Natural Language Processing

Book: Speech and Language Processing, by M. Jurafsky, & J. Martin, New York: Prentice-Hall (2000).

What is natural language processing?

- **Process information contained in natural language text**

  - Computational Linguistics(CL),
  - Human Language Technology(HLT),
  - Natural Language Engineering(NLE)

NLP for Machines

- Analyse, understand and generate human language just like humans do.
- Applying computational techniques to language domain.
- To explain linguistic theories to use the theories to build systems that can be of social use.
- Started off as a branch of Artificial Intelligence.
- Borrows from Linguistics, Psycholinguistics, Cognitive Science and Statistics.

NLP Applications

Automatic Summarization

Intelligently shortening long pieces of text

Named entity recognition

Locate and classify named entities pre-defined categories such as the organizations; person names; locations etc.

Speech recognition

Enables computers to recognize and transform spoken language into text - dictation - and, if programmed, act upon that recognition - e.g. in case of assistants like Google Assistant Cortana or Apple's Siri

Sentiment analysis

To identify, for instance, positive, negative and neutral opinion form text or speech widely used to gain insights from social media comments, forums or survey responses

Topic segmentation

Automatically divides written texts, speech or recordings into shorter, topically coherent segments and is used in improving information retrieval or speech recognition

Production-Level Applications

A computer program in Canada accepts daily weather data and
automatically generates weather reports in English and French

Over 1,000,000 translation requests daily are processed by the Babel Fish system available through Altavista

A visitor to Cambridge, MA can ask a computer about places to eat using only spoken language. The system returns relevant information from a database of facts about the restaurant scene.

Prototype-Level Applications

Computers grade student essays in a manner indistinguishable from human graders

An automated reading tutor intervenes, through speech, when the reader makes a mistake or asks for help

A computer watches a video clip of a soccer game and produces a report about what it has seen

A computer predicts upcoming words and expands abbreviations to help people with disabilities to communicate

Why NLP ?

- A hallmark of human intelligence.
- Text is the largest repository of human knowledge and is growing quickly
- Computer programmes that understood text or speech.

History of NLP

- In 1950, Alan Turing published an article titled “Machine and Intelligence” which advertised what is now called the Turing test as a subfield of intelligence.
- Some beneficial and successful Natural language system were developed I the 1960s were SHRDLU, a natural language system working in restricted “blocks of words” with restricted vocabularies was written between 1964 to 1966.

Components and Process of NLP

- Components of NLP
- Linguistics and Language
- Steps of NLP
- Techniques and Methods

Components of NLP

- **Natural Language Understanding**: taking some spoken/typed sentence and working out what it means.
- **Natural Language Generation**: taking some formal representation of what you want to say and working out a way to express it in a natural(human) language.

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

Components of NLP

Natural Language Generation:

- Producing output in the natural language form some internal representation.

Different level of synthesis required:

- Deep planning(what to say)
- Syntactic generation

Note: NL Understanding is much harder than NL Generation. But, still both of them are hard.

Linguistics and language

Linguistics is the science of language.

Its study includes:

- Sounds which refers to phonology
- Word formation refers to morphology
- Sentence structures refers to syntax
- Meaning refers to semantics
- Understanding refers to pragmatics

Steps of NLP

1. Morphological and Lexical Analysis
2. Syntactic Analysis
3. Semantic Analysis
4. Discourse Integration
5. Pragmatic Analysis

Morphological and Lexical Analysis

- The lexicon of a language is its vocabulary that includes its words and expressions.
- Morphology depicts analysing, identifying and description of structure of words.
- Lexical analysis involves dividing a text paragraphs, words and the sentences.

Syntactic Analysis

- Syntax concerns the proper ordering of words and its affects on meaning
- This involves analysis of the word in a sentence to depict the grammatical structure of the sentence.
- The words are transformed into structure that shows how the words are related to each other
- For example: “`the girl the go to the school`”. This would definitely be rejected by the English syntactic analyser.

Semantic Analysis

- Semantics concerns the (literal) meaning of word, phrases and sentences.
- This abstracts the dictionary meaning or the exact, meaning from the context.
- The structures which are created by the syntactic analyser are assigned meaning.

For example: “`colourless blue idea`”. This would be rejected by the analyser as colourless blue do not make any sense to gether.

