CSME 206A – Natural Language & Speech Processing **Spring Semester**

Lecture 9: Introduction to Natural Language Processing-II

Lecturer: K.R. Chowdhary *: Professor of CS*

9.1 Theoretical Issues

The theoretical issues in NLP are, commonly divided into three area [hirschberg1988natural]:

- *Syntax*, the study of sentence structure,
- *Semantics*, the study of context-independent meaning, and
- *Pragmatics* and *discourse*, the study of context-dependent meaning.

The study of syntax is about how the sentences of a language can be composed from smaller units – words and phrases – and how the constraints on this process can be expressed and represented. From a practical point of view, the constraints impose restrictions in generating a sentence such as (see Fig. 9.1),

(i). "The cat is on the mat,"

such that it will not generate *Is cat the mat on*.

The semantics is concerned with associating meanings with these structures, mapping between words and phrases and the real world or the models of it. This allows us (or should allow the machine) to interpret sentences or to generate them by specifying the conditions under which they are true (i.e., their truth conditions).

The knowledge of pragmatics/discourse help us how to interpret sentences in the larger discourse context of speaker and listener and what belief they have on each other, and based on what has already been said, and the surroundings under which their exchange occurs.

9.1.1 Syntax

The syntax in NLP is concerned for construction of *grammars* and the specification of *parsing algorithms*. Grammars are specifications of what constitute a legal string in a language. The grammar used is phrase structure grammar, which is set of rewrite rules of the form $A \rightarrow \alpha$, where the nonterminal A is replaced with a collection of terminals and non-terminals α. The nonterminals correspond to phrases, and terminals corresponds to words of the language.

With such grammar G, a parsing algorithm can determine whether a string of input is accepted by the grammar G, i.e., whether strings constitutes a legal sentence of the language whose grammar is G.

Most parsers also assign structures (or parse trees) to legal strings in terms of the grammar being used, a typical case is shown in Fig. 9.1. Parse trees allow NLP systems to distinguish between the various meanings of sentences. This distinction can be shown by attaching the Prepositional Phrase (PP) under different nodes in the parse tree directly under the sentence node S, or under the VP (Verb-Phrase) node, Fig. 9.1 shows the PP attached under the Verb phrase.

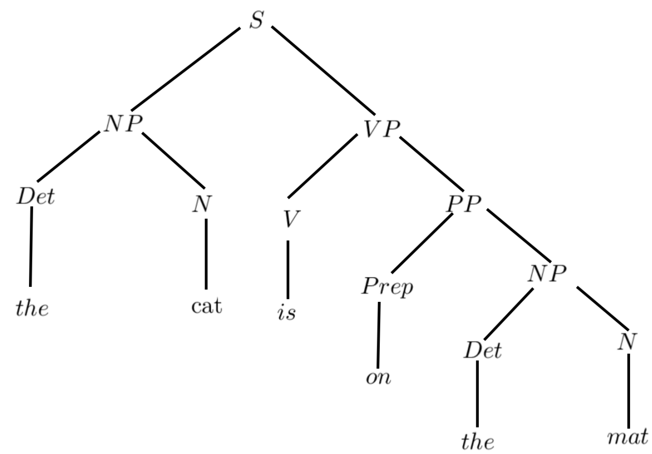

Figure 9.1: With a parse-tree, NL systems can determine sentence meaning

From theoretical point of view, grammars and parsing strategies are of interest as far as human linguistics is concerned.

9.1.2 Semantics

The semantics is concerned on questions like:

- What type of representation is best suited to Natural Language?
- How do we associate the components of a given representation with real-world counterparts, or how we can associate a model of real world?

The content of a lexicon and its structure, or of a lexicon or dictionary is related to the construction of the grammars they access. The study of word meanings, how they are related to one another, and the basis for lexical choice is concerned with *lexical semantics*. For example, a lexicon should be able to capture the fact that the words *kill* and *assassinate* describe similar actions but are not interchangeable. A lexicon should also be able to represent the relationship between compounds such as *pet owner* or *polo player* and their sub-parts, so that not every acceptable variation receives a separate lexical entry, and compounds such as "pet player" and "polo owner" will be disallowed.

