Communication, Perceiving and Acting

Natural Language Processing

- Natural Language Processing (NLP) is the process of computer analysis of input provided in a human language (natural language), and conversion of this input into a useful form of representation.
- The field of NLP is primarily concerned with getting computers to perform useful and interesting tasks with human languages. The field of NLP is secondarily concerned with helping us come to a better understanding of human language.
- • The input/output of a NLP system can be:
  - – **written text**
  - – **speech**
- • We will mostly concerned with written text (not speech).
- • To process written text, we need:
  - – **lexical, syntactic and semantic knowledge about the language**
  - – **discourse information and real world knowledge**
- • To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

Cont'd…

- There are two components of NLP.
- • **Natural Language Understanding**
  - – Mapping the given input in the natural language into a useful representation.
  - – Different level of analysis required:
- ***morphological analysis,***
- ***syntactic analysis,***
- ***semantic analysis,***
- ***discourse analysis***, …
- • **Natural Language Generation**
  - – Producing output in the natural language from some internal representation.
  - – Different level of synthesis required:
- ***deep planning*** (what to say),
- ***syntactic generation***
- • NL Understanding is much harder than NL Generation. But, still both of them are hard.

Cont'd…

- The difficulty in NL understanding arises from the following facts:
- Natural language is extremely rich in form and structure, and **very ambiguous**.
  – How to represent meaning,
  – Which structures map to which meaning structures.
- One input can mean many different things. Ambiguity can be at different levels.
  – Lexical (word level) ambiguity -- different meanings of words
  – Syntactic ambiguity -- different ways to parse the sentence
  – Interpreting partial information -- how to interpret pronouns
  – Contextual information -- context of the sentence may affect the meaning of that sentence.
- • Many input can mean the same thing.
- • Interaction among components of the input is not clear.

Cont'd…

- The following language related information are useful in NLP:
- • **Phonology** – concerns how words are related to the sounds that realize them.
- • **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.
- • **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- • **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.
- • **Pragmatics** – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.
- • **Discourse** – concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.
- • **World Knowledge** – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

Cont'd…

- Ambiguity
- **I made her duck.**
- • How many different interpretations does this sentence have?
- • What are the reasons for the ambiguity?
- • The categories of knowledge of language can be thought of as ambiguity resolving components.
- • How can each ambiguous piece be resolved?
- • Does speech input make the sentence even more ambiguous?
  - – Yes – deciding word boundaries
- • Some interpretations of : **I made her duck.**

Cont’d…

- 1. I cooked *duck* for her.
- 2. I cooked *duck* belonging to her.
- 3. I created a toy duck which she owns.
- 4. I caused her to quickly lower her head or body.
- 5. I used magic and turned her into a *duck*.
- • duck – morphologically and syntactically ambiguous:
- noun or verb.
- • her – syntactically ambiguous: dative or possessive.
- • make – semantically ambiguous: cook or create.
- • make – syntactically ambiguous:
- – Transitive – takes a direct object. => 2
- – Di-transitive – takes two objects. => 5
- – Takes a direct object and a verb. => 4

Cont’d…

- Ambiguities are resolved using the following methods.
- • *models* and *algorithms* are introduced to resolve ambiguities at different levels.
- • **part-of-speech tagging** -- Deciding whether duck is verb or noun.
- • **word-sense disambiguation** -- Deciding whether make is create or cook.
- • **lexical disambiguation** -- Resolution of part-of-speech and word-sense ambiguities are two important kinds of lexical disambiguation.
- • **syntactic ambiguity** -- her duck is an example of syntactic ambiguity, and can be addressed by probabilistic parsing.

Models to represent Linguistic Knowledge

- We will use certain formalisms (*models*) to represent the required linguistic knowledge.
- **State Machines --** FSAs, FSTs, HMMs, ATNs, RTNs
- **Formal Rule Systems --** Context Free Grammars, Unification Grammars, Probabilistic CFGs.
- **Logic-based Formalisms --** first order predicate logic, some higher order logic.
- **Models of Uncertainty --** Bayesian probability theory.