Discourse Integration

- Sense of the context
- The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it.

For example: the word “`it`” in the sentence “`she wanted it`” depend upon the prior discourse context.

Pragmatic Analysis

- Pragmatics concerns the overall communicative and social context and its effect on interpretation.
- It means abstracting or deriving the purposeful use of the language in situations.
- Importantly those aspects of language which required world knowledge.
- The main focus is on what was said is reinterpreted on what it actually means.

For example: “close the window?” should have been interpreted as a request rather than an order.

Natural Language Generation

- NLG is the process of constructing natural language outputs from non-linguistic inputs

- NLG can be viewed as the reverse process of NL understanding

- A NLG system may have following main parts:
  i) Discourse Planner: what will generated, which sentences
  ii) Surface Realizer: realises a sentence from its internal representation
  iii) Lexical Selection: selecting the correct words describing the concepts

Techniques and methods

Machine Learning

- The learning procedures used during machine learning
- Automatically focuses on the most common cases
- Whereas when we write rules by hand it is often not correct at all
- Concerned on human errors

Techniques and methods

Statistical inference

- Automatic learning procedures can make use of statistical inference algorithms
- Used to produce models that are robust(means strength) to unfamiliar input e.g. containing words or structures that have not been seen before
- Making intelligent guesses.

Techniques and methods

Input database and Training data

- System based on Automatically learning rules can be made more accurate simply by supplying more input data or source to it.
- However, system based on hand-written rule can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task.

Natural language vs. Computer language

- Ambiguity is the primary difference between natural and computer language.
- Formal programming language are designed to be unambiguous. They can be defined by a grammar that produces a unique parse for each sentence in the language.
- Programming languages are also designed for efficient(deterministic) parsing.
- They are deterministic context-free languages(DCLFs)

Future of NLP

- Make computers as they can solve problems like humans and think like humans as well as perform activities that humans cannot perform and making it more effiecient than humans.
- As natural language understanding or readability improves, computers or machines or devices will be able to learn from the information online and apply what they learned in the real world.

Difficulties in NLP

- “The person who done it- it’s their fault”
- “The man who knew him went left”
  “The man who knew he went left”
- Mother was baking.
  The apple pie was baking.
  Mother baked an apple pie.
  An apple pie was baked by mother.

Lexical ambiguity

- “Time flies like an arrow.”
- “I saw that gas can explode.”
- “They should have scheduled meeting.”
- “Visiting relatives can be annoying.”
- **Note: ambiguity due to class of word.**

Lexical ambiguity

- “The pitcher fell and broke”
- “The water is in the pitcher.”
- “John drank a pitcher.”
- “John drank a tall pitcher while watching the baseball game.”
- “She approached the bank.”
- **Note: ambiguity due to multiple definitions.**

Syntactic ambiguity

- “John saw the woman in the park with a telescope.”

“Gardens paths”

- The sentence "The horse raced past the barn door fell down" is not ambiguous, but processing it certainly causes structural ambiguity problems. Its ambiguity is said to be local rather than global since it can be resolved by the end of the sentence.\
- Such sentences are called garden path sentences, possibly because they lead one down the garden path in a quest for understanding. Here are some more examples:
  - The artist painted on eh wall was black.
  - John told the man the dog bit Jane was hungry.
  - The horse raced down the garden path meandered.

“Gardens paths” (2)

- Using the context-free grammar formalism, the underlying model for this phenomenon is a grammar segment of the form:
  1. A—xy
  2. B—yz
  3. C—xB
- Given the input sentence xyz, the xy part is first interpreted as an A and then the z is left dangling since Az is unparsable. The processor has to back up and reanalyze the xy, grouping the y with the z of the x.
- Computers can easily be programmed to handle this, either to an extent that is arbitrarily limited by using look-ahead techniques or to a virtually unlimited extent by backtracking. But people have trouble with garden path sentences because they do not typically do backtracking an can handle only very limited amount of parallel processing to look-ahead. The limit is commonly believed to be three. This means a person can keep three syntactic constitutes hovering unanalyzed in his or her head and can parse three levels of embedded phrases.