Semantic form as well as content receives much attention in NLP. Usually, the semantic components of Natural Language systems map between syntactic structure and predicate argument structure (or the logical form) of a sentence. For example, the sentence "The cat is on the mat", might be simply represented as *on(cat, mat)*. But, whether this mapping should be defined in terms of some "logical" language, such as FOPL (first order predicate logic), or the other forms such as, frame or script are betetr choice. The conclusion is that there is need of semantic phenomena that must be represented.

One such phenomena is, scope ambiguity, i.e., what is scope of a particular word in the given sentence. For example, the sentence,

(ii). All doors will not open,

will have different truth conditions depending on the scope of *all* and *not*. It may mean either, “no doors will be open”, or “some doors will open”. This depends on whether, *not* has scope over the *all*, or the *all* has effect over the *not*.

Another example is,

(iii). John only introduced Mary to Sue.

This has following more than one interpretations.

- John did noting except introducing Mary to Sue.
- John introduced only Mary to Sue, and none else was introduced to her.
- John introduced Mary to Sue, but Mary was not introduced to any one else.

This is due to different scopes of *only*.

The other problems in Natural Language are representation of:

- Natural Language processes and substances,
- mental attitude such as beliefs, desires, and intentions, and
- situations, events, and actions.

9.1.3 Pragmatics

The semantic meaning is often identified with context-independence, and pragmatic-meaning is defined as context-dependence. However, this division of labor often breaks down. The phenomena commonly accepted as subjects of pragmatic study include: presuppositions, conversational implications, indirect speech acts, and discourse phenomena such as the various means of communicating attentional and intentional structure in discourse.

Presuppositions are aspects of utterance meaning that a speaker seems to assume to be true by making the utterance rather than asserting such to be the case. For example, a speaker who utters.

(iv). The king of Skyland is bald,

might be said to presuppose the existence of a king of Skyland.

The sentence:

(v). Some birds fly,

it conveys that as far as speaker knows, it is not the case that all the birds fly. Suppose, we learn that all birds fly, but we will not say that speaker lied. Thus, the "some" may extend to "all."

In case of indirect speech acts, there is yet another context dependent meaning. Consider the following sentence:

(vi) Can you sit up?

This may convey either the direct yes-no, or request that a hearer perform some action of sitting up. In case a visitor asks this to a patient in hospital, it would be an inquiry, while if a doctor is asking to a patient, it would be a request.

Each of these types of pragmatic meaning are the subject of investigation in Natural Language and are incorporated into Natural Language question-answering systems.

9.1.4 Discourse

More general studies of discourse or text generation and understanding have focused on identifying discourse structures, scripts, schemas, or rhetorical relations. While some of these proposals have permitted text generation in well-defined domains, ability to generalize from domain to domain, dependence on vast amounts of encoded *world knowledge*, and difficulties in specifying just what an *elaboration schema* is and how it can be recognized or inferred have plagued all such work.

Recent work proposes that discourses have three structures, *linguistic*, *attentional*, and *intentional*. The attentional structure is the structure of what speaker and hearer are attending to during the discourse; it includes the topic structure of the discourse as well as the information structure items mentioned in the discourse whether these items represent "new" or "old" information.

The *intentional* structure is the structure of what speakers intend to accomplish during the discourse. We can infer both structures, at least partly, from the *linguistic structure* (the sentences speakers utter). It has been proposed in various researches that some ways intonational features of these utterances can communicate both *attentional* and *intentional* structure [].

