University of Malta
BSc(IT) HonsYear III

CSA305: Natural Language Processing

Lecture 1

Introduction

- General Course Information
- Nature of NLP
  - Related Disciplines
  - Typical Applications

Department of Computer Science and AI 2001/02
Lecturer: Michael Rosner

Course Information

- Practical Assignment (25% of marks = 12.5hrs)
- Books

  M. Covington: NLP for Prolog Programmers - Prentice Hall 1994

  F. Pereira and S. Shieber, Prolog and Natural Language Analysis CSLI lecture Notes No. 10, 1987.

  D. Jurafsky & J. Martin, Speech and Language Processing, Prentice-Hall 2000.
- Web Site
  http://www.cs.um.edu.mt/~mros/cs305
- Me: mros1@um.edu.mt

What is NLP?

Natural language processing concerns the development of computational models of aspects of human language processing, such as:

- Reading and interpreting a textbook.
- Writing a letter.
- Holding a conversation.
- Translating a document.
- Searching for useful information.

Such models are useful in order to write computer programs to perform useful tasks involving language processing, and in order to develop a better understanding of human communication. To some extent these goals are complementary. A better undertanding of human communication may result in better tools.

Nature of NLP

NLP is Interdisciplinary

- Theoretical Linguistics
- Computational Linguistics
- Artificial Intelligence

This list is not exhaustive.

Each field has a contribution to make.

Theoretical Linguistics: Typical Questions

- Typical Questions:
  - What is a language?
  - What is knowledge of language?
  - How can it be finitely characterised?
  - What linguistic forms are there?
  - How does linguistic form constrain meaning?
  - How is knowledge of language acquired given limited exposure.

> The perspective offered by formal language theory enables the answers to many of these questions to be formulated in a precise way. This was one of Chomsky's insights when he wrote *Syntactic Structures* in 1957. It has since become the dominant perspective in theoretical linguistics

Theoretical Linguistics: Tools and Methods

- **Empirical Studies:** Studies of frequency of occurrence, conditional probabilities (e.g. that word X occurs in the vicinity of word Y), intuitions, grammaticality judgements.

> Every competent speaker of any natural language has acquired an immensely rich and systematic body of unconscious knowledge which can be investigated by consulting speakers' intuitive judgements.

it is unlikely that Alfred will be elected
it is improbable that Alfred will be elected

Alfred is unlikely to be elected
Alfred is improbable to be elected*

- **Formal Language Theory:** provides usable definitions of grammatical knowedge.
- **Transformational Grammar:** for handling identity of meaning between non-identical sentences.

Transformational Grammar

Chomsky's Problem

Linguists wanted to explain how the sentences

John loves Mary
Mary is loved by John

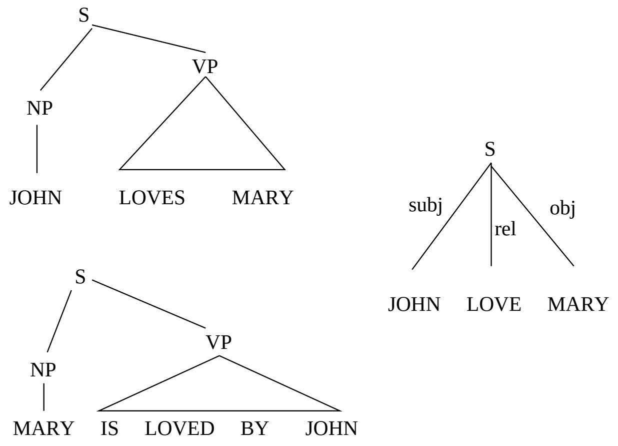
are synonymous (same meaning), despite having different *surface structure*, i.e. roles of subject and object are inverted.

Chomsky's answer

Both sentences are generated from the same underlying "deep structure" in which "deep subject" is John and "deep object" is Mary for both sentences.

These considerations led to a model for NL grammar that employs **two levels** of syntactic representation.

Deep and Surface Structures



SURFACE DEEP

Processing Deep and Surface Structure

DEEP TREE STRUCTURE

↓ tree-tree transformations ↑

SURFACE TREE STRUCTURE

flattening ↓ ↑ parsing

SENTENCE

Problems with Computational Models of Transformational Grammar

Model is essentially generative. To use the model for analysis, generative processes have to be inverted. This is not necessarily straightforward.