“Gardens paths” (3)

- Less extreme cases of local ambiguity occur with verbs like "have," which are sometimes auxiliary verbs and sometimes main verbs. After the first three words of each of the following sentences, one cannot tell whether it is a command or a question.
  - Have the people do it!
  - have the people done it!
- If the last words were omitted from the following sentences, they would still be complete sentences: reaching the last words causes the preceding phrase to be reanalyzed as reduced relative clauses.
  - Is the book on the shelf red?
  - Is the number of people over 40 odd?

Discourse analysis

- The rest of a discourse can resolve ambiguities that are global on the sentence level
- At the discourse level, two particular linguistic connection phenomena are also handle:
  - ellipsis
  - anaphora.

Ellipsis

- Ellipsis is the omission of a word or words from a sentence, rendering it syntactically, but not semantically, incomplete. Not all cases require context.
  - "Stop that" is always short for "You stop that.“
  - "John has five dollars and Jane nine."
- Some sentences are almost completely elicited and hence totally depend on context, such as "Why?“
- Example of dialog:
  - John: Who just walked by?
  - Jane: A tall blonde man.
- The implicit verb phrase for the isolated noun phrase may arise from a context at large rather than a previous statement, as in "The next train to Nashville," when said to someone is a railway information booth.

Anaphora

- Anaphora is a matter of abbreviation rather than omission. The referent is generally a previous expression. The abbreviated form is usually a noun phrase, either a pronoun or a definite noun phrase, such as "that" in "Stop that," but it can also be an adjective or adverb, as in "such things" or "do so.“
- A natural language system needs reasoning capability to find the possible referents and then select on of them. This process is facilitated by keeping track of the current focus of the discourse. The focus is the entity with which the discourse is most concerned at any particular time. It can shift unpredictably and there can minor foci.
- One effect of the syntactic distinction in the active/passive pair of sentences "Mother baked an apple pie" and "An apple pie was baked by Mother" is that in the first sentence. Mother is more in focus than the pie, whereas in the second the opposite is true. Tracking methods vary with the type of discourse—narrative, directions, argument, or conversation.

Anaphora (2)

- As with modifier attachment, proximity is a major consideration in determining referents, but it certainly does not suffice. For example, in "Mother cleaned the house, baked a pie, sat in a chair, and ate it," the correct referent is the closest edible one. In the following dialogue, the first pronoun ("that") refers to the most recent possible referent ("one") refers to the previous referent ("the answer")
  - John: The answer is one
  - Jane: That is wrong— it is two.

  As a more subtle example, consider:
  - I just found a kitten and I have a cat so I am going to give it away.

Anaphora (3)

- The knowledge that tells us seniority is being honored comes from living in a society where pets are treated a certain way. It is not the kind of knowledge that could be easily be encoded in semantic markers. Compare the last sentence to "I just won anew car and I have an old car so I'm going to give it away.“
- Syntactic considerations alone sometimes eliminate possible referents. Although the pie owner and eater may or may not be the same person in the first sentence of the following pair, they cannot be in the second sentence:
  - John ate his pie.
  - He ate John's pie.

Anaphora (4)

- The next example shows that syntax might play no role whatsoever. The referent of "she" is unclear in the fist sentence and very clear, though different, in the following two:
  - Jane gave Joan the candy because she was nice.
  - Jane gave Joan the candy because she was hungry.
  - Jane gave Joan the candy because she wasn't hungry.
- "They" and "it" have the same referent in the following example, despite the fact that they differ in number and hence are syntactically incomplete:
  - Mother picked an apple
  - They are good sources of pectin.
  - She will make a pie with it.
- Thus even knowing precisely what the focus is may not pinpoint it. Although the apple is the only thing in focus, it could be as a type of fruit or as a specific piece of fruit. The difficulties of determining the referents of ellipsis and anaphora are obviously great.

Pragmatics

- Often the referent of anaphora or ellipsis is something that was never previously stated but merely implied. In "The next train to Nashville" and "I just found a kitten and I have a cat so I am going to give it away," the referents could not be established from the discourse alone but required broader contexts. The extra knowledge used was of a pragmatic nature.
- Extensive knowledge about the subject matter may be necessary to resolve references. Basic concepts used include connections between parts of objects, actions, and events. Thus, in the following text, we infer that the definite noun phrase "the apples" refers to an ingredient of the pie mentioned in the previous sentence:
  - Mother is going to make a pie.
  - She is washing the apples now.