9.2 Ambiguity in Natural Language

Natural Language is full of ambiguity, and it has its own merits, of course along with challenges of resolving the meaning many times. The principal difficulty in processing Natural Language is this pervasive *ambiguity* found at all levels of the problem. For example, all Natural Languages involve:

- *Simple lexical ambiguity* (e.g. "duck" can be a noun [the animal] or a verb [to avoid something thrown]).
- *Structural* or *syntactic ambiguity* (e.g. in "I saw the man with a telescope," the telescope might be used by subject "I" for the viewing or might be held by the man being observed).
- *Semantic ambiguity* (e.g. "go" as a verb has well over 10 distinct meanings in any dictionary). The difference with "duck" in above example is, that "duck" can be ambiguous as it can be noun or verb. But, for "go", even though its types does not change, it can have different senses.

- *Pragmatic ambiguity* (e.g. "Can you lift that rock?" may be a yes/no question or a request to lift the rock).
- *Referential ambiguity* (e.g. in "Jack met Sam at the station. He was feeling ill." It is not clear who is ill, although the remaining of the sentence might suggest a preferred interpretation).

When, all the above mentioned forms of ambiguity interact, it may be an extremely complex interpretation process. It is the prevalence of ambiguity that distinguishes Natural Languages from precisely defined artificial languages, such as logic and programming languages. It also makes most of the techniques developed in programming language grammars, such as parsing, and the semantics, ineffective for NLP unless significantly modified.

Many a times the word boundaries are overlapped[1] and the sentence understood are totally different. At the next level, the syntax of the language help us to decide how the words are being combined to make larger meanings. Hence, when there is sentence "the dealer sold the merchant a dog," it is important to be clear about what is sold to whom. Some of the common examples which requires far more background knowledge to understand them, are given below:

> I saw the Golden gate bridge flying into San Francisco.
>
> (Is the bridge flying?)
>
> I ate dinner with friend.
>
> I ate dinner with fork.
>
> Can companies litter the environment.
>
> (Is this a statement or question?)

Finally, assuming that we have overcome the problem at the previous levels (of sentences), we must create internal representation, and then, somehow use the information in an appropriate way. This is the level of semantics and pragmatics. Here too, the ambiguity is prevalent. Consider the following more examples.

Jack went to store. He found the milk in aisle three.
He paid for it and left.

Here the problem is deciding whether "it" in the sentence refers to "aisle", "milk", or even the "store".

The most important part in the above is, what is internal representation, so that these ambiguities in understanding the sentence do not occur, so that a machine understands the sentence the way human understands it.

9.2.1 Ambiguity and overlapping Subsystem

Many words in English appear as different parts of speech depending on context. This is true even of the most common words, such as *to*, which can be a preposition or the auxiliary of an infinitival verb[2]. And even

[1] scope of the word in effecting interpretation of other words.
[2] Any verb that is preceded by the word "to love", "to eat", "to run", "to believe", "to wonder", is called infinitive verb.

with a single part of speech, many words have multiple meanings. Ambiguity arises from many sources when words are combined in a sentence according to the grammatical rules of the language. *Structural ambiguity* (the ambiguity that arises from syntax) alone can be staggering. A ten-word sentence typically has several distinct structural analyses, and can have more than one hundred.

Several distinct modules of linguistic knowledge (the phrase structure, predicate-argument structure, noun-phrase reference, and informational structure) contribute to the resolution of ambiguity. No single component has responsibility for resolving all ambiguity. Current research focuses on developing parsers in which each component will say what it can and no more. Each one of the subsystems,

- phrase structure,
- predicate-argument structure,
- reference, and
- informational structure,

has its own organizational principles and makes its own contributions to the description of a sentence.

Look again at the example,

(i). The cat is - on - the mat.

It is possible to construct an analogous sentence that has the same basic constituents but different lexical material.

(ii). *My favorite show is - on - the first Monday of the month,*

Note that sentence (i) and (ii) both have same structure: *noun-phrase verb phrase - noun phrase.*

This sentence means roughly the same as (iii) and (iv):

(iii).*The first Monday of the month, my favorite show is on.*

(iv). *My favorite show is on, the first Monday of the month.*

The structural difference between (i) and (ii) is that in (ii), the final noun phrase "the first Monday of the month", is not an object of the preposition, as in (i), but, it is a time modifier of the entire sentence.

