Lecture 1
Introduction to NLP

CS 6320

Outline

- Introduction to NLP
- NLP Resources

Definition

- NLP is a technology that creates and implements computer models for the purpose of performing various natural language tasks. It is used for building NL interfaces to databases, machine translation, and others.
- NLP is playing an increasing role in curbing the information explosion on Internet, Government and corporate America.

Related areas

- NLP is a difficult, and largely unsolved problem. One reason for this is its **multidisciplinary** nature:
  - **Linguistics** : How words, phrases, and sentences are formed.
  - **Psycholinguistics** : How people understand and communicate using human language.
  - **Computational linguistics:** Deals with models and computational aspects of NL (e.g. algorithms).

Related areas

- **Philosophy**: relates to the semantics of language; notion of meaning, how words identify objects. NLP requires considerable knowledge about the world.
- **Computer science**: model formulation and implementation using modern methods.
- **Artificial intelligence**: issues related to knowledge representation and reasoning.
- **Statistics:** many NLP problems are modeled using probabilistic models.
- **Machine learning:** automatic learning of rules and procedures based on lexical, syntactic and semantic features.
- **NL Engineering**: implementation of large, realistic systems. Modern software development methods play an important role.

Applications of NLP

- **Text - based applications**:
  - Finding documents on certain topics (document classification)
  - Information retrieval: search for key words or concepts,
  - Information extraction: extract information related to key words,
  - Complete understanding of texts: requires a deep structure analysis,
  - Translation from a language to another,
  - Summarization,
  - Knowledge acquisition.
- **Dialogue - based applications** (involve human - machine communication):
  - Question - answering
  - Tutoring systems
  - Problem solving.
- **Speech processing**

Basic levels of language processing 1/2

- **Phonetic** - how words are related to the sounds that realize them. Essential for speech processing.
- **Morphological Knowledge** - how words are constructed : e.g friend, friendly, unfriendly, friendliness.
- **Syntactic Knowledge** - how words can be put together to form correct sentences, and the role of each play in the sentence. e.g.:

  John ate the cake.
- **Semantic Knowledge** - Words and sentence meaning:

  They saw a log.

  They saw a log yesterday.

  He saws a log.

Basic levels of language processing 2/2

- **Pragmatic Knowledge**- how sentences are used in different situations(or contexts).

  Mary grabbed her umbrella.

  a) It is a cloudy day.

  b) She was afraid of dogs.

- **Discourse Knowledge** - how the meaning of words and sentences is affected by the preceding sentences; pronoun resolution.

  John gave his bike to Bill.

  He didn't care much for it anyway.

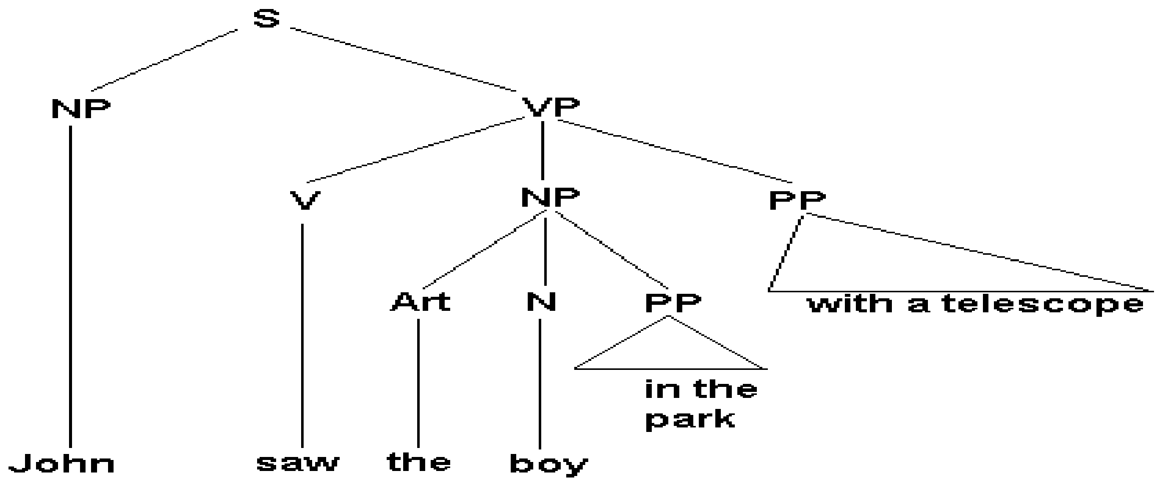
- **World Knowledge** - the vast amount of knowledge necessary to understand texts. Used to identify beliefs, goals.
- **Language generation** - have the machine generate coherent text or speech. Needs planning.