Pragmatics (2)

- Establishing the referent in "I just found a kitten and I have a cat so I am going to give it away," on the other hand, involved knowledge that was conceptually more complicated and much more subjective.
- Even systems that deal with simple objective knowledge domains should be equipped with extra knowledge about their domains. That way they can avoid situations like the following. An insurance data base query system that seemed to understand gender distinctions when asked about male policy holders was asked a question about male insurance agents. In an attempt to be helpful, it responded: "Insurance agents don't have sex-only customers do."
- Real understanding goes beyond facts to ascertaining goals. Goal inferencing was applied in interpreting "The next train to Nashville," and its application is attempted in the following situation. A person who attempted to phone a theatre but reached a taxi company instead did not understand the initial greeting and inquired, "Metropolitan Theatre?" The response was "Which one?", indicating that the inquiry was interpreted as a request for a ride to the theatre, for that was the only way it made sense to the hearer.

Pragmatics (3)

- The general nature of a response depends on the statement's underlying form, which is related to but not necessarily the same as its superficial mood. In "Do you know what time it is?" we saw that an imperative can masquerade as an interrogative. Conversely, declarative statements sometimes should be interpreted as commands or questions, for example, "I forgot how to tie this" or "I thought you were going to have left but now." The conditional interrogative can be misleading. "Would you pass the pie?" is a request, whereas "Would you like some pie?" is an offer. \
- Modern approaches to natural language processing have emphasized semantics and pragmatics at the expense of syntax. First the concept of syntactic case was broadened to encompass semantics. Case grammars capture the distinction between the syntactically identical "Mother made the pie with a new apple" and "Mother made the pie with a new recipe" by assigning the instrumental case to "recipe" and the material case to "apple." They also explain the puzzle of "Mother and the apple pie were baking"; its ungrammatically is due to the conjoining of two different semantic cases.

Pragmatics (4)

- Conceptual dependency theory practically eliminated syntactic considerations and used a small set of semantic primitives that describe relationships to represent meanings. It led to a trend of incorporating world knowledge into increasingly complex data structures based on frames. A frame is a cluster of properties associated with an object of an event.
- When generalized to a sequence of events or an involved situation, frames are known as scripts. Scripts for common occurrences get filled in with the standard details unless given contrary information. Thus a restaurant script would have a default recording of this typical chain of events: being seated, getting a menu, ordering, being served, eating, getting a bill, and paying.

Pragmatics (5)

- If a system is told that John went to Friendly's and ordered a hamburger and then asked, "What did John eat?", it would demonstrate the inference that he had eaten the hamburger he'd ordered. But if told that John went to Friendly's and ordered a hamburger then left, it would say he hadn't eaten and may also be able to answer the question "Why was John arrested?" provided it had other scripts that relate arrests to money, Gauging the significance of an omission to determine whether it should be filled in requires both domain knowledge and language knowledge.
- The frame devices effectively endow the computer system with a background of human experiences, providing it with default contexts for resolving ambiguity and referents as well as encoding expectations. However, they do not capture interaction generalizations. For example, completely separate scripts are needed for different types of purchasing situations.

Pragmatics (6)

- Since meaning does not just depend on a shared knowledge base of objective descriptions of the world but also on subjective aspects of the response, such as belief systems and current cognitive processing, a natural language system also needs a model of the user. User modeling is harder than representing any quantity of world knowledge because it's a matter of representing mental processes that aren't understood. Ultimately a dynamic user model, capable of readjusting its expectations, is needed to model interpersonal aspects of communication.
- It is not clear that user models are respectable and, if they are, the representations still may not model human understanding. Even the necessary objective knowledge may not be representable by a formal system, let alone one that can be computerized. Representing language by pieces of formal structures is akin to representing images by dots, and it's well known how difficult it is to recognize an image from a close-up view of the visual patterns. Until cognitive processes are better understood, the approach to incorporating pragmatics into natural language systems must be pragmatic itself.

Difficulties of NLP (Conclusions)

- Ambiguity
- Usage of context of different levels
  - Ellipsis
  - Anaphora
- Idioms and metaphors
- Usage of extralinguistic knowledge
  - About domain
  - About users, in particular, mimics and features of articulation during dialog
  - About world