Algorithms to Manipulate Linguistic Knowledge

- We will use *algorithms* to manipulate the models of linguistic knowledge to produce the desired behavior.
- Most of the algorithms we will study are **transducers** and **parsers**.
  - – These algorithms construct some structure based on their input.
- Since the language is ambiguous at all levels,
- these algorithms are never simple processes.
- Categories of most algorithms that will be used can fall into following categories.
  - – state space search
  - – dynamic programming

Natural Language Understanding

The steps in natural language understanding are as follows:

Words

Morphological Analysis

Morphologically analyzed words *(another step: POS tagging)*

Syntactic Analysis

Syntactic Structure

Semantic Analysis

Context-independent meaning representation

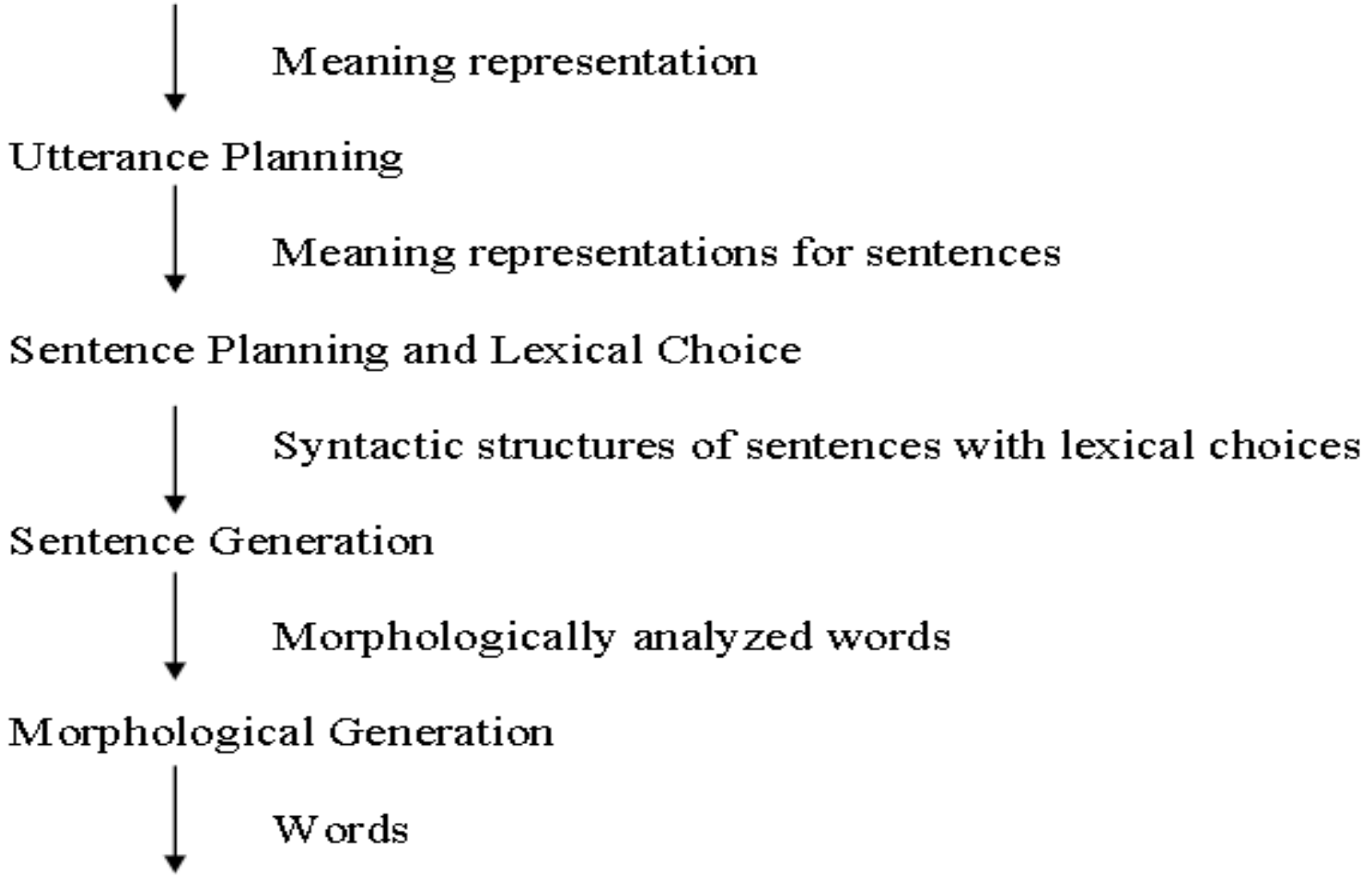
Discourse Processing

Final meaning representation

Parsing

- **Natural Language Generation**

The steps in natural language generation are as follows.



