harris's Transformations and Discourse Analysis Project (TDAP), which was implemented between June 1958 and July 1959 at the University of Pennsylvania (Harris, 1962). (*continued*)

Some Brief History

The Two Camps: 1957–1970

– The second line of research was the new field of artificial intelligence.

- In the summer of 1956 John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester brought together a group of researchers for a two-month workshop on what they decided to call artificial intelligence (AI).
- Although AI always included a minority of researchers focusing on stochastic and statistical algorithms (include probabilistic models and neural nets), the **major focus of the new field was the work on reasoning and logic** typified by Newell and Simon's work on the Logic Theorist and the General Problem Solver.
- At this point **early natural language understanding systems were built**.
  - These were simple systems that worked in single domains mainly by a combination of pattern matching and keyword search with simple heuristics for reasoning and question-answering.
  - By the late 1960s more formal logical systems were developed.

Some Brief History

The Two Camps: 1957–1970

- The stochastic paradigm took hold mainly in departments of statistics and of electrical engineering.
  - By the late 1950s the Bayesian method was beginning to be applied to the **problem of optical character recognition**.
  - Bledsoe and Browning (1959) built a Bayesian system for text-recognition that used a large dictionary and computed the likelihood of each observed letter sequence given each word in the dictionary by multiplying the likelihoods for each letter.
  - Mosteller and Wallace (1964) applied Bayesian methods to the problem of authorship attribution on *The Federalist* papers.
  - The 1960s also saw the rise of the first serious testable psychological models of human language processing based on transformational grammar, as well as the first on-line corpora: the Brown corpus of American English, a 1 million word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.), which was assembled at Brown University in 1963–64 (Kučera and Francis, 1967; Francis, 1979; Francis and Kučera, 1982), andWilliam S. Y.Wang's 1967 DOC (Dictionary on Computer), an on-line Chinese dialect dictionary.

Some Brief History

Four Paradigms: 1970–1983

- The next period saw an explosion in research in SLP and the development of a number of **research paradigms** that still dominate the field.
- The **stochastic** paradigm played a huge role in the development of *speech recognition* algorithms in this period,
  - particularly the use of the *Hidden Markov Model* and the metaphors of the noisy channel and decoding, developed independently by Jelinek, Bahl, Mercer, and colleagues at IBM's Thomas J. Watson Research Center, and by Baker at Carnegie Mellon University, who was influenced by the work of Baum and colleagues at the Institute for Defense Analyses in Princeton.
  - AT&T's Bell Laboratories was also a center for work on speech recognition and synthesis; see Rabiner and Juang (1993) for descriptions of the wide range of this work.

Some Brief History
Four Paradigms: 1970–1983

- The **logic-based** paradigm was begun by the work of Colmerauer and his colleagues on Q-systems and metamorphosis grammars (Colmerauer, 1970, 1975),
  - the forerunners of Prolog, and Definite Clause Grammars (Pereira andWarren, 1980).
  - Independently, Kay's (1979) work on functional grammar, and shortly later, Bresnan and Kaplan's (1982) work on LFG, established the importance of *feature structure unification*.

Some Brief History
Four Paradigms: 1970–1983

- The **natural language understanding** field took off during this period,
  - beginning with Terry Winograd's SHRDLU system, which simulated a robot embedded in a world of toy blocks (Winograd, 1972a).
    - The program was able to accept natural language text commands *(Move the red block on top of the smaller green one)* of a hitherto unseen complexity and sophistication.
    - His system was also the first to attempt to build an extensive (for the time) grammar of English, based on Halliday's systemic grammar.
  - Winograd's model made it clear that the problem of parsing was well-enough understood to begin to focus on semantics and discourse models.
  - Roger Schank and his colleagues and students (in what was often referred to as the *Yale School*) built a series of language understanding programs that focused on human conceptual knowledge such as scripts, plans and goals, and human memory organization (Schank and Albelson, 1977; Schank and Riesbeck, 1981; Cullingford, 1981; Wilensky, 1983; Lehnert, 1977).
  - This work often used network-based semantics (Quillian, 1968; Norman and Rumelhart, 1975; Schank, 1972; Wilks, 1975c, 1975b; Kintsch, 1974) and began to incorporate Fillmore's notion of *case roles* (Fillmore, 1968) into their representations (Simmons, 1973).
- The logic-based and natural-language understanding paradigms were unified on systems that used predicate logic as a semantic representation, such as the LUNAR question-answering system (Woods, 1967, 1973).