Examples of NLP difficulties 1/4

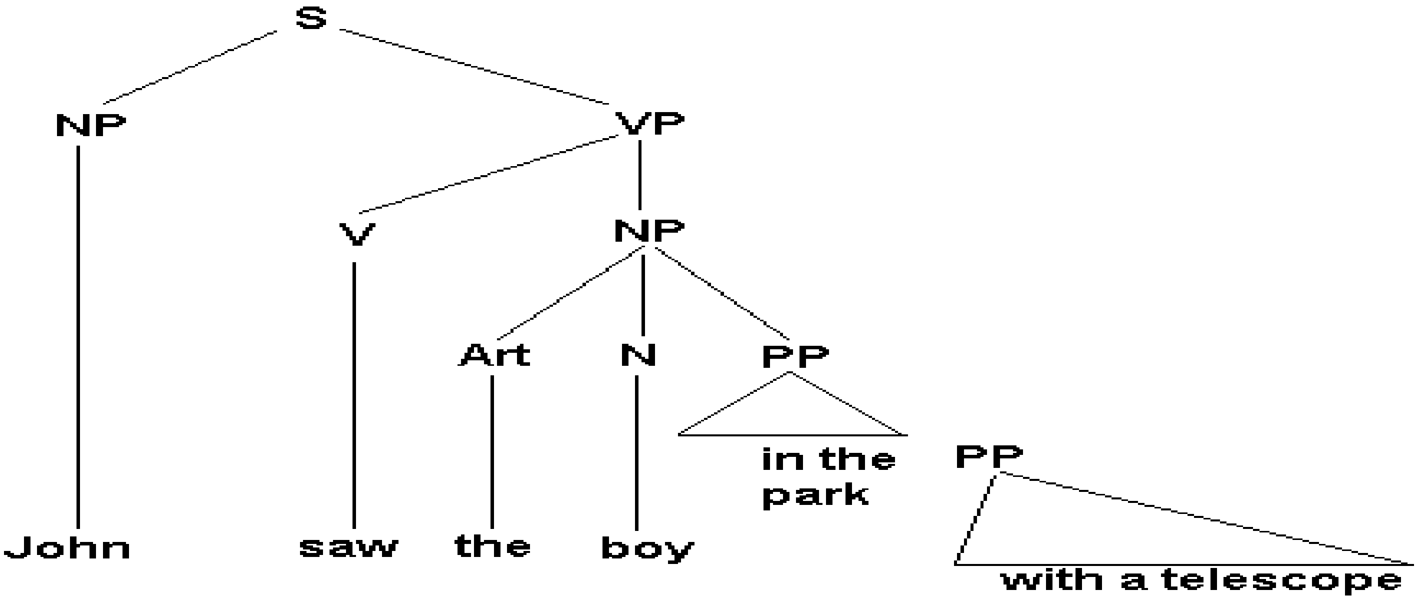
A major difficulty is *lexical ambiguity*. There are three types:

- **Structural ambiguity**- when a sentence has more than one possible parse structures; e.g. attachment :

John saw the boy in the park with a telescope.



Examples of NLP difficulties 2/4



Examples of NLP difficulties 3/4

- **Syntactic ambiguity**- when a word has more than one part of speech:

  Rice flies like sand.

  Note that these syntactic ambiguities lead to different parse structures. Sometimes it is possible to use grammar rules (like subject verb agreement) to disambiguate:

  Flying planes are dangerous.

  Flying planes is dangerous.

- **Semantic ambiguity**- when a word has more than one possible meaning (or sense):

  John killed the wolf.

  John killed the project.

  John killed that bottle of wine.

  John killed Jane. (at tennis , or murdered her)

Example of NLP difficulties 4/4

- **Ambiguities of a sentence**:

  Example:
  I made her duck.

  Possible interpretations:

1. I cooked waterfowl for her.
2. I cooked waterfowl belonging to her.
3. I created the (plaster ?) duck she owns.
4. I caused her to quickly lower her head or body
5. I wave my magic wand and turned her into undifferentiated waterfowl.

Computational Aspects of NLP

- Language processing is symbolic

  words, concepts, events, actions, ideas
- Language processing is discrete
- Language processing is module sequential

  tokenizer, POS, syntactic parser, NER, semantic parser, coreference
- Language processing is compositional

  letters, words, phrases, sentences, paragraphs, documents
- Language processing is sparse

  infinite possible combination of words, yet only some appear in text

State of the art in NLP Research 1/2

- **NL Publications**
  - ACL, NAACL, EACL
    - Conferences
    - Journals
  - AAAI - every year proceedings.
  - IJCAI - every second year proceedings.
  - SemEval

- **Natural Language Engineering** (journal).

