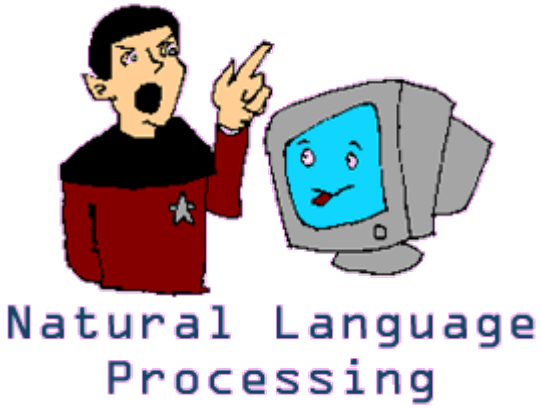
Natural Language Processing
An Introduction



B. Senthil Kumar

Sr. Lecturer, CSE

SSN College of Engineering

Overview

- Natural Language and Speech Processing – Introduction
  - Dialogue Systems
  - MT and QA
- Knowledge in Speech and Language Processing
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics
  - Discourse
- Ambiguity
- Models and Algorithms

Natural Language and Speech Processing

- Different fields coming together, looking at speech and language processing from different perspectives
  - Speech recognition and synthesis [Electrical Engineering]
  - Human language technology
  - Natural language processing [Computer Science]
  - Computational linguistics [Linguistics]

Goal

- How can machines recognize and generate text and speech?
  - Human language phenomena
  - Theories, often drawn from linguistics
  - Algorithms & Applications
- To get computers to perform useful tasks
  - Involving human language
  - Enabling human-machine communication
  - Useful processing of text or speech

Dialogue System

- Programs that converse with humans in natural language – conversional agents
- Language input (automatic speech recognition, natural language understanding)
- Language output (dialogue, speech synthesis)

MT & QA

- The goal of machine translation is to automatically translate a document from one language to another
- Algorithms and mathematical tools needed
- Question-Answering Systems: Instead of just typing keywords, a user might ask complete questions
  - What does “database” mean?
  - What year was Abraham Lincoln born?
  - How many states were in the United States that year?
  - How much Chinese silk was exported to England by the end of the 18th century?

Knowledge Needed

- Phonetics and Phonology – the study of sounds in language
- Morphology – knowledge of the meaningful *components of words*
- Syntax – knowledge of the *structural relationships between words*
- Semantics – knowledge of *meaning*
- Pragmatics – knowledge of the *relationship of meaning to the goals and intentions of the speaker*
- Discourse – knowledge about *linguistic units* larger than a single utterance

- Goal: *Discover what the speaker meant?*

Morphology

- Producing and recognizing the variations of individual words
- The way the word *breaks down into component parts* that carry meaning like singular or plural
- Example: dish, dishes, dishwasher
  - recognizing that *dishes* is plural
  - milk is to milkman -> infer dish is to dishwasher

Morphology is the branch of linguistics that studies **patterns of word formation** within and across languages, and attempts to **formulate rules** that model the knowledge of the speakers of those languages

Syntax

- The sequence of words does not make any sense
  - Ex: I'm I do, sorry that afraid Dave I can't
- Word Order: The knowledge needed to order and group words
  - John hit Bill
  - Bill was hit by John
  - Bill, John hit
- Constituent Structure: *Enraged Cow Injures Farmer With Ax*
  - [Enraged Cow] [Injures] [Farmer With Ax]
  - [Enraged Cow] [Injures] [Farmer] [With Ax]

Syntax is the branch of linguistics that studies the **principles and rules for constructing sentences in natural languages**

Semantics

- The study of meaning
  - Ex: How much Chinese silk was exported to Western Europe by the end of the 18th century?
- Lexical semantics – the meaning of all the words
  - Export or silk
  - Europe, century, end
- Compositional semantics:
  - What exactly constitues *Western Europe* as opposed to Eastern or Southern Europe
  - What does *end* mean when combined with *the 18th century*

Pragmatic

- The kind of actions that speakers intend by their use of sentences is pragmatic or dialogue knowledge
  - Request: Brad, open the door
  - Statement: Brad, the door is open
  - Information question: Brad, is the door open?