Some Brief History
Four Paradigms: 1970–1983

- The **discourse modeling** paradigm focused on four key areas in discourse.
  - Grosz and her colleagues introduced the study of **substructure in discourse**, and of **discourse focus** (Grosz, 1977a; Sidner, 1983),
  - a number of researchers began to work on **automatic reference resolution** (Hobbs, 1978),
  - and the **BDI** (Belief-Desire-Intention) framework for logic-based work on speech acts was developed (Perrault and Allen, 1980; Cohen and Perrault, 1979).

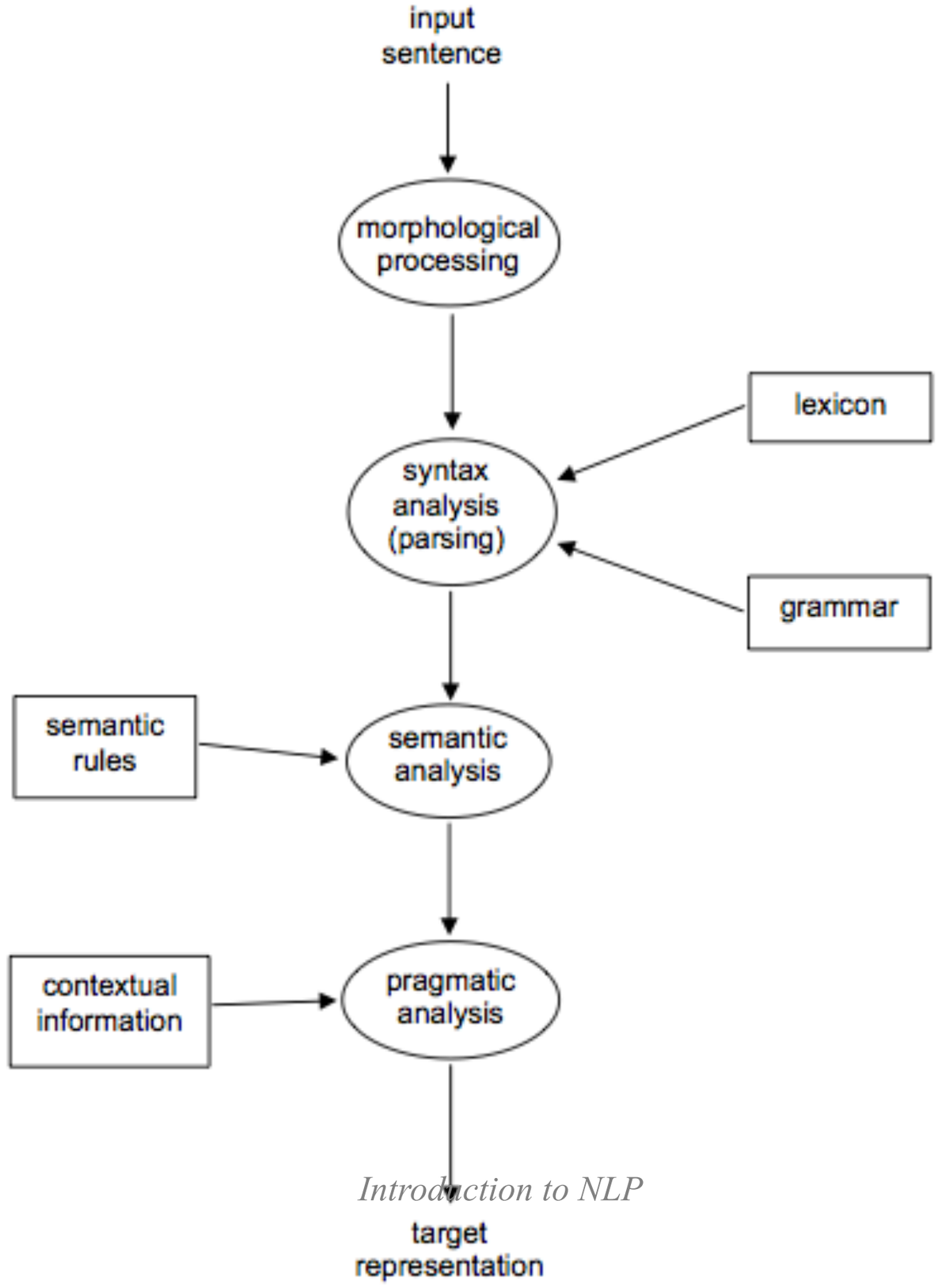
Some Brief History
Empiricism and Finite State Models Redux: 1983–1993

- This next decade saw the return of two classes of models which had lost popularity in the late 1950s and early 1960s, partially due to theoretical arguments against them such as Chomsky's influential review of Skinner's *Verbal Behavior* (Chomsky, 1959b).
  - The first class was finite-state models, which began to receive attention again after work on finite-state phonology and morphology by Kaplan and Kay (1981) and finite-state models of syntax by Church (1980).
  - The second trend in this period was what has been called the "return of empiricism"; most notably here was the rise of probabilistic models throughout speech and language processing, influenced strongly by the work at the IBM Thomas J. Watson Research Center on probabilistic models of speech recognition.
    - These probabilistic methods and other such data-driven approaches spread into part-of-speech tagging, parsing and attachment ambiguities, and connectionist approaches from speech recognition to semantics.
- This period also saw considerable work on natural language generation.

