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

Lecture 10: Introduction to Natural Language Processing

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

10.1 Introduction

Developing a program that understands natural language is a difficult problem. Number of natural languages are large, they contain infinitely many sentences. Also there is much ambiguity in natural languages. Many words have several meanings, such as can, bear, fly, orange, and sentences have meanings different in different contexts. This makes creation of programs that understands a natural language, a challenging task.

10.2 Challenges of NLP

Many times the word boundaries are mixed 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 are:

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

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

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", "three", "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 and a machine understands the way a human understands the sentences.

10.3 Applications

There is huge amounts of data in Internet,at least 20 billions pages, and increasing in accelerated way. Applications for processing such large amounts of texts require NLP expertise in programs. Some potential applications are:

- Classify text into categories
- Index and search large texts
- Automatic translation
- Speech understanding: Understand phone conversations
- Information extraction: Extract useful information from resumes
- Automatic summarization: e.g., condense a book into 1 page
- Question answering
- Knowledge acquisition
- Text generations / dialogues

10.3.1 Some Applications details

Information Extraction: This is an application where one extracts information from unstructured text, and creates a database like relational database, which can be later queried.

"Firm XYZ is a full service advertising agency specializing in direct and interactive marketing. Located in Bigtown CA, Firm XYZ is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automotive account. Experience in online marketing, automotive and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables . . . Compensation: $50,000-80,000$ Hiring Organization: Firm XYZ."

Given the above text, the extracted information may be:

Attribute	Value
INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Bigtown, CA.
COMPANY	Firm XYZ
SALARY	$50,000-80,000$

10.4 Computational Linguistics

A simple sentence consists a subject followed with predicate. A word in a sentence acts a part of speech (POS). For English sentence, the parts of speech are: nouns, pronouns, adjectives, verb, adverb, prepositions, conjunctions, and interjections. Noun tells about names, where as the verb talks of action. Adjectives and adverbs are modifying the nouns and verbs, respectively. prepositions are relationships between nouns and other POS. Conjunctions joins words and groups together, and interjections express strong feelings.

Most of us understand both written and spoken language, but reading is learned much later, so let us start with spoken language. We can divide the problem into three areas - acoustic-phonetic, morphological-syntactic, and semantic-pragmatic processes as shown in figure 10.1.

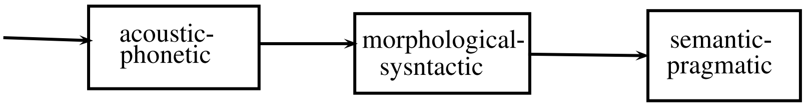

Figure 10.1: The three levels of linguistic analysis.

10.4.1 Levels of knowledge in language understanding

A language understanding program must have considerable knowledge about the structure of the language including what the words are and how they combine into phrases and sentences. It must also know meaning of the words, how to contribute meaning of the sentence and to the context in which they are being used. In addition, the program must have general world world knowledge and knowledge about how the humans reason.

The components of the knowledge needed to understand the language are following:

- **Phonological:** Relates sounds to the words we recognize. Phoneme is smallest unit of sound, and the phones are aggregated into word sounds.
- **Morphological:** This is lexical knowledge, which relates to word construction from basic units called morphemes. A morpheme is the smallest unit of meaning, for example, the construction of *friendly* from *friend* and *Ly.*
- **Syntactic:** It is knowledge about how the words are organized to construct meaningful and correct sentences.
- **Pragmatics:** It is high level knowledge about how to use sentences in different contexts and how the contexts effects the meanings of the sentences.
- **World:** It is useful in understanding the sentence and carry out the conversation. It includes the other persons beliefs and goals.

The figure 10.2 shows the stages of analysis in processing Natural language.

Exercises and Review Questions

1. What are the challenges of NLP?
2. Give one example of following ambiguities:
   (a) Phonetic
   (b) Syntactic

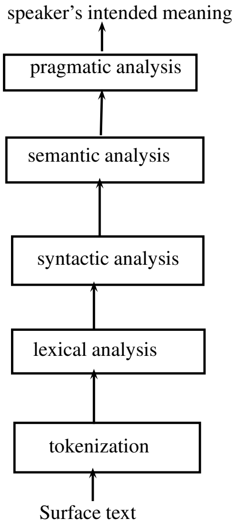

Figure 10.2: Stages in Natural Language Processing.

   (c) Pragmatic

3. What are the applications of NLP?
4. How the Information Extraction is different from Information Retrieval?
5. Explain the difference between syntax, semantic, and pragmatic analysis.
6. Write an algorithm for performing tokenization of a steam of text.

References

[1] D. Jurafsky and J. Martin, "Speech and Language Processing," *Pearson India*, 2002.