- **Information Retrieval/Extraction**

State of the art in NLP Research 2/2

- **Machine Readable Dictionaries** (MRD) WordNet, LDOCE.
- **Large corpora**:
  - Penn Treebank—contains 2-3 months of Wall Street Journal articles (~ .5 million words of English, POS tagged and parsed),
  - Brown corpus,
  - SemCor.

NLP Resources

- WordNet
- Extended WordNet (XWN)
- FrameNet
- POS Tagger
- Syntactic Parse
- Treebank
- SemCor
- Stanford Core NLP
- SQuAD (Stanford QA data set)
- Deep Learning software packages
- ***http://www.hlt.utdallas.edu/~moldovan/CS6320.20/resources.html***

WordNet 1/7

- A lexical database for the English language
- Developed by the Cognitive Science Laboratory at Princeton University (professor George A. Miller)
- Its design was inspired by current psycholinguistic theories of human lexical memory
- User friendly interface
- Library of C functions: allows you to access the synsets and the relations between them directly from your programs

WordNet 2/7

- Parts of speech covered by WordNet:
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs
- The fundamental unit of WordNet is the SYNSET (synonym set)
  - each synset represents one underlying lexical concept
- Different relations link the synonym sets

WordNet 3/7

- WordNet 3.0 (latest release)
- A synset example:

  (car, auto, automobile, machine, motorcar)
- “Car” has 5 senses in WN.
- Each synset has at least one definition and there could be some sentences using the words of the synset.
- For the previous synset, we have:
  - Definition: 4-wheeled motor vehicle; usually propelled by an internal combustion engine.
  - Sentence: “He needs a car to go to work.”

WordNet 4/7

- Relations between synsets:
  - Synonymy
  - Hypernymy (superordination)
  - Hyponymy (subordination)
  - Holonymy (whole to part relation)
  - Meronymy (part to whole relation)
  - Antonymy
  - Troponymy (particular way to do something)

WordNet 5/7

- **Synonymy relation**:
  - (motor vehicle, automotive vehicle)
  - Definition: a self propelled wheeled vehicle that does not run on rails.
- **Hypernymy relation**:
  - (vehicle)
  - Definition: a conveyance that transports people or objects.
- **Hyponymy relation**:
  - (ambulance)
  - Definition: a vehicle that takes people to and from hospitals.

WordNet 6/7

- **Holonymy relation:**
  - (bicycle wheel)
  - Definition: the wheel of a bicycle
  - Has the holonym:
  - (bicycle, bike, wheel)
  - Definition: has two wheels; moved by foot pedals
- **Meronymy relation**:
  - (bicycle wheel)
  - Definition: the wheel of a bicycle
  - Has the meronym:
  - (spoke, radius)
  - Definition: a radial member of a wheel joining the hub to the rim.

WordNet 7/7

- **Antonymy relation:**
  - (sweet)
  - Definition: having a pleasant taste (as of sugar)
  - Has the antonym:
  - (sour)
  - Definition: having a sharp biting taste.
- **Troponymy relation**:
  - (dream)
  - Definition: experience while sleeping.
  - Has the troponym:
  - (fantasize)
  - Definition: have fantasies.

eXtended WordNet 1/4

- Provides several important enhancements (over WordNet 2.0) intended to remedy the present limitations of WordNet
- WordNet 2.0 glosses are syntactically parsed, transformed into logic forms and content words are semantically disambiguated
- eXtended WordNet is an ongoing project at the Human Language Technology Research Institute (http://www.hlt.utdallas.edu), The University of Texas at Dallas
- second release- the next release scheduled for the end of 2004

eXtended WordNet 2/4

- For each WordNet 2.0 gloss, eXtended WordNet associates three types of information:
  - its parse tree
  - its logic form
  - each noun, verb, adjective and adverb of the gloss is semantically disambiguated (with respect to WordNet 2.0)

- Exploits the rich information contained in the definitional glosses

- Increases the connectivity between synsets

- Provides computer access to a broader context for each concept

eXtended WordNet 3/4

- Consists of four XML files--one for each part of speech:

  Noun
  Verb
  Adjective
  Adverb

- The eXtended WordNet may be used as a Core Knowledge Base for applications such as:
  - Question Answering
  - Information Retrieval
  - Information Extraction
  - Summarization
  - Natural Language Generation
  - Inferences
  - other knowledge intensive applications

eXtended WordNet 4/4

- The glosses contain a part of the world knowledge since they define the most common concepts of the English language

FrameNet 1/7

- Frames and Understanding

  Hypothesis: People understand things by performing mental operations on what they already know. Such knowledge is describable in terms of information packets called **frames.**

FrameNet 2/7

The core work of FrameNet

- Characterized frames
- Find words that fit the frames
- Develop descriptive terminology
- Extract sample sentences
- Annotate selected examples
- Derive “valence” descriptions

FrameNet 3/7

Sample Event Frame:

Commercial Transaction

Initial state:

Vendor has Goods, wants Money

Customer wants Goods, has Money

Transition:

Vendor transmits Goods to Customer

Customer transmits Money to Vendor

Final State:

Vendor has Money

Customer has Goods

FrameNet 4/7

Meaning and Syntax

- The various verbs that evoke this frame introduce the elements of the frame in different ways.
  - The identities of the buyer, seller, goods and money
- Information expressed in sentences containing these verbs occurs in different places in the sentence depending on the verb.