Some Brief History
The Field Comes Together: 1994–1999

- By the last five years of the millennium it was clear that the field was vastly changing.
  - First, probabilistic and data-driven models had become quite standard throughout natural language processing.
    - Algorithms for parsing, part-of-speech tagging, reference resolution, and discourse processing all began to incorporate probabilities, and employ evaluation methodologies borrowed from speech recognition and information retrieval.
  - Second, the increases in the speed and memory of computers had allowed commercial exploitation of a number of subareas of speech and language processing, in particular
    - speech recognition and spelling and grammar checking.
    - Speech and language processing algorithms began to be applied to Augmentative and Alternative Communication (AAC).
  - Finally, the rise of the Web emphasized the need for language-based information retrieval and information extraction.

Levels of natural language processing.



Problems in NLP

- Two problems in particular make the processing of natural languages difficult and cause different techniques to be used than those associated with the construction of compilers etc for processing artificial languages.
- These problems are
  - (i) the level of ambiguity that exists in natural languages and
  - (ii) the complexity of semantic information contained in even simple sentences.

Typically language processors deal with large numbers of words, many of which have alternative uses, and large grammars which allow different phrase types to be formed from the same string of words. Language processors are made more complex because of the irregularity of language and the different kinds of ambiguity which can occur. The groups of sentences next slides are used as examples to illustrate different issues faced by language processors.

Problems in NLP

- “The old man the boats. “
- In the sentence "The old man the boats" problems, such as they are, exist because the word "old" can be legitimately used as a noun (meaning a collection of old people) as well as an adjective, and the word "man" can be used as a verb (meaning take charge of) as well as a noun. This causes ambiguity which must be resolved during syntax analysis. This is done by considering all possible syntactic arrangements for phrases and sub-phrases when necessary.
- The implication here is that any parsing mechanism must be able to explore various syntactic arrangements for phrases and be able to backtrack and rearrange them whenever necessary.

Problems in NLP

- “Cats play with string.
- * Cat play with string.“

- The problem with the second sentence in this group is that there is no numeric agreement between the subject and the verb (one is a singular form, the other plural). Grammars must be expressive enough to specify checks for such anomalies and also specify actions which should take place if they occur. Mechanisms to signal failure in processing such cases are useful. For example, when combining semantics for "colourless" and "green" in a phrase like "the colourless green car" a signal of failure marks a sub-phrase as ill-formed and prevents it being considered any further. In the case of problems with the numeric agreement between subject and verb it may be more appropriate to signal a warning. A warning marks a sub-phrase as potentially-flawed but does not reject it outright.

Problems in NLP

- “I saw the racing pigeons flying to Paris.
- I saw the Eiffel Tower flying to Paris.“