- Difficult to formulate grammar to obtain surface tree (unflattening) that does not also assign spurious trees i.e. could never be produced by transformations.
- Discovering sequence of transformations that give rise to surface tree is non trivial, since transformations are generally not reversible.
  - Deletions.
  - Optional Transformations
  - Cycles
  - Other dependencies between transformations.

Computational Linguistics

- **Typical Questions**

  How can linguistic theory be made concrete enough to test?

  How can we represent grammatical and lexical knowledge efficiently?

  Given a grammar and a lexicon, how is the structure of sentences *actually* identified?

  What are the properties of particular grammar formalisms?

  What applications can be based on linguistic theory?

- **Tools and Methods**

  Analysis and Generation Algorithms

  Grammar Formalisms

  Statistical Models based on Empirical Studies of Corpora

Grammar Formalisms Context Free Grammar

- CFG for Arithmetic Expressions

```
AE      ::= Simple | Complex
Simple  ::= Integer | Real | Boolean
Complex ::= (AE BinOp AE)
BinOp   ::= + | - | * | /
```

- CFG for Sentences of English

$S \rightarrow NP\ VP$
$NP \rightarrow DET\ N$
$VP \rightarrow V$
$VP \rightarrow AUX\ V$

Artificial Intelligence

- **Typical Questions**

  What role does language play in the operation of an intelligent agent?

  How can we situate a theory of language use within a general theory of intelligence?

  How does language function as a communicative activity for shared problem solving?

- **Tools and Methods**

  Knowledge Representation

  Formal Reasoning

  Plan Generation and Recognition

  Agent Technology

Examples of NLP Applications

Applications can be classified in different ways: e.g. medium/modality; depth of analysis; degree of interaction.

- Text-based applications
- NL Understanding
- Dialogue Systems
- Multimodal

Text-Based Applications

Processing of written texts such as books, newspapers, reports.

- Finding appropriate documents on certain topics from a text database.
- Extracting information from messages, articles, web pages etc.
- Translating documents from one language to another.
- Text summarisation

N.B. Not all such applications require NLP.

Keyword based techniques can suffice for identifying particular subject areas, e.g. legal, financial etc.

NL Understanding

Other kinds of request require a deeper level of analysis.

Find me all articles concerning car accidents involving more than two cars in Malta during the first half of 1998

Here the system must extract enough information to determine whether the article meets the criterion defined by the query.

A crucial characteristic of an understanding system is that it can compute some representation of the information that can be used for later inference.

A crucial question for an NLP system is: *how much* understanding is necessary to achieve the purpose of the system?

Dialogue-Based Applications

Dialogue-based applications involve man-machine communication.

- NL database query systems
- Automated customer services (e.g. banking services).
- General NL-mediated problem solving systems.

Some of the differences between dialogue and text-based systems

- Language used is less formal.
- System needs to act proactively in order to maintain smooth conversation.
- Use of acknowledgements, clarification subdialogues

Multimodal Applications

Involve two or more modalities of communication

- Text
- Speech
- Gesture
- Image

Text → speech

Speech → text

Multimodal document generation

Spoken translation systems

Spoken dialogue systems

Successful NLP Depends on ...

- Explicit Criteria of Success
- Putting limits on the need for outside knowledge and experience.
- Availability of cheap computing power. Previously, computing power was so expensive that people would expect perfection. This is no longer the case.
- Encoding knowledge of how language works. Still a long way to go here.

Exercises

Give several reasons why weather reports are especially easy for computers to process

Give an example of an English sentence whose meaning is quite unclear in isolation but perfectly clear when heard in the appropriate context.

Course Overview

- **Introduction/Revision**
  - Nature and History of NLP
  - Concepts from Linguistics
  - Prolog Fundamentals
- **Syntax**
  - Grammar of English
  - DCG Grammar Formalism
  - Parsing Algorithms
- **Semantics and Representation**
  - Logical Form
  - Semantics and Representation
  - Linking Syntax and Semantics