Pragmatics is a subfield of linguistics which studies **the ways in which context contributes to meaning**

Discourse

- Makes use of knowledge about how words like that or pronouns like it or she refer to previous parts of the discourse
  - How many states were there in the United States *that year*?
  - Examine the earlier sentence that mentioned about the year
  - For QA, examine the previous questions that were asked

In semantics, discourses are linguistic units composed of several sentences;

in other words, conversations, arguments, or speeches

Pragmatic – the influence of context

- Scene 1: Egmore Railway station, Chennai
  - John: Parry's Corner?
  - Passerby: Ground floor, 3rd counter
- Scene 2: Ticket counter, Egmore Railway station
  - John: Parry's Corner?
  - Clerk: Rs.4.00
- Scene 3: Information Booth, Egmore Railway station
  - John: Parry's Corner?
  - Clerk: 4.25 PM, Platform 2

Pragmatic – the influence of context

- Scene 4: On the Train
  - John: Parry's Corner?
  - Passenger: Change at Park Railway Station
- Scene 5: On the next train, vicinity of Beach Station
  - John: Parry's Corner?
  - Passenger: Opposite to Beach Railway Station

Ambiguity

- Almost in every level ambiguity is introduced, and one of the main tasks in NLP is to resolve such ambiguities
- Input is said ambiguous in multiple, alternative linguistic structures can be built for it

I made her duck =

1. I cooked waterfowl for her
2. I cooked waterfowl belonging to her
3. I created the (plastic?) duck she owns
4. I caused her to quickly lower her body
5. I waved my magic wand and turned her into a waterfowl

Ambiguity

- *duck* and *her* are morphologically or syntactically ambiguous in part-of-speech
- duck --> a verb or noun, her --> dative pronoun or possessive pronoun
- *make* is sematically ambiguous, i.e., make --> *create* or *cook*
- *make* is syntactically ambiguous
  - transitive – taking single direct object (2)
  - ditransitive – taking two objects (5)
  - direct object and a verb – object (her) got caused to perform the verbal action (duck) (4)

I made her duck =

1. I cooked waterfowl for her
2. I cooked waterfowl belonging to her
3. I created the (plastic?) duck she owns
4. I caused her to quickly lower her body
5. I waved my magic wand and turned her into a waterfowl

Ambiguity

- To decide whether duck is a verb or noun --> part-of-speech tagging
- To decide whether make means *create* or *cook* --> word sense disambiguation
- Resolution of part-of-speech and word sense disambiguation --> lexical disambiguation
- Deciding whether *her* and *duck* are part of the same entity (1&4) or are different entity (2) --> syntactic disambiguation

Models and Algorithms

- Taken mainly from Computer Science, Mathematics, and Linguistics
- State Machines and Automata : State machines are formal models that consist of states, transition among states and an input representation

  Ex: deterministic and non-deterministic finite-state automata, finite-state trasducers
- Formal Rule Systems: regular grammars and regular relations, context-free grammars, feature-augumented grammars
- State machines and formal rule systems are the main tools used when dealing with knowledge of phonology, morphology and syntax

Models and Algorithms

- Logic : first order logic (predicate calculus)
- Probability Theory: ability to solve ambiguity problems
- *“given N choices for some ambiguous input, choose the most probable one”*
- Each of the other models can be augumented with probabilities
  - Ex: Markov Models (HMM)
- Processing a language with any models involves a search through a space of states representing hypothesis about an input
- In speech processing --> search through a space of phone sequences for the correct word

Models and Algorithms

- In parsing --> search through a space of trees for the syntactic parse of an input sentence
- In machine translation --> search through a space of translation hypothesis

  for translation of a sentence into another language

The Turing Test

- Alan Turing (1950): Empirical test for Artificial Intelligence
- A human interrogator asks questions to a human and to a machine through a teletype, and tries to find out who is the human and who is the machine.

Q: Please write me a sonet on the topic of the Fouth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause for 30 seconds) 105621.

ELIZA

- Weizenbaum (1966): Program imitating the responses of a psychotherapist
- Used simple pattern matching, without any deeper knowledge of the world or of the conversation
- http://www-ai.ijs.si/eliza/eliza.html

User: You are like my father in some ways.
ELIZA: What resemblance do you see?
User: You are not very aggresive but I think you don't want me to notice that.
ELIZA: What makes you think I am not very aggressive?
User: You don't argue with me.
ELIZA: Why do you think I don't argue with you?

Reference

Speech and Language Processing, Daniel Jurfsky and James H. Martin

http://www-ai.ijs.si/eliza/eliza.html

Thank You!