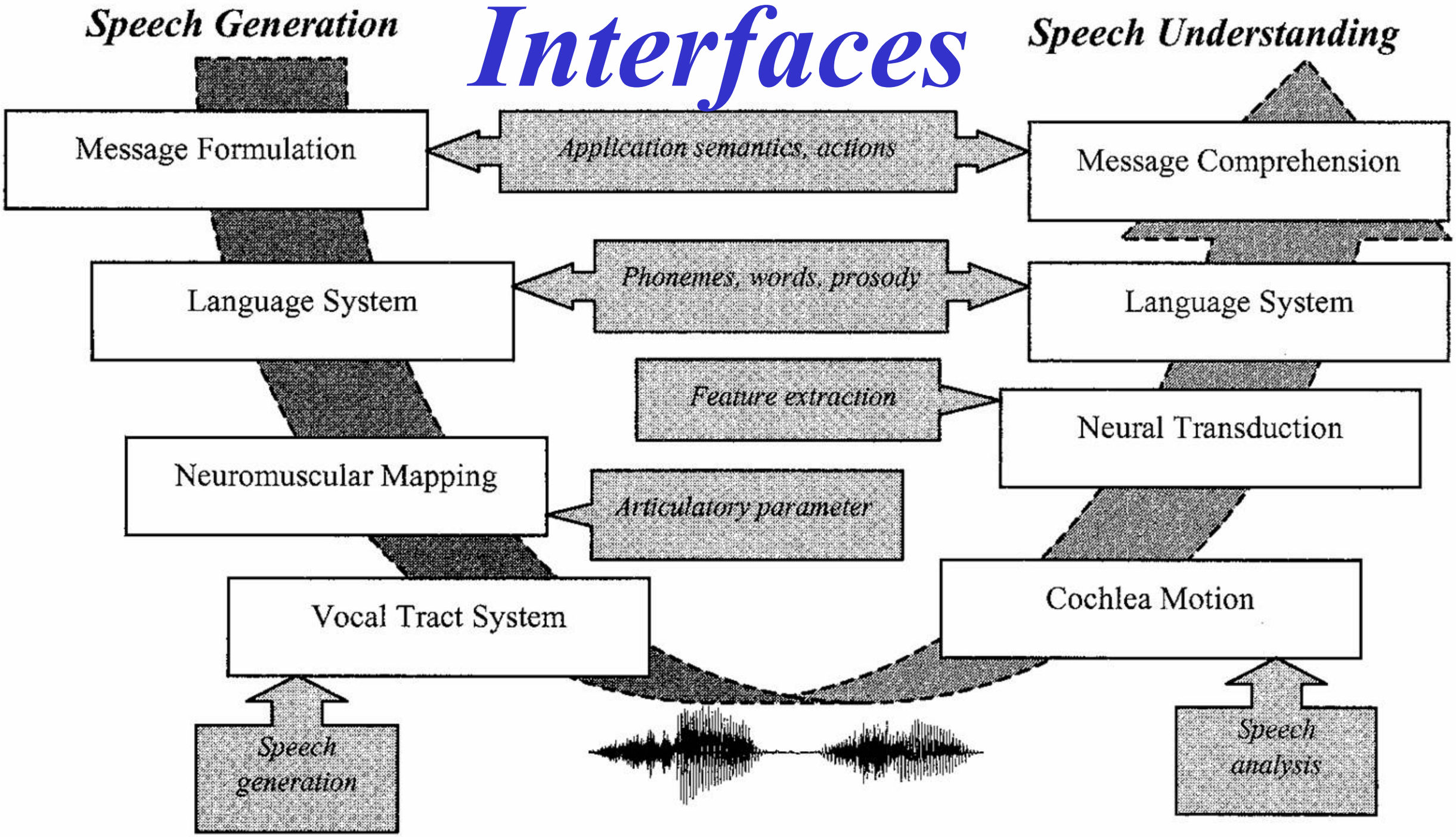
- Assuming these sentences are taken in isolation so there is no previous dialog which introduces a racing pigeon named "the Eiffel Tower"...
- The sensible way to interpret the meaning of the second sentence is "While I was flying to Paris I saw the Eiffel Tower in its usual position - firmly rooted to the ground". What prevents the second sentence being restructured in the same way as the first is the inconsistency with objects like the Eiffel Tower and activities like flight. Sensible semantic rules must detect this inconsistency and perhaps halt progress on any sub- phrase which implies the Eiffel Tower is involved in a flying activity.
- A semantic rule in this case might reason as follows:
  - (i) the set of objects capable of flight is {humans, birds, bats, aeroplanes}
  - (ii) the Eiffel Tower is defined as a structural-object
  - (iii) structural-objects are not in the set of objects capable of flight so signal "ill-formed semantics" and halt progress on this rule.

Problems in NLP

- “The boy kicked the ball under the tree.
  The boy kicked the wall under the tree.“

- The implication in the first sentence is that the activity performed by the boy causes the ball to move to a position which is under the tree. The kicking activity has a meaning of "move, using the boy's foot as an instrument to cause that movement". In the second sentence the wall is assumed not to have changed position. The activity which took place was one of "strike, using the foot as an instrument". The apparent disambiguation in this case can take place if it is known that balls are mobile objects (and are often moved using a foot as an instrument) and walls are static objects.

Spoken Language Interfaces



Speech Processing Systems

Types and Characteristics

- Speech Recognition vs. Speaker Recognition (Voice Recognition; Speaker Identification)
- speaker-dependent vs. speaker-independent
- training?
- unlimited vs. large vs. small vocabulary
- single word vs. continuous speech

Speech Recognition Phases

- acoustic signal as input
- signal analysis - spectrogram
- feature extraction
- phoneme recognition
- word recognition
- conversion into written words

Speech Signal Analysis

Analog-Digital Conversion of Acoustic Signal

Sampling in Time Frames ("windows")

- frequency = 0-crossings per time frame

→ e.g. 2 crossings/second is 1 Hz (1 wave)

→ e.g. 10kHz needs sampling rate 20kHz

- measure amplitudes of signal in time frame

→ digitized wave form

- separate different frequency components

→ FFT (Fast Fourier Transform)

→ spectrogram

- other frequency based representations

  → LPC (linear predictive coding),

  → Cepstrum