FrameNet 5/7

She bought some carrots from the greengrocer for a dollar.

She paid a dollar to the greengrocer for some carrots.

She paid the greengrocer a dollar for the carrots.

FrameNet 6/7

FrameNet Product

- For every target word,
- describe the *frames* or conceptual structures which underlie them,
- and annotate example sentences that cover the ways in which information from the associated frames are expressed in these sentences

FrameNet 7/7

FrameNet Entities and Relations

- Frames
  - Background
  - Lexical
- Frame Elements (Roles)
- Binding Constraints
  - Identify
- ISA (x:Frame, y:Frame)
- SubframeOf (x:Frame, y:Frame)
- Subframe Ordering
  - precedes
- Annotation

TreeBank 1/4

- TreeBank web page:

  Treebank = a bank of linguistic trees

  - Large corpus of syntactic and semantic annotated texts
  - Penn Treebank Project has been developed at University of Pennsylvania
  - POS (Part Of Speech) tagged
  - Parsed trees
  - Corpora:
    - Wall Street Journal
    - The Brown Corpus
    - Switchboard
    - ATIS

TreeBank 2/4

- An example of POS tagged text:

  [Mr./NP Volk/NP]
  ,/,
  [55/CD years/NNS]
  old/JJ ,/, succeeds/VBZ
  [Duncan/NP Dwight/NP]
  ,/,
  [who/WP]
  retired/VBD in/IN
  [September/NP]
  ./.

TreeBank 3/4

- The parse tree for previous text:

```
((S
   (NP (NP Mr. Volk)
          ,
          (ADJP (NP 55 years) old)
           ,)
   (VP succeeds
          (NP (NP Duncan Dwight)
                 ,
                 (SBAR
                   (WHNP who)
                   (S (NP T)
                        (VP retired
                              (PP in
                                    (NP September)))))))    .))
```

TreeBank 4/4

- This information is very useful for learning algorithms
- The hand tagged text is a good source of correct examples for training the algorithms
- The parsing algorithms use TreeBank for training and measuring their accuracy
- Brill's tagger was trained on TreeBank
- Charniak's parser was trained on TreeBank
- Statistical purposes

POS Tagger

- The most popular Part-Of-Speech tagger is Brill's tagger
- Implements a simple rule based tagger using transformation-based learning
- INPUT: tokenized text -> OUTPUT: POS tagged text (each word gets attached its POS)
- Accuracy in the upper 90s percents
- Can be trained on your own corpus
- Can be augmented with new rules, making possible its tuning for a certain domain
- Has been used in eXtended WordNet development (an enhanced version)
- State-of-the-art POS tagger

POS Tagger (example)

INPUT: I saw the boy in the park with the telescope .

OUTPUT: I/PRP saw/VBD the/DT boy/NN in/IN the/DT park/NN with/IN the/DT telescope/NN ./.

Syntactic Parser

- The most popular syntactic parser in NLP community is Charniak's parser
- A maximum-entropy-inspired parser
- INPUT: plain text -> OUTPUT: parse trees (each sentence gets its parse tree)
- Accuracy around 90%
- Has been trained on TreeBank
- Has been used in eXtended WordNet development
- State-of-the-art syntactic parser

Syntactic Parser (example)

INPUT:

<s> I saw the boy in the park with the telescope . </s>

OUTPUT:

```
(S1 (S (NP (PRP I) )
          (VP (VBD saw)
                (NP (DT the) (NN boy) )
                (PP (IN in)
                      (NP (NP (DT the) (NN park) )
                            (PP (IN with)
                                  (NP (DT the) (NN telescope) ) ) ) ) )      (. .)
    )
```

SemCor

- Developed by Miller in 1993.
- Contains 37176 sentences from 352 newspaper articles on different topics.
- A corpus tagged with WordNet senses.
- An XML-like format: for each word has attached its POS, its lemma and its WordNet sense.
- The different versions of SemCor depend on the WordNet version used for the senses.

SemCor

- Word Sense Disambiguation (WSD).
- Learning algorithms for WSD could use this corpus as a training text.
- Projects that need data annotated with WordNet senses may use this corpus.