Machine Learning

Lecture 17

Learning in Natural Language Processing (NLP).

Problems and main concepts of NLP.

Programs based on NLP

- Question-Answering Systems
- Control by command in Natural Language
- Readers from text to speech
- Translators
- Search of information by query in Natural Language
- OCR – Optical Characters Recognition

Main areas of NLP

- Understanding of NL
- Generation of NL
- Analyzing and synthesis of speech

Kinds of learning in NLP

- Learning in ALICE-like dialog systems
- Learning to recognize grammatically correct sentences
- Categorization of texts
- Learning to search documents by query in Natural Language
- Speech recognition
- Learning to association between images and description of image in NLP

Features of Natural Language – reasons of difficulty of simulation of its understanding

- Knowledge of the subject matter of a sentence is clearly required. The meaning of a sentence depends not only on the things it describes, but also in both aspects of its causality; what caused it to be said and what result is intended by saying it. In other words, the meaning of a sentence depends not only on the meaning of a sentence itself, but on who says it and when, where, how, why, and to whom it is said.
- Precise shades of meaning vary with context and that meanings of certain words are always relative. Comparative modifiers such as "light" and "heavy" belong to this category; we interpret them according to what they modify. We assume, for example, that a light computer is heavier than a heavy book.
- Idioms and metaphors ("walking on thin ice“, "walking on water“, “to eat dog”)
- The cognitive process of understanding is itself not understood. First we must ask what it means to understand a sentence. The answer usually given is to make a model of its meaning. But this answer just generates another: What does meaning mean? Rather than delve into the meaning of meaning as philosophers have been doing for centuries, we approach this as 20th century computer scientist and seek a more practical answer.

Features of Natural Language – reasons of difficulty of simulation of its understanding (2)

- The appropriateness of a response depends on the situation. For example, suppose a woman tells a natural language interface to a train schedule database that she needs to take the first train to Nashville. A response consisting of the departure time and track of the next available train indicates that the system completely understood what she said. But if she tells it to boyfriend, who knows her mother is in the Nashville hospital, she would think he wasn't at all understanding if he responded with railway information.
- As another example, consider the sentence: "Do you know what time it is?" The response to this yes/no question should be based on it semantic equivalence to the imperative: "Please tell me what time it is." You may think an unamplified affirmative response would be perfectly appropriate—that is the question that is inappropriate—but the following examples illustrate the ludicrousness of always basing responses on literal interpretations.

Features of Natural Language – reasons of difficulty of simulation of its understanding (3)

- It is technically correct to answer "Yes" to the question: "Is there any water in the refrigerator?" when the only water present is frozen into ice is in the cells of the celery. Questions of the form: "Do you want this or not?" could always be answered affirmatively by interpreting "or" as the logical inclusive disjunctive, for the choices given exhaust the possibilities. We certainly don't want computer systems to respond in these ways any more than we want people to
- Different representations of the same sentence are appropriate in different circumstances. In the preceding example, the train data base should use a very simple structure of facts, whereas the boyfriend must make use of nonfactual, extralinguistic knowledge of undetermined structure. The complexity of meaning representations required for the general cases is one of the chief difficulties of natural language understanding.

Levels of language

- Words, parts of words (lexical level, morphology)
  - – Structure of words
- Phrases, sentences (Syntax, syntactic level)
  - – Structure of phrases and sentences
- Sense, meaning of phrases (Semantics, semantic level)
  - – The meaning here is that associated with the sentential structure, the juxtaposition of the meanings of the individual words
- Sense, meaning of sentences (Semantics, discourse level)
  - - Its domain is intersentenial, concerning the way sentences fit into the context of a dialog text
- Sense as goals, wishes, motivations and so on (Pragmatics)
  - – Deals with not just a particular linguist context but the whole realm of human experience

Example

- Following sentences are unacceptable on the basis of syntax, semantics, and pragmatics, respectively:
  - John water drink.
  - John drinks dirt.
  - John drinks gasoline.
- Note that the combination of "drink" and "gasoline" is not unacceptable, as in "People do not drink gasoline" or the metaphorical "Cars drink gasoline.“
- It is traditional for linguists to study these levels separately and for computational linguists to implement them in natural language systems as separate components. Sequential processing is easier and more efficient but far less effective than an iterated approach.

Difficulties

- Traditional grammars dealt primarily with syntax. The most popular kind of grammar in computational linguistics is the context-free grammar. Since most structured computer languages have context-free grammars, efficient context-free grammar parsing algorithms have been developed from compiler design work.
- Although ungrammatical sentences are unparsable, they are not necessarily unmeaningful. In many ways, syntax is irrelevant to understanding. Communication is rarely impeded by a lack of agreement in number or tense, for example, as in "The person who done it—it's their fault."

Example

- However, the role of syntax can be crucial. There is absolutely no other way to distinguish "The man who knew him went left" from "The man who knew he went left." Of the following four sentences, the first two are syntactically similar but should be interpreted very differently by a natural language system, while the last two, which are quite different in form, should transform to exactly the same internal meaning representation:
- Mother was baking.
- The apple pie was baking.
- Mother baked an apple pie.
- An apple pie was baked by mother.
- Context-free grammars do not account for such phenomena; transformational grammars do, but all attempts to parse them have resulted in combinational explosion.
- Once a meaning representation scheme has been selected, there is still the problem of how to map the input sentences to it. The mapping procedure is especially complicated because a single sentence can have many meanings, and many different sentences can have the same meaning. The former phenomenon, which presents the greater difficulty, is known as ambiguity.