9.2.2 Sentence structure Analysis

It is clear from examples above that grammar deals with organization of phrases and their types; however, the grammar does not deal with their meaning and reference. But, the structural analysis we find for (i) and (ii) depend on the meaning and reference of the noun-phrase constituents. How can we capture this fact without explicitly representing this ambiguity?

The solution has two aspects. First, we decompose the notion of attachment into a phrase-structure component and a predicate-argument component. In phrase structure, attachment or domination is interpreted to mean "is somewhere in the phrase." Thus, we can say that the sentence dominates the final noun phrase in both (i) and (ii). This is consistent with the prepositional phrase dominating the noun phrase in (i). In predicate-argument structure[3], attachment means "is an argument of" or "is an immediate constituent of." Both these components must be involved in describing the structure of a sentence.

In Example (i), the syntax can find a prepositional phrase *on* and a noun phrase following it. Yet it can remain unsure about whether the noun phrase is the complement of *on.* The phrase structure analyzer only says what it is certain about. Essentially, this representational shift succeeds in delaying the decision about the structure by encoding the ambiguity without explicitly representing it.

Thus, the phrase-structure analyses of both (i) and (ii) are the same: the phrase-structure analyzer can only say that there is a prepositional phrase "on" that is dominated by the sentence and the following that prepositional phrase there is a noun phrase that is dominated by the sentence. The predicate-argument component has to determine whether the noun phrase is an argument of the preposition. It does this on the basis of having established the referent of the noun phrase, and having looked at the adjacency of the phrases (along with other considerations)(see Fig. 9.2).

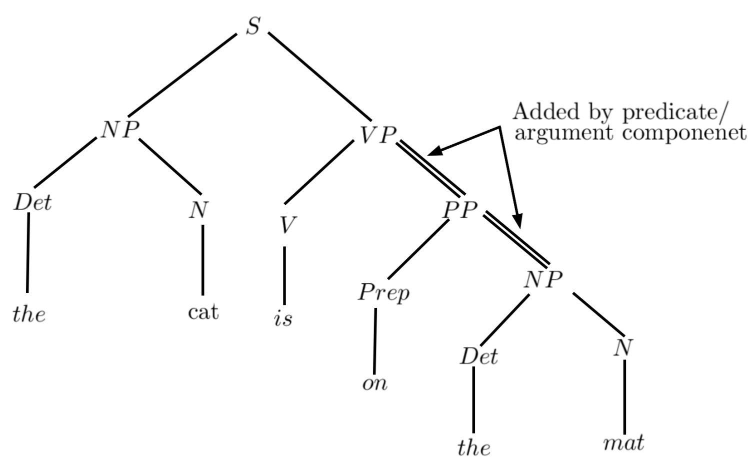

Figure 9.2: Parse-tree after predicate-argument analysis

The idea of treating parsing as description and using *domination* as a special predicate has wide applications. It solves long standing problems in parsing, including prepositional phrase attachment and conjunction of phrases with *and*.

9.3 Applications of NLP

Developing a program that understands Natural Language is a difficult problem. Number of Natural Languages are large in number, they contain infinitely many sentences. Also there is much ambiguity in Natural Languages. Many words have several meanings, such as discussed in the previous section, the sentences have meanings different in different contexts. This makes creation of programs that understands a Natural Language, a challenging task.

[3]An example of predicate argument structure is "on(cat, mat)"

There is huge amounts of data in Internet, in terms of hundreds of billions pages, and increasing in accelerated way. Applications for processing such large amounts of texts require NLP expertise in programs. Some potential applications of NLP are:

- Classify text into categories
- Index and search large texts
- Information Retrieval
- Automatic language translation
- Natural Language generation
- Speech understanding: Understand phone conversations
- Information extraction
- Automatic summarizing: e.g., condense a book into 1 page
- Question answering
- Knowledge acquisition
- Text generations/dialogues