Module 13

Natural Language Processing

Version 1 CSE IIT, Kharagpur

13.1 Instructional Objective

- The students should understand the necessity of natural language processing in building an intelligent system
- Students should understand the difference between natural and formal language and the difficulty in processing the former
- Students should understand the ambiguities that arise in natural language processing
- Students should understand the language information required like like
    - Phonology
    - Morphology
    - Syntax
    - Semantic
    - Discourse
    - World knowledge
- Students should understand the steps involved in natural language understanding and generation
- **The student should be familiar with basic language processing operations like**
    - Morphological analysis
    - Parts-of-Speech tagging
    - Lexical processing
    - Semantic processing
    - Knowledge representation

At the end of this lesson the student should be able to do the following:

- **Design the processing steps required for a NLP task**
- **Implement the processing techniques.**

Lesson 40

Issues in NLP

Version 1 CSE IIT, Kharagpur

13.1 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, semantic knowledge about the language**
  - **discourse information, real world knowledge**
- To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.

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.

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.

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.

13.1.1 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.**

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

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.

13.1.2 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.

13.1.3 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

13.2 Natural Language Understanding

The steps in natural language understanding are as follows: