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

Lecture 8: Introduction to Natural Language Processing-I

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

8.1 Introduction

Natural Language processing (NLP) refers to computer systems that analyze, attempt to understand, or produce one or more human languages, as its output. The input might be text, spoken language, or from keyboard input. The task might to translate to another language, to comprehend and represent the content of text, to build a database or generate summaries, or to maintain a dialogue with a user as part of an interface for database/information retrieval. This chapter addresses issues about Natural Language which have been solved as well as those not solved yet, the key areas of Natural Language processing, common applications of NLP, theoretical issues of syntax, semantics, and ambiguities in Natural Language.

It is extremely difficult to make a system to actually "understand" a language. All we can actually test is whether a system appears to understand language by successfully performing its task. The *Turing test*, proposed by Alan M. Turing (1950) has been the classical model. In this test, the system must be indistinguishable from a human when both answer arbitrary interrogation by a human over a computer terminal. This test has the unfortunate property that, while it sets the ultimate goal, it provides for no intermediate evaluation of work along the way, or no evaluation in phased manner.

A growing concern in NLP is with developing more sensitive models of evaluation that can measure progress, given the current performance levels. The usual approach is to develop evaluation tests within limited domains to test specific capabilities. For example, in the area of Natural Language interfaces for data query, statistical performance measures can be determined based on test sets of human-generated questions collected in protocols that use another human to simulate the system. It remains an area of active concern, however, as to how more complex systems that can handle extended dialogue can be evaluated.

Below given is explanation of certain terms and the status of problems solved /pending, as well as the challenges there in.

Mostly solved: Spelling check, text categorization, named entity recognition.

Syntax and ambiguity: I saw the man with a telescope.

Who had the telescope? (I or man)

Semantics: Study of meaning of words, and how these combine to form the meaning of sentences

Following are the domains related to semantics, many of them are solved through context:

- Synonymy: Fall and Autumn.

- Hypernymy and hyponymy (is a): Animal and Dog.
- Homonymy: fall (is verb and also a season)
- Autonymy: Big and Small.
- The astronomer loves the *star*. (The star in the sky or a celebrity?)

Pragmatics: It is about Social use of language, i.e., the study of how language is used to accomplish goals, and the influence of context on meaning – understanding the aspects of a language which depends on situation and world knowledge. The examples are:

Give me the salt! (Command or request.)

Could you please give me the salt? (Question or request.)

Discourse: It is study of linguistic units larger than a single statement.

John reads a book. He borrowed it from his friend. ('He' stands or friend or john?)

Discourse analysis: Alice understands that you like your mother, but *she* ..

(Does *she* refer to Alice or your mother?)

8.2 Some characteristics of Natural Languages

The Natural Languages are very powerful, far more complex in syntax and semantics than the computer languages. It is important to understand their unique properties so that we are better equipped to understand the processing of Natural Language(s) through machine [briscoe2011introduction].

Following are the unique properties of Natural Languages:

Productivity Animal communication appears to be restricted to a finite set of calls. Vervet monkeys[1] have 3 alarm calls for 'look out there's a snake/leopard/eagle' which induce different defensive behavior in the troop (up tree / away from tree/under tree). But human languages allow an infinite range of messages with finite resources.

Discreteness /Duality Words and morphemes are comprised of phonemes. Words and morphemes have (referential or grammatical) meanings, but phonemes do not. For example, /*pat*/ and /*bat*/ are different words distinguished by the phonemes /*p*/ and /*b*/ which also distinguish /*pad*/ and /*bad*/ but /*p*/ and /*b*/ alone do not have a meaning. The plural morpheme (+*s*) can be suffixed to three of these words, but is realized as either /*s*/ and /*z*/ – so-called allomorphs of the plural morpheme. An inventory of 40 or so phonemes provides a much bigger inventory of words, even given phonotactic restrictions on the combination of phonemes into syllables.