Steps in Language Understanding and Generation

- **Morphological Analysis**
- • Analyzing words into their linguistic components (morphemes).
- • Morphemes are the smallest meaningful units of language.
- cars car+PLU
- giving give+PROG
- geliyordum gel+PROG+PAST+1SG - I was coming
- • Ambiguity: More than one alternatives
- flies flyVERB+PROG
- flyNOUN+PLU
- adam adam+ACC - the man (accusative)
- adam+P1SG - my man
- ada+P1SG+ACC - my island (accusative)

Parts-of-Speech (POS) Tagging

- • Each word has a part-of-speech tag to describe its category.
- • Part-of-speech tag of a word is one of major word groups (or its subgroups).
  - – **open classes** -- noun, verb, adjective, adverb
  - – **closed classes** -- prepositions, determiners, conjuctions, pronouns, particples
- • POS Taggers try to find POS tags for the words.

- • duck is a verb or noun? (morphological analyzer cannot make decision).

- • A POS tagger may make that decision by looking the surrounding words.
  - – Duck! (verb)
  - – Duck is delicious for dinner. (noun)

Lexical Processing

- • The purpose of lexical processing is to determine meanings of individual words.
- • Basic methods is to lookup in a database of meanings – **lexicon**
- • We should also identify non-words such as punctuation marks.
- • Word-level ambiguity -- words may have several meanings, and the correct one cannot be chosen based solely on the word itself.
  - – bank in English
- • Solution -- resolve the ambiguity on the spot by POS tagging (if possible) or pass-on the ambiguity to the other levels.

Syntactic Processing

- • **Parsing --** converting a flat input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.

- • There are different parsing formalisms and algorithms.
- • Most formalisms have two main components:
  - – **grammar --** a declarative representation describing the syntactic structure of sentences in the language.
  - – **parser --** an algorithm that analyzes the input and outputs its structural representation (its parse) consistent with the grammar specification.

- • CFGs are in the center of many of the parsing mechanisms. But they are complemented by some additional features that make the formalism more suitable to handle natural languages.

Cont'd…

- Semantic Analysis
- • Assigning meanings to the structures created by syntactic analysis.

- • Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.

- • Semantic can play an import role in selecting among competing syntactic analyses and discarding illogical analyses.
  - – I robbed the bank -- bank is a river bank or a financial institution

- • We have to decide the formalisms which will be used in the meaning representation.

Cont'd…

- Knowledge Representation for NLP
- Which knowledge representation will be used depends on the application -- Machine Translation, Database Query System.
- • Requires the choice of representational framework, as well as the specific meaning vocabulary (what are concepts and relationship between these concepts -- ontology)

- • Must be computationally effective.
- • Common representational formalisms:
  - – first order predicate logic
  - – conceptual dependency graphs
  - – semantic networks
  - – Frame-based representations

Discourse

- • Discourses are collection of coherent sentences (not arbitrary set of sentences)
- • Discourses have also hierarchical structures (similar to sentences)

- • **anaphora resolution** -- to resolve referring expression
  - – Mary bought a book for Kelly. **She** didn't like **it**.
    - • **She** refers to Mary or Kelly. -- possibly Kelly
    - • **It** refers to what -- book.
  - – Mary had to lie for Kelly. **She** didn't like **it**.
- Discourse structure may depend on application.
- – Monologue
- – Dialogue
- – Human-Computer Interaction

Applications of Natural Language Processing

- • Machine Translation – Translation between two natural languages.
  - – See the Babel Fish translations system on Alta Vista.
- • Information Retrieval – Web search (uni-lingual or multi-lingual).

- • Query Answering/Dialogue – Natural language interface with a database system, or a dialogue system.

- • Report Generation – Generation of reports such as weather reports.

- • Some Small Applications –
  - – Grammar Checking, Spell Checking, Spell Corrector

Machine Translation

- • Machine Translation refers to converting a text in language A into the corresponding text in language B (or speech).

- • Different Machine Translation architectures are:
  - – interlingua based systems
  - – transfer based systems
- • Challenges are to acquire the required knowledge resources such as mapping rules and bi-lingual dictionary? By hand or acquire them automatically from corpora.
- • Example Based Machine Translation acquires the required knowledge (some of it or all of it) from corpora.