Lexical ambiguity

- "Time flies like an arrow." Each of the first three words could be the main verb of the sentence, and "time" could be a noun or adjective, "flies" could be a noun , and "like" could be a preposition.
- Thus the sentence could have various interpretations other than the proverbial one. It could be a command to an experimenter to perform temporal measurements on flies the same way they are done on arrows. Or it could be a declaration that a certain species of fly has affection for a certain arrow.
- Some less artificial examples are: "I saw that gas can explode" (either explosive incident was witnessed or an explosive property was demonstrated), "They should have scheduled meetings" and "Visiting relatives can be annoying.“
- Those examples all involve word class ambiguity. A simpler type of lexical ambiguity involves multiple meanings of a word within the same class. "The pitcher fell and broke" is syntactically incomplete or semantically invalid if a system happened to select the baseball related definition of "pitcher."

Lexical ambiguity (2)

- Since so many words have multiple definitions, it is important for a system to have some criteria for distinguishing the appropriate one at an early stage of analysis. One way to accomplish this is by supplementing the dictionary definitions with semantic markers—general semantic properties (such as animate, abstract, location, mobile) whose usage is guided by contextual clues.
- Suppose the two entries for "pitcher" were so marked, one with the containment property for liquids and the other a s baseball-related and human. Then given the sentence: The water is in the pitcher,” a system would select the former definition due to the presence of the preposition "in" and it might even be able to understand the eclipsed "John drank a pitcher." Of course, we could still confuse it with "John drank a tall pitcher while watching the baseball game." Unfortunately, relying on semantic markers to perform lexical disambiguation in general requires a quantity and a specificity that makes them as unwieldy as the word definitions themselves
- Resolving lexical ambiguity often requires a context larger than the sentence. In reading the isolated sentence, "She approached the bank," there is no way to know whether the bank is a lake ridge or a financial building. However, previous sentences might contain helpful information, such as that she was wearing a ski mask or she was a boat.

Syntactic ambiguity

- Syntactic ambiguity is structural ambiguity.
- A very common type of structural ambiguity is due to modifier placement, as in the following innocuous-looking example: "John saw the woman in the park with a telescope." Each of the two prepositional phrases, "in the park" and "with the telescope," could be modifying either "saw" or woman," and the second one could also be modifying the first's noun, "park.“
- From the various ways of combining these possibilities, five synaptic structures result. The interpretation corresponding to structure IV, for example, is that John is in the park and the telescope is the park but John is seeing the woman, who may or may not be in the park, with his naked eye
- Part of the multiple ambiguity involved is due to the choice of the word "telescope," for it s both an a object used for seeing and one that is found both in parks and with people. If we replace "telescope" with "fountain," only structures II and IV make sense; substituting "cat" for "telescope" rules out at least I and III, whereas substituting "baby" definitely rules out all but V.

Syntactic ambiguity (2)

- Since the number of possible structures increases exponentially with the number of modifier phrases, it becomes necessary to eliminate the unlikely ones at an early stage of processing. In the absence of contrary information, the tendency is to try to attach the modifier to the closest constituent first. The following joke, in which the modifier is an adverb, plays on this tendency
- John: I want to go to bed with Marilyn Monroe again tonight.
- Jane: Again?
- John: Yes, I've had this desire before.

Syntactic ambiguity (3)

- Nominal compounds, in which nouns may be used as adjectives, entail a similar type of modifier ambiguity. Our knowledge that electric pencils don't need sharpening helps us parse "electric pencil sharpener," but "dangerous animal trainer" and "metal shelf bracket" could each be interpreted either way. And without carpentry experience, there is no way to know whether a wood screw would screw wood.
- A semantic analog of this problem affects the structure of the deeper meaning representation. For example, consider the difference between "knowledge engineer" and "blonde engineer"; "knowledge" modifies "engineer" and "engineer " modifies the implicit noun "person," whereas "blonde" modifies "person" directly.

Syntactic ambiguity (4)

- One of the most difficult technical issues for natural language systems to deal with is conjunction scope. For example, in the phrase "old men and women," the women are also supposed to be old only if "old" is outside the scope of "and." Context-free grammars that deal with conjunctions in general require them to be binary operators, so a nested pair conjoins has two possible structures
- If the conjunctions are the same, these could be semantically equivalent, as in "I'll have cake and pie and cookies." But consider the less greedy:
- I'll have bread or toast an tea.
- I'll have toast or tea and sugar.

Syntactic ambiguity (5)

- Systems also need a way to account for the inappropriateness of conjoining the sentences "Mother was baking" and "The apple pie was baking" to produce "Mother and the apple pie were baking.“
- Negation and quantifier scope engender further confusion. These phenomena are particularly problematic in expert systems, which use such logical terms a lot. The command "List the trains that service every city" could be interpreted to yield a list for each city or a single list consisting of their intersection. On the other hand, when a parent tells a child "Everyone does not do that," the parent could be taking advantage of ambiguity to seem to be making a stronger statement
- Subtler situations occur with vaguer quantifiers. Compare "Not many people voted for him" to "Many people didn't vote for him." It is very hard to distinguish cases semantically. In neither case is the election outcome apparent, but that's our linguistic system

“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