[1] Vervet monkey or somply vervet is an old world monkey of native Africa.

Syntax Human languages are not just bags of words with no further structure. The organization of words into sentences is conveyed partly by word structure (endings/inflectional suffixes in English) and arrangement/order. So "Kim loves Sandy" does not mean the same thing as "Sandy loves Kim" and "* loves Kim Sandy" does not convey much at all. In "They love each other," love has a different form because it is agreeing with a plural subject rather than a 3rd person singular subject.

In order to gain further insight into the function of syntax, consider what a language without syntax. Such a language would be just a vocabulary and a sentence would be any set of words from that vocabulary. Now, imagine that this language has English as its vocabulary. A 'sentence' in this imaginary language will be some what like this:

```
   the    hit(s)
with         tramp(s)
sharp    poor       rock(s)  some
boys     cruel
```

There is no clue which words should be interpreted with which others in this sentence, so there are many possible interpretations which can be 'translated' into real English, as in (a, b), in the following.

a) The cruel boys hit(s) some poor tramp(s) with sharp rock(s).

b) The cruel, sharp tramp(s) with rock hit some poor boys.

How many more possible interpretations can you find? Without syntax, sentences would be very ambiguous indeed and, although context might resolve some of these ambiguities in everyday communication.

Grammar and Inference Linguists tend to use the term *grammar* in an extended sense to cover all the structure of human languages: *phonology*, *morphology*, *syntax* and their contribution to meaning. However, even if you know the grammar of a language, in this sense, you still need more knowledge to interpret many utterances. Below given are some sentences, which are not straight forward to interpret. Out of these phrases and sentences, can you contextualise them to give a different meanings and explain how the context resolves the ambiguities?

She smiled,

I didn't.

Who?

The farmer killed the duckling in the barn.

Everyone in this room speaks one language.

Every student thinks he is the cleverest person in his class.

Can you open the gate?

While the grammatical knowledge required to encode or decode messages in a particular language have bounds, however, the inferences can be easily extracted from utterances. Consider the kinds of knowledge required to make sense of the following dialogues between speaker A and B:

A: The phone's ringing. B: I am in the bath.

A: John bought Volkswagen. B: Must be costly.

We need to know all sorts of culturally specific and quite arbitrary things to interpret these sentences. E.g., normal location of phones, the word 'bath' stands for bathroom, the car brands, and what plausible inferences can be made on these dialogues.

Displacement Most animals' communication is about the here and now (recall Vervet monkey calls, though the bee dance, indicating direction and distance of food sources, is sometimes said to be a partial exception) but human language allows communication about the past, the future, the distant and the abstract, as well as the here and now and the perceptually manifest.

Cultural Transmission Animal communication systems are very largely innate – Vervet monkeys are genetically programmed to make 3 calls, although some aspects of the meaning and sound are tuned up by experience. Human language is very largely learned (that's why there are 6000 or so attested languages with widely differing grammatical systems and vocabulary). However, in many ways first language acquisition differs from learning, say, to swim or do sums – it is very reliable under widely differing conditions, does not require overt tuition, and there is not that much variation in the core grammatical skills of all adult humans.

Human children only consistently fail to learn fluent language if entirely denied access to any sample until they are in their teens. There is much wider variation between individuals and between children and adults in acquisition of passive (understood) and active (produced) vocabulary. Vocabulary learning is an ongoing process throughout life and is supported by teaching aids like dictionaries in literate cultures, whilst first language, grammatical acquisition appears to be largely complete before puberty.

Speak/Sign /Write Animal languages always use a single modality: manual gestures, e.g., dances, oral sounds, clicks, etc. Humans can acquire or even create natural sign languages if denied access to spoken language. Human languages also often have a written form, though the latter is significantly less 'natural' and literacy is only acquired (by most individuals) if explicitly taught over a sustained period.

Variation and Change Human languages, unlike animal communication systems, vary considerably through time and space (within-species birdsong being the partial exception). Of the 6000 attested languages we know about, about 1000 are spoken in Papua New Guinea (an area about the size of state Rajasthan, in India). There have probably been 100,000-500,000 human languages depending on when language first emerged (mostly un-documented, prehistoric, and extinct, of course). Languages have constantly (dis)appeared as a result of population movements, and the birth and collapse of societies. However, the current rate of language death far exceeds that of creation.

For each language spoken by a population of any size, there are many dialects associated with different regions and/or social classes. New words and novel grammatical constructions are constantly entering languages and old ones are constantly decaying. It is impossible to predict with certainty whether an innovation will spread or decay, although afterwards it is possible to document with some accuracy what did happen (historical linguistics), and some social situations (e.g. creolisation[2], population movement) cause partly predictable rapid and radical change. Dialectal variation is often a function of social groups' self-identity, so often the explanation of change or variation is in terms of social change, movement or interaction of individuals between groups, etc (called sociolinguistics).

[2]A mother tongue that originates from contact between two languages.

8.3 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 8.1.

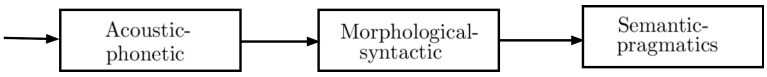

Figure 8.1: The three levels of linguistic analysis

8.3.1 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:

- *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 bearing word, 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:* This knowledge is useful in understanding the sentence and carry out the conversation. It includes the other persons beliefs and goals.

The Fig. 8.2 shows the stages of analysis in processing a Natural Language.

The concepts of phonological and morphological are separately discussed in text. Here, we will be give introduction about syntax, semantics, pragmatics, prosody, and discourse.

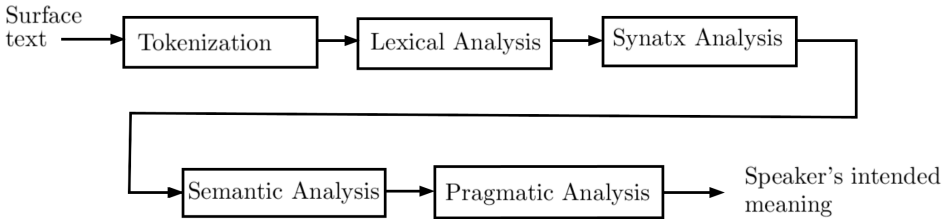

Figure 8.2: Stages in Natural Language Processing

8.3.2 Syntax

Syntax concerns the way in which words can be combined together to form (grammatical) sentences; e.g., "revolutionary new ideas appear infrequently", is grammatically correct in English; "colorless green ideas sleep furiously" is grammatical but nonsensical, while "**ideas green furiously colorless sleep*" is ungrammatical too[3]. Words combine syntactically in certain orders in a way which mirrors the meaning conveyed; e.g., "John loves Mary" means something different from "Mary loves John." The ambiguity of John gave her dog biscuits stems from whether we treat her as an independent pronoun and dog biscuits as a compound noun or whether we treat her as a demonstrative pronoun modifying dog. We can illustrate the difference in terms of possible ways of bracketing the sentence.

(John (gave (her) (dog biscuits)))

vs.

(John (gave (her dog) (biscuits)),

which is parsing the sentence, and to find out which parse is more probable.

8.3.3 Semantics

Semantics is about the manner in which lexical meaning is combined morphologically and syntactically to form the meaning of a sentence. Mostly, this is regular, productive and rule-governed; e.g., the meaning of: "John gave Mary a dog" can be represented as,

(*some (x) (dog x) and (past-time (give (john, mary, x)))*),

but sometimes it is idiomatic as in the meaning of "John kicked the bucket," which can be (past-time (die (john))). To make this notation useful we also need to know the meaning of these capitalized words and brackets too. Because the meaning of a sentence is usually a productive combination of the meaning of its words, syntactic information is important for interpretation – it helps us work out what goes with what – but other information, such as punctuation or intonation, pronoun reference, etc, can also play a crucial part.

[3]Linguists use asterisks to indicate 'ungrammaticality', or illegality given the rules of a language.

8.3.4 Pragmatics

Pragmatics is about the use of language in context, where context includes both the *linguistic* and *situational* context of an utterance; e.g., if I say, "Draw the curtains", in a situation where the curtains are open, this is likely to be a command to someone present to shut the curtains (and vice versa if they are closed). Not all commands are grammatically in imperative mood; e.g., "Could you pass the salt?", is grammatically a question but is likely to be interpreted as a (polite) command or request in most situations. Pragmatic knowledge is also important in determining the referents of pronouns, and filling in missing (elliptical) information in dialogues; e.g., "Kim always gives his wife his wages." "Sandy does so too."

8.3.5 Prosody

Besides the phonemes that carry the textual content of an utterance, prosodic information gives valuable support to understand a spoken utterance. In short, prosody is the rhythm, stress and intonation of continuous speech, and is expressed in *pitch*, *loudness* and *formants*. Prosody is an important mean of conveying non-verbal information.

There are two aspects of prosody: 1. Concrete aspect, that defines prosody in physical term, 2. Abstract aspect, which defines prosody as influence to linguistic structure. The concrete aspect is concerned with phenomena that involves the acoustic parameters of pitch, duration, and intensity, while abstract aspect is concerned with phenomena that involve phonological organization at levels above the segment. Prosody in speech has both, measurable manifestations and underlying principles. Hence, the following definition is found appropriate:

Definition 8.1 Prosody. *Prosody is a systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production.*□

Its realization involves both segmental features of speech, and serves to convey not only linguistic information, but also paralinguistic and non-linguistic information.

Individual characteristics of speech are generated in the process of speech sound production. These segmental and suprasegmental features arise from the influence of linguistic, paralinguistic, and nonlinguistic information. This explains the difficulty of finding clear and unique correspondence between physically observable characteristics of speech and the underlying prosodic organization of an utterance. Following are some definitions of above terms.

1. Linguistic Information: It is symbolic information that is represented by a set of discrete symbols and rules for their combination i.e. it can be represented explicitly by written language, or can be easily and uniquely inferred from the context.

2. Paralinguistic Information: It is information added to modify the linguistic information. A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles which are under conscious control of the speaker.

3. Non-linguistic Information: It is physical and emotional factors, like gender, age, happiness, crying, which cannot be directly controlled by the speaker. These factors are not directly related to (para-) linguistic contents, but influence the speech anyway.

4. Prosodic characteristics: These are typically expressed in several types of features, which can serve as basis for automatic recognition. The most prominent of those features are duration, loudness, pitch and glottal characteristics. Following is the brief description of these.

Duration The utterances of speech can be elongated or shortened; the relative length carries prosodic information. Usually, it is found that short non-verbal fill-words shows affirmation, whereas elongated fill-words express disagreement.

Power The signal power or loudness of an utterance is another important prosodic feature. In German and English, the intensity often marks or emphasizes the central information of a sentence. Without this information, spontaneous speech could be ambiguous and easily misunderstood. The loudness is measured by the intensity of the signal energy.

Pitch At the bottom of the human vocal tract are vocal cords, called *glottis*. For unvoiced speech, the glottis remain open, while for the voiced speech it opens and closes periodically. The frequency of opening is called the *fundamental frequency* or pitch. It can be calculated from the spectrum of a given speech and its contour over the utterance reveals several information. For example, in *Mandarin Chinese*, the F_0 carries phonetic/lexical information, and in English or German, the pitch specifies a question by a final fall-rise pattern.

Glottal Characteristics The physiological voice characteristics of an individual also contribute to convey non-verbal information. The glottis is the vocal cord area of the human articulatory system and is most commonly known for creating voicing in pronunciation by opening